THE UNIVERSITY OF CHICAGO

INFERENCES ABOUT THE EVOLUTIONARY HISTORY OF GEOGRAPHICALLY STRUCTURED POPULATIONS: THE INTERSECTION OF POPULATION GENETICS, BIOGEOGRAPHY AND SYSTEMATICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON EVOLUTIONARY BIOLOGY

BY
DAN DAVISON

CHICAGO, ILLINOIS
AUGUST 2006

John Davison

1914 — 2005

ABSTRACT

A central goal of evolutionary biology is to understand the evolution of biological diversity in terms of population genetic processes. Since the youngest portions of lineages are characterised by geographical replacement of populations, this research, which lies at the intersection of population genetics, biogeography and systematics, involves using genetic and phenotypic data to make statistical inferences about the evolutionary history of geographically structured populations. I start by discussing this research program in Amazonian forest birds, and presenting new data on geographical genetic and phenotypic variation. These data are typical of those for many organisms, in that the Amazonian lowlands appear to be occupied by a spatial mosaic of geographically representative populations which in some regions appear to be reproductively isolated by intrinsic or extrinsic barriers to gene flow, or both. A common approach in population genetics is to assume that the data have been sampled from a stochastic process at stationarity and then to query the relative rates at which processes such as coalescence, mutation, recombination and dispersal are occurring. However, understanding recent diversification places greater emphasis on non-equilibrium models, and chapters 2 and 3 explore some issues in statistical inference for non-equilibrium models of the history of structured populations. In chapter 2 I investigate inferences based on completely linked data such as those from the mitochondrial genome and find that estimates of rates of gene flow are upwardly biased and that there is little power to distinguish a recent barrier to gene flow from the alternative of ongoing equilibrium gene flow. An outstanding problem is the lack of statistical methods for studying recombining data such as nearby SNPs and sequences from the nuclear genome. In chapter 3 I develop an approximate likelihood-based approach to inference from these types of data under a model in which a panmictic ancestral population has split into two panmictic daughter populations.

ACKNOWLEDGEMENTS

For many things: Shannon Hackett, Jonathan Pritchard. For helpful conversations: Dick Hudson, John Bates, Tom Schulenberg, the Pritchard and Przeworski labs, Trina Roberts, Jody Hey and especially Graham Coop. For help with fieldwork in Brazil: Alexandre Aleixo, Maria Luisa Videira Marceliana, José 'Pepe' Tello, Nan Pimentel, Otávio, Pesão, Martin Davison, Will Davison, Fred Davison: For the above and more: Mary Noble. I would also like to thank the individuals and institutions responsible for collection and preservation of the samples used in chapter 1, and the free software community for the software used in the course of writing this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The use of genetic data to make inferences about recent evolution has been a major concern of evolutionary biology since the 1960s, when it was first demonstrated that there was substantial genetic variation in natural populations. The basic idea can be summarised as follows:

1. Evolution proceeds by transmission from parents to offspring of genetic material, some of which is involved in determining the phenotype.

2. Alterations to the genetic material occur during this transmission, and there is variation between the genomes of interbreeding sexual organisms, as well as between those of more distantly related organisms.

3. The pattern of genetic variation in a particular group of organisms (the data) is affected by unknown aspects of their biology and their evolutionary history in which we are interested (the model).

4. The transmission of genetic material from generation to generation, and the alterations to the material which occur, are random processes, and therefore the connection between model and data is probabilistic.

5. Therefore we can learn about the model by collecting data on genetic variation and using the methods of statistical inference.

There is currently a very high level of excitement in this area of biology (population genetics), resulting from technological advances in data collection and computing, and theoretical advances, mainly since the 1980s. In particular the collection of genetic data has been revolutionised: it is now possible to obtain much larger quantities of data much more quickly and much more cheaply than before. Three properties of genomes should be emphasised at this point: (i) genomes are large, (ii) the process

of recombination over many generations acts to break down correlations that would otherwise result between sites on the same chromosome (explained in more detail in section 3.3), and (iii) the entire genome experiences the same population history, although variation close to functional regions of the genome may additionally be affected by selection. The point is that genomic variation data potentially contain very large amounts of information about the recent evolution of the populations from which they were sampled. From a statistical point of view, genomewide variation data represent many independent draws from the unknown model which is of interest — statistical inference problems are commonly addressed with far less information than is potentially available in population genetics.

This potential for highly informative data sets is now becoming actuality. For Humans (*Homo sapiens*), the International HapMap Project distributes a data set for a sample of 270 individuals from four ethnic groups (The International HapMap Consortium 2005), which currently contains genotypes at over 1.6 million single nucleotide polymorphisms (SNPs). Another large human data set is that analysed by Bustamante *et al.* (2005) and Nielsen *et al.* (2005) in which over $20,000$ protein-coding genes were sequenced in 39 humans and one Chimpanzee (*Pan troglodytes*). In conjunction with a project to produce a genome sequence for the Domestic Dog (*Canis familiaris*) Lindblad-Toh (2005) report the identification of over 2.5 million SNPs, and also sequenced $22,000$ genomic fragments in four Grey Wolves (*C. lupus*) and a Coyote (*C. latrans*). Nordborg *et al.* (2005) analyse a data set comprising 876 sequence fragments from the flowering plant (*Arabidopsis thaliana*) in 96 individuals distributed across several continents. If variable sites have already been ascertained in the study organism, 'microchip' technology is now available with which genotypes at a few hundred thousand predetermined loci in an individual may be determined simultaneously.

This dissertation focuses on the genomewide effects of recent evolutionary history, as opposed to the locus-specific effects of selection on functional variation. However, before largely abandoning questions of organismal function, the equally exciting

prospects for the use of population genetic data in understanding phenotypes deserve mention. Data sets of the sort referred to above open up the possibilities of characterising the extent of different sorts of selection in the evolution of the genome as a whole (e.g. Bustamante *et al.* 2005), and of identifying genomic locations in which evidence for selection seems particularly strong (e.g. Nielsen *et al.* 2005, Voight *et al.* 2006), ideally resulting in a catalogue of genetic changes whose functional effects are well-understood from molecular biological, physiological and ecological perspectives. Additionally, it is now possible to measure expression levels in an individual at many genes simultaneously, opening up the possibility of studying expression variation together with genomic variation. Finally, an important use of genomewide variation data is to attempt to locate genetic variation that is impacting complex, and in particular pathological, phenotypes via studies of association between genetic and phenotypic variation (see e.g. Risch & Merikangas 1996).

Population genetics, biogeography and systematics

The central theme of this dissertation is the ongoing need for better integration of population genetics with the fields of biogeography and systematic biology. This is an exciting prospect and I consider it to be a very high priority in all three disciplines. Choosing appropriate terminology for this subject is difficult. Clearly, when systematists consider evolutionary relationships among closely related populations, and population geneticists study the evolutionary histories of structured populations, the two fields are approaching each other very closely. Population structure (defined in section 1.1.1) may be caused by a variety of shared properties of individuals, such as membership of social groups. However, because of the overwhelming importance of spatial separation of populations in biological diversification (e.g. Mayr 1942, Coyne & Orr 2004), it is spatial population structure that is of key importance in the intersection of population genetics and systematic biology.

The most closely related sexual populations cannot coexist in sympatry either because they are not intrinsically reproductively isolated, and therefore would form a

single interbreeding population, or alternatively because their ecologies are insufficiently distinct. Some allopatric populations which might coexist do not, because the flux of individuals or propagules between their respective ranges is too low. The area of intersection of population genetics and systematics is the most recent period of evolution, and these basic considerations mean that the youngest portions of lineages are characterised by geographical replacement of populations — there is typically a one-to-one mapping between continuous non-overlapping areas of space and mutually exclusive subsets of the system of populations under consideration.

The literature in evolutionary biology has been confused by attempts to use the word 'species' to refer to collections of populations with various properties (hypothesised, or observed). In order to refer to the subject matter in a more agnostic fashion it is helpful to have a term for such related collections of geographical representative populations. The usual term is 'zoogeographic species', but the implicit restriction to animals seems spurious and since the literature has already suffered as a result of varied usage of the word 'species', I prefer to avoid using a term that includes it. Instead, I will use the phrase 'system of geographical representative populations' (SGRP).

Whether or not a collection of populations strictly replace each other geographically is almost always unknown, and I will use the term SGRP to refer to situations in which there are very few data regarding the extent of sympatry. As an example of the shortcomings of using the term in this way, a widely-distributed lineage may be partitioned into populations which largely replace each other geographically, but which are sympatric in a restricted area. Referring to the lineage as a SGRP ignores the local sympatry, which may be of considerable evolutionary significance. The most common definition of the word 'species' is the biological species concept (BSC) (e.g. Mayr 1942). Although intrinsic reproductive isolation is of profound significance in evolution, the BSC is inadequate for describing biological diversity in space. In particular, the relevance of intrinsic reproductive isolation of disjunct populations between which there is no flux of individuals or propagules is very obscure. Spatially

separated populations exist at various levels of genomic and phenotypic differentiation and have had various evolutionary histories, and there is no reason to expect it to be simple to describe the products of the evolutionary process. The word 'species' ceased to be non-problematic in the 19th century when *de novo* creation of every differing biological form ('species') by a deity ceased to be a tenable position in biology, and the only uncontroversial solution is to not use the word and to attempt to say more precisely what is meant. In order to refer to a taxon classified at the level of 'species' (as opposed to an evolving biological entity), I will follow Hey (2001) in using the term 'species taxon'.

Although Mayr was well-known for his advocacy of the BSC, he wrote extensively on the problems of geographic variation. Mayr & Diamond (2001) is a comprehensive study of the systematics and biogeography of an entire regional avifauna, based not on the BSC but on the SGRP (zoogeographic species) concept. The knowledge of low-level systematics gained without genetic data in this and many other regional systems is impressive. However, as Mayr & Diamond (2001) anticipate for their study system, and as has been demonstrated in the last 20 years, the introduction of statistical population genetic analyses of genomic variation within SGRPs radically advances our understanding of low-level systematics and recent evolutionary history.

The first chapter, *'Subspecific' taxa in Amazonia: distributions of mitochondrial lineages and plumage types in suboscine passerines along a major Amazonian river*, describes geographic variation in three SGRPs that belong to the suboscine radiation of passerine birds. These SGRPs extend across the Amazonian lowlands, and there is considerable spatial variation in plumage characters which is known to be discontinuous across certain major rivers. I find perfect association between patterns of geographic variation in mitochondrial DNA and the groups identified by the plumage variation. The data indicate reproductive isolation between geographically replacing populations, and in at least one region the populations are separated by a river that seems too narrow for this to be the result merely of a low flux of dispersing individuals. However, the significance of phenotypic and genetic differentiation of samples sepa-

rated by continuous forest in the major Amazonian interfluvia will remain uncertain until the intervening areas are better sampled.

The next two chapters concern statistical inference of the recent evolutionary history of structured populations. The population structure could be of any sort but is most naturally viewed as geographic population structure, and these inference problems therefore lie at the intersection of population genetics, biogeography and systematic biology. In *The use of completely linked genetic data for inference of population history*, I investigate the power to detect a recent barrier to gene flow between two populations that were previously connected by gene flow. In *A new approximate likelihood for genetic data under a model of population splitting* I introduce a new method for approximate likelihood-based inference from recombining genetic data under a model in which a panmictic ancestral population has split into two panmictic daughter populations. These chapters focus on simple two-population models and the current state of the field is such that if a comprehensive survey of genetic variation in the Amazonian SGRPs studied in chapter 1 were made, inference would be limited by the lack of suitable statistical methods. This is especially so for models of non-equilibrium population history that are most relevant at the interface of population genetics and systematic biology, and it is especially so for linked genetic data which are particularly informative about such models. There is therefore much important work to be done, and some suggestions for future research are made in the concluding chapter.

References

Bustamante, C., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M., Glanowski, S., Tanenbaum, D., White, T., Sninsky, J., Hernandez, R., Civello, D., Adams, M., Cargill, M. & Clark, A. (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–7.

Coyne, J. A. & Orr, H. A. (2004). *Speciation*. Sinauer.

Hey, J. (2001). *Genes, Categories and Species: the Evolutionary and Cognitive Causes of the Species Problem*. Oxford University Press, Oxford, U.K.

Lindblad-Toh, K. e. a. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–19.

Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press, New York, U.S.A.

Mayr, E. & Diamond, J. (2001). *The birds of Northern Melanesia*. Oxford University Press.

Nielsen, R., Bustamante, C., Clark, A., Glanowski, S., Sackton, T., Hubisz, M., Fledel-Alon, A., Tanenbaum, D., Civello, D., White, T., J Sninsky, J., Adams, M. & Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**, e170.

Nordborg, M., Hu, T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N., Shah, C., Wall, J., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M. & Bergelson, J. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**, e196.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–7.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Voight, B., Kudaravalli, S., Wen, X. & Pritchard, J. (2006). A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

CHAPTER 1

'SUBSPECIFIC' TAXA IN AMAZONIA: DISTRIBUTIONS OF MITOCHONDRIAL LINEAGES AND PLUMAGE TYPES IN SUBOSCINE PASSERINES ALONG A MAJOR AMAZONIAN RIVER

ABSTRACT

Spatial variation in phenotype among geographically representative populations is common on large continents, especially in the tropics. This phenomenon is the basis for a major line of enquiry in evolutionary biology which seeks to explain current spatial variation in terms of recurrent population genetic processes such as drift, gene flow and selection, and unique historical events. It is also at the centre of the confusion surrounding the word 'species'. I outline the phenomenon as it occurs in Amazonian forest birds of the South American endemic suboscine passerine radiation, and present new data on spatial variation in plumage types and mitochondrial DNA in the drainage basin of the Rio Tapajós in Brazil, within three such systems of geographically representative populations. I discuss these data in the context of previous studies which have sampled at broader spatial scales. The new data suggest that phenotypically distinct populations of the antbird *Hylophylax poecilinota*, individuals of which bear very distantly related mitochondrial genomes, approach each other closely in one area without syntopy, occupying opposite banks of a river that is on the order of 50 - 100 m wide. Widely scattered specimen data and, more recently, data on mitochondrial DNA variation, indicate that similar situations may be common among Amazonian suboscines. However sampling in the region of putative contact zones has so far been restricted to opposite-bank samples separated by river channels of 300 m or more. The data presented here thus provide what is to my knowledge the strongest evidence of an intrinsic barrier to gene flow between such geographically representative populations — something which may be much more common among Amazonian forest birds than current taxonomy suggests.

1.1 Introduction

The Amazon basin, with its vast system of forest and wetland ecosystems, is frequently said to be one of the most biologically diverse areas on Earth. A good case can be made for many of the ways in which that statement could be interpreted. For example, in many higher taxa of sexual organisms (say birds, or beetles), the number of non-interbreeding populations coexisting in sympatry — i.e. found together in some area that is small relative to the dispersal capabilities of the organisms — can be very high in Amazonia compared to other areas of the world. In sympatry, reproductive isolation between the different populations is clearly of profound significance and it is natural to treat the reproductively isolated populations as the basic units of any study of the community. However, the topic of this dissertation is biological diversity at larger spatial scales.

At spatial scales that are large relative to typical dispersal distances of the organisms or their propagules, one expects spatial heterogeneities in rates of gene flow. These may result from environmental heterogeneities, or from the evolution of intrinsic pre- or post-mating barriers to gene flow between populations in different areas. Additionally, at such spatial scales populations are often disjunct, and recent gene flow between them is very low or nil because they are separated by an uninhabited zone. The evolution of intrinsic barriers to gene flow between allopatric populations has the property that its effects may persist if the populations later come into sympatry, and is thus of special significance in evolution. However, the current distribution of neutral and non-neutral genetic and phenotypic variation is the result of a history of gene flow which involves both intrinsic genetic and extrinsic spatial and environmental phenomena. Those seeking to understand the evolution of biological diversity

in terms of population genetic processes are therefore concerned with the history of reproductive isolation and barriers to gene flow understood in a wide sense to include both intrinsic and extrinsic reproductive isolation (e.g. Wiens 2004).

Once that objective is agreed upon, it is clear from the continuous, quantitative nature of the phenomena of interest that the situation in sympatry is unique — at larger spatial scales the biota no longer necessarily comprises indivisible biological units. The problem is not merely that it is difficult to identify them, but that in general they do not exist at all. The insistence of many authors (e.g. Coyne & Orr 2004) that the biological species concept is a sufficient basis for the study of biological diversity in situations other than sympatry is therefore inappropriate. This viewpoint requires some terminological adjustments in order to proceed with a discussion of the evolution of geographically structured populations. In order to avoid using terms like 'species' whose meaning is unclear without an accompanying definition, it is common to refer to 'widespread taxa' or 'widespread lineages'. However, this is unnecessarily vague. After all, in Amazonia, Aves (birds) is an example of both, and some means is required of referring to the most recent period of evolution when that is what is meant.

The youngest portions of lineages, whose study lies at the intersection of population genetics, biogeography and systematics, are characterised by geographical replacement of populations — i.e. at any one location, the lineage is represented by only one population. I will use the phrase 'system of geographical representative populations' (SGRP) to refer to a putatively closely-related collection of populations for which available data are consistent with them having that property, at least throughout most of their range. I intend my usage of SGRP to be very similar to that of

'zoogeographic species' by other authors (e.g. Mayr & Diamond 2001) (which term suffers from the implicit restriction to animals and the potentially confusing presence of the word 'species').

Many SGRPs extend widely across Amazonia and exhibit geographic variation in phenotype (see Haffer 1974, 1985, 1997b, and references therein). Understanding the evolutionary processes which have given rise to this diversity is a fascinating and important challenge in population genetics, biogeography and systematics. In this chapter I present new data from the SGRPs *Myrmoborus myotherinus*, *Hylophylax poecilinota* and *Glyphorynchus spirurus*. All belong to the suboscine clade of passerine birds. Suboscine taxa are found in the Old World tropics, but the overwhelming majority are neotropical. The New World suboscines are monophyletic (except for the single SGRP *Sapayoa* in the Pacific lowlands which is related to the Old World suboscines) (Chesser 2004) , and therefore appear to have diversified *in situ*. This endemic South American radiation comprises two major clades: the Tyranni contains the traditional (but not all monophyletic) taxa Tyrannidae (tyrant flycatchers), Pipridae (manakins) and Cotingidae (cotingas), and the Furnarii contains the traditional taxa Furnariidae (ovenbirds), Dendrocolaptidae (woodcreepers), Thamnophilidae (antbirds), Conopophagidae (gnateaters), Formicariidae (ground antbirds) and Rhinocryptidae (tapaculos).

The taxonomic treatment of suboscine SGRPs in Amazonia is variable. Many are classified at the level of 'species'. Some of these are classified monotypically — a single subspecies taxon is recognised which is coincident with the species taxon, as a result either of phenotypic similarity throughout the range, or because work recognising phenotypic variation taxonomically has not been published. An exam-

ple are flycatchers assigned to the species taxon *Ramphotrigon ruficauda*. More frequently the entire SGRP is classified at the level of 'species', but 'subspecies' are recognised in different regions. These may show pronounced variation in plumage or vocalisations (the many suboscine examples include the antbirds *Hylophylax poecilinota*, *Myrmoborus myotherinus*), or relatively little variation (e.g. the woodcreeper *Glyphorynchus spirurus*). Alternatively, certain subsets of populations within a single SGRP may be classified as different 'species' (the SGRP is then referred to as a 'superspecies'). Examples include manakins belonging to the genus *Lepidothrix*. Note that in this case, a list of 'species' will make no distinction between 'species' that are *part of* a larger SGRP, and 'species' that *are* a SGRP, obscuring matters of biological interest such as why the former 'species' do not coexist with nearby populations that are part of the same SGRP. Sometimes the SGRP is classified as a 'species' which occurs sympatrically with congeners, and sometimes the SGRP is a genus itself (e.g. the antbird genus *Rhegmatorhina* or, more trivially, any monotypic genus). Outside the suboscines but still within Amazonian birds, the trumpeter family Psophiidae is an SGRP classified at a level above genus.

It is not always obvious *a priori* which local populations are members of which SGRP (this is one of the tasks of 'basic systematics' or 'alpha taxonomy'), and in any case situations are common in which the SGRP concept becomes less clear (e.g. when there is sympatry in a limited portion of the range). However it is striking that a wide range of taxonomic treatments are applied above to biological situations sharing the same essential features: (i) the SGRP is represented by a single population only in any one locality, and (ii) there is some degree of geographic phenotypic variation within the SGRP. In the vast majority of cases the available data on each SGRP are limited to museum collections and notes from field observations, both of which are distributed

sparsely throughout the area. Geographic distributions, and the extent of sympatry, are therefore known imprecisely. An understanding of the biota from an evolutionary point of view requires statistical inferences to be made about current levels of gene flow across space and about the histories of these structured populations. It is not possible to make such inferences adequately from museum specimens and field observations, and molecular genetic approaches are the obvious way forward. Although the diversity of known forms would be incomprehensible without previous taxonomic work, it is clear that the required population genetic work should be minimally influenced by preconceptions about gene flow derived from current notions of taxonomic rank.

The problem of low-level taxonomic arbitrariness is particularly acute in Amazonia as a result of the prevalence of geographic variation. At least three reasons may be put forward for this: (i) large area: similar habitat extends more-or-less continuously over very large areas and thus populations may have very large ranges, permitting the maintenance of geographic variation in organisms of relatively low vagility, (ii) habitat subdivision: the known distributions of phenotypic variants frequently seem to be coincident with major rivers, giving rise to the hypothesis that the rivers are involved in the origin and/or maintenance of the phenotypic diversity (e.g. Wallace 1852, Snethlage 1913, Sick 1967, Capparella 1987, but see Haffer 1997b) and (iii) temporal stability at large spatial scales: unlike, for example, areas that were glaciated during the Pleistocene, it seems that extinction and recolonisation have operated at relatively small spatial scales over the relevant time period, allowing pronounced geographical variation to arise and persist in many lineages independently.

This geographic variation in heritable components of the phenotype immediately raises a collection of related and important questions. Limited gene flow between

phenotypically differentiated groups is clearly implied: in cases where the variation appears to be discontinuous, do the groupings defined by the phenotypic labels in fact represent completely isolated populations? If not, non-neutral processes may be at work maintaining the phenotypic differentiation whose elucidation would be broadly of interest within evolutionary biology and ecology. Even in cases in which the phenotypic variation is less discrete, any inconsistency with the 'background' pattern of neutral variation would suggest non-neutrality of the phenotypic traits. However, our interest is not limited to variation in phenotypic traits that is easily perceived by humans. For every SGRP we would like to identify subsets of populations within which gene flow is substantial and between which it is low or currently nil, whether apparently phenotypically differentiated or not. Furthermore we would like to consider the biota as a whole, and ask whether historical events can be identified that have affected multiple independent lineages simultaneously, and whether generalisations can be made about the role of geological and landscape history in Amazonia on biotic diversification.

These sorts of questions suggest research programs in Amazonian ornithology and beyond, in which empirical surveys of neutral population genetic variation play a central role by permitting inferences to be made about the contemporary structure and evolutionary histories of SGRPs. In the next section I discuss some key issues in this research program. This discussion provides the context for the data sets discussed in the remainder of this chapter, and for chapters 2 and 3 of this dissertation, which investigate some of the population genetic issues in greater depth.

1.1.1 Population genetic study of SGRPs

If population A1 in locality A is considered to be the geographical representative of population B1 in locality B, when in fact population B2 is much more closely related to A1, the results will be very misleading. Any population genetic investigation of geographic variation is preceded by a phase in which the proper comparisons are established, or at least sensible hypotheses regarding them are formed. This work is based on museum specimens and field observations, and traditionally involves the formal assignment of species and subspecies epithets to populations in particular areas. (In phenotypically very obscure organisms such as bacteria, this work is instead based on population genetics). This is sometimes referred to as 'alpha taxonomy' or 'basic systematics', and it is both essential to and frequently neglected by other areas of modern evolutionary biology. Ornithology is fortunate in that birds are relatively easy to observe, and a large amount of good work in basic systematics has been performed during the last 200 years. Therefore even in Amazonia, although new populations are still being discovered, the sort of mistake outlined above is unlikely to be made.

Flying organisms have relatively strong dispersal capabilities. However, the Amazonian birds studied in this chapter are non-migratory inhabitants of the lower strata of upland ('terra firme') forest. They avoid other types of vegetation and infrequently cross gaps in the forest. These sorts of properties of the organismal biology result in a decrease in genetic similarity with distance which is often referred to as isolation by distance (IBD). This is a form of population structure. Population structure means that the organisms do not exist as a single panmictic population; rather, the individuals can be labelled in some way such that individuals with the same (or similar) labels

are on average more closely related than individuals with different (or dissimilar) labels. In the case of IBD the labelling of the individuals reflects their geographical location. Population structure with that property is referred to as geographical population structure (or simply 'geographical' or 'spatial' structure). Chapters 2 and 3 concern simpler types of population structure in which there are two panmictic populations, connected by past migration (chapter 2) or common ancestry from a single panmictic population (chapter 3). In those cases the population structure might not be geographic — for example the 'populations' could be social groups in the same spatial location. In contrast, IBD refers to distances in space. By concentrating different genetic variants in different regions of space, IBD permits the entire population to maintain higher levels of genetic diversity than it would if it were panmictic. Some of this genetic variation may be functional, and therefore IBD may affect phenotypic variation and play an important role in adaptive evolution.

It is simplest to consider the effects of IBD on unlinked genetic variation (see section 3.3). Focus on a particular locus with two alleles, and let X be the population frequency of one of the alleles (the frequency of the other is $1 - X$). One reasonable mathematical model of the situation supposes that X varies continuously in space, so that the population allele frequency at a point s is $X(s)$. The data are finite samples of alleles and so underlying population frequencies like $X(s)$ are unknown and are random variables in the model. Under IBD the frequencies at points s and s' are expected to be similar if s and s' are nearby: that is, the covariance of $X(s)$ and $X(s')$ decreases with increasing separation of s and s'. The allele frequency is then said to be 'spatially autocorrelated'.

Linked genetic data potentially contain much more information about the genealogy in the sampled region of the genome. IBD results in positive correlations between pairwise coalescence times (and therefore numbers of nucleotide differences between haplotypes) and spatial separation. However, as a result of the genealogy, pairwise differences are not all independent and there is additional information about population structure in the genealogical relationships of the sampled chromosomes. It is straightforward to simulate linked data under models of a collection of discrete populations which mimic spatial structure by permitting gene flow between nearby populations only. However the stochastic processes that, conditional on the spatial location of samples, generate genealogies under models of continuous space are quite complicated and are an area of active theoretical research (Barton & Wilson 1996, Wilkins 2004) that has not yet resulted in useful procedures for statistical inference.

In an ideal model of IBD, genetic similarity is expected to vary continuously in space (note that I am talking about the underlying population variation; not about variation in a sample). In contrast, the basic systematic survey of the Earth's biota is primarily concerned with identifying cases of discontinuous variation, in particular those that reflect complete reproductive isolation. If two samples are taken from two different locations in a large population evolving under pure IBD, those samples will be genetically differentiated to some extent, and that may be mistaken for evidence of discontinuous spatial variation. Therefore developing statistical methods to identify spatial heterogeneities in gene flow is an important challenge in the integration of population genetics with biogeography and systematics.

Further objectives of this research program include accounting for patterns of phenotypic diversity and assessing their functional relevance, understanding the mecha-

nisms that initiate and maintain reproductive isolation between parapatric populations, understanding the mechanisms that prevent or permit sympatric coexistence of geographical representatives, identifying sets of codistributed lineages that appear to have been simultaneously affected by particular historical biogeographic events, and developing a coherent approach to low-level taxonomy. The focus here is on birds. Although new low-level bird taxa are described every year from Amazonia, these generally belong to one or other known SGRP, and it is fair to say that no taxon at a comparable level is as well understood from a systematic and biogeographic perspective in Amazonia as birds. The research program outlined above must be complemented by intensive field surveys and research in basic systematics, and this is particularly true for less well-known taxa.

Ultimately, this research will permit a low-level taxonomic system that reflects the new information about population history. Despite evolutionary biologists' awareness that the complexities of recent evolution cannot be adequately summarised by a simple taxonomic system, some attempt to do so is nevertheless essential for efficient communication, most importantly for conservation purposes. In principle therefore the population genetic research program will lead to a better basis for conservation decision-making. Unfortunately the process of habitat destruction is occurring at a very high rate relative to the process of transformation of technical research findings into conservation action, and it is worth emphasising that no research in evolutionary biology whatsoever is required to justify immediate habitat protection.

1.1.2 The SGRPs studied in this chapter

Over the last two decades, this research program has been started, and the first surveys of population genetic variation in Amazonian bird SGRPs have been com-

pleted (Capparella 1987, 1988, Hackett & Rosenberg 1990, Hackett 1993, Hackett & Lehn 1997, Bates *et al.* 1999, Bates 2000, Marks *et al.* 2002, Bates 2002, Bates *et al.* 2004, Aleixo 2004, Armenta *et al.* 2005, Cheviron *et al.* 2005). Although genetic differentiation may be weak or absent among birds with high dispersal capabilities such as the ara caris (small toucans in the genus *Pteroglossus*) studied by Hackett & Lehn (1997), the overall picture for birds of the forest interior is of strong genetic differentiation between populations in different regions which is frequently coincident with phenotypic variation. However, these studies have three major limitations:

1. Samples are distributed sparsely within the SGRP with unsampled areas of tens of thousands of square kilometres.

2. Sample sizes per location are small — frequently between 1 and 5.

3. Apart from some early surveys of electrophoretic variation in allozymes, all the studies are of mitochondrial DNA. This issue is discussed at length in section 3.3 and in chapter 2; briefly, the concern is that a single genealogy underlies all the sites and thus the data are highly correlated.

4. Model-based statistical procedures for making the inferences that are of interest from such linked genetic data are unavailable.

I discuss these limitations at greater length in section 1.4. The data presented in this chapter are an improvement with respect to limitation 1, but are still subject to limitations 2, 3 and 4.

Representative populations of the antbirds *M. myotherinus* and *H. poecilinota* occupy *terra firme* forest throughout the Amazonian lowlands and adjacent Andean foothills, although interestingly *M. myotherinus* is absent from the Guianan region

(north of the Amazon and east of the Rio Negro). Birds of both SGRPs forage on or close to the forest floor and are highly territorial. The small woodcreeper *G. spirurus* occupies the same Amazonian forests, and also extends north along Andean valleys in Colombia, and occurs west of the Andes in the Chocó and throughout Central America to Mexico, and in forests along the Atlantic coast of Brazil. It uses its elongated and strengthened tail feather shafts to forage vertically on tree trunks (like other woodcreepers), and ranges more widely in the forest, frequently following mixed groups of understory passerines.

M. myotherinus and *H. poecilinota* show striking geographic variation in plumage and, at least in *M. myotherinus*, in vocalisations also. This geographic variation is reflected in Peters (1934-1987) by the recognition of 8 and 7 subspecies taxa respectively. Additionally, birds that in plumage resemble a dark *M. myotherinus* occur in a restricted area of Amazonian Peru and are recognised as a different species taxon, *M. melanurus*. Whereas *M. myotherinus* tends to occupy *terra firme*, these occupy seasonally flooded (*várzea*) forest and their vocalisations differ. However, in contrast to the congruence between plumage types and reciprocally monophyletic mitochondrial clades that is typically observed in Amazonian suboscines, initial results indicate sharing of mitochondrial haplotypes between *M. melanurus* and nearby *M. myotherinus* (D. Davison unpublished results). Geographic variation in *M. myotherinus* is discussed by Haffer & Fitzpatrick (1985), and in *H. poecilinota* by Isler *et al.* (in prep.). Peters (1934-1987) recognises a total of 10 subspecies of *G. spirurus*, of which 6 are Amazonian. The phenotypic distinctions between these subspecies are however much more subtle than is true of most of the *M. myotherinus* and *H. poecilinota* subspecies.

1.1.3 The study area

Achieving an understanding of the evolutionary basis of these continent-wide patterns of phenotypic variation will require many logistically-challenging field studies and collections of specimens and tissue samples, and innovative assembly and analysis of large population genetic data sets. In this chapter I make a start by focusing on geographic variation among the geographical representative populations of these three taxa in a limited portion of their range.

The study area (figure 1.1) is the drainage basin of the Rio Tapajós, and its two major tributaries — the Rios Teles Pires and Juruena. The two tributaries receive waters from the *cerrado* (savannah) -covered Central Brazilian highlands and descend northwards into Amazonian forest in fairly deeply cut and stable channels, without meandering and the associated ox-bow lake formation. They unite $\sim$ 600 km north of the southern extent of humid forest, turn north-east, and flow the remaining $\sim$ 600 km to the Amazon as the Rio Tapajós. At collecting site TPLN the Teles Pires is approximately ?300 m wide, and at collecting site JL the Juruena is a similar width. The Tapajós is around ?1 km wide for much of its course but widens substantially over the last 200 km, and is 4-5 km wide at its mouth into the Amazon.

The Tapajós- Teles Pires axis represents a zone of substantial biogeographic discontinuity and there are many examples of Amazonian bird SGRPs with phenotypically differentiated subgroups on either side of the lower Rio Tapajós and/or the Teles Pires (see Haffer 1997b). The biogeographic discontinuity is also evidenced by SGRPs such as *Thamnomanes saturninus / ardesiacus*, which has no right bank representative. The biogeographic situation is more complicated than delimitation of distributions by a major river because, as Haffer (1997b) emphasises, in many SGRPs (including

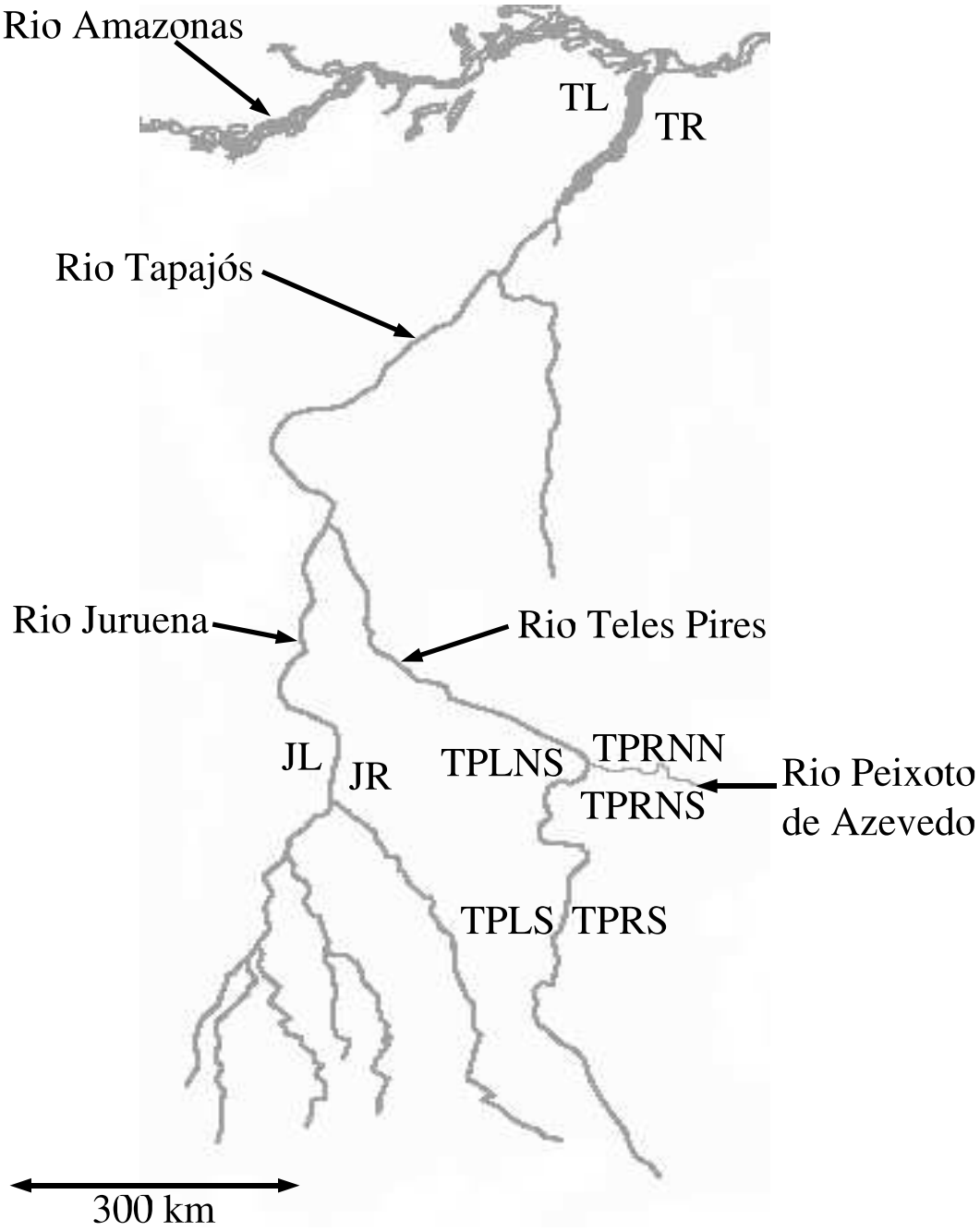


Figure 1.1: **Map of the study area showing collection areas**

H. poecilinota and *M. myotherinus*) the right bank representative is found on the left bank of the lower Tapajós, and there are consequently apparent zones of contact in continuous forest in the Madeira - Tapajós interfluvium. The Juruena seems to represent less of a biogeographic discontinuity, although in the *Rhegmatorhina* SGRP the species taxa *hoffmansi* and *gymnops* do appear to be separated by the Juruena at the sites visited (pers. obs.).

1.2 Methods

The results in this chapter concern the geographical distributions of plumage types, and sequence variation in the mitochondrial ND2 gene in the study area. These results are based on my own field observations, and collections of museum specimens and tissue samples, made during June-August 2002 and 2003, supplemented by information from specimens in the Museu Paraense Emílio Goeldi (MPEG) in Belém, Brazil, tissue samples from several collections in the U.S.A. and mitochondrial sequences obtained previously by Marks *et al.* (2002) (*G. spirurus*) and Bates *et al.* (2004) (*H. poecilinota*).

At each field site I used mist nets to capture understory birds indiscriminately. From each bird I removed a growing ('pin') feather from the wing or tail and stored the growing tip (which contains much live tissue) in an EDTA-based buffer at ambient temperature, thus creating a collection of tissue samples from understory birds of quite wide taxonomic scope. I subsequently released these birds. At each site, I also collected up to five birds of each of the three focal study taxa and prepared them in the field as museum study skins. These were obtained from the mist nets or with a shotgun. One advantage of the latter method is that collection of several closely

related birds within a single territory can be avoided without time-consuming relocation of mist nets. These specimens are stored in the ornithological collections of the MPEG. Field collections were performed under permits number 02001.004116/2002-11 and 02001.002036/2003-2, and tissue samples were exported under export licence number 0112651, issued by Brazilian federal government agencies IBAMA and CGEN. All tissue samples collected are stored in the tissue collections of the MPEG, and of the Universidade Federal do Pará, the latter under charge of Dr Wilsea Figueiredo. Subsamples of the tissue samples used for the work described here are also stored in the tissue collections of the Field Museum of Natural History, Chicago. PCR-based amplification and sequencing of the amplification product followed standard protocols which are not described here. Analyses of the alignments were performed in R (R Development Core Team 2006).

1.3 Results

1.3.1 Mitochondrial sequence divergence across the lower Tapajós

Table 1.2 describes mitochondrial genetic differentiation and sequence divergence between the left and right banks of the lower Rio Tapajós in seven different SGRPs. The data here are all 347 and 117 codons of the mitochondrial ND2 and ND3 genes respectively, in small samples collected on both banks. These are a random sample of Amazonian bird SGRPs from neither phylogenetic nor ecological/life history perspectives. All are suboscine passerines: two woodcreepers (*Dendrocincla merula* and *Glyphorynchus spirurus*), four thamnophilid antbirds (*Hylophylax poecilinota*, *Hypocnemis cantator*, *Myrmoborus myotherinus*, *Myrmotherula leucophthalma*) and *Schiffornis turdinus* (related to certain 'cotingas'). And all inhabit the understory of

terra firme forest and are therefore expected to have low rates of dispersal across the $\sim$ 4 km wide lower Tapajós.

These data suggest that in these SGRPs, mitochondrial genomes in left and right bank populations are reciprocally monophyletic, and that sequence divergence between these clades varies between about 1% and 5%. Due to the small sample sizes and the fact that the samples were collected from only a single location on each bank (sites TL and TR), it is possible that the 'left bank' sequences do in fact occur on the right bank, and *vice versa*. However the concordance of the results across the seven independent SGRPs, together with the existence of concordant plumage (and vocalisation) differences in some of them, argue against that possibility.

SGRP	ND2						ND3					
	n	bp	S	1^{st}	2^{nd}	3^{rd}	n	bp	S	1^{st}	2^{nd}	3^{rd}
Myrmoborus myotherinus	35	1035	50	0.22	0.04	0.73	0	-	-	-	-	-
Hylophylax poecilinota	39	570	50	0.16	0.06	0.78	0	-	-	-	-	-
Glyphorynchus spirurus	36	336	36	0.18	0.11	0.71	0	-	-	-	-	-
Dendrocincla merula	4	1041	10	0.10	0.00	0.90	4	351	6	0.17	0.00	0.83
Hypocnemis cantator	5	1041	61	0.18	0.13	0.69	13	351	16	0.25	0.06	0.69
Myrmotherula leucophthalma	9	1041	17	0.18	0.12	0.71	9	351	3	0.00	0.00	1.00
Schiffornis turdinus	6	1041	27	0.22	0.11	0.67	6	351	11	0.18	0.09	0.73

Table 1.1: **Summary of all ND2 and ND3 alignments**
Columns headed 1^{st}, 2^{nd} and 3^{rd} give the proportion of polymorphic sites at the respective codon position.

SGRP	plumage differences	ND2 n	ND2 pd	ND2 K_{st}	ND3 K_{st}	ND3 pd	ND3 n
Dendrocincla merula	no	(1, 3)	0.9	-	-	1.5	(1, 3)
Glyphorynchus spirurus	no	(9, 4)	3.2	.84	-	-	-
Hylophylax poecilinota	no	(14, 3)	2.3	.85	-	-	-
Hypocnemis cantator	yes	(4, 1)	5.6	-	.90	4.2	(10, 3)
Myrmoborus myotherinus	no	(14, 2)	0.8	.66	-	-	-
Myrmotherula leucophthalma	yes	(7, 2)	1.5	.88	1.00	0.9	(5, 4)
Schiffornis turdinus	no	(3, 3)	2.1	.85	.96	2.9	(3, 3)

Table 1.2: **Across-river comparisons**

K_{st} is the statistic of Hudson *et al.* (1992). It is analogous to F_{st}, ranging between 0 (no differentiation) and 1 (complete differentiation). When calculating K_{st}, the average within-group pairwise distances were computed using weights proportional to the sample sizes. pd is the average percent nucleotide difference for all comparisons of one sequence from the left bank and one from the right bank. n is the number of samples (left, right) sequenced for each gene. Missing values result from absence of ND3 sequence (*G. spirurus*, *H. poecilinota*, *M. myotherinus*) and samples of size one on either bank, in which case K_{st} is undefined.

1.3.2 Mitochondrial sequence variation throughout the study area

Table 1.2 compared mitochondrial ND2 and ND3 sequences across the lower Tapajós for seven SGRPs. For three of these (the 'focal' SGRPs *M. myotherinus*, *H. poecilinota* and *G. spirurus*), I have obtained ND2 sequences from samples collected throughout the study area at the collecting sites marked on the map (figure 1.1). I sequenced all 347 codons of the ND2 gene, but for *H. poecilinota* and *G. spirurus* I report only on variation at the sites that were also sequenced by Bates *et al.* (2004) and Marks *et al.* (2002) respectively. Summary statistics describing these alignments, including Watterson's (1975) estimator of $4N_eu$ (θ_w) and the average number of pairwise nucleotide differences (π) (both per base pair), are given in table 1.3. Under panmixia both θ_w and π have expectation equal to $4N_eu$. Note that for all three SGRPs (when not broken down by plumage type) $\pi > \theta_w$ (i.e. Tajima's (1989) D is positive). This indicates an excess of sites with mutations at intermediate frequencies over the panmictic expectation, and in this case is caused by the geographic structure in these data. For the plumage types *M. m. sororius* *H. p. griseiventris*, $\pi \approx \theta_w$ suggesting little population structure in the samples studied here. However for the plumage types *M. m. ochrolaema* and *H. p. nigrigula* $\pi > \theta_w$, suggesting that there is structure even within these plumage types, perhaps unsurprisingly as both are found on both banks of the wide lower Tapajós.

Under the assumption that the sequences do not recombine, all sites share the same genealogy. And if at every variable site the variation is the result of a single mutation, and if the polarity of the nucleotide changes is known, then the entire alignment is equivalent to a tree diagram with mutations marked on their corresponding branches, and with multifurcations where there are no mutations to indicate the true

Table 1.3: **Description of alignments of mitochondrial ND2 sequences**

Taxon	n	bp	S	h	θ_w	π
H. poecilinota	39	556	50	17	0.021	0.032
H. p. griseiventris	16	556	7	6	0.004	0.003
H. p. nigrigula	23	556	23	11	0.011	0.013
M. myotherinus	35	1040	50	15	0.012	0.017
M. m. sororius	11	1040	4	5	0.001	0.001
M. m. ochrolaema	24	1040	18	10	0.005	0.006
G. spirurus	36	335	24	11	0.017	0.024

relationships. To investigate the conformity of the data from the three focal SGRPs to these infinite-sites-no-recombination assumptions, for each alignment I included an outgroup sequence, and identified the largest set of sites that could have evolved via unique mutations on a single genealogy, under the assumption that the outgroup sequence bore the ancestral allele at sites that were polymorphic within the ingroup. For *M. myotherinus*, *H. poecilinota* and *G. spirurus* there were respectively 1, 4 and 7 incompatible sites out of a total of 50, 50 and 24 segregating sites. Note that the infinite-sites-no-recombination model seems to fit the data from *M. myotherinus* and *H. poecilinota* significantly better than those from *G. spirurus*. I used the software `genetree` (e.g. Bahlo & Griffiths 2000) to display the alignments as trees, after deletion of these sites (figure 1.2). Construction of these trees from polarised infinite-sites-no-recombination-compatible data is merely an alternative way of displaying the alignments and does not represent an inference.

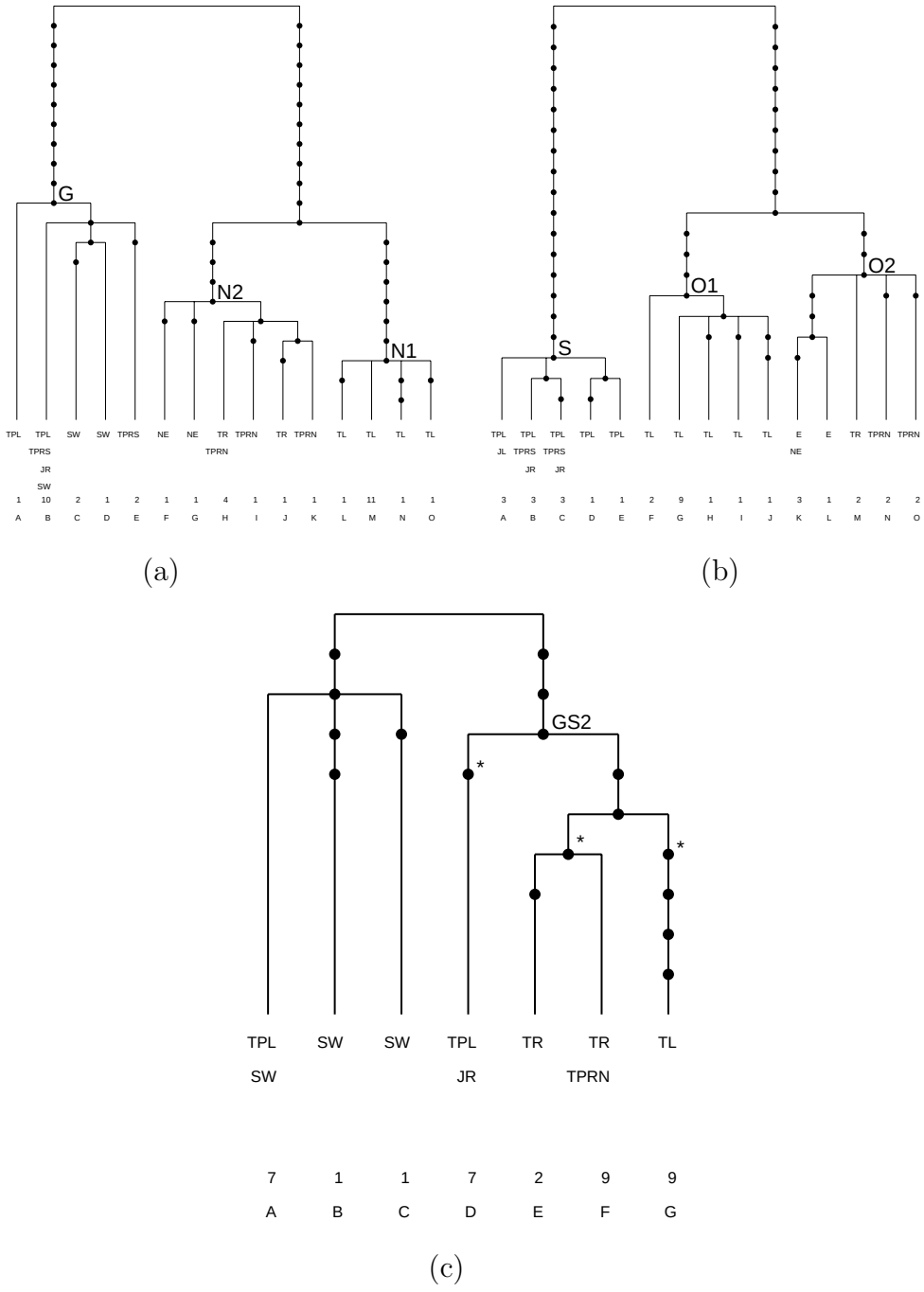


Figure 1.2: **Tree representations of alignments of mitochondrial ND2 sequences**
Sites violating infinite-sites-no-recombination assumptions are excluded. Black circles represent mutations. Clades discussed in the text are marked. (a) *H. poecilinota* (570 bp), (b) *M. myotherinus* (1041 bp), (c) *G. spirurus* (336 bp)

1.3.3 Distributions of plumage types and mitochondrial clades

In this section I describe what is known about the geographical distributions of plumage types in the study SGRPs in the study area. I also describe the distribution of some major clades of mitochondrial DNA. In a sense, this aspect of the genetic data is easy to describe because mutations defining the major mitochondrial clades are perfectly associated with the labelling with respect to plumage types (in *M. myotherinus* and *H. poecilinota* at least). In other words, the data on plumage and mitochondrial DNA for *M. myotherinus* and *H. poecilinota* are consistent with the following two hypotheses.

1. Every bird belonging to the SGRPs *M. myotherinus* and *H. poecilinota* and inhabiting forests in the study area is of one of the plumage types *M. m. ochrolaema*, *M. m. sororius*, *H. p. nigrigula*, *H. p. griseiventris*.

2. For each of these four plumage types there is an associated clade in the genealogy of the mitochondrial genome, and every individual bird of that plumage type bears mitochondria belonging to that clade.

Since this is so, I will label the major clades with the initial letter of the subspecies label with which they appear to be coincident. The mapping between plumage types and mitochondrial clades is given in table 1.4. The existence of these clades in the genealogy of the mitochondrial genome is of course technically an inference about which there is uncertainty. However the data conform closely to infinite-sites-no-recombination assumptions and the clades are supported by several mutations, and in this chapter I do not consider their existence to be in doubt.

Table 1.4 also gives p-values from a permutation-based Mantel test (see Smouse *et al.* 1986) of correlation between matrices of pairwise geographic distances and

numbers of nucleotide differences. In addition to the geographic population structure represented by the major mitochondrial clades and plumage types, the results of the Mantel tests indicate geographic population structure within the widespread O2 and N2 clades in *M. myotherinus* and *H. poecilinota* (respectively *ochrolaema* and *nigrigula* east of the Tapajos - Teles Pires) and within both the major *G. spirurus* mitochondrial clades.

plumage type	mt clade	ND2 haplotypes	n	p
M. m. ochrolaema	O1	F,G,H,I,J	14	.76
	O2	K,L,M,N,O	10	.00
M. m. sororius	S	A,B,C,D,E	11	.95
H. p. nigrigula	N1	L,M,N,O	14	.76
	N2	F,G,H,I,J,K	9	.06
H. p. griseiventris	G	A,B,C,D,E	15	.09
G. spirurus	GS1	A, B, C	9	.00
	GS2	D, E, F,G	27	.00

Table 1.4: **Mapping between major mitochondrial clades and plumage types** n, number of samples; p, p-value of Mantel test for correlation between matrices of pairwise geographic distances and nucleotide differences

M. myotherinus

Known populations of *M. myotherinus* in the study area belong to one of two plumage types which correspond to the subspecies taxa *M. m. ochrolaema* and *M. m. sororius*. (Although I use the subspecies name *sororius*, in their revision of morphological variation in this SGRP, Haffer & Fitzpatrick (1985) recommend assigning *sororius* populations to the nominate subspecies *M. myotherinus myotherinus* which occupies adjacent western and southwestern Amazonia south of the Amazon.) The differences are pronounced in females and much less so in males. (This phenomenon

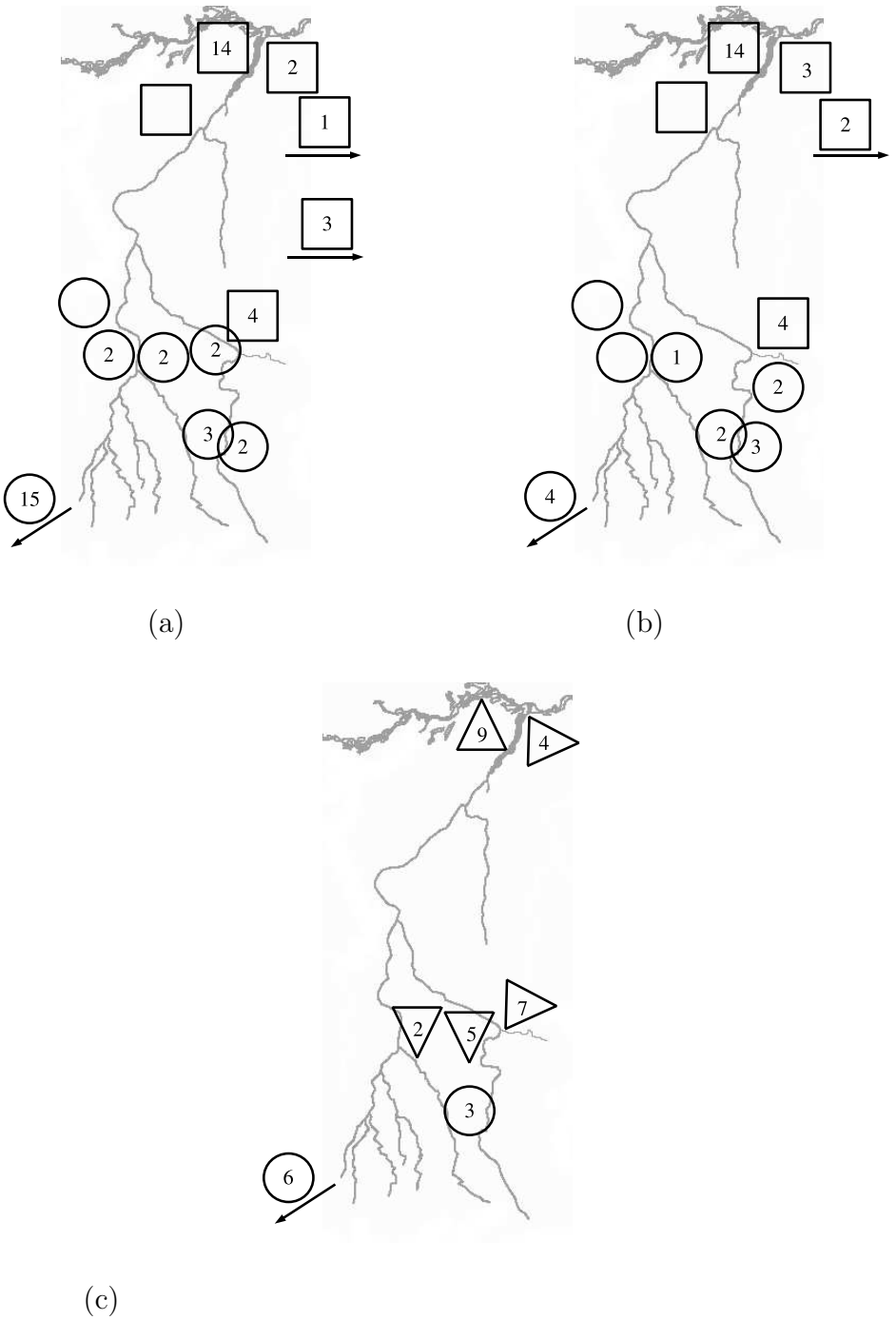


Figure 1.3: **Plumage types and mitochondrial clades at sampling localities.** Numbers inside symbols are numbers of mitochondrial sequences sampled; empty symbols represent museum specimens or observations only. (a) *M. myotherinus ochrolaema* (squares; mt clades O1/O2 left/right of the Tapajós- Teles Pires) and *sororius* (circles; mt clade S), (b) *H. poecilinota nigrigula* (squares; mt clades N1/N2 left/right of the Tapajós- Teles Pires) and *griseiventris* (circles; mt clade G), (c) *G. spirurus* mt clades GS1 (circles) and GS2 (triangles; the three types of triangle represent the three subclades of GS2 indicated with asterisks in figure 1.2)

of greater geographic variation in female plumage, called 'heterogynism' by Hellmayr (1929), is seen in several Amazonian antbirds whereas the converse is not.) In *M. m. ochrolaema* the females are a rich ochraceous colour below and (variably) show a 'necklace' of black spots; *M. m. sororius* is much whiter below and lacks the necklace. Birds on both banks of the lower Tapajós (sites TL and TR) are of the *ochrolaema* form (mt clades O1 and O2 respectively) and are not visually distinguishable. Birds of plumage type *ochrolaema* extend south along the left bank of the Tapajós to around 5° S in the Parque Nacional do Tapajós (MPEG specimens, no genetic data). Further south on the left bank, the next known records are of plumage type *sororius* (MPEG specimens, no genetic data) at around 9° S on the Rio Aripuanã. Birds of type *sororius* with mt clade S occur around 10° S on both banks of the Juruena and on the left bank of the Teles Pires (sites JL, JR, TPLN), whereas at this latitude the right bank of the Teles Pires is occupied by birds of type *ochrolaema* (mt clade O2). Further north in the Tapajós-Xingú interfluvium birds are of type *ochrolaema* where known, and mt clade S extends to the right bank of the Xingú. On the upper Teles Pires at 11.6° S birds of type *sororius* (mt clade S) occupy both banks. Thus a transition between *ochrolaema* (mt clade O2) and *sororius* (mt clade S) occurs on the right bank of the Teles Pires between Rio Cristalino at 9.6° S and the soy farming town of Sinop at 11.6 ° S. Another transition between *ochrolaema* and *sororius* occurs on the left bank of the Tapajós or lower Teles Pires between 5 ° S (*ochrolaema*) and 9° S (*sororius*), although genetic data are available only from around 3° S (mt clade O1) and 10° S (mt clade S).

H. poecilinota

A strikingly similar situation exists in *H. poecilinota*. The SGRP as a whole is very variable in plumage, especially the females. Two plumage types occur in the study area: *H. p. nigrigula* and *H. p. griseiventris*. Differences include that males of *nigrigula* have a black throat whereas it is grey in *griseiventris*, and that females of *nigrigula* have grey cheeks whereas these are rufous in *griseiventris*. Visually indistinguishable birds of the *nigrigula* type occur on both banks of the lower Tapajós, and seem to have reciprocally monophyletic mitochondrial genomes with 2.3% ND2 sequence divergence. *nigrigula* extends south along the left bank at least as far as Parque Nacional do Tapajós at $\sim 5^\circ$ S (MPEG specimens, no genetic data). Further south, the next specimens are of type *griseiventris* again at about 9° S on the Rio Aripuanã (MPEG specimens, no genetic data), and *griseiventris* (mt clade G) occurs around 10° S on both banks of the Juruena and the left bank of the Teles Pires (sites JL, JR, TPLN; specimens but no sequence currently available from JL), whereas, again, at this latitude the right bank of the Teles Pires is occupied by the northeastern form *nigrigula* (mt clade N2). On the upper Teles Pires in the area of Sinop, *griseiventris* (mt clade G) occurs on both banks. In this case however, I was able to locate a zone of contact between the two forms: at 10.1° S along the right bank tributary Rio Peixoto de Azevedo birds of type *griseiventris* only were encountered on the south bank, and of type *nigrigula* only on the north bank. A single site was visited on each bank, at which no more than three individuals were observed, so syntopy of the two forms in this area is not ruled out. The Peixoto de Azevedo is around 50-100 m wide at the point visited (see figure 1.4). Novaes & Lima (1992) also recorded *nigrigula* along Rio Peixoto de Azevedo. Their collecting station (close to the BR-163 at Fazenda São Jose) is on the north bank of the river approximately 40 km to

the east of my north bank site and so the available data, although rather limited, are consistent with the hypothesis that the two plumage types are separated by the Peixoto de Azevedo in this region. Further north in the Tapajós-Xingú interfluvium birds are of type *nigrigula*.

Figure 1.4: **Rio Peixoto de Azevedo at site 16.**

G. spirurus

In this SGRP, Peters (1934-1987) again recognises two subspecies taxa in the study area which are described as being delimited by the Tapajós. However I was unable to perceive differences between these taxa either in the field or in museum specimens, and therefore describe only the distribution of mitochondrial clades. The situation is comparable to those previously described for *M. myotherinus* and *H. poecilinota*: the

deepest subdivision in the tree is between a northern clade (clades 1, 2 and 4) and a southern clade (clade 3), with considerable (3.2%) sequence divergence across the lower Tapajós. However in this case the northern clade occurs on the left bank of the middle Teles Pires and on the Juruena (sites JR and TPLN), whereas the southern clade was found only on the upper Teles Pires near Sinop around 11.6 ° S, as well as at sites outside the study area to the southwest.

1.4 Discussion

I have presented data on variation in mitochondrial ND2 sequences and plumage in the drainage basin of the Rio Tapajós in three suboscine passerine SGRPs that occur throughout the Amazonian lowlands. In order to assess the implications of these data, I start at a broader spatial scale and briefly describe what is known about phenotypic and genetic variation throughout Amazonia in these and some related SGRPs.

1.4.1 Broad scale geographic variation

The focus here is on birds inhabiting the understory of *terra firme* Amazonian forest that belong to the endemic suboscine passerine radiation. Within these birds, there are many examples of SGRPs that extend across much of Amazonia and in which phenotypic differences between birds in different areas have been noted. By SGRP I refer to a system of populations that are considered to be each other's geographical representatives, irrespective of their current taxonomic treatment. Examples of SGRPs therefore include collections of species taxa such as manakins in the genus *Lepidothrix*, as well as collections of subspecies taxa such as those comprising the three focal SGRPs in this study. It is essential to realise that data on geographic

variation in Amazonian SGRPs come from a small number of locations sparsely distributed within very large organismal ranges. Therefore the geographical distribution of any particular phenotype, just as that of any particular allele or genealogical clade, is unknown and can only be inferred from the sparse point data.

Three studies have analysed mtDNA variation in suboscine SGRPs with samples of 60-80 individuals from several locations distributed widely across large areas of Amazonia (Marks *et al.* 2002, Aleixo 2004, Cheviron *et al.* 2005). Marks *et al.* (2002) study variation in one of the focal SGRPs in this chapter (*G. spirurus*), Aleixo (2004) studies a different woodcreeper SGRP in the genus *Xiphorhynchus* (comprising species taxa *spixii* and *elegans*) and Cheviron *et al.* (2005) study variation among western Amazonian members of the *Lepidothrix* manakin SGRP (species taxon *coronata*). In all three cases the mitochondrial data show an extremely high degree of geographical structure. In all cases there is plenty of variation and the authors estimate trees relating the haplotypes. As with any tree, mutually exclusive subtrees ('major clades') may be identified. Much of the literature on geographic population genetic structure concerns cases where such major clades are widely distributed, but at markedly different frequencies in different regions. In contrast, in these three studies the study area may be partitioned into mutually exclusive continuous subregions within which the sampled individuals have mitochondrial genomes belonging to *one major clade only*, which is unique to that subregion. Because of the geographically sparse sampling, many such partitions of space can be drawn on the map that have that property. However, for several major rivers, opposite-bank samples suggest that the geographical ranges of two clades of mitochondrial genomes are bounded by the river at that point. In other areas, birds at sampling locations separated by continuous forest bear mitochondrial genomes belonging to different major clades. But at no single

location are mitochondrial genomes belonging to different major clades detected. In the *Xiphorhynchus* SGRP, labelling the samples with respect to major mitochondrial clade generally identifies the same groups as labelling the samples with respect to the plumage variation which is the basis for species- and subspecies-level taxonomy, thus indicating strong linkage disequilibrium between variation in the mitochondrial genome and that in at least one region of the nuclear genome. However this is not always the case in *Lepidothrix coronata* and *G. spirurus*.

The mitochondrial data described here also conform to the above description, which is emerging as a general result for forest understory suboscines in Amazonia. Specifically, the results described here agree with the findings of Marks *et al.* (2002), Aleixo (2004) and Cheviron *et al.* (2005) in the following respects.

1. In no case did I detect mitochondrial genomes belonging to more than one of the major clades defined in table 1.4 at a single location.

2. Continuous partitions of the study area may be drawn within which all samples belong to a single, unique, mitochondrial clade.

3. Major clades are found to be apparently delimited by the lower Tapajós and middle Teles Pires, and are also found at locations that are separated by continuous forest on the left bank of the Tapajós.

4. In *M. myotherinus* and *H. poecilinota*, the partition of the sample defined by the major clades O/S and N/G is the same as the partition of the sample defined by the plumage types *ochrolaema*/*sororius* and *nigrigula*/*griseiventris*.

The finding of a one-to-one mapping between mitochondrial clades and subregions of the study area has prompted many authors to treat the clades in the genealogy of

the mitochondrial genome as if they corresponded to isolated lineages of sexual organisms. The relation between the mitochondrial genealogy and the geography in these cases is indeed striking. However it is necessary to bear in mind the possibility that some of the groupings of individuals identified by the mitochondrial genealogy have no such significance, and that genealogies at other loci (or an imaginary repetition of the mitochondrial genealogical process) would identify other groupings. Despite frequently contrary appearances there is much awareness of this in the 'phylogeographic' literature (e.g. Templeton *et al.* 1995, Irwin 2002, Kuo & Avise 2005).

On the basis of the known patterns of mitochondrial variation, some models for the recent evolutionary history of these organisms may be ruled out. For example it is inconceivable that these SGRPs are in a transient phase resulting from a single recent colonisation of the Amazon basin from a single source population, because such a process would surely have resulted in a pronounced spatial trend in levels of diversity, and would surely not have resulted in spatially replacing clades of mitochondrial genomes with such high levels of sequence divergence. However, there is a large and diverse collection of plausible models. In order to make the discussion possible, I will imagine the possibilities to be arrayed between two extremes. At one extreme, each SGRP comprises a mosaic of spatially contiguous populations, each of which has been completely isolated from all other such populations for an amount of time that is substantial on the time scale defined by its effective size — i.e. a lot of drift has occurred in each isolated population during this period. I will refer to this model as a mosaic of isolated populations (MIP). In this case, when samples are taken from the SGRP as a whole, the population structure will result in very strong linkage disequilibrium (LD), and genealogies at unlinked loci may in fact contain monophyletic subtrees corresponding to the samples from each isolated mosaic component. In the

MIP model, the isolated populations need not have evolved *in situ* but may have, for example, originated in forest refugia during a drier period and subsequently expanded to occupy their current range. The analyses of Marks *et al.* (2002), who describe their mitochondrial genealogy as a "phylogeny of *Glyphorynchus spirurus* populations", are predicated on the assumption that the evolutionary history of the *G. spirurus* SGRP lies towards this end of the spectrum.

At the other extreme is a model of pure isolation-by-distance (IBD) at equilibrium. In that case spatially structured genealogies will arise, and depending on the spatial configuration of samples there may even be a one-to-one mapping between genealogical clades and subregions (Barton & Wilson 1996, Irwin 2002). However, such phenomena tell us nothing about reproductive isolation of populations inhabiting different subregions, because no such isolation exists. Given what is known about genetic and phenotypic variation, the IBD model is clearly inaccurate at the largest spatial scales in forest understory Amazonian suboscines. However, the biology of these birds suggests that isolation-by-distance will always occur to some extent, and in between the two extremes lies an enormous variety of models in which various local extinctions and range expansions may have occurred and in which there may be partially isolated populations exchanging genes at a lower-than-usual rates with neighbouring populations, and recently isolated populations in which weak drift has less obviously revealed their isolation.

Elucidating the details of these evolutionary histories is a fascinating and important project for evolutionary biology. It is also an extremely ambitious one. The sampling of individuals, space and genomic loci in this study dictates that its aims must be simpler. Even in birds, the initial surveying and cataloguing of extant forms is incom-

plete, and I suggest that the current priority in the study of widely distributed SGRPs is identification of reproductively isolated subgroups, and in general characterisation of the extent to which the MIP model is accurate. When a handful of mitochondrial sequences are available at two distant locations reciprocal monophyly may result, even at multiple unlinked loci, under either the MIP or IBD model (Barton & Wilson 1996, Irwin 2002). However, under the MIP model some alleles or haplotype clades would be discovered to have spatially coincident boundaries if their true spatial distributions were known, whereas this is not expected under IBD. Therefore in order to distinguish between these models sampling should ideally be dense in the vicinity of putative contact zones.

1.4.2 Fine scale geographic variation

This study has extended the emerging understanding of geographical variation in plumage and mtDNA in Amazonian suboscines by sampling at a finer spatial scale than previous studies. The data sets of Marks *et al.* (2002), Aleixo (2004) and Cheviron *et al.* (2005) do indicate locations in which birds of different plumage types, bearing mitochondria belonging to different clades, are found in close proximity. These are invariably samples located on opposite banks of wide rivers, such as the Rio Amazonas in eastern Peru near the confluence of the Rio Napo, or the Rio Teles Pires near Alta Floresta (areas TPRNN and TPLN). Bates *et al.* (2004) have further investigated levels of mitochondrial sequence divergence between small samples of forest understory birds on opposite banks of the Teles Pires at this point and found greater than 2% divergence across the river in 5 out of the 8 suboscines studied (which include *H. poecilinota* and *G. spirurus*). The river in this area is 100-300 m wide and, since there is no evidence for historical alterations in the course of

the Teles Pires, it is possible that it has been an effective barrier to dispersal of forest interior birds for several million years. Similarly, in the western Amazon basin the Rio Amazonas is a wide zone of shifting open water channels and islands with seasonally flooded forests and may have been an effective barrier to the dispersal of *terra firme* inhabitants over a comparable time scale. It is therefore conceivable that the observed mitochondrial and phenotypic differentiation, and the apparent lack of spatial overlap between mitochondrial clades and plumage types, is the result of a low flux of individuals across these zones, rather than of intrinsic pre- or post-mating barriers to gene flow coupled with ecological incompatibilities preventing syntopy.

In contrast, the discovery of a zone of parapatry between phenotypically distinct populations bearing distantly related mitochondrial genomes without an intervening barrier to dispersal would be good evidence for an intrinsic barrier to gene flow. Specimen data indicate that many such zones of contact between phenotypically differentiated subsets of SGRPs occur in continuous forest within Amazonian interfluvia (e.g. Haffer 1997b). Furthermore, the surveys of mtDNA variation are consistent with parapatric contact between geographically replacing lineages of mitochondrial genomes in the same areas. The several examples in the Madeira-Tapajós interfluvium include the transitions between *H. p. nigrigula* (mt clade N1) and *H. p. griseiventris* (mt clade G), and between *M. m. ochrolaema* (mt clade O1) and *M. m. sororius* (mt clade S). However the transect sampling that would be required to establish the spatial limits of the plumage types and mitochondrial clades in these areas has not yet been performed, and the lack of roads, and rapids on the Tapajós-Teles Pires, mean that the field work would be time-consuming and expensive. To my knowledge, the contact between *H. p. nigrigula* (mt clade N2) and *H. p. griseiventris* (mt clade G) along the 50-100 m wide Rio Peixoto de Azevedo represents the closest Amazonian

example to parapatry without a barrier to gene flow. In that case it seems implausible to me that the river prevents frequent dispersal of individuals between banks and, if confirmed, lack of syntopy of plumage types and mitochondrial clades would indicate intrinsic barriers to gene flow with insufficient ecological differentiation for coexistence.

Where spatial sampling is sufficient, the Mantel tests indicate isolation by distance within the subsets of populations defined by the plumage types and major mitochondrial clades. Isolation by distance is also suggested by inspection of figure 1.2 in that ancestral haplotypes such as *H. poecilinota* haplotypes B and H, *M. myotherinus* haplotype A and *G. spirurus* haplotypes A and F are more widespread than the haplotypes which appear to be their descendents. This pattern is also clearly evident in the data from *Xiphorhynchus* woodcreepers in Aleixo (2004, fig. 2B). However because sample sizes are so small, an alternative explanation is that the spatial restriction of the descendent haplotypes is illusory and that the ancestral haplotypes were sampled at several locations because they are at higher frequencies. The strong positive correlation between sites in alignments of mitochondrial sequences make them unsuitable for inference of fine scale geographic population structure and it will be exciting to further investigate population structure in Amazonian organisms with data sets of autosomal variation, ideally in conjunction with improved spatial sampling. Such data sets could be of microsatellites or AFLPs (Amplified Fragment Length Polymorphisms), or could comprise large amounts of sequence data from several regions of the nuclear genome.

1.4.3 Similarity of patterns across independent SGRPs

The similarity of the results for the three SGRPs studied is striking. In *H. poecilinota* and *M. myotherinus* the deepest split in the mitochondrial genealogy is into northern and southern clades which are perfectly associated with plumage variation. The deepest split in the northern clade is into a clade found on the left bank of the lower Tapajós (southern limit unknown), and a clade found on the right bank of the Tapajós - Teles Pires, east to the right bank of the Xingú and south to the north bank of the Peixoto de Azevedo. The southern clade is found south of the Peixoto de Azevedo (contact possible but unconfirmed in *M. myotherinus*), on both banks of the Juruena (presumably so, based on plumage, in *M. myotherinus*) and extends far into southwestern and western Amazonia in Rondônia, Bolivia and Peru south of the Amazon (including one individual of *M. melanurus* in *M. myotherinus*). In *G. spiurus* the deepest split is between southern and northern clades, but the results differ in that the northern clade is found between the Teles Pires and the Juruena. In all three cases, as in the other suboscines *D. merula*, *M. leucophthalma*, *H. cantator* and *S. turdinus*, there is strong differentiation across the lower Tapajós, and in the three focal SGRPs it is clear that the mitochondrial genomes on the left bank of the lower Tapajós are most closely related to those on the right bank rather to those further south in the Madeira - Tapajós interfluvium. There is no evidence for differentiation across the Juruena in these SGRPs, although species taxa in the *Rhegmatorhina* SGRP do seem to be separated by the Juruena at the sites visited.

If, as is often assumed, the mutation rate in these mitochondrial genes is between 0.01 and 0.02 per million years then the oldest portions of these genealogies are Pliocene in age. It is likely that the organismal ranges have changed over that time

and that the oldest portions of the genealogy reflect such changes (Barton & Wilson 1996). Under certain models of equilibrium population structure, the most ancient portion of the genealogical process behaves like a time-rescaled panmictic coalescent (Wakeley 1999, Wilkins 2004). The spatial-topological similarities of the oldest portions of the genealogies observed across independent SGRPs are inconsistent with such models and suggest that the lineages have been affected in similar ways by historical biogeographic events.

There is a fairly large literature on possible processes of biological diversification in Amazonia (reviewed by Haffer 1997a). One popular hypothesis is that allopatric differentiation of forest organisms occurred during periods when the forest contracted into refugia during the Pleistocene or before (Haffer 1969). If that were so, then the zones of contact observed between the northern and southern lineages would represent secondary contact. An alternative hypothesis is that the development of major rivers and their floodplains caused vicariance, and that current zones of contact along rivers represent primary contact. It is commonly asserted in the phylogeographic literature (e.g. Moritz *et al.* 2000) that such alternative models make simple predictions for the topologies of mitochondrial genealogies. However the genealogical process in spatially structured populations depends on the details of recent population structure, population history and the locations of samples in a complex way, and convincing inferences will be based on data from several SGRPs and many independent loci with good spatial sampling and will require novel statistical techniques at the intersection of population genetics, biogeography and systematics.

While considering the pattern of sister clades of mitochondrial genomes across the lower Tapajós, it is worth mentioning the intriguing hypothesis of Willis (1969) that

the Madeira and Tapajós rivers previously joined the Amazon to the west of their current confluences. I found that mitochondrial genomes on the left bank of the lower Tapajós were most closely related to those on the right bank, and the plumage similarity of left and right bank populations with putative parapatry of populations with distinct plumage in continuous forest in the Madeira - Tapajós interfluvium is seen not only in *H. poecilinota* and *M. myotherinus*, but in many other SGRPs (Haffer 1997b). These findings are consistent with the change in course of the lower portion of the Tapajós proposed by Willis (1969), but are of course also consistent with other biogeographic histories.

1.4.4 Future research

Although the work described here is an improvement with respect to fine scale spatial sampling, the sparse spatial distribution of field observations and specimens, and small sample sizes available for population genetic analyses leave many questions unanswered about the geographical distribution of phenotypic and genetic variation in these SGRPs and the implications for their evolutionary history. In particular, the data are consistent with various parapatric contacts between populations with distinct plumage and bearing distantly related mitochondrial genomes, but these hypotheses were not confirmed. Ideally one would like to establish whether reproductively isolated populations indeed meet in these areas, whether there is evidence for recent gene flow between them, whether the contacts are primary or secondary in nature, whether there is syntopy of the different forms, and whether there are ecological differences between them that may facilitate coexistence.

In the SGRPs studied here, and in others in which phenotypically distinct populations approach each other in the Madeira - Tapajós interfluvium, it would be of great

interest to sample along a roughly northeast - southwest transect in that interfluvium (or, more feasibly, along the left bank of the Tapajós). Additionally, since this study has identified the Rio Peixoto de Azevedo as separating the two forms of *H. poecilinota* (and is consistent with the same being true in *M. myotherinus*), the upper stretches of that river are a possible location for future fine-scale studies of the population genetic and ecological basis of the apparent parapatry. However there may be very little suitable habitat in that area, and such studies should be performed soon. In conjunction with larger sample sizes and higher density spatial sampling, inferences about the evolutionary history of geographically structured SGRPs in Amazonia will be greatly enhanced by variation data from the nuclear genome.

Acknowledgements

Fieldwork in Brazil was made possible by the help of Alexandre Aleixo and Maria Luisa Videira Marceliana of the Museu Paraense Emílio Goeldi in Belém, and of the Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis (IBAMA) and the Conselho de Gestão do Patrimônio Genético (CGEN). José 'Pepe' Tello, Wilsea Figueiredo, Nan Pimentel, Otavio, Pesão, Luis, Martin Davison, Will Davison, Fred Davison and Mary Noble assisted in the field. Further samples were loaned from the collections of the Field Museum of Natural History, the Smithsonian National Museum of Natural History, the American Museum of Natural History, and the Academy of Natural Sciences. I am grateful to all those involved in collecting and preserving this material. Laboratory work was conducted in the Pritzker Laboratory for Molecular Systematics and Evolution. This research was funded by the Chapman Fund of the American Museum of Natural History, and the Center for Latin American Studies and the Committee on Evolutionary Biology at the University of Chicago.

1.5 References

Aleixo, A. (2004). Historical diversification of a *terra-firme* forest bird superspecies: a phylogeographic perspective on the role of different hypotheses of Amazonian diversification. *Evolution* **58**, 1303–1317.

Armenta, J. K., Weckstein, J. D. & Lane, D. F. (2005). Geographic variation in mitochondrial DNA sequences of an Amazonian nonpasserine: the Black-spotted Barbet complex. *The Condor* **107**, 527–536.

Bahlo, M. & Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol* **57**, 79–95.

Barton, N. & Wilson, I. (1996). Genealogies and geography. In *New uses for new phylogenies* (edited by P. H. Harvey, A. J. L. Brown & J. M. Smith). Oxford University Press.

Bates, J., Hackett, S. & Goerck, J. (1999). High levels of mitochondrial DNA differentiation in two lineages of antbirds (*Drymophila* and *Hypocnemis*). *Auk* **116**, 1093–1106.

Bates, J. M. (2000). Allozymic genetic structure and natural habitat fragmentation: Data for five species of Amazonian forest birds. *Condor* **102**, 770–783.

Bates, J. M. (2002). The genetic effects of forest fragmentation on five species of Amazonian birds. *Journal of Avian Biology* **33**, 276–294.

Bates, J. M., Haffer, J. S. & Grismer, E. S. (2004). Avian mitochondrial DNA sequence divergence across a headwater stream of the Rio Tapajós, a major Amazonian river. *Journal of Ornithology* **145**, 199–205.

Capparella, A. P. (1987). *Effects of riverine barriers on genetic differentiation of Amazonian forest undergrowth birds.* Ph.D. thesis, Louisiana State University.

Capparella, A. P. (1988). Genetic variation in tropical birds: implications for the speciation process. In *Acta XIX Congressus Internationalis Ornithologici.* (edited by H. Ouellet). XIX International Ornithological Congress: Ottawa.

Chesser, R. T. (2004). Molecular systematics of New World suboscine birds. *Molecular Phylogenetics and Evolution* **32**, 11–24.

Cheviron, Z. A., Hackett, S. J. & Capparella, A. P. (2005). Complex evolutionary history of a Neotropical lowland forest bird (*Lepidothrix coronata*) and its implications for historical hypotheses of the origin of Neotropical avian diversity. *Molecular Phylogenetics and Evolution* **36**, 338–357.

Coyne, J. A. & Orr, H. A. (2004). *Speciation*. Sinauer.

Hackett, S. J. (1993). Phylogenetic and biogeographic relationships in the neotropical genus *Gymnopithys* (Formicariidae). *Wilson Bull* **105**, 301–315.

Hackett, S. J. & Lehn, C. A. (1997). Lack of genetic divergence in a genus (*Pteroglossus*) of neotropical birds: the connection between life history characteristics and levels of genetic divergence. *Ornithological Monographs* **48**, 267–280.

Hackett, S. J. & Rosenberg, K. V. (1990). A comparison of phenotypic and genetic differentiation in South American antwrens (formicariidae). *Auk* **107**, 473–489.

Haffer, J. (1969). Speciation in amazonian forest birds. *Science* **165**, 131–137.

Haffer, J. (1974). Avian speciation in tropical South America. *Publ. Nuttall Ornithol. Club* **14**.

Haffer, J. (1985). Avian zoogeography of the neotropical lowlands. *Ornithological Monographs* **36**, 113–146.

Haffer, J. (1997a). Alternative models of vertebrate speciation in Amazonia: an overview. *Biodiversity and Conservation* **6**, 451–476.

Haffer, J. (1997b). Contact zones between birds of southern Amazonia. *Ornithological Monographs* **48**, 281–305.

Haffer, J. & Fitzpatrick, J. W. (1985). Geographic variation in some Amazonian forest birds. *Ornithological Monographs* **36**, 147–168.

Hellmayr, C. E. (1929). On heterogynism in formicarian birds. *J. Ornithol.* **77**, 41–70.

Hudson, R. R., Boos, D. D. & Kaplan, N. L. (1992). A statistical test for detecting geographic subdivision. *Mol Biol Evol* **9**, 138–151.

Irwin, D. E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution* **56**, 2383–94.

Isler, M., Bates, J., Isler, P. & Hackett, S. (in prep.) .

Kuo, C. & Avise, J. (2005). Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees. *Genetica* **124**, 179–86.

Marks, B. D., Hackett, S. J. & Capparella, A. P. (2002). Historical relationships among neotropical lowland forest areas of endemism as determined by mitochondrial DNA sequence variation within the Wedge-billed Woodcreeper (Aves : Dendrocolaptidae : *Glyphorynchus spirurus*). *Molecular Phylogenetics and Evolution* **24**.

Mayr, E. & Diamond, J. (2001). *The birds of Northern Melanesia*. Oxford University Press.

Moritz, C., Patton, J. L., Schneider, C. J. & Smith, T. B. (2000). Diversification of rainforest faunas: an integrated molecular approach. *Annual Review of Ecology and Systematics* **31**, 533–563.

Novaes, F. C. & Lima, M. F. C. (1992). As aves do Rio Peixoto de Azevedo, Mato Grosso, Brasil. *Revista Brasileira de Biologia* **7**, 351–381.

Peters, J. L., ed. (1934-1987). *Checklist of birds of the world*, volume 1-15. Harvard University Press.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sick, H. (1967). Rios e enchentes na Amazônia como obstáculo para a avifauna. *Atas do Simpósio Sôbre a Biota Amazônica* **5**, 495–520.

Smouse, P. E., Long, J. C. & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* **35**, 627–632.

Snethlage, E. (1913). Uber die Verbreitung der Vogelarten in Unteramazonian. *J.Ornithol.* **61**, 469–539.

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

Templeton, A., Routman, E. & Phillips, C. (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–82.

Wakeley, J. (1999). Nonequilibrium migration in human history. *Genetics* **153**, 1863–1871.

Wallace, A. R. (1852). On the monkeys of the Amazon. *Proc. Zool. Soc. London* **20**, 107–110.

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

Wiens, J. J. (2004). What is speciation and how should we study it? *Am. Nat.* **163**, 914–923.

Wilkins, J. (2004). A Separation-of-Timescales Approach to the Coalescent in a Continuous Population. *Genetics* **168**, 2227–44.

Willis, E. O. (1969). On the behavior of five species of *Rhegmatorhina*, ant-following antbirds of the Amazon basin. *Wilson Bulletin* **81**, 363–395.

1.6 Appendix A: Collecting localities

Table 1.5: **Collecting localities**

site	area	river	bank	location	North	East
1	TL	Tapajós	L	upper R. Arapiuns, mun. Santarém, PA	-2.99	-55.84
2	TL	Tapajós	L	right bank R. Maró, near Lago da Panela, mun. Santarém, PA	-2.85	-55.68
3	TL	Tapajós	L	right bank R. Maró, near Sitio Esperança, mun. Santarém, PA	-2.82	-55.70
4	TL	Tapajós	L	left bank R. Maró, near Lago da Panela, mun. Santarém, PA	-2.86	-55.67
5	TL	Tapajós	L	left bank R. Maró, near Sitio Esperança, mun. Santarém, PA	-2.85	-55.69
6	TR	Tapajós	R	Base Sucupira, FLONA do Tapajós, km 117 BR-163, mun. Santarém, PA	-3.00	-55.00
7	TR	Tapajós	R	mun. Santarém, PA	-2.50	-55.00
8	JL	Juruena	L	Faz. São Nicolau, mun. Cotriguaçu, MT	-9.83	-58.25
9	JR	Juruena	R	Faz. Vale Verdi, mun. Nova Bandeirante, MT	-10.26	-58.29
10	TPLN	Teles Pires	L	Faz. Rio da Mata, mun. Carlinda, MT	-9.99	-55.58
11	TPLN	Teles Pires	L	mun. Alta Floresta, MT	-9.60	-55.90
12	TPLS	Teles Pires	L	Sitio Chale Kapp, mun. Sinop, MT	-11.59	-55.67
13	TPRNN	Teles Pires	R	right bank R. P. de A., Faz. Agro Florestal, mun. Novo Mundo, MT	-10.15	-55.36
14	TPRNN	Teles Pires	R	mun. Alta Floresta, MT	-9.60	-55.80
15	TPRNN	Teles Pires	R	R. Cristalino, mun. Alta Floresta, MT	-9.60	-55.80
16	TPRNS	Teles Pires	R	left bank R. P. de A., Faz. Pontal, mun. Nova Guarita, MT	-10.10	-55.51
17	TPRS	Teles Pires	R	Faz. Missioneira, mun. Sinop, MT	-11.59	-55.67
18	TPRS	Teles Pires	R	Praia do Cortado, mun. Sinop, MT	-11.75	-55.70

All localities are in Brazil. 'PA', state of Pará; 'MT', state of Mato Grosso; 'Faz.', Fazenda (farm/ranch); 'mun.', Município (local administrative area); FLONA, Floresta Nacional; R. P. de A., Rio Peixoto de Azevedo

1.7 Appendix B: Tissue samples

Table 1.6: **Tissue samples**

taxon	collection	tissue	site	haplotype	genbank	MPEG
Gspi	DED	135	6	E		56054
Gspi	DED	141	6	F		
Gspi	DED	144	6	F		56052
Gspi	DED	166	1	G		56175
Gspi	DED	167	1	G		56176
Gspi	DED	172	1	G		56174
Gspi	DED	176	1	G		56177
Gspi	DED	193	2	G		56179
Gspi	DED	218	3	G		56182
Gspi	DED	277	12	A		57217
Gspi	DED	279	12	A		57218
Gspi	DED	281	12	A		
Gspi	DED	305	13	F		57219
Gspi	DED	309	10	D		57220
Gspi	DED	327	9	D		57222
Gspi	DED	332	9	D		57223
Gspi	FMNH-JH	6	14	F	AY097015	
Gspi	FMNH-JH	20	14	F	AY097016	
Gspi	FMNH-JH	22	14	F	AY097017	
Gspi	FMNH-JH	23	14	F	AY097018	
Gspi	FMNH-JH	274	11	D	AY097011	

taxon	collection	tissue	site	haplotype	genbank	MPEG
Gspi	FMNH-JH	311	11	D	AY097012	
Gspi	FMNH-JH	312	11	D	AY097013	
Gspi	FMNH-JH	366	11	D	AY097014	
Gspi	LSUMNS	35366	14	F		
Gspi	LSUMNS	35591	4	G		
Gspi	LSUMNS	35599	4	G		
Gspi	MP	37	1	G		
Gspi	WM	390	7	E		
Hpoe	DED	136	6	H		
Hpoe	DED	137	6	J		56091
Hpoe	DED	147	6	H		56087
Hpoe	DED	162	1	M		56222
Hpoe	DED	164	1	M		56223
Hpoe	DED	173	1	N		56224
Hpoe	DED	180	1	M		56225
Hpoe	DED	183	1	M		56226
Hpoe	DED	201	2	L		56228
Hpoe	DED	202	2	M		56229
Hpoe	DED	219	3	M		56230
Hpoe	DED	222	5	M		56232
Hpoe	DED	226	3	M		56236
Hpoe	DED	240	18	B		57261
Hpoe	DED	249	17	B		57258
Hpoe	DED	251	17	B		57259

taxon	collection	tissue	site	haplotype	genbank	MPEG
Hpoe	DED	272	12	B		57264
Hpoe	DED	286	16	E		57266
Hpoe	DED	287	16	E		57267
Hpoe	DED	300	13	H		57271
Hpoe	DED	301	13	K		57272
Hpoe	DED	322	9	B		57275
Hpoe	DEDP	64	12	B		
Hpoe	FMNH-JH	12	15	H	AY612508	
Hpoe	FMNH-JH	51	15	I	AY612507	
Hpoe	FMNH-JH	452	11	A	AY612510	
Hpoe	MP	3	1	M		
Hpoe	MP	6	1	M		
Hpoe	MP	58	1	M		
Hpoe	MP	61	1	O		
Mmyo	DED	154	6	M		56244
Mmyo	DED	155	6	M		56097
Mmyo	DED	158	1	G		56216
Mmyo	DED	165	1	H		56218
Mmyo	DED	171	1	F		56219
Mmyo	DED	177	1	G		56217
Mmyo	DED	182	1	F		56220
Mmyo	DED	198	2	G		56211
Mmyo	DED	199	2	G		56210
Mmyo	DED	209	2	G		56212

taxon	collection	tissue	site	haplotype	genbank	MPEG
Mmyo	DED	213	2	I		56213
Mmyo	DED	215	2	G		56214
Mmyo	DED	216	5	G		56215
Mmyo	DED	244	17	B		57251
Mmyo	DED	260	17	C		57252
Mmyo	DED	268	12	A		57253
Mmyo	DED	278	12	D		57254
Mmyo	DED	283	12	B		
Mmyo	DED	314	8	A		57243
Mmyo	DED	319	8	A		57245
Mmyo	DED	336	9	B		57247
Mmyo	DED	338	9	C		57249
Mmyo	FMNH	392067	14	N		
Mmyo	FMNH	392068	14	O		
Mmyo	FMNH	392069	14	N		
Mmyo	FMNH	392070	14	O		
Mmyo	FMNH	392071	11	C		
Mmyo	FMNH	392072	11	E		
Mmyo	MP	4	1	G		
Mmyo	MP	101	2	J		
Mmyo	MP	111	5	G		

CHAPTER 2

THE USE OF COMPLETELY LINKED GENETIC DATA FOR INFERENCE OF POPULATION HISTORY

ABSTRACT

The existence of non-equilibrium population structure in the form of a recent barrier to gene flow is commonly inferred in an informal manner from surveys of geographic variation in mitochondrial sequences. However, the features of data which lead to this inference may also result under equilibrium models, and there is a need for formal statistical approaches to this inference problem. I study likelihood ratio tests of these possibilities under the assumption that the genealogy at a non-recombining locus is completely known, using a new importance sampling algorithm to estimate the probability density associated with the observation of a particular genealogy under a structured coalescent model with a recent barrier to gene flow. With information from a single locus under a simple two-population model, I find that power to detect a barrier which would result in strong drift is very low for realistic rates of ancestral gene flow. Additionally, I demonstrate that the known-genealogy maximum likelihood estimator of the rate of gene flow under the two population equilibrium model is upwardly biased. I discuss the implications of these results for inference of population history from completely linked data.

2.1 Introduction

In the last two decades there have been many studies in which polymorphism data are collected from regions of the mitochondrial genome, with a view to making inferences about the history of the sampled populations. When samples are available from several geographical locations, this area of population genetics tends to be referred

to as 'phylogeography' (see e.g. the review by Avise 2000, or any issue of *Molecular Ecology* in recent years). Such studies tend to be concerned with identifying population structure (see section 3.1) and forming conclusions about the population history that has given rise to that structure. In particular, a frequent conclusion is that two populations (taxa) are isolated, i.e. that there is no current gene flow between them. There are several reasons for the popularity of mitochondrial DNA for these studies. These include its ease of amplification and high mutation rate relative to typical nuclear loci. Because the effective population size for the mitochondrial genome is one-quarter that of autosomal loci, drift is expected to occur four times more quickly; whether this is an advantage or not depends on the question. Additional reasons for its popularity stem from the fact that, because the mitochondrial genome does not recombine, all $\sim 16,000$ nucleotide sites share the same underlying genealogy — i.e. they are 'completely linked'. In this chapter I study the problem of using completely linked data to make inferences about isolation and gene flow.

Coalescent theory implies that *all* the information about population history contained in genomic (including mitochondrial) variation data is information about the genealogies that relate the sampled chromosomes in the sampled regions of the genome. Despite the smaller effective size, alignments of mitochondrial sequences tend to have many polymorphic sites, and therefore any branch in the mitochondrial genealogy that is of appreciable length stands a chance of being hit by several mutations. Therefore the principal attraction of such completely linked data is that one may obtain good information about one genealogy, which may be put to use in inference of population history. Much recent work on inference in models of structured populations has focused on this type of data (e.g. Beerli & Felsenstein 1999, Bahlo & Griffiths 2000, Nielsen & Wakeley 2001, de Iorio & Griffiths 2004, Ewing *et al.* 2004,

Hey & Nielsen 2004).

A potential problem here is the correlation in the data. As long as the population allele frequencies at some site are unknown, the *alleles* sampled at that site are positively correlated, wherever it is located in the genome. In the nuclear genome, patterns of allelic variation at *sites* that are sufficiently far apart on a chromosome, or on different chromosomes, are independent (or 'exchangeable', if you consider the parameters of the model to be random). Such sites are said to be 'unlinked'. Therefore from the frequentist point of view the variance of any estimator can be made arbitrarily small by sampling more independent sites (the size of the nuclear genome obviously imposes a limit). Patterns of variation at different sites in mitochondrial DNA are conditionally independent, given the genealogy of the mitochondrial genome. But since the genealogy is unknown they are positively correlated, even when the population is unstructured. Therefore the variance of any estimator based on mitochondrial data cannot be made arbitrarily small, even by sequencing the entire mitochondrial genome in every individual in the population. Even if the genealogy of the mitochondrial genome for the whole population were known exactly, it would only be a sample of size one from the population history model that generates population genealogies. This probabilistic structure of completely linked data such as mitochondrial DNA raises the question of whether the data are sufficiently informative. In particular it may be necessary to address issues of experimental design such as: what is the power to reject a particular null hypothesis in favour of plausible alternative hypotheses?

The above should not however be interpreted as an argument for devoting resources to obtaining genotypes exclusively at unlinked sites. The information that linked sites contain about the local genealogy is potentially very valuable with respect to

inference of population history. The problem is that there are few independent regions of the nuclear genome within which recombination is known to be completely absent. In general nearby sites are 'loosely linked': they do contain information about the local genealogy, but as a result of recombination this genealogy changes along the chromosome in a correlated fashion. Statistical methods that capture the information about local genealogies while properly accounting for recombination do not yet exist for models of population history other than constant-sized panmixia. Chapter 3 describes one attempt to make progress on this front.

Another attraction of mitochondrial DNA is that there is frequently sufficient polymorphism in mitochondrial DNA that a tree of the haplotypes can be estimated. This is superficially similar to the situation in molecular phylogenetics in which a single haplotype is sampled from each of several relatively distantly related populations (taxa). In that case, as with mitochondrial DNA sampled from a single population, all the sites share the same genealogy, and so there is strong linkage disequilibrium (LD) in the data. However, in molecular phylogenetics this results not necessarily from physical linkage, but from the extreme form of population structure (i.e. the tree-like ancestral relationships between the sampled populations (taxa)).

The problem with this approach is that the genealogy is itself the object of interest only in an ideal molecular phylogenetic scenario wherein lineages are strictly bifurcating, and all lineages have experienced a lot of drift; otherwise it is the statistical distribution from which the genealogies are drawn which one wishes to learn about. However, it is understandable that a molecular phylogenetic philosophy should seep into some population genetic studies, since there is no sharp line between the two. At one extreme, under the canonical population genetic model of a completely un-

structured (panmictic) population, the topologies of the genealogies at the sampled loci contain no information whatsoever; but under structured models the topologies are informative about the structure (see e.g. Wakeley 2004). The topologies are especially of interest in more complex, hierarchically structured models such as might be considered when samples are available from diverse geographical locations. These are not normally studied by theoretical and statistical population geneticists, but the reason is that the simpler models are hard enough to study, not that the more complex ones fall outside population genetics. The point here is that 'phylogeography' is population genetics. Studies that are usually referred to simply as 'population genetics' are just trivial cases of phylogeography in which population structure is of less concern. However the term 'population genetics' has primacy and maintaining the separate term 'phylogeography' fosters several unhealthy beliefs including: (i) that the genealogy at any particular locus is itself of interest, (ii) that populations are related by bifurcating trees, and more generally (iii) that the problems faced, and the methods to be employed, are somehow different from those of population genetics. These issues are critical: the sociological situation described by Hey & Machado (2003) is highly undesirable.

Even studies that might superficially be considered as 'phylogenetic' are better considered as subdivided population genetics with a lot of drift and small sample sizes. An example at the phylogenetic end of the spectrum of possible sampling situations is the clade comprising humans (*Homo* etc), chimpanzees (*Pan*) and gorillas (*Gorilla*). The human and chimpanzee genomes are reciprocally monophyletic, other than at a handful of aberrant loci (presumably gorillas too, but insufficient polymorphism data are currently available). Thus one could sample a single homologous sequence from each taxon and attempt to estimate the genealogy (tree) at that locus. How-

ever, even assuming a strictly bifurcating model, the distribution of coalescence times in the ancestral population would contain information about the effective ancestral population size. As large amounts of data from the genomes of these organisms have become available, it has become clear that not only the coalescence time, but even the topology can vary — at some loci the chimp and human sequences are not the first to coalesce — and this variation can also be used to make inferences about the ancestral populations (Chen & Li 2001, Patterson *et al.* 2006). Thus even under a strictly bifurcating model of population history, non-negligible variation in topologies and branch lengths results from insufficient drift along the external or internal branches of the population tree, and it is therefore inappropriate to focus on estimation of trees at particular loci. Only phylogeneticists who strictly avoid such situations can justifiably never think of their trees as random variables (and then only non-Bayesians).

The need for an approach to studying the history of geographically structured populations that is more statistical than is typical of 'phylogeographic' studies has been widely recognised (e.g. Cavalli-Sforza & Piazza 1975, Felsenstein 1982, Templeton *et al.* 1995, Barton & Wilson 1996, Beerli & Felsenstein 2001, Knowles & Maddison 2002, Hey & Machado 2003, Guillot *et al.* 2005). This chapter is 'statistical' in the sense that I focus on the parameters of the distribution from which the genealogies are drawn, rather than the genealogies themselves. The aim is to study the effects of the correlation in completely linked data and its implications for experimental design. Although the potential for insufficient information about population history in completely linked data has been pointed out many times, there has been relatively little investigation of the statistical properties of estimators and hypothesis tests based on completely linked data under models of population history other than panmixia. In particular, it would be useful to be able to specify null and alternative models and

ask whether completely linked data generated under the alternative are likely to result in rejection of the null, and if so what would be appropriate sample sizes and other features of the experimental design. Since the informativeness of the data is in question, one would like to answer such questions for a likelihood-based test that uses all the information in the data, rather than for one based on summary statistics. The problem is that making likelihood-based estimates is usually difficult and computationally expensive, and in any case the necessary software exists only for a few models.

In this chapter I take an alternative approach. An estimator based on completely linked data is a deterministic function of data, which are created by a random mutational process occurring along the branches of a random genealogy. Therefore both the variability of the genealogical process, and the variability of the mutational process given the genealogy, contribute to the variance of the estimator. In contrast to the genealogical variance component, the mutational variance component can be reduced by sampling more polymorphic sites and thereby obtaining more information about the genealogy. The minimum possible variance of the estimator (maximum information) would be achieved if the genealogy were known without error, in which case the estimator would be a deterministic function of the genealogy and the variance of the estimator would be due only to the variability of the genealogical process.

Here I study the informativeness of completely linked data about models of population history by studying the statistical properties of estimators based on the genealogy itself, and the power of hypothesis tests based on such estimators. The advantage is that information is not lost by summarising the data, while the computational burden of using polymorphism data to make likelihood-based estimates is avoided.

One disadvantage is that it raises a different statistical question of how much more informative the genealogies are than realistic data sets. Therefore one way of viewing the results is as an indication of the 'upper bound' on the informativeness of completely linked data: if the sample sizes and other aspects of the sample are the same, it is impossible for a 'real' estimator based on completely linked data to have a lower variance than the corresponding 'known-genealogy' estimator, and it is impossible for a real hypothesis test to be more powerful than the known-genealogy likelihood ratio tests investigated here.

Because I focus on the issue of distinguishing pure isolation from ongoing gene flow, the models of population history feature gene flow. It is important to distinguish the information contained in a 'genealogy' (the topology and the times at which the coalescence events occurred), from other information about the ancestry of the sample, such as the times at which ancestral lines 'migrated' from one population to another and the populations in which the coalescence events occurred. The significance of this distinction is that sequence data contain information directly about the genealogy, and only indirectly (via the model) about the other information: in theory, by sampling sufficiently many polymorphic sites the uncertainty about the genealogy can be made arbitrarily small, but uncertainty about the location of the ancestral lines of the genealogy would still remain. I will refer to the complete specification of the locations of all the ancestral lines at all points in time back to the most recent common ancestor (MRCA) of the sample as the 'history' of the sample — note that the genealogy is contained within the history. Although the terminology is non-standard, this distinction between 'genealogies' and 'histories' recurs throughout the chapter.

A 'likelihood' in this chapter is then not the usual probability of an alignment of polymorphism data as a function of model parameters, but instead the probability (or rather, probability density) of a genealogy as a function of those parameters. Because the model features gene flow, it is not straightforward to calculate the probability density of a genealogy (let alone find the maximum). The approach I take here is to estimate the genealogy-based likelihood using importance sampling. Additionally, I compare results obtained by assuming that the history is known, in which case it is straightforward to calculate likelihoods and find the maximum likelihood estimates.

The structure of the chapter is as follows. In section 2.2 I describe the population genetic model and the likelihood ratio tests for isolation. In section 2.3.1 I derive expressions for the known-history likelihoods and maximum likelihood estimators, and in section 2.3.2 I describe how the known-genealogy likelihoods and maximum likelihood estimates are estimated. Results concerning the properties of the estimators and the power of the hypothesis tests are presented in section 2.4 and in section 2.5 I discuss their implications for the use of completely linked data to estimate rates of gene flow and test for reproductive isolation.

2.2 The model and the hypothesis test

The model of population history considered in this chapter (figure 2.1) is motivated by the typical situation wherein the existence of some population structure is not in question. What is in question is whether this population structure is the result of an equilibrium between drift and gene flow, or whether there is evidence for a recent period of isolation. The model supposes that two populations (*sensu* section 3.1) of equal effective size N_e exchanged migrants at equal rates since indefinitely

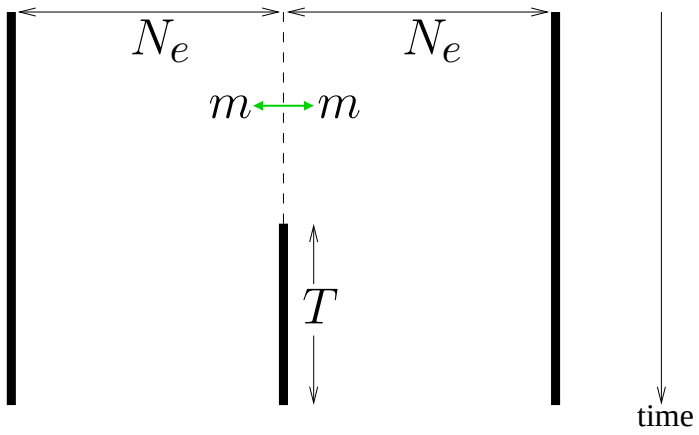


Figure 2.1: **The model of population history**

far in the past — at least since the time of the most recent common ancestor of the sample — until some time T generations ago when gene flow ceased. In one possible forwards-in-time model, the next generation in each population is created by sampling $2N$ chromosomes with replacement. On each of the $2N$ draws, with probability $1 - m$ a chromosome is sampled from the same population and with probability m it is sampled instead from the other population. This continues until T generations ago when migration ceases and the next generation is thereafter formed simply by sampling from the previous generation in that population.

This forwards-in-time evolutionary process is one of several that result in the following coalescence process, when the ancestry of a particular sample from the population is viewed backwards in time. During the 'initial' T generations, the coalescence process proceeds independently in each population according to the standard neutral coalescence process under constant-sized panmixia. At T generations ago, ancestral lines start to 'migrate' between the two populations with probability m each generation. Coalescences can of course only occur between lines that are currently in the same population. Neutral population genetic theory indicates that genealogies

only contain information about the *relative* rates of processes such as drift and migration, and it is standard to measure time in units of $2N_e$ generations (since this determines the rate of drift), and therefore introduce the parameters $M = 4N_e m$ and $\tau = T/2N_e$. M is equal to twice the rate at which each ancestral line migrates divided by the pairwise rate of coalescence, and τ can be thought of as the amount of drift that has occurred since gene flow ceased. Suppose that there are currently n_1 and n_2 ancestral lines in the two populations at some point in the history. On the new time scale coalescence events occur in each population at rates $\binom{n_1}{2}$ and $\binom{n_2}{2}$. If the current time is older than τ, migration from population 1 to 2 occurs at rate $n_1 M/2$, and migration in the opposite direction occurs at rate $n_2 M/2$.

This is a simple model but is has not been studied in the literature to my knowledge. I will refer to it as a 'barrier model', and sometimes make statements like 'the barrier has been in place for T generations', although the cause of the lack of gene flow may be intrinsic or extrinsic to the organismal biology. Since gene flow only occurs prior to the barrier, I will refer to M as the 'ancestral gene flow' rate (relative to the rate of coalescence) and refer to that part of the history as the 'gene flow period', and the more recent period as the 'barrier period'. The stochastic process describing the ancestry of the sample during the gene flow period is a 'structured coalescent' (e.g. Notohara 1990, Hudson 1991, Nordborg 2001) in which the initial sample configuration is determined by the (random) number of ancestral lines in each population at the end of the barrier period. Note that this model differs from the popular and simpler model of isolation wherein an unstructured population splits into two daughter populations. This difference is discussed further in section 2.5.2.

Interest centres on rejecting the null hypothesis that there is current gene flow between these populations. Therefore the hypothesis tests considered here are of

$$H_0 : \tau = 0 \quad \text{vs.} \quad H_1 : \tau > 0, \tag{2.1}$$

where H_0 and H_1 refer respectively to the null (no isolation) and alternative (isolation) hypotheses.

A sample of chromosomes is available, and it is known which population each chromosome was sampled from. I will write the vector of model parameters as $\phi = (M, \tau)$. The known-genealogy likelihood-ratio test statistic is

$$\Lambda = \frac{L(\widehat{M}, \widehat{\tau})}{L(\widehat{M}_0, \tau = 0)} = \frac{p(G; \widehat{M}, \widehat{\tau})}{p(G; \widehat{M}_0, \tau = 0)}. \tag{2.2}$$

Here, $(\widehat{M}, \widehat{\tau})$ is the joint maximum likelihood estimate (MLE) of the parameter vector, i.e. the value of (M, τ) for which the observed data (G) are most probable under the model. Since the branch lengths of G are continuous random variables, $p(G; M, \tau)$ is a probability density (not a probability). $\widehat{M}_0$ is the MLE of M subject to the constraint imposed by H_0 that $\tau = 0$. With this definition, $\Lambda > 1$ (i.e. a positive value of the log likelihood-ratio λ) is evidence against the null hypothesis. The known-history version of the test is the same, but with G replaced by the history H.

2.3 Methods

2.3.1 Computing the known-history likelihood

In this section I give closed-form expressions for the likelihood, maximum-likelihood estimators and the likelihood ratio statistic under the assumption that the history is known and treated as data.

A history comprises a random number $n - 1 + J$ of intervals, of which $n - 1$ end with coalescences and J (random) end with migrations (n is the total number of chromosomes sampled, as above). Each interval is characterised by the configuration of ancestral lines during that interval — in this two-population case this is a vector of length two recording the number of ancestral lines in each population during that interval. The stochastic process generating histories is known as the structured coalescent. It is a continuous time Markov process with state space equal to the set of possible configurations, which starts in the sampling configuration and proceeds through a random sequence of configurations until the MRCA is reached. The probability density of a history H as a function of M and τ is therefore easy to obtain from basic properties of continuous time Markov processes as outlined below.

The history can be viewed as comprising the sequence of configurations visited (the 'jump chain') together with the durations of the intervals spent in these configurations (the 'waiting times'). Consider two consecutive configurations A and B. In the structured coalescent there will be just one event that could have resulted in the move from configuration A to configuration B. For example, if B differs from A in having one fewer line in population 2 and one more line in population 1, then the transition $A \to B$ must have occurred as a result of a migration event (backwards in time) out

of population 2 into population 1. The probability of the realised transition is given by the rate of the type of event which actually occurred relative to the total rate of all possible events, so in this example it would be equal to

$$\frac{n_2 \frac{M}{2}}{n_1 \frac{M}{2} + \binom{n_1}{2} + n_2 \frac{M}{2} + \binom{n_2}{2}}.$$

The waiting time spent in configuration A before the jump to configuration B has an exponential density with rate parameter equal to the total rate of all possible events given configuration A (i.e. the denominator of the above expression). The probability of the realised jump chain is given by the product of all the corresponding transition probabilities, the probability density of the sequence of waiting times given the jump chain is given by the product of all the corresponding exponential densities, and the probability density of the history is equal to the product of those two quantities. This results in

$$p(H; M, \tau) = \left(\frac{M}{2}\right)^J \exp\left\{-\left(Q_c + Q_m\right)\right\}, \tag{2.3}$$

where Q_c and Q_m are respectively the total rates of possible coalescence and migration events, integrated over the duration of the history. I.e.

$$Q_c = \sum_i t_i \left[\binom{n_1^{(i)}}{2} + \binom{n_2^{(i)}}{2}\right], \tag{2.4}$$

where the sum is over the intervals in the history and t_i is the duration of interval i. Each ancestral line migrates at rate $M/2$, but only during the portion of the history rootward of τ. Writing $D(\tau)$ for the total length of all the branches rootward of τ,

$$Q_m = \frac{M}{2} D(\tau). \tag{2.5}$$

Recall that the objective is to find the value of (M, τ) that jointly maximises the probability density of the observed history, and also to find the value of M which maximises it given that $\tau = 0$. Rather than making a more mathematical argument, note that it is easy to see what these values must be. In any history, viewed backwards in time from the time of sampling, one observes some number of intervals terminated by coalescence events (or perhaps none), and then an interval terminated by a migration event. The observation of an initial sequence of intervals terminated by coalescence (and not migration) events is most probable under a model in which migration events during that time were in fact impossible. Therefore the maximum likelihood value of τ is older than the end of those intervals. The time between τ and the 'first' migration event has some exponential density (depending on M and the configuration of the ancestral lines at that time) and therefore the probability density is highest if that amount of time is very short. Therefore the MLE of τ is immediately 'before' the 'first' migration event, whatever the value of M. Since migration events occur as a Poisson process with rate $M/2$ along the branches rootward of τ, the MLE of the rate is given as usual by dividing the number of events by the time for which the process was observed. Writing t_m for the time of the first migration event, and J for the number of migration events observed in the history, the known-history joint MLEs of τ and M under the alternative hypothesis are

$$\widehat{\tau} = t_m \qquad \text{and} \qquad \widehat{M} = 2\frac{J}{D(\widehat{\tau})}, \tag{2.6}$$

and the known-history MLE of M under the null hypothesis is

$$\widehat{M_0} = 2\frac{J}{D(0)}, \tag{2.7}$$

where $D(0)$ is the total length of all the branches of the genealogy.

Alternatively, substituting the quantities (2.5) and (2.4) into equation (2.3) and taking logs, the log-likelihood of the parameters is

$$l(M,\tau) = J\log\left(\frac{M}{2}\right) - Q_c - \frac{M}{2}D(\tau). \tag{2.8}$$

Since $D(\tau)$ is a decreasing function of τ, this is maximised by the largest value that τ can take, which is the time of the 'first' migration event, and so this is the MLE of τ for all values of M. Differentiating with respect to M and solving for the maximum shows that the MLE of M is $2J/D(\widehat{\tau})$.

The likelihood ratio is

$$\begin{aligned}\Lambda =& \frac{L(\widehat{M},\widehat{\tau})}{L(\widehat{M}_0,0)}\\ =& \frac{\left(\frac{2J}{D(t_m)}\right)^J \exp\left\{-Q_c - \frac{2J}{D(t_m)}D(t_m)\right\}}{\left(\frac{2J}{D(0)}\right)^J \exp\left\{-Q_c - \frac{2J}{D(0)}D(0)\right\}}\\ =& \left(\frac{D(0)}{D(t_m)}\right)^J\end{aligned} \tag{2.9}$$

Thus an observed history provides evidence for a barrier if (i) much of the total length of the genealogy is 'before' the first migration event, and (ii) many migration events occurred in the history.

2.3.2 Estimating the known-genealogy likelihood

The previous section presented expressions for the MLEs of M and τ, and the likelihood ratio statistic, in terms of features of the history which were supposed to

be known, such as the number of migration events, the time of the first migration event, and the proportion of the total length of the genealogy which lies on either side of this first migration event. However, although in principle a very large quantity of completely linked polymorphism data might result in very little uncertainty about the topology and branch lengths of the genealogy, considerable uncertainty will always remain about the other aspects of the history. Therefore the known-history assumption is highly unrealistic.

Of greater interest are the statistical properties of the MLEs and likelihood ratio statistic in the case where only the genealogy is supposed to be known, and all other aspects of the history are not known. Unfortunately closed-form expressions such as equation 2.3 are not known for the probability density of a genealogy under the structured coalescent, except for trivial cases such as a sample of size 2 (e.g. Nath & Griffiths 1993, Wakeley 1996b). It is perhaps tempting to think that the problem could be solved by summing over the unknown locations of the coalescence events in the tree, in the same way as Felsenstein's (1981) 'pruning algorithm' sums over unknown alleles at internal nodes of a tree. This is however not possible. The reason is essentially that whereas, conditional on the genealogy, ancestral lines mutate (neutrally) in independence of each other, they do not migrate in independence of each other. For example, the observation that two ancestral lines do not coalesce in the recent past may make it more plausible that they were in different populations during that time.

Estimating the likelihood: naive Monte Carlo

The problem of computing the likelihood (i.e. the probability density of the genealogy) under a model with migration has the structure of a missing data problem (see

3.1): if, in addition to the genealogy, the history of movements among populations of the ancestral lines of the genealogy were known, then the probability density of the resulting structure (the 'history') would be straightforward to calculate, as described in section 2.3.1.

Let a 'path' ν be the location of the ancestors of the sample, at all times back until the MRCA (see figure 2.2). The true path that gave rise to the genealogy is not known, but the model does specify the prior probability density $p(\nu;\phi)$ of any path ν. Therefore the likelihood can be computed as the probability density of the genealogy given the path, averaged over the distribution on paths:

$$p(G;M,\tau) = \int p(G|\nu)p(\nu;M,\tau)d\nu \tag{2.10}$$

It is also simple to simulate paths from the prior since, under the prior, each ancestral line migrates independently between populations according to the migration rate matrix. A naive Monte Carlo estimator of the likelihood is

$$p(G;M,\tau)\hat{=}\ \frac{1}{B}\sum_{b=1}^{B} p(G|\nu^{(b)}), \quad \nu^{(b)} \sim p(\nu;M,\tau) \tag{2.11}$$

(cf equation 3.2). In words, simulate a large number B of paths, for each one record the probability density of the genealogy given that path, and compute the average of those quantities.

It is necessary to incorporate in the simulations the fact that the ancestral lines coalesce at the coalescence events observed in G. An algorithm for obtaining one term in the sum in equation 2.11 is given below (algorithm 1). The algorithm is given for the general case in which the number of populations is arbitrary and the effect of

the isolating event is to alter the migration rate matrix from $\mathbf{M_1}$ (isolating barrier present) to $\mathbf{M_0}$ (equilibrium migration). In the simple case of two populations,

$$\mathbf{M_1} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{M_0} = \begin{pmatrix} -M/2 & M/2 \\ M/2 & -M/2 \end{pmatrix}$$

In the description of the algorithm, i indexes the $n-1$ observed coalescence events in the genealogy, t_i is the time at which the i^{th} coalescence event occurs, and t records the current time as the simulation proceeds from the present back to the MRCA. Each node in the genealogy (including the tips) is assigned a distinct label and the same label is given to the line ancestral to that node. Let α_i and β_i be the labels of the two ancestral lines that coalesce at coalescence event i, and let γ_i be the label of their parental node.

Algorithm 1 (naive Monte Carlo)

1. Set $t \leftarrow 0$, $i \leftarrow 1$, $\mathbf{M} \leftarrow \mathbf{M_1}$ and place the n sampled lines in the populations from which they were sampled.

2. If $t < \tau$ set $t \leftarrow \min\{\tau, t_i\}$, otherwise set $t \leftarrow t_i$

3. For each ancestral line, simulate its movement from its current location until time t, independently according to $\mathbf{M}$

4. If $t = \tau$

   set $\mathbf{M} \leftarrow \mathbf{M_0}$ and repeat from 2;

   otherwise, if lines α_i and β_i are in different populations

   terminate the simulation and record $p(G|\nu) = 0$ for this path;

   otherwise, if $i = n - 1$

terminate the simulation, compute and record $p(G|\nu)$ for this path;

otherwise

remove line α_i, change the label of line β_i to γ_i, set $i \leftarrow i+1$ and repeat from 2.

$p(G|\nu)$ is the probability density associated with the observation that coalescence events occurred exactly as observed in G, given the locations of the ancestral lines specified in ν. In the two population case, using the notation introduced in section 2.3.1, this is equal to e^{-Q_c}. In general it is given by

$$p(G|\nu) = \exp\left\{-\sum_i t_i \sum_j \binom{n_j^{(i)}}{2}\right\}. \tag{2.12}$$

Figure 2.2a shows an example of a path that might be simulated according to this scheme. In this example, the path is compatible with the genealogy — at each of the two coalescence events the coalescing lines are together in the same population. Since the coalescence rate is zero when the two lines are apart and 1 when they are together, the probability density of G given the path is

$$p(G|\nu) = e^{-(t_1+t_3)}.$$

Figure 2.2b shows an example of a path that is incompatible with the genealogy — at the time of coalescence of lines 1.1 and 1.2 those lines are not in the same population and therefore $p(G|\nu) = 0$ for this path. Unfortunately, for realistic sample sizes, and especially when there are more than 2 populations, this sort of situation will occur in the vast majority of paths simulated according to a scheme like algorithm 1. Most simulated paths therefore do not contribute to the sum in equation 2.11

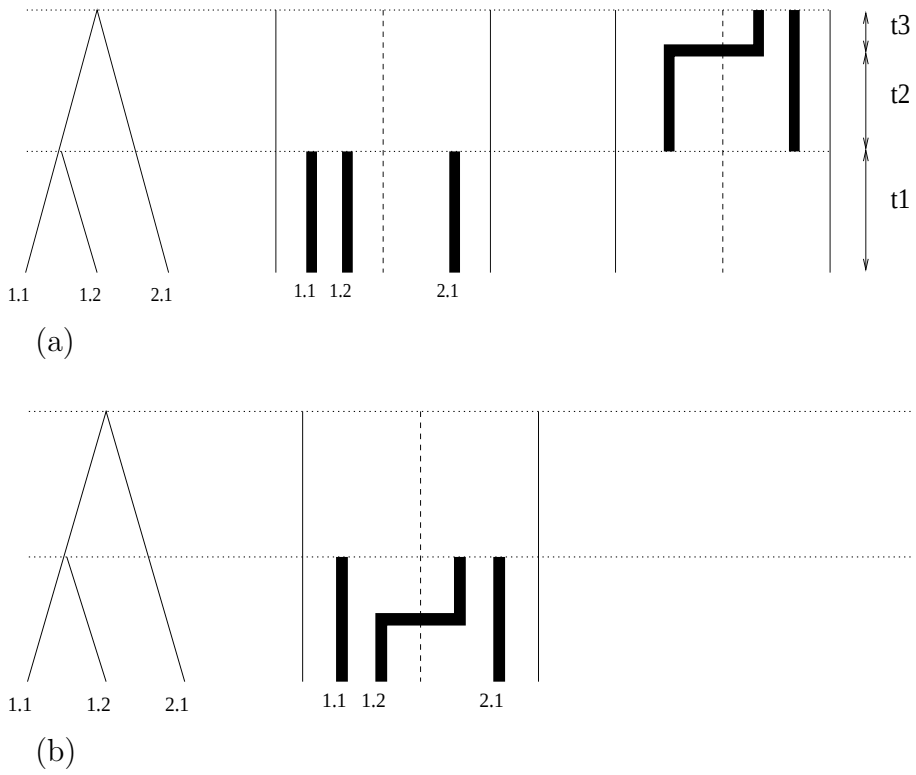


Figure 2.2: **Simulating paths in order to estimate the likelihood.** A genealogy is shown to the left of panels (a) and (b). Chromosomes 1.1 and 1.2 are sampled from population 1, and chromosome 2.1 is sampled from population 2. The two panels illustrate two paths simulated in the course of estimating the probability density of the genealogy. The first coalescence event in the genealogy is between lines 1.1 and 1.2. Note that the coalescence events could occur in either population. In (a) the first lines to coalesce are found to be in the same population at the time of their coalescence. The simulation is then continued assuming that they coalesced at that time and, in this case, the two remaining lines are again together at the time of the final coalescence. See text for calculation of $p(G|\nu)$ and the importance weight for this path. In (b), the coalescing lines are in different populations at the time of their coalescence; the path is therefore incompatible with the genealogy, this simulation is terminated and contributes a zero term to the summations in equations (2.11) and (2.14).

and, for practical values of B, the variance of the naive Monte Carlo estimator will be too high for it to be useful. This problem is familiar from section 3.4.1 in which the possibility of estimating likelihoods for polymorphism data by simulating high-

dimensional missing data from its prior distribution was discussed. The standard solution (and the approach adopted here) is essentially to not simulate the missing data from the prior, but instead to bias the simulations towards missing data that is more consistent with the observed data, and to account correctly for the way in which this bias is introduced. Markov chain Monte Carlo and importance sampling both employ this basic idea.

Estimating the likelihood: importance sampling

The approach taken in this chapter is to estimate the likelihood using importance sampling (IS), which is a standard way of reducing the variance of the estimator in situations like this. Note that for any probability density $q(\nu;\phi)$ equation 2.10 can be trivially rewritten as

$$p(G;M,\tau) = \int p(G|\nu)\frac{p(\nu;M,\tau)}{q(\nu;M,\tau)}q(\nu;M,\tau)d\nu; \tag{2.13}$$

the only requirement for this to make sense is that all paths which are possible under the prior density $p(\nu;M,\tau)$ are also possible under the new density $q(\nu;M,\tau)$. Just as equation 2.10 can be estimated by equation 2.11, equation 2.13 can be estimated by

$$p(G;M,\tau)\hat{=}\frac{1}{B}\sum_{i=1}^{B} p(G|\nu^{(i)})\frac{p(\nu^{(i)};M,\tau)}{q(\nu^{(i)};M,\tau)}, \quad \nu^{(i)} \sim q(\nu;M,\tau). \tag{2.14}$$

In words, simulate a large number B of paths, but from a distribution q that is different from the prior distribution (i.e. different from the distribution used in algorithm 1). For each path, compute the probability density of the genealogy given the path as before, and also compute the ratio of the probability density of the path under

the prior to its probability density under q. The estimate of the likelihood is the average of the product of these two quantities. Notice that it is intuitively plausible that equation 2.14 estimates the same quantity as equation 2.11 because, although the paths are being simulated from a different distribution, the contribution of paths that have high probability under the new distribution is downweighted and that of paths that are improbable under the new distribution is upweighted.

Note that, mathematically at least, the parameter vector (M, τ) that parameterises the proposal density (the 'driving value') need not be the same as the parameter values at which the likelihood is being estimated. Therefore one could simulate a large number of paths using a single driving value, and then attempt to use this one set of simulated paths to estimate the likelihood at a series of parameter values, perhaps with the aim of decreasing the computational cost. The alternative is to repeat the simulations at each point in parameter space using new driving values at each point. Even for the simplest problems I found that these two ways of estimating the likelihood agreed only over a relatively small region in the vicinity of the driving value, and it was evident that attempting to estimate the likelihood surface using a single driving value would introduce complications which were unnecessary for the aims of this chapter. Therefore whenever estimating the likelihood at some point (M, τ), I used the same value of (M, τ) in the proposal simulations.

The new distribution q is referred to as the proposal density. The choice of q is arbitrary, but the optimal choice is that which results in the lowest variance estimator for some fixed B. If it were possible to choose the proposal density to be the posterior density $p(\nu|G)$ of paths conditional on G, then the estimator would equal the likelihood with zero variance. Clearly this is not possible, since knowing this conditional

density implies knowing the probability whose estimation is the objective. However, loosely speaking, the more similar the proposal density is to this optimal density, the lower will be the variance of the estimator. Stephens & Donnelly (2000) and Fearnhead & Donnelly (2001) use this observation to design IS proposal distributions for the problem of estimating likelihoods for polymorphism data.

Conditional on an observed genealogy, the probability distribution of the path is altered in a complicated way such that paths that are more consistent with the genealogy have higher weight, and paths that are incompatible with the genealogy have no weight at all. It is difficult to intuit what the key features of this unknown conditional distribution are that should be captured in an efficient proposal density. However it seems reasonable to expect that the unknown conditional distribution should have the following property (under certain circumstances at least). Suppose that the simulation of the path has just reached step 3 of algorithm 1, and that the next event back in time is the coalescence of lines α_i and β_i, t time units from now. If, given the matrix of migration rates, few migration events are expected over the time t then a migration event that brings α_i and β_i into closer proximity will have increased probability relative to the prior than an event which moves them apart.

In order to estimate the known-genealogy likelihoods I used importance sampling with a proposal density that was intended to have this feature. The algorithm to obtain one term of the sum in equation 2.14 is the same as algorithm 1 except for the following two differences.

1. **Biased migration in the proposal** At step 3, when simulating the movement of lines α_i and β_i, the migration rates are altered in the following way. Suppose that the rate of migration to a particular adjacent population is $M/2$ under

the prior. Under the proposal density, if the migration decreases the separation of α_i and β_i its rate is increased to λM, whereas if it increases their separation its rate is decreased to $(1-\lambda)M$. The bias parameter λ depends on the probability that the line migrates prior to its coalescence (i.e. prior to the next coalescence event). Writing t for the time to the next coalescence event, I used $\lambda = 0.5(1-e^{-aMt}) + be^{-aMt}$. The parameters a and b control respectively the dependence of the bias on the probability of migrating prior to coalescence, and the maximum amount of bias. I found that $a = 1$ and $b = .95$ resulted in a low variance estimator of the likelihood for the sample sizes and numbers of populations used in this chapter.

2. **Computing the importance weight** Whereas in algorithm 1 the term contributed to the sum in equation 2.11 is simply $p(G|\nu)$, it is now necessary to multiply this contribution by the 'importance weight' $p(\nu; M, \tau)/q(\nu; M, \tau)$. These quantities are straightforward to calculate because the movements of each ancestral line during each interval are independent continuous time Markov processes with rates given by the migration rate matrices, and the biasing procedure described above. The resulting importance weight is

$$\frac{p(\nu; M, \tau)}{q(\nu; M, \tau)} = \prod_i \frac{M_i}{M_i^*} \exp\{-t_i(q_i^* - q_i)\},$$

where i indexes intervals in the history, t_i is the duration of interval i, M_i and M_i^* are the rates at which the particular event that ended interval i occurred at under the prior and under the proposal respectively, and q_i and q_i^* are the total rate at which migration events occurred during interval i under the prior and under the proposal respectively. For example, the importance weight associated

with the path illustrated in figure 2.2a would be

$$\begin{aligned}\frac{p(\nu; M, \tau)}{q(\nu; M, \tau)} = &\exp\left\{-t_1\left(2(1-\lambda_1)M + \frac{M}{2} - 3\frac{M}{2}\right)\right\} \\ &\times \frac{\lambda_2 M}{\frac{M}{2}} \exp\left\{-t_2\left(2\lambda_2 M - 2\frac{M}{2}\right)\right\} \\ &\times \exp\left\{-t_3\left(2(1-\lambda_3)M - 2\frac{M}{2}\right)\right\},\end{aligned} \tag{2.15}$$

where $\lambda_i = 0.5(1 - e^{-Mt_i}) + 0.95e^{-Mt_i}$.

It is worth emphasising that the biasing procedure in the proposal distribution is a fairly crude attempt to mimic the unknown conditional distribution of paths given the genealogy. For example, no attempt is made to incorporate the effects of the current configuration of ancestral lines among populations when simulating migration events and so improbable histories wherein many ancestral lines occupy the same population for a long time without coalescing may be proposed. Nevertheless, I found that likelihoods were estimated with much lower variance than the naive Monte Carlo scheme for sample sizes up to 16 distributed among two populations.

2.3.3 Using the likelihoods

In this chapter I focus on two aspects of the known-genealogy and known-history likelihoods: the distribution of the MLE of M under models of equilibrium migration, and the power of the likelihood ratio test for isolation described in section 2.2. A model is specified by the number of populations, the sample configuration and the values of the parameters M and τ. For a particular combination of these quantities, I simulated histories from the corresponding structured coalescent density in the usual

way (see e.g. Takahata 1988, Hudson 1991). For a particular simulated history, the known-history MLEs of M and τ and the likelihood ratio statistic can be calculated using the expressions in section 2.3.1.

The known-genealogy versions of these quantities are harder to obtain. Section 2.3.2 describes how to estimate the likelihood at any point (M, τ) in parameter space given a genealogy. However that leaves the problem of finding the MLE $\widehat{M_0}$ under the null hypothesis and finding the joint MLE $(\widehat{M}, \widehat{\tau})$ under the alternative hypothesis. I found these values by evaluating the likelihood at several different values of M (null) or at points of a two-dimensional grid (alternative), and treating the maximum likelihood thus encountered as the likelihood at the MLE. If the maximum occurred at either end of the line of points, or on an edge of the grid, I evaluated the likelihood at further points in that direction until an internal maximum seemed to have been found. Once an internal maximum was found I repeated the likelihood evaluations with a finer grid centred on the maximum to obtain the final estimate. I found that $B = 40,000$ IS draws resulted in estimates of the likelihood with suitably low variance.

2.4 Results

2.4.1 Properties of the maximum likelihood estimators

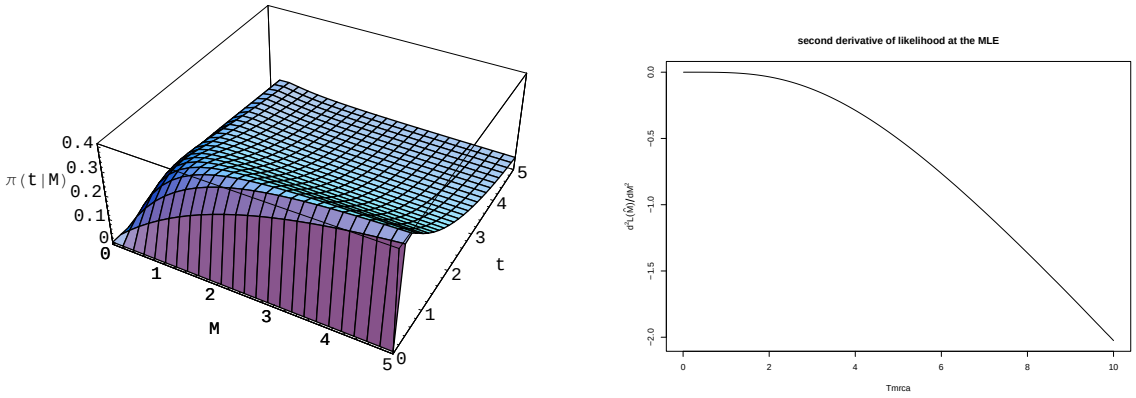
Sample size $n = 2$

Since the known-history MLE of the migration rate is equivalent to dividing the number of events observed in a one-dimensional Poisson process by the time for which the process was observed, it is obviously unbiased. It is less obvious whether to expect

bias in the known-genealogy MLE of the migration rate. The simplest case is that of $n = 2$ lines sampled under a two population model. In that case an exact expression for the likelihood (the probability density of the coalescence time t) can be found (Nath & Griffiths 1993, Wakeley 1996b), and this is illustrated for the case when the two lines are sampled in different populations in figure 2.3a. The observation of an ancient coalescent time supports a low migration rate, whereas a recent coalescence time supports high migration rates. However, the likelihood surface becomes flat over high values of M when the coalescence is more recent indicating that there is no information about exactly how high the migration rate is. In fact, as $t \to 0$, the second derivative of the likelihood surface at the MLE tends to zero (figure 2.3b). Statistical properties of the maximum likelihood estimator in the $n = 2$ case are illustrated in figure 2.4. For low values of M, the median of the estimator lies above the true value, while at high values it lies below. Because occasional recent coalescences result in very high estimates of M, the means of the estimates lie well above the truth (i.e. the estimator is biased) at the lower values of M.

Sample size $n > 2$

The approach of Wakeley (1996b) enables an exact expression for the likelihood to be found for $n = 2$ or 3, but for larger n, and especially for more populations, the number of possible configurations of the ancestral lines among the populations rapidly becomes unmanageable. However, real sample sizes are larger than 2, and in order to study properties of the estimators in such cases an alternative method of evaluating the likelihood is necessary. Here I estimate the likelihood of a genealogy using the importance sampling scheme described in section 2.3.2, and search for the maximum likelihood values of M or (M, τ) using a simple grid search. Figure 2.5 illustrates



(a) (b)

Figure 2.3: **The likelihood function with one sample from each of two populations**
(a) Conditional density of coalescence time of 2 chromosomes under the 2 population equilibrium migration model, given the migration rate M. $n = (1, 1)$. As $M \to \infty$ the process converges on that of a single population with twice the effective size. Thus, for example, as $M \to \infty$ and $t \to 0$, the density approaches 0.5. (b) Second derivative of the likelihood function at the MLE: sample $= (1, 1)$.

distributions of the known-genealogy and known-history estimators resulting under the assumption of equilibrium migration (i.e. $\tau = 0$), for data sets (histories) of 8 samples from each population simulated under equilibrium migration with a range of true values of M. The estimated bias and root mean square error are given in table 2.1. It is evident that the known-genealogy MLE is upwardly biased, and, as expected, much more variable than the known-history MLE.

For some genealogies, the known-genealogy likelihood surface for M has its maximum at a high value (e.g. $\widehat{M} > 10$). In such cases the likelihood surface is usually almost flat over a large range of large values of M, and a very precise estimate of $\widehat{M}$ is not therefore required to obtain a reasonable estimate of the likelihood ratio. Since the computational cost of estimating the likelihood increases with M (because it is necessary to simulate many migration events in each proposed history), if $\widehat{M}$

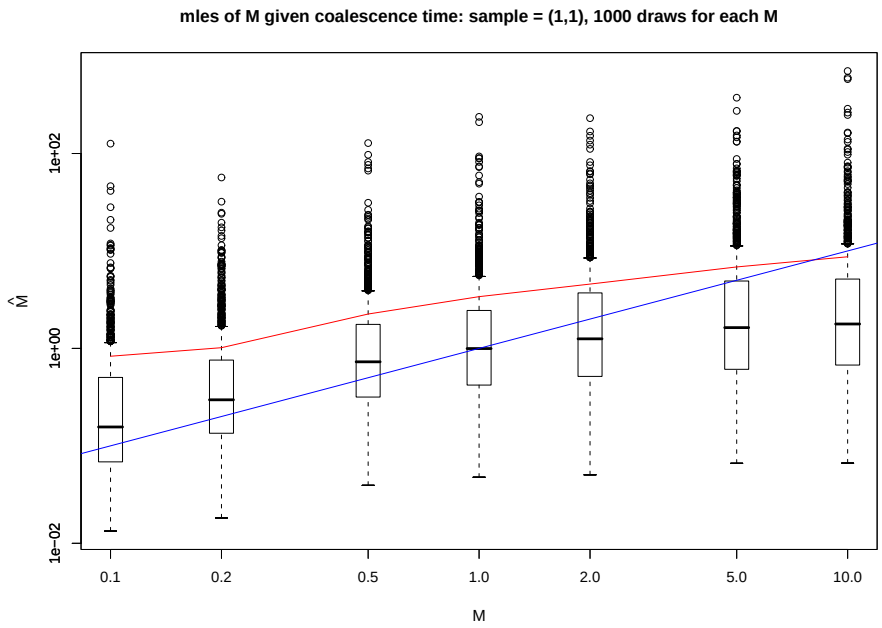


Figure 2.4: **Distributions of the known-genealogy MLE of M, for a sample of size 2, with one sample per population.**
The red line passes through the mean of the estimates, and the blue line is $y = x$. Note that both axes have a logarithmic scale. MLEs found by maximising the exact expression for the likelihood using `optimise()` in *R*.

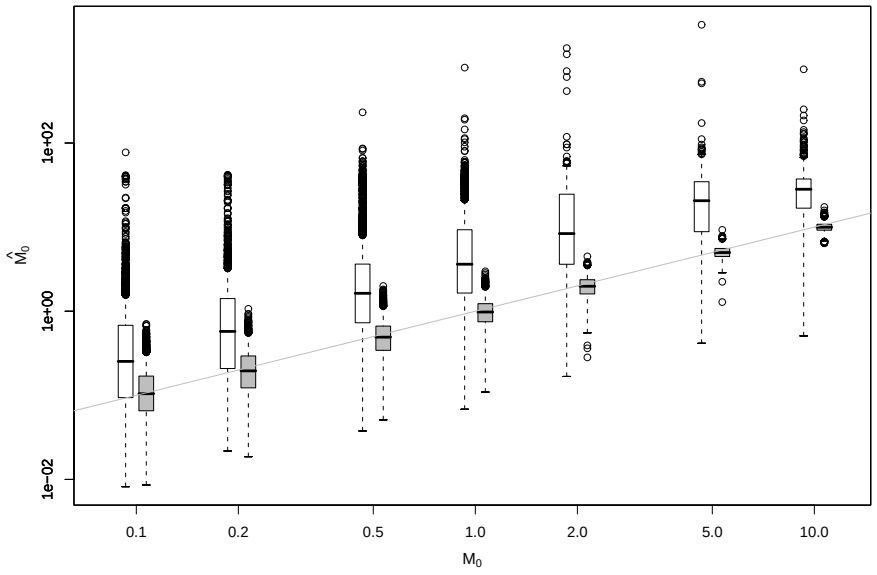


Figure 2.5: **Distributions of known-genealogy and known-history estimators of** M.
True M values along x-axis and estimates along y-axis (log-log scale). Sample $= (8, 8)$.

	bias		RMSE	
true M	KG	KH	KG	KH
0.1	0.77	0.03	3.25	0.10
0.2	1.40	0.03	4.37	0.14
0.5	3.99	0.03	9.93	0.25
1.0	8.12	0.01	20.17	0.37
2.0	15.79	0.00	55.85	0.57
5.0	21.59	0.01	87.07	0.87
10.0	17.56	0.01	28.01	1.29

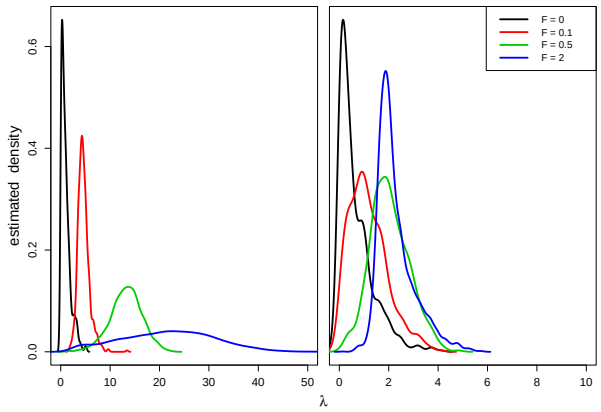
Table 2.1: **Statistical properties of known-genealogy and known-history estimators of M**
Sample $= (8, 8)$, 1 locus.

appeared to be high, I did not permit the grid search to continue until a precise estimate of M had been obtained. The largest estimates of M are therefore very crude, and the associated estimate of the likelihood at the MLE less so. As a result, the bias and variability of the known-genealogy estimator are underestimated at the high true values of M.

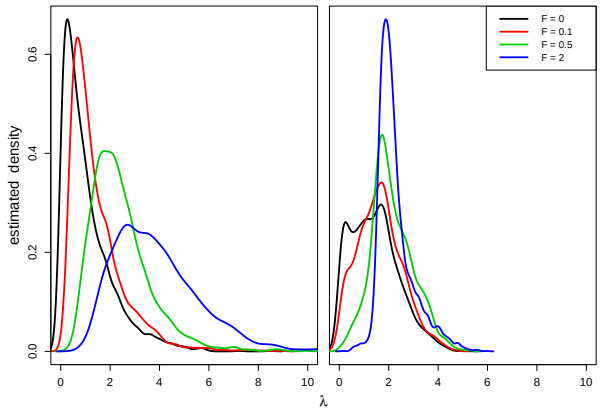
2.4.2 Power of the hypothesis tests

I investigated the relationship between τ and the power of the known-history and known-genealogy hypothesis tests described in section 2.2 for various sample configurations, and values of the migration parameter M. For each combination of parameter values this involved simulating a large number of histories (genealogies) and, for each one computing (estimating) the value of the likelihood ratio test statistic, as described in section 2.3.1 (2.3.2).

Figure 2.6 plots the distribution of the log likelihood ratio statistic λ in the known-history and known genealogy cases with a sample of size 4 taken from each population for several values of τ and two different migration rates. The density of λ under



(a) $M = 10.0$



(b) $M = 1.0$

Figure 2.6: **Distributions of known-history (left) and known-genealogy (right) log likelihood ratio statistic**
Sample $= (4, 4)$

the null hypothesis is shown in black, and the coloured lines show its density for values of $\tau > 0$. Consider first the known-history tests (left panels of figure 2.6). Any genealogy simulated under the alternative hypothesis ($\tau > 0$) features an initial period during which no lines migrate. If the migration rate is high, this initial period without migration is improbable under the null hypothesis. Therefore the likelihood ratios for genealogies generated under the alternative tend to be large, and these

large likelihood ratios are rare under the null hypothesis. Thus there is little overlap between the distribution of λ for $\tau = 0$ and its distribution for $\tau > 0$ when the migration rate is high ($M = 10$; top left panel of figure 2.6). When the migration rate is lower ($M = 1.0$; bottom left panel of figure 2.6), the initial period without migration is not so improbable, the likelihood ratios resulting for $\tau > 0$ tend to be much smaller, and the distributions are more similar.

In order to assess significance, one would like to compare each observed value of the test statistic to the sampling distribution of the test statistic under the null hypothesis. However, the migration rate M is not specified by the null hypothesis, and in practice it would be unknown. In addition, the sampling distribution of the test statistic is not known from analytical theory. A standard way to proceed in this situation is 'parametric bootstrapping'. Here one would estimate the value of M under the null hypothesis, and then compare the observed test statistic to the distribution obtained by simulating many values of the test statistic from this estimated null model.

When $\tau > 0$, incorrectly fitting the null model will result in an underestimate of M. Intuitively, this is because the initial 'lack of migration' which τ is causing is wrongly ascribed to the effects of a low migration rate. The left column of figure 2.7 illustrates this phenomenon in the known-history case. From figure 2.6 it is clear that a consequence of this underestimation is that the test based on estimating M under the null model will be conservative: the observed value of the test statistic will be compared to a distribution that is more similar to the distribution under the alternative than it would be were the true value of M known. Thus estimating M under the null results in a decrease in power.

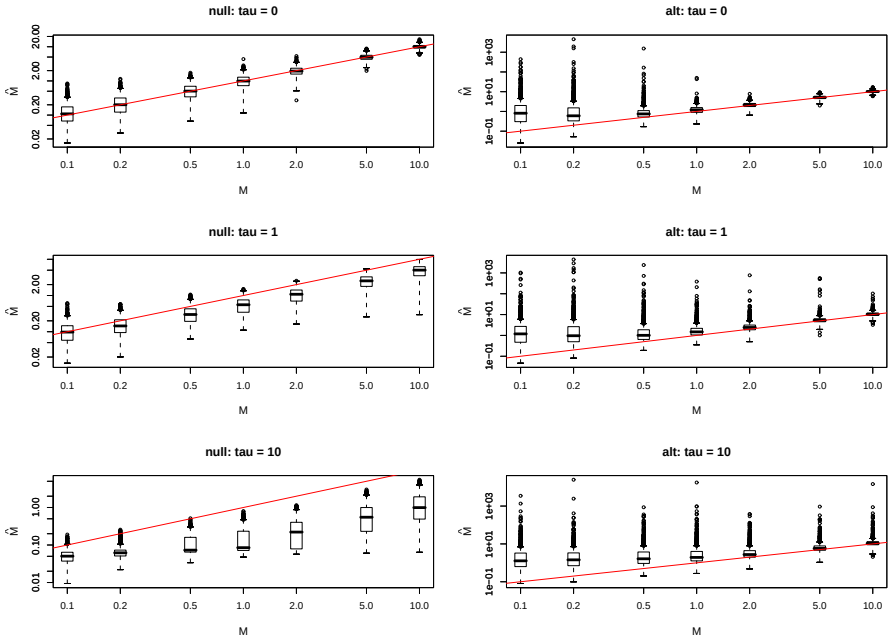
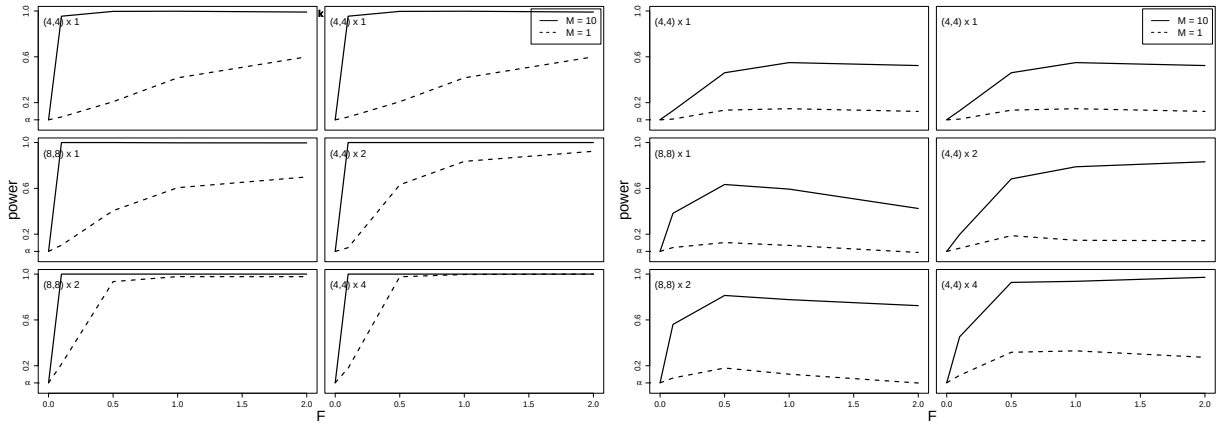


Figure 2.7: **Distributions of known-history $\widehat{M}$.**
True M values along x-axis and estimates along y-axis (log-log scale). Each row corresponds to a different true value of τ. The left column shows $\widehat{M}_0$, computed assuming that $\tau = 0$, whereas the right column shows $\widehat{M}$, computed assuming that $\tau = t_m$, i.e. its MLE. The line $y = x$ is shown in red. Note that incorrectly fitting the null model results in an underestimate of M, the magnitude of which increases with τ. Fitting the alternative model (correctly or not) results in an overestimate of M, because the estimate of τ is always larger than the true value of τ, although the difference is small when the migration rate is high.



(a) known-history (b) known-genealogy

Figure 2.8: **Effect on power of the number of chromosomes and number of independent loci sampled**

For the known-genealogy tests (right panels of figure 2.6), the migration events themselves are not observed. While large values of λ do tend to result when τ is large, they are also common under the null hypothesis. Loosely speaking, under the equilibrium migration model with $M = 1.0$, many genealogies are generated which rather strongly suggest the non-equilibrium isolation model. Although the known-genealogy distribution of λ is seen to be centred on larger values for larger values of τ, the variance of the four distributions illustrated varies in a rather unpredictable manner. This seems to reflect some dependence of the accuracy of estimation of the likelihood ratio on the location of the MLEs, perhaps introduced by undesirable properties of the algorithm used to locate maxima on the (noisy) estimated likelihood surface.

Figure 2.8 illustrates the effect on power of varying the migration rate, the sample size and the number of independent loci. In this figure a test statistic was deemed significant if it fell within the 5% upper tail of the distribution obtained under the null using the true value of the migration rate. As discussed above, this procedure results

in a more powerful test than if the migration rate under the null were estimated from the data, as it would be in practice. When the migration rate is high ($M = 10.0$; solid lines), the known-history test has high power against the most recent isolation time investigated ($\tau = 0.1$), even for a sample size of four in each population with the history known at just one locus (top left panel of figure 2.8a). The corresponding known-genealogy power (top left panel of figure 2.8b) is much lower, and is not greatly improved by doubling the sample size. With a high migration rate, the known-genealogy test has high power comparable to that of the known history test only when the genealogy is known at 4 independent loci, and then only to detect isolation times of N_e generations or more. When the migration rate is lower ($M = 1.0$; broken lines) even the known-history test fails to have high power unless histories at 2 or more independent loci are known. In the known-genealogy case with $M = 1.0$ neither a sample of size 16 at two independent loci, nor a sample of size 8 at four loci, results in sufficiently high power to detect isolation for it to be attractive to attempt to do so experimentally.

As discussed above, power in figure 2.8 is higher than would result in the more realistic situation wherein the migration rate under the null is estimated from the data. Figure 2.9 illustrates the effect on known-history power of estimating the migration rate. The unknown M test has the somewhat counterintuitive property that the power is not an increasing function of τ. Instead the power peaks at an intermediate value of τ and then declines to an asymptotic value. This seems to be the result of the positive relationship between τ and the degree of upward bias of $\widehat{M_0}$ illustrated in figure 2.7. When τ is large, looking backwards in time, only a single lineage remains on each side of the barrier at time τ. Consequently there are few migration events in the history, the estimate of M is much lower than the true value,

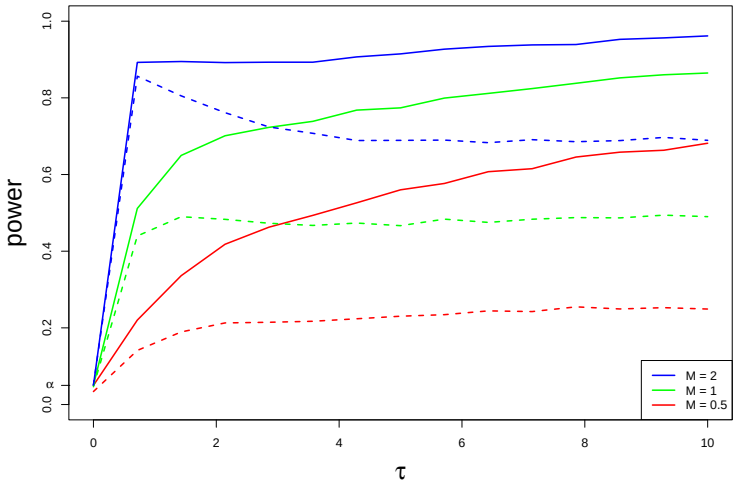


Figure 2.9: **Effect of estimating the migration rate on power of the known-history test**
M known (solid line), M estimated (dashed line)

and the test statistics are compared to a null distribution which is much more similar than it ought to be to the distribution under the alternative.

Code in R (R Development Core Team 2006) (with dependent code in C implementing the simulations and importance sampling scheme) to perform these analyses is available from the author.

2.5 Discussion

Much of the literature on using linked genetic data for inference of population history does so without explicit reference to a stochastic model of the process from which the data were sampled. The methods of parametric statistical inference developed during the twentieth century are in principle appropriate here, but such studies do not follow this paradigm. This is, understandably, especially true of studies in which

DNA sequences are obtained from diverse geographic locations, because of the extra difficulty of modelling such data. Part of the reason is that, because of the linkage, an alignment of haplotypes cannot be summarised in a lower-dimensional form without losing much of the information in the data. The full likelihood treatments outlined in chapter 3 section 3.4.1 would avoid such information loss, but their correct implementation for any particular model requires devotion of considerable time to a somewhat eccentric computer programming problem. Furthermore, although *a priori* plausible models of population history will vary widely from study to study, especially when samples are available from diverse geographic locations, full likelihood methods tend to assume fairly simple models and offer little flexibility. Nevertheless, as Pritchard *et al.* (2000) point out, "in situations where these assumptions are deemed unreasonable, alternative models should be built."

The main consequence of this departure from standard statistical practice is that no formal assessment is made of the uncertainty about the matters of interest that remains after the data have been analysed. It is therefore difficult to assess the results of a single study, and difficult to compare and use the results of different studies. This is especially problematic in the analysis of completely linked data since the correlation in the data implies that wide confidence intervals, or high posterior uncertainty, may be a worry.

This situation may be contrasted with that concerning inference of population history from unlinked genetic data. In this case the data can be summarised in a fairly low dimensional form (typically, counts of alleles in different populations at each locus), and the technical aspects of the statistical problems tend to lie closer to the mainstream. As a consequence, much work has been done on developing

statistical methods for inference of population structure and history from unlinked data (e.g. Pritchard *et al.* 2000, Chikhi *et al.* 2001, Nicholson *et al.* 2002, Corander *et al.* 2003, Falush *et al.* 2003, Holsinger & Wallace 2004). In these works, as in this chapter, population structure is discrete: the models feature systems of panmictic populations, which may be explicitly connected by direct ancestry or by gene flow. Statistical models for unlinked data have also been developed which model variation in allele frequency across continuous space (see Wasser *et al.* 2004, Vounatsou *et al.* 2000) and that incorporate the sample locations in continuous space into the prior on cluster membership when inferring population structure (Guillot *et al.* 2005).

Of course, as discussed in section 3.3, limiting oneself to unlinked data is a severe limitation, and the difficulty and desirability of analysing linked data has not gone unrecognised. Recently much effort has been put into developing full likelihood approaches to making inferences about structured populations from completely linked data (e.g. Kuhner *et al.* 1995, Beerli & Felsenstein 1999, Bahlo & Griffiths 2000, Nielsen & Wakeley 2001, Drummond *et al.* 2002, de Iorio & Griffiths 2004, Hey & Nielsen 2004, Ewing *et al.* 2004). These methods represent very useful advances whenever it is reasonable to treat the data (or a subset of the data) as having been sampled from the models implemented. One consequence of the computational requirements of these methods is that relatively few studies of the statistical properties of the likelihood-based estimators under these models have been made; for more complex models of population history for which likelihood-based methods are lacking, no such studies have been made.

In this chapter I have attempted to fill this gap to some extent by investigating the statistical properties of an estimator of the rate of migration relative to drift and

of hypothesis tests for isolation, given completely linked population genetic data. I have proceeded under the assumption that the genealogy is completely known, with branch lengths measured in units of $2N_e$ generations. The reason for this approach is that it represents the maximum possible information available from completely linked data, the results are independent of details of the mutational process, and it simplifies the problem of evaluating the likelihoods. It is however still non-trivial to evaluate the likelihoods. In principle this involves integrating over the (infinitely-many) possible sequences of migration events in the ancestry of the sample, and I developed a new importance sampling algorithm to approximate this integration. In addition I have compared results obtained under the assumption that the sequence of migration events in the ancestry of the sample is known.

2.5.1 Estimating M under equilibrium

The known genealogy estimator of $M = 4N_em$ under equilibrium migration has a very high variance and tends to greatly exceed the true value (figure 2.5). Even if the bias in the estimator were accounted for, any single estimate would be consistent with a wide range of true values of M. The results obtained here thus suggest that it is impossible to estimate $M = 4N_em$ well using completely linked data such as mitochondrial genomes, at least with sample sizes similar to those investigated (up to 8 from each population). Hudson *et al.* (1992) studied the statistical properties of two estimators of M based on summaries of the data: one estimator uses the numbers of pairwise differences between haplotypes, and the other is based on reconstructing the tree and ancestral migration events via a parsimony algorithm (Slatkin & Maddison 1989). Beerli & Felsenstein (1999) and Ewing *et al.* (2004) have implemented full likelihood based approaches for estimating M from completely linked data, and found

that it tends to be overestimated and that there is little relatively little information in the data about the parameter. I have implemented a Metropolis-Hastings Markov chain Monte Carlo algorithm to sample from the posterior density of M under an equilibrium model given completely linked sequence data, and investigated frequentist properties of a simple Bayes estimator (the posterior mode with a uniform prior on M). Its distribution was also centred well above the true values of M (Davison unpublished).

In studies of population history the main concern may be testing equilibrium versus non-equilibrium models, identifying zones across which differentiation is particularly high or dating times of population divergence. In such cases the precise values of parameters such as M could be argued to be little interest themselves, and the poor performance of these likelihood-based estimators of M might therefore not seem particularly troubling at first sight. However, statistical models that permit the desired questions to be answered will always have to model the effects of limited dispersal, and the poor statistical properties of the completely linked data estimator of M under equilibrium, even in this maximum information limit, suggests that fitting more complex models will require considerable care. Furthermore, those studying the evolution of reproductive incompatibility among sympatric or parapatric organisms propose a research program involving identifying loci at which gene flow is unusually low by the standards of the rest of the genome (e.g. Wu 2001). The upwards bias and high variance of the estimator at a single locus indicate that this will be a harder statistical problem than one might hope.

2.5.2 Testing for isolation

Because of its connection to the concept of diversification ('speciation') in evolutionary biology, there is a substantial existing literature on making inferences about isolation of populations from population genetic data (e.g. Watterson 1985, Wakeley 1996a, Nielsen & Wakeley 2001, Nicholson *et al.* 2002, Rosenberg & Feldman 2002, Leman *et al.* 2005, and chapter 3). This has largely focused on models in which a sample is taken from each of two unstructured populations, which are assumed to have descended from a single unstructured ancestral population. I will refer to such models as 'split models' in order to distinguish them from the 'barrier models' studied in this chapter — the difference is that in the barrier model, the ancestral population is structured. In the split models, as in the barrier model, interest centres on estimating the time since the separation of the populations in units which depend on the historical effective sizes of the populations.

These two-population models are clearly very restrictive — in principle it would be of great interest to study multiple descendent populations, and the correlations between them that are induced by differing amounts of shared ancestry (the 'population tree'), differing amounts of gene flow, or some combination of these (Felsenstein 1982). Nicholson *et al.* (2002) study a model of unlinked data from multiple descendent populations, and discuss how it might be generalised to study such correlations, and Nielsen *et al.* (1998) describe a model featuring a 'population tree'. Nevertheless, it is helpful to first focus on the two-population situation studied in this chapter and in the majority of the existing literature.

In the simplest case of the split model, the daughter populations are completely isolated — there is no gene flow between them subsequent to their separation. In

this case one is essentially estimating a parameter which corresponds to the 'amount of drift' that has occurred since separation (F in chapter 3; see Nicholson *et al.* (2002)). Notice that in these models, any population structure is assumed *a priori* to be the result of isolation, rather than limited but ongoing gene flow. In general however the existence of population structure is rarely in doubt, especially when study populations extend over an area which is large relative to the dispersal capabilities of the organisms (or their propagules). Therefore, the real challenge is not merely to perceive some structure and fit it to a model of isolation (although, as in chapter 3, that might be a helpful first step in developing methods), but rather to learn about the role of recurrent processes (gene flow) and unique historical events (splits/barriers) in creating the observed structure.

In more general terms, one can always fit a statistical model to observed data; this activity on its own tells one nothing about the degree to which the model is an accurate representation of the truth. One way to proceed beyond merely imposing a model on data is to assess how *well* the model fits (in statistical jargon, to examine the 'residuals'). Nicholson *et al.* (2002) provide an example of this approach when fitting their model of isolation. Preferably however, one would study models representing a variety of plausible alternatives and use the data to characterise ones uncertainty about them. The traditional way to do so is to specify an alternative model (hypothesis) which may feature several unknown parameters, and also to specify a null model in which some of those parameters are fixed at particular values, representing a simpler explanation of the data. Because the null model is a special case of the alternative, the maximum likelihood under the alternative will be at least as large as that under the null; the question is whether the difference is such that the alternative should be preferred.

The barrier model studied in this chapter represents one attempt to provide such a framework, and I am aware of three others in the literature. The most widely-used is the 'isolation ('split' in my terminology) with migration' (IM) model of Wakeley (1996a) and Nielsen & Wakeley (2001), which is discussed below. In addition Ciofi *et al.* (1999) developed a method for unlinked data that compares the fit of the data to a model of equilibrium migration with that to a model of pure isolation. Their work is based on the multinomial-dirichlet likelihood for multiallelic unlinked data (beta-binomial in the diallelic case) developed by Balding & Nichols (1995) (see also Rannala & Hartigan (1996), Beaumont (2001), Balding (2003)).

Alan Templeton and colleagues have developed an ambitious non-parametric procedure to make a wide range of inferences about population structure and non-equilibrium situations on the basis of linked data (Templeton *et al.* 1995, Templeton 1998) which has been very popular in the 'phylogeographic' community, where the lack of parametric statistical procedures is felt most acutely. The method is based on using information on sampling locations to summarise an estimated haplotype network and assessing significance of the resulting statistics by permutation tests. However, at various points in this procedure subjective decisions are required on the part of the user, and as a consequence it is not straightforward to write a computer program to compute the test statistics and assess their significance, and in fact none is currently publicly available. The lack of automation and idiosyncratic nature of the 'inferences' makes assessment of the statistical properties of this method very difficult (but see Panchal & Beaumont (in prep.)).

The 'isolation with migration' (IM) model (Wakeley 1996a, Nielsen & Wakeley 2001, Hey & Nielsen 2004) is not the same as the one which is studied here. The basic IM

model can be outlined as follows: looking backwards in time, for some unknown period τ, two populations of unknown effective size exchange migrants at an unknown rate. Prior to that they form a single panmictic population of unknown effective size. If the migration rate in the IM model is fixed at zero, it becomes similar to the barrier model but with the difference that in the IM model the ancestral population is unstructured. Essentially the difference between the two models is that the IM model features an unstructured ancestral population, and a recent structured era of unknown length with migration at unknown rate; the barrier model features a structured ancestral population with migration at unknown rate, and a structured recent era of unknown length with zero migration.

In this chapter I have taken the view that the null model ought to be the model of equilibrium between gene flow and drift, and that the question is whether the data warrant the conclusion that there has additionally been a recent barrier to gene flow and therefore that the system is not at equilibrium. This choice reflects the fact that frequently we do not doubt that limited gene flow is resulting in population structure: what we are actually interested in is whether or not historical events have perturbed that equilibrium. It is interesting to contrast this with the IM framework, under which two equilibrium models can be fitted — panmixia ($\tau = 0$) and equilibrium migration ($\tau = \infty$). The non-equilibrium alternative in the IM framework assumes that the ancestral population is panmictic, and therefore fitting the IM model to data from the barrier model would presumably result in overestimation of τ. Which model is more appropriate in any particular case depends on the populations sampled. For example, in previously glaciated northern Europe, currently allopatric or parapatric populations have recently colonised their current range and may descend from the same refugial population during the last glaciation, in which case the assumption of

an unstructured ancestral population may be a good one. In contrast, in Amazonia populations may have been geographically structured since prior to the MRCA at most or all loci.

Two equal-sized populations connected by symmetric migration at a very high rate behave like an unstructured population with effective size equal to the sum of the two effective sizes. Therefore the IM model, the barrier model and other statistical procedures for fitting models of equilibrium migration and isolation, are different ways of investigating the following question. Has gene flow between the two populations occurred homogeneously in time (null), or are the data better explained by a recent period of low migration (alternative)? The motivation for the question is obvious: rejection of the null hypothesis implies that something has happened to reduce gene flow between the two populations. In allopatry and parapatry this could be a reduction in the rate of movement of individuals or propagules between them, and in parapatry and sympatry it could be that other pre- or post-mating barriers to gene flow have arisen or increased in strength. The results in this chapter indicate that in many situations, when using completely linked data such as mitochondrial genomes, there is little power to reach these important conclusions.

2.6 References

Avise, J. (2000). *Phylogeography*. Harvard University Press.

Bahlo, M. & Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol* **57**, 79–95.

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* **63**, 221–230.

Balding, D. J. & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**, 3–12.

Barton, N. & Wilson, I. (1996). Genealogies and geography. In *New uses for new phylogenies* (edited by P. H. Harvey, A. J. L. Brown & J. M. Smith). Oxford University Press.

Beaumont, M. (2001). Conservation genetics. In *Handbook of Statistical Genetics* (edited by D. Balding, M. Bishop & C. Cannings), chapter 29, pages 779–809. Wiley.

Beerli, P. & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–73.

Beerli, P. & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A* **98**, 4563–8.

Cavalli-Sforza, L. L. & Piazza, A. (1975). Analysis of evolution: evolutionary rates, independence and treeness. *Theor. Pop. Biol.* **8**, 127–165.

Chen, F.-C. & Li, W.-H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456.

Chikhi, L., Bruford, M. W. & Beaumont, M. A. (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* **158**, 1347–1362.

Ciofi, C., Beaumont, M. A., Swingland, I. R. & Bruford, M. W. (1999). Genetic divergence and units for conservation in the komodo dragon *Varanus komodensis*. *Proc. Roy. Soc. Lon. B* **266**, 2269–2274.

Corander, J., Waldmann, P. & Sillanpaa, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.

Davison, D. (unpublished) .

de Iorio, M. & Griffiths, R. C. (2004). Importance sampling on coalescent histories ii: subdivided population models. *Adv. Appl. Probab.* **36**, 434–454.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320.

Ewing, G., Nicholls, G. & Rodrigo, A. (2004). Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics* **168**, 2407–2420.

Falush, D., Stephens, M. & Pritchard, J. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–87.

Fearnhead, P. & Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–318.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376.

Felsenstein, J. (1982). How can we infer geography and history from gene frequencies? *J. Theor. Biol.* **96**, 9–20.

Guillot, G., Estoup, A., Mortier, F. & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.

Hey, J. & Machado, C. (2003). The study of subdivided populations — new hope for a difficult and divided science. *Nat. Rev. Genet.* **4**, 535–543.

Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–60.

Holsinger, K. E. & Wallace, L. E. (2004). Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (Orchidaceae). *Mol Ecol* **13**, 887–894.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. *Oxford Surveys in Evoutionary Biology* **7**, 1–44.

Hudson, R. R., Slatkin, M. & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589.

Knowles, L. & Maddison, W. (2002). Statistical phylogeography. *Mol. Ecol.* **11**, 2623–2635.

Kuhner, M., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421–30.

Leman, S., Chen, Y., Stajich, J., Noor, M. & Uyenoyama, M. (2005). Likelihoods from summary statistics: recent divergence between species. *Genetics* **171**, 1419–36.

Nath, H. B. & Griffiths, R. C. (1993). The coalescent in two colonies with symmetric migration. *J Math Biol* **31**, 841–851.

Nicholson, G., Smith, A., Jónsson, F., Gústafsson, O., Stefánsson, K. & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society B* **64**, 695–715.

Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. & Slatkin, M. (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677.

Nielsen, R. & Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–96.

Nordborg, M. (2001). *Handbook of Statistical Genetics*, chapter Coalescent theory. Wiley.

Notohara, M. (1990). The coalescent and the genealogical process in geographically structured population. *J Math Biol* **29**, 59–75.

Panchal, M. & Beaumont, M. (in prep.) .

Patterson, N., Richter, D., Gnerre, S., Lander, E. S. & Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature* Advanced online publication.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rannala, B. & Hartigan, J. A. (1996). Estimating gene flow in island populations. *Genet Res* **67**, 147–158.

Rosenberg, N. A. & Feldman, M. W. (2002). The relationship between coalescence times and population divergence times. In *Modern developments in theoretical population genetics: the legacy of Gustave Malécot* (edited by M. Slatkin & M. Veuille), chapter 9. Oxford University Press.

Slatkin, M. & Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613.

Stephens, M. & Donnelly, P. (2000). Inference in molecular population genetics. *Philosophical Transactions of the Royal Society Series B* **354**, 1–31.

Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genet Res* **52**, 213–222.

Templeton, A. (1998). Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Mol Ecol* **7**, 381–97.

Templeton, A., Routman, E. & Phillips, C. (1995). Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140**, 767–82.

Vounatsou, P., Smith, T. & Gelfand, A. E. (2000). Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics* **1**, 177–189.

Wakeley, J. (1996a). Distinguishing migration from isolation using the variance of pairwise differences. *Theor Popul Biol* **49**, 369–386.

Wakeley, J. (1996b). Pairwise differences under a general model of population subdivision. *Journal of Genetics* **75**, 81–89.

Wakeley, J. (2004). Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. In *The evolution of population biology* (edited by R. S. Singh & M. K. Uyenoyama), chapter 10, page 193. Cambridge University Press.

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B. & Stephens, M. (2004). Assigning African elephant DNA to geographic region of

origin: applications to the ivory trade. *Proc Natl Acad Sci U S A* **101**, 14847–14852.

Watterson, G. A. (1985). The genetic divergence of two populations. *Theoretical Population Biology* **27**, 298–317.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of evolutionary biology* **14**, 851.

CHAPTER 3

A NEW APPROXIMATE LIKELIHOOD FOR GENETIC DATA UNDER A MODEL OF POPULATION SPLITTING

ABSTRACT

I describe a new approximate likelihood for recombining genetic data under a model in which a panmictic ancestral population has split into two panmictic daughter populations, from which the samples were taken. The approximate likelihood is based on the 'PAC likelihood' of Li and Stephens (2003 *Genetics* 165:2213). It may be used for efficient approximate likelihood-based analyses of unlinked data, but the most important application is to loosely linked haplotype data for which statistical models explicitly featuring non-equilibrium population structure have so far been unavailable. I place the work in context with an initial discussion of statistical aspects of some different types of population genetic data that are available, and of likelihood-based methods in population genetics. The method may be applied to both resequenced data and to SNP data, and I demonstrate its application to simulated SNP data sets and to SNP data from human populations in Amazonia and Siberia. The simulations indicate that the new PAC likelihood successfully makes use of information contained in haplotype structure, giving rise to a more efficient estimator of the amount of drift since the split than one which ignores the linkage. However biases are observed when the daughter populations have drifted by unequal amounts, and these are problematic when analysing the human data as a result of strong drift associated with colonisation of the Americas. Although the details of the implementation described here will need to be refined, this work demonstrates that the PAC approach may be used for efficient approximate likelihood-based inference from recombining data in models of more complex population histories than constant-sized panmixia.

3.1 Introduction

A central aim of population genetics is to develop effective methods for learning about the structure and evolutionary history of populations. The importance of this challenge stems from the fact that these phenomena, which affect patterns of variation genome-wide, are not only of intrinsic interest themselves, but also represent critical modelling assumptions when using the data to answer questions about evolutionary forces acting on particular regions of the genome. Thus a statistical problem in population genetics is of considerable importance indirectly in medical and molecular evolutionary genetics, as well as more directly in the inference of recent evolutionary history, low-level systematics, historical biogeography, landscape history, conservation biology and in areas of ecology that adopt a broad spatial or explicitly historical perspective.

The collection of models of population structure and history which might be considered plausible *a priori* in any particular instance is potentially unlimited and much work on this model selection problem in general remains to be done, despite a large literature spanning several decades. However, when attention is restricted to certain models, or a collection of closely related models, there has been substantial progress. Two fundamental distinctions are helpful when formulating inference problems in population genetics. One is between models which explain the data as having been sampled from a population genetic process at equilibrium, and models in which the population genetic process has been perturbed by some event, and has yet to reach its new equilibrium. The other is between models in which all the sampled individuals are exchangeable, and those in which they are not, as a result of some form of population structure.

Population structure means that the data have not been sampled from a single panmictic population; rather, some labelling of the individuals exists such that individuals with the same (or similar) labels are on average more closely related than individuals with different (or dissimilar) labels. The labelling might be strongly influenced by geographic location, or by some other property of the individuals. Each group defined by the labels represents a larger collection of individuals, from most of which typically no data are available, which I will call a 'population'. It is often the case that some degree of population structure is anticipated, or evident from informal examination of the data, especially when samples are available from different geographic locations.

This chapter focuses on the simple case in which the existence of two populations is hypothesised. The hypothesis includes a specification of which of the sampled individuals belongs to each population. Questions of interest concern whether or not the hypothesis is correct, and if so, what the cause of the structure is. A simple equilibrium explanation for such structure is that movement of genetic material between the populations occurs at a low rate, homogeneously through time. I will refer to this as 'equilibrium migration'. A simple non-equilibrium explanation is that there has been no recent movement of genetic material between the two populations, but that the two populations contain descendents of a single unstructured ancestral population (see figure 3.1). I will refer to this as 'isolation without gene flow', or simply 'isolation'.

In the context of these two-population models the importance of the equilibrium versus non-equilibrium distinction is clear: under isolation the two populations henceforth evolve in complete independence; a novel mutation that arises in one of the

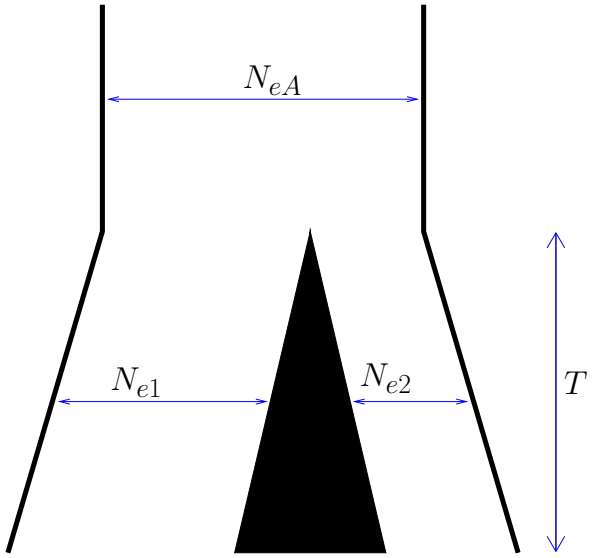


Figure 3.1: **The isolation without gene flow model.**

populations may increase in frequency to fixation in that population, but may not spread, even if it would be strongly selected in both populations. Thus the two populations can follow independent adaptive evolutionary paths. The same is not true under equilibrium migration: even a low rate of migration is likely to lead eventually to the presence of a positively selected mutation in both populations. Additionally, if the reason for the low or absent gene flow is not reproductive incompatibility of the organisms in sympatry but rather is due to some extrinsic factor such as a physical barrier to dispersal, then whether gene flow is nil or merely low can be important in determining whether the populations evolve a genetic, and hence permanent, basis for reproductive incompatibility.

The statistical problems presented by this situation, in which the existence of two populations, and population assignment of the sampled individuals, are hypothesised, include:

1. Can a model of a single unstructured population be rejected? ('testing panmixia')

2. What is to be preferred as an alternative model, and what are the values of the model parameters? ('model selection')

Problem 1 is the simpler, but although it is straightforward to simulate neutral population genetic data under a wide variety of models, the hypothesis test is not trivial because the null hypothesis (panmixia) has at least one unknown parameter (the relative rates of mutation and drift). Problem 2 is harder. A variety of approaches to these problems have been proposed in the literature, which depend on the nature of the data that are available.

In this chapter I describe a new statistical method for estimating the parameters of the isolation without gene flow model. In contrast to existing methods, the new method can make use of information about haplotype structure (see section 3.3.2) in the presence of recombination. In section 3.2 I describe the model of population history more precisely. In order to motivate the new method introduced in this chapter, and to discuss its novel aspects, section 3.3 discusses the sorts of data that are available, and the ways in which they alter the statistical problem. In section 3.4.1 I describe the new method. Section 3.5 investigates the properties of the new method by applying it to data sets simulated under known population histories, and also applies it to recently-obtained single nucleotide polymorphism data from human populations in Siberia and Amazonia. In section 3.6 I discuss the results and their significance for the inference problem outlined above.

3.2 The model of population history

In the model (figure 3.1) an ancestral population T generations ago was at equilibrium under neutral mutation, recombination and drift, with effective size N_{eA}. At that time it split into two daughter populations which subsequently evolved under those three processes without gene flow between them. A quantity of central interest is the time at which this separation occurred. Neutral population genetic data contain information about this time in the form of information about the amount of drift that has occurred in each population since their separation. The amount of drift in the two daughter populations are parameters of the model which will be denoted F_1 and F_2. If the daughter populations have had constant effective diploid sizes N_{e1} and N_{e2}, then $F_1 = T/2N_{e1}$ and $F_2 = T/2N_{e2}$. In this case, $1/2N_{ei}$ can be thought of as being the instantaneous 'rate of drift' in population i, and F_i as equal to this instantaneous rate multiplied by the length of time T over which drift has occurred. If the effective population sizes have not remained constant then neither has the rate of drift and the relationship between the two F parameters and the time T will depend on the history of the effective sizes through time. The model of Nicholson *et al.* (2002) for unlinked data features analogous parameters to the F_i, and that work should be consulted for a much more careful description of their interpretation, as well as their relation to summary statistics such as F_{st}. When studying isolated populations one would like to know the time since their separation in generations (and therefore years). However, in the absence of external information about the effective sizes, or about per-generation probabilities of mutation or recombination, it is possible only to learn about the F parameters; not about T or the N_es separately.

The results in this chapter (but not the actual implementation) assume a restricted model in which the larger of the effective sizes of the two daughter populations is the same as the ancestral effective population size. Suppose that a sample of DNA sequences is available from some region of the genome in which each generation the probability of recombination and mutation are respectively r and u. Recombination and mutation events occur uniformly throughout the genomic region. In this case the model is (making the large population approximations that are standard in coalescent theory) completely specified by F_1 and F_2, and parameters $\rho = 2N_{e1}r$ and $\theta = 2N_{e1}u$ which specify respectively the rates of recombination and mutation relative to the rate of coalescence (drift) in the ancestral (and larger daughter) population (note that these parameters are often defined as twice their value here). I will sometimes use a different parameterisation in which F without a subscript refers to F_1 and α is the ratio F_1/F_2 of the amounts of drift experienced by the two populations. If the daughter populations have had constant effective sizes since the split then α is equal to the ratio N_{e2}/N_{e1} of these sizes.

3.3 Population genetic data

There are various types of population genetic data, and they have a more complicated probabilistic structure than the independent and identically distributed (i.i.d.) data which are the subject of classical statistics. All types of population genetic data comprise the alleles present at a sample of loci in a sample of individuals. I assume that the number of individuals to be sampled from each hypothesised population is fixed in advance, and these are then sampled at random within the respective populations. What differs between different data types, and has consequences for the statistical analysis, is the way in which the loci are chosen. Although nuclear loci

are diploid in most organisms, I assume that the sampled populations are at Hardy-Weinberg equilibrium, which has the consequence that all n copies of a chromosome in a sample of size $n/2$ individuals are exchangeable. Therefore from now on I will refer to sampling chromosomes rather than individuals.

At a single nucleotide site, the sampled chromosomes are related by some unobservable genealogy; this genealogy must therefore be treated as a random variable. I will make the 'infinite sites' assumption that the mutation rate per site is small and so the probability of more than one mutation occurring on the genealogy at a single site may be neglected. Therefore the site will be polymorphic if a mutation occurs on the genealogy, and the pattern of occurrence of the derived and mutant alleles at that site among the sampled chromosomes is determined by the branch of the genealogy on which the mutation occurs. Fundamentally, population genetic data contain information about the genealogies relating the sampled chromosomes at the sampled loci. Since the prior probability distribution on genealogies depends on the model of population history and the values of the model parameters, the data contain information about those models and parameters (see e.g. Felsenstein 1992, Nordborg 2001, Stephens 2001). Basically, a model has a high likelihood if it tends to generate genealogies on which the data have high probability.

3.3.1 Unlinked data

Because of the underlying genealogy, the *alleles* sampled at a single site are not independent of each other. However, when two *sites* are 'unlinked' the genealogies (and therefore the patterns of variation) at them *are* independent. Sites are 'unlinked' if they are on separate chromosomes, and also if they are on the same chromosome but sufficiently far apart for recombination to eliminate the correlation between them.

Under a model of a single unstructured population, unlinked data can be summarised without loss of information by the (unfolded) 'sample frequency spectrum'. Given that a sample of size n is polymorphic for two alleles at some site, those alleles may be present in any number of copies from 1 to $n-1$. The unfolded version records, for each possible allele count, the number of loci at which the derived allele has that count. The folded version is appropriate when allele polarity is unknown, and pools counts in bins i and $n-i$ in the unfolded version. Under the two population model, since the chromosomes within each population are exchangeable, the data may be summarised by recording, for each possible ordered pair of allele counts, the number of sites showing that pair of counts in the two populations. Figure 3.2 illustrates joint probability distributions on these ordered pairs of allele counts that result after four different amounts of drift since the population separation. I will refer to this data summary as the 'bivariate sample frequency spectrum'.

In order to understand the information contained in data about the parameters of the isolation model, consider a random genealogy which describes the relationships at some locus of a sample of chromosomes from the two daughter populations. For simplicity, assume that the daughter and ancestral populations have all been of constant and equal effective size N_e. When $F = T/2N_e$ is large, many coalescence events will occur in the daughter population, and few ancestral lines will survive to the ancestral population. Considering the entire population forwards in time, this corresponds to the observation that a mutation that was segregating at some frequency in the ancestral population is likely to be at quite different frequencies in the daughter populations at the time of sampling (figure 3.2c), and may well have been fixed or lost in one or other or both (figure 3.2d). In this case one says that there has been a large amount of drift since the split. For large F, most of the polymorphic sites may

result from mutations that have occurred since the split and are segregating in one population only (figure 3.2d).

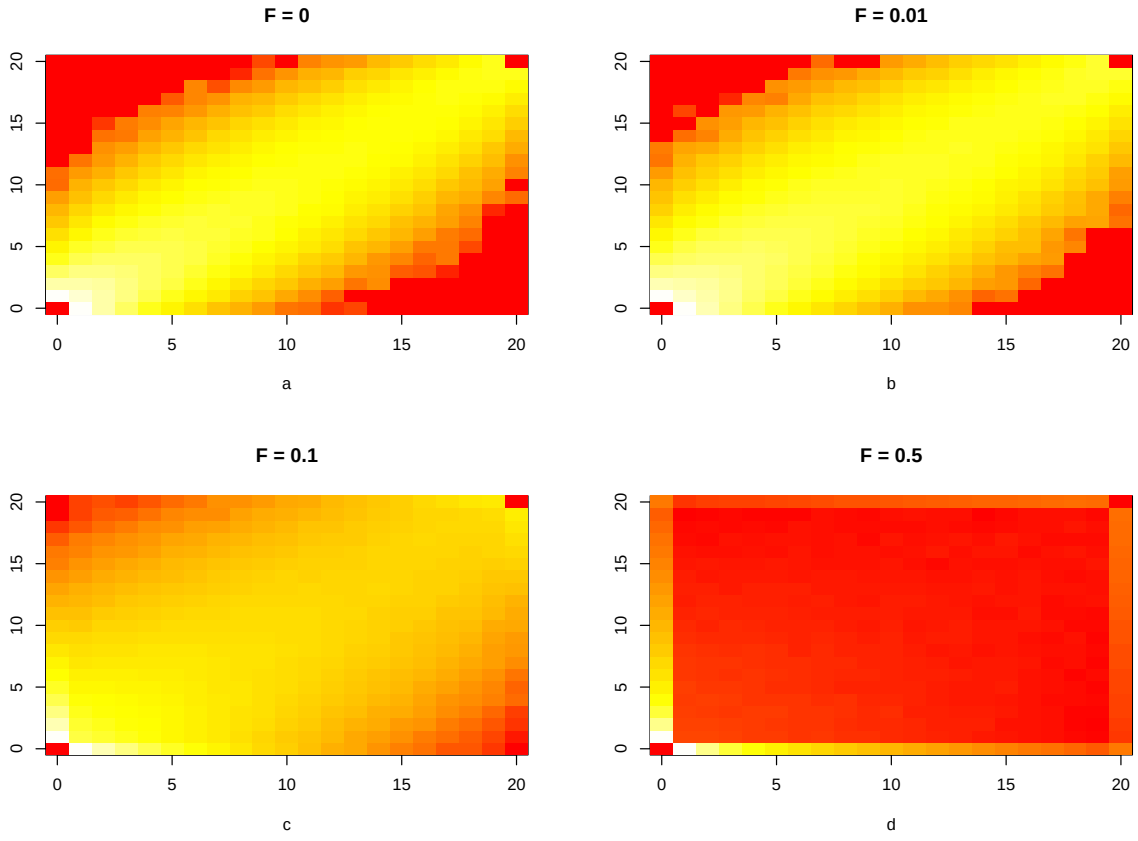


Figure 3.2: **Joint probability distributions on pairs of derived allele counts under isolation without gene flow.**
The colour of cell (i, j) represents the probability with which there are i copies of the derived allele in the sample from one population and j copies in the sample from the other population, conditional on polymorphism in at least one of the populations. Brighter colours represent higher probability. The probabilities were estimated by simulation. The sample size is 20 alleles from each population. The amount of drift F is equal in the two daughter populations. Note that when the samples are drawn from an unstructured population ($F = 0$), the allele counts in the two populations are highly correlated. As the amount of drift increases, this correlation decreases. When $F = 0.5$, the majority of the mutations are segregating in one population only.

Summary statistics such as F_{st} capture information about the extent to which allele frequencies differ between the populations, and methods have been developed to address various aspects of the model selection problem (problem 2 above) given unlinked data, based on summary statistics (Watterson 1985, Gaggiotti & Excoffier 2000, e.g.) or on full data likelihood (Nielsen & Slatkin 2000, Nielsen *et al.* 1998, Beaumont 2001, Nicholson *et al.* 2002, e.g.). Furthermore, since the summarised data are fairly low-dimensional, the likelihood of any point in parameter space can be estimated simply by performing many Monte Carlo simulations. However, collecting unlinked data does not make use of the full potential of genomic variation to inform us about population history.

3.3.2 Completely linked and loosely linked data

Although sample frequency spectra do contain valuable information under many models of population history, for the model selection problem more detailed information about each genealogy would be very valuable. In particular, one would like to learn about the distribution of coalescence times for the three classes of pairwise comparisons of chromosomes (within each population, and between populations). Informally, one can imagine that the distribution of between-population coalescence times would be quite informative about the model selection question. Looking backwards in time, under pure isolation, between-population coalescence events are impossible prior to the ancestral population. When the ancestral lines reach the ancestral population they become exchangeable and every pair coalesces at a rate inversely proportional to the effective size of the ancestral population. Therefore, under isolation without gene flow, the distribution of between-population coalescence times is equivalent to the distribution of times of these ancestral coalescence events. In contrast, under

isolation with a low but non-zero rate of gene flow, between-population coalescences derive from a mixture of two distributions: in addition to the coalescences in the ancestral population, there is a new component deriving from occasional instances of a lineage migrating and coalescing in the other population. For the same ancestral effective size, the variance of the between-population coalescence times will be higher, although the shape of the distribution will depend on the rate of migration. And with the same migration rate under the equilibrium migration model, the variance will be still higher, and the rate of between-population coalescence will not reflect a sudden pooling of ancestral lines into a single panmictic population.

In this chapter I do not address this general model-selection problem, but instead focus on estimation of the parameters of the isolation without gene flow model. Under this model, it is true of *all* coalescence events between lines with descendents in different populations that they occur at some time $F + t_A$, where t_A is the (random) coalescence time in the ancestral population, in units of $2N_{e1}$. Therefore, if a statistical model for t_a is assumed (e.g. the panmictic equilibrium coalescent), learning about the times of such between-population coalescence events corresponds to learning about F. In order to obtain this sort of more detailed genealogical information under infinite-sites assumptions, polymorphism data at several nearby sites are required. In principle, such data contain information about genealogies deriving from both mutation and recombination events in the history of the sample. One informal way of describing the information available about the parameters of the isolation model is that between-population comparisons of chromosomes contain information about the quantities $uT = \theta F$ (from mutation) and $rT = \rho F$ (from recombination), while within-population comparisons contain information about θ, ρ and α. The statistical model is responsible for extracting estimates of the individual parameters

from these sources of information.

Completely linked data

In the mitochondrial genome, and parts of the mammalian Y chromosome, there is no recombination and so all sites share a single genealogy. Although there is obviously nothing to be learned about the genealogy from recombination events, the fact that many sites share the same genealogy means that the data can be quite informative about the presence of particular branches in the genealogy and their relative lengths. In particular, when comparing two chromosomes from different populations, the information about their coalescence time derives from the numbers of mutations falling on the two branches.

Much recent work on statistical inference under models of isolation and migration has focused on this situation (Beerli & Felsenstein 1999, Bahlo & Griffiths 2000, Nielsen & Wakeley 2001, Hey & Nielsen 2004). Fortunately for the prospects of inference in population genetics, almost the entire genome does not satisfy the assumption of these methods that there is no recombination and hence only a single genealogy (complete linkage). Restricting oneself to those regions that do means restricting oneself to a tiny fraction of the total information content of the genome.

One pragmatic way forward is to obtain sequence from several unlinked and relatively short genomic fragments. The termini of the regions can be pruned until procedures such as those of Hudson & Kaplan (1985) and Myers & Griffiths (2003) show no evidence of recombination, and the resulting fragments analysed under the assumption of complete linkage. Although this is a reasonable strategy in the absence of statistical methods that model recombination, there are several drawbacks.

Perhaps most serious is that, unless the rate of mutation greatly exceeds that of recombination, the potential for learning about genealogical structure in a particular genomic region is limited by the few segregating sites that can be analysed. Furthermore, since the procedure is essentially a search for relatively long stretches of genome without recombination, the regions selected will tend to have short genealogies and an ascertainment bias in the selection of loci results (Hey & Nielsen 2004). Lastly, as described in the next section, in a similar way to mutation events, recombination events themselves convey information about branch lengths which is discarded in these analyses.

Loosely linked data

I will refer to data from a region of the genome in which recombination occurs, but is insufficient to destroy the background LD (see Falush *et al.* 2003), as 'loosely linked'. There are two common types of loosely linked data. Single nucleotide polymorphism (SNP) data sets comprise allelic variation at sites that are known in advance to be polymorphic (at least in some other sample). In this case, unsampled sites lie between the sampled sites. Resequencing data sets comprise the allelic variation at every site in some contiguous block of sites.

In loosely linked data, mutations contribute information about the presence and lengths of branches in the genealogy, as in completely linked data. However, because of recombination, it cannot be assumed that a branch present at one site in the alignment persists throughout the alignment. Although this makes analysis more complicated, it has the far-reaching consequence that correlation is eliminated between more distant sites. Therefore data sets combining many loosely linked sites can combine the benefits of both unlinked data (many independent samples from

the genealogical process) and completely linked data (good information about local genealogies). Furthermore, there is extra information about the genealogy deriving from the lengths of chromosomal regions over which genealogical branches persist.

To understand the extra information about F contributed by the recombination process, fix attention on a particular pair of chromosomes, one from each of the daughter populations, and suppose that at a randomly chosen point along the chromosome they are observed to coalesce at time $t > T$. Now look along the chromosome to the right. At some point, a recombination event will occur on one or other of the two branches below the coalescence event. Beyond that point the two lines will no longer coalesce with each other at time t. In the same way as mutations, recombination events on the two branches below the coalescence occur along the chromosome as a Poisson process with rate $2rt$ (until the first one occurs). Since the two chromosomes were sampled from different populations, t is greater than T, and the expected distance to this change in genealogy is therefore less than $1/2rT$, i.e. the distance is expected to be shorter the larger T is (and the higher the per-generation recombination probability is). Thus the length of chromosome over which the coalescence of two lines from different populations persists without alteration contains information about $rT = \rho F$.

A sample of haplotypes from a single unstructured population (such as either of the daughter populations) also contains information about $\rho = 2N_e r$ alone, where N_e is the effective size of the population. Consider the genealogy of those haplotypes at some site and imagine a recombination event somewhere in the genealogy. Immediately to the left of the site, the branch on which the recombination occurred will subtend some clade of chromosomes (or perhaps a single chromosome); but as a

result of the recombination the position in the genealogy of that clade may be quite different to the right of the site. In that case, the haplotypes belonging to the clade may suddenly come to resemble a different set of haplotypes from those that they resembled to the left of the recombination point, and the fact of the recombination will be evident in data. The frequency with which such patterns are encountered along the chromosome contains information about ρ.

It should be clear from the preceding paragraphs that a statistical method that makes efficient (in the statistical sense) use of polymorphism data from loosely linked data for inference about population history would be highly desirable. Such a method does not exist, and the aim of this chapter is to suggest one possible approach, and to assess the potential of this approach by describing its implementation under the isolation without gene flow model and investigating properties of the resulting estimators.

Ideally a statistical method for loosely linked data would use all of the information that the data contain about local genealogies while correctly accounting for the correlation between sites resulting from limited recombination. That is a difficult task, and it may be hard to design summary statistics that make efficient use of the data and then to specify confidence intervals associated with parameter estimates. Sufficient statistics (i.e. summaries of the raw data that retain all the information about the parameters of the model) under models of population genetics incorporating recombination are not known, and so if it were possible to base inference on the full data likelihood then this would be the preferred (most efficient) solution. Unfortunately when recombination is included in the model this has proven challenging and computationally intensive, even for the simplest such model (panmixia; Kuhner *et al.* 2000,

Nielsen 2000, Fearnhead & Donnelly 2001).

Haplotype phase Note that when data comprise genotypes at linked diploid loci in unrelated outbreeding organisms it is not known which allele resides on which copy of the chromosome (maternal or paternal), and therefore the sequence of alleles along a single chromosome (the haplotype of that chromosome) is unknown. Nevertheless, this chapter makes the assumption that this 'haplotype phase' is known. In practice the phase may be estimated by a separate statistical procedure (e.g. Stephens *et al.* 2001). It is also possible to design the genotyping experiment in such a way that there is more information about the phase: for example one may sample individuals with known pedigree information (e.g. The International HapMap Consortium (2005) who sampled individuals in parent-offspring trios); one may sample experimental populations that have been inbred until they are homozygous at all loci (this approach is common in studies of *Drosophila*); or one may clone a single copy of the chromosomal region prior to genotyping (e.g. Jennings & Edwards 2005).

'Haplotype structure' I will use the term 'haplotype structure' to refer to all the information about local genealogies in completely linked and loosely linked data that is not contained in unlinked data. In completely linked data the haplotype structure information derives from the occurrence of multiple mutations on branches in the genealogy that are shared across all sites. In loosely linked data this source of information is augmented by information about branch lengths deriving from recombination events as described below.

The method described here is based on the approximate likelihood scheme of Li & Stephens (2003). In the original description, the authors apply their method to

the problem of inferring the population-scaled recombination rate parameter ρ, and variation along the chromosome in this parameter. Subsequent extensions and further applications have included its use to infer haplotypic phase (Stephens & Scheet 2005), rates of gene conversion (Gay in prep., Hellenthal & Stephens 2005) and diversifying selection (Wilson & McVean 2006).

3.4 A new approximate likelihood for the isolation model

3.4.1 Likelihood-based methods in population genetics

Given some data that are assumed to have been sampled from a model with parameters ϕ, a 'likelihood' function expresses the probability of the data as a function of the model parameters. Here, the sampled data are an alignment of n haplotypes $h = (h_1, \ldots, h_n)$. The i^{th} haplotype $h_i = (h_{i1}, \ldots, h_{iL})$ is a specification of the allelic state at sites $1, \ldots, L$ and has an associated known label $z_i \in \{1, 2\}$ (fixed in advance) specifying the daughter population from which that haplotype was sampled. The model parameters are (F_1, F_2, ρ, θ) and will be represented by the symbol ϕ. Therefore the likelihood $L(\phi)$ is the joint sampling probability of the n haplotypes as a function of ϕ:

$$L(\phi) = p(h; \phi) = p(h_1, h_2, \ldots, h_n; \phi).$$

Closed-form mathematical expressions for the likelihood are known only for the special case of parent-independent mutation, and then only for models of constant-sized panmictic populations (e.g. Ewens 1972). As a result, the literature on statistical population genetics has concentrated on finding informative summaries of the full data, and more recently on developing computational procedures for estimating or

approximating the likelihood, which is the approach taken in this chapter. The difficulty of writing down a closed-form expression for the likelihood is the result of the probabilistic structure of the data outlined in section 3.3: the data are correlated at a single site as a result of a random underlying genealogy, and they are correlated along the chromosome as a result of random alterations to this genealogy made by recombination.

The computational approaches to estimating the likelihood are based on identifying some unknown aspect G of the data, given which one *can* calculate the probability of the data. Under a model specifying the probability distribution of G, the likelihood can be estimated as the probability of the observed data given G, averaged over the distribution of G (e.g. Felsenstein 1992):

$$L(\phi) = p(h;\phi) = \int p(h|G;\phi)p(G;\phi)dG \tag{3.1}$$

This is a standard approach in statistics in general, where G is sometimes referred to as 'missing data'. For example, in population genetics, if the population allele frequencies were known, then the data at a single site would have a multinomial distribution with parameters given by those allele frequencies. Therefore it is possible to estimate the likelihood by approximating the integral over the unknown allele frequencies (e.g. Pritchard *et al.* 2000). Alternatively, for completely linked data G might be the genealogy of the sample (e.g. Kuhner *et al.* 1995, Nielsen & Wakeley 2001) or the sequence of mutation and coalescence events in the history of the sample (e.g. Griffiths & Tavare 1994, Stephens & Donnelly 2000). For loosely linked data, G might specify the genealogy at each site (Kuhner *et al.* 2000, Nielsen 2000), or the sequence of mutation, coalescence and recombination events in the history of the

sample (Fearnhead & Donnelly 2001).

A simple method for solving (3.1) is Monte Carlo integration, in which the integral is approximated by the average value of $p(h|G)$ for a random sample of values of G drawn from the model:

$$L(\phi)\hat{=}\frac{1}{B}\sum_{i=1}^{B} p(h|g^{(i)}), \qquad g^{(i)} \sim p(G;\phi), \tag{3.2}$$

However for population genetic data, the missing data G usually has to be quite high-dimensional (e.g. a bifurcating topology with branch lengths and labelled tips) (Stephens & Donnelly 2000). As a consequence the vast majority of the terms in (3.2) would contribute negligibly to the sum, and for a practical number B of draws, the variance of the estimate of the likelihood would be too high to be useful. The studies cited above approximate the integral using either importance sampling (IS) or Markov chain Monte Carlo (MCMC), which are more computationally efficient elaborations of (3.2). Because (3.2) relies on sampling a large number of values of G from the model by simulation, such methods have greatly increased in importance with the rise in computing power over the last few decades. However, IS and MCMC methods can be time-consuming to code and debug, the resulting code is normally a rather inflexible implementation of a particular model, they can still be very computationally intensive, and it can be difficult to know when enough samples have been taken. In particular, full likelihood methods for loosely linked data (Kuhner *et al.* 2000, Nielsen 2000, Fearnhead & Donnelly 2001) have suffered from various combinations of these problems and despite representing useful advances from a theoretical point of view they are not currently widely-used for data analysis.

This chapter takes an alternative approach suggested by Li & Stephens (2003) (hereafter LS), in which the joint probability of the sampled haplotypes is decomposed into a product of conditional probabilities. These are approximated in a way which can be viewed as specifying low-dimensional missing data, and the resulting approximation to (3.1) can be computed efficiently and, in contrast to Monte Carlo methods, deterministically.

3.4.2 PAC likelihoods

The joint probability of the n haplotypes can of course be rewritten as a product of conditional probabilities:

$$p(h_1, h_2, \ldots, h_n; \phi) = p(h_1)p(h_2|h_1; \phi)p(h_3|h_1, h_2; \phi) \ldots p(h_n|h_1, \ldots, h_{n-1}; \phi).$$

The probability $p(h_{k+1}|h_1, \ldots, h_k; \phi)$ is the sampling probability of the $(k+1)^{\text{th}}$ haplotype conditional on the k observed so far, and is central to the following discussion. Except for the case of the infinite alleles model under panmixia these are unknown, and LS substituted approximate expressions for them, giving rise to an approximate likelihood that they called a 'product of approximate conditionals' (PAC) likelihood. The dependence on ϕ indicated by the semicolon will be present throughout and will be left implicit in order to simplify the notation. Using hats to indicate the approximate conditional (AC) sampling probabilities, the PAC likelihood is

$$L_{\text{pac}}(\phi) = \hat{p}(h_2|h_1)\hat{p}(h_3|h_1, h_2) \ldots \hat{p}(h_n|h_1, \ldots, h_{n-1}). \tag{3.3}$$

The unconditional sampling probability of the first haplotype does not feature in this product because it does not depend on the model of population history, and therefore

does not need to be included in the likelihood for the purposes of this chapter. Because of the approximations, this product depends on the ordering of the haplotypes and so LS proposed using the average value of the product over several random orderings. Note that once a procedure for computing the AC probability $\hat{p}(h_{k+1}|h_1,\ldots,h_k)$ for arbitrary k is decided upon and implemented, computing the PAC likelihood is straightforward. I will refer to the $(k+1)^{\text{th}}$ haplotype as the 'new' haplotype, and use a subscript asterisk for quantities relating to the new haplotype in order to make the notation more readable.

3.4.3 A PAC likelihood for unlinked data under isolation

The approximate likelihood introduced here extends the approach of LS by introducing an approximation to the conditional sampling distribution $\hat{p}(h_*|h_1,\ldots,h_k)$ under the isolation without gene flow model. Consider initially the simple case in which all sites are unlinked. In that case the likelihood for the alignment of haplotypes may be computed as the product of the likelihoods for individual sites. Therefore, temporarily consider notation such as h_* and h_i to refer to alleles at some site rather than entire haplotypes; the problem is then reduced to computing the approximate probability of a new allele h_* given the alleles $h_1,\ldots,h_k$ observed so far.

The AC sampling probability in LS is based on the observation that, unless a mutation has occurred, the new allele will be the same as one of the alleles observed so far. LS introduced an unobserved quantity $X \in \{1,2,\ldots,k\}$ which specifies which of the previous k alleles h_* 'copies': with some (high) probability of fidelity h_* will be the same as the allele on haplotype X, otherwise it will differ. Although there is no formal correspondence, one may think of 'copying' as corresponding to 'coalescing

recently with', and the copying fidelity probability as corresponding to the probability that no mutation occurs since the coalescence event.

The fidelity of copying is determined by the 'emission probability' $u(h_*|h_x)$, which is the probability of observing allele h_* given that allele h_x was copied. Since the identity of h_x is unobserved, the AC probability in LS is computed by averaging the emission probability over a prior probability distribution on the unknown copied allele:

$$\hat{p}(h_*|h_1,\ldots,h_k) = \sum_{x=1}^{k} u(h_*|h_x)\frac{1}{k}. \tag{3.4}$$

The AC probability in LS was designed with a model of an unstructured population in mind and so this expression features the quantity $1/k$, reflecting a uniform prior probability distribution on the missing data X. In contrast, under the isolation model the prior probability on X is uniform only in the unstructured case $T = 0$, in which case the new model reduces to that in LS (with some minor differences).

For $T > 0$ the model is structured, and so it is appropriate that h_* has a higher prior probability of copying an allele from the same population than one from the other population. The approach taken in this chapter (due to G. Coop) is to follow the analogy between 'copying' and coalescence fairly faithfully, by introducing a new dimension $S \in \{A, D\}$ to the missing data, which specifies whether the 'copying' (coalescence) event occurs in the ancestral population ($S = A$), or in the daughter population ($S = D$). Thus the missing data are now a pair $(X, S) \in \{1,\ldots,k\} \times \{A, D\}$ specifying who is copied and when. I will refer to S as the 'level' at which copying occurs.

It is worth emphasising that the choice of the missing data in the model is a critical modelling decision, which was motivated by the isolation without gene flow model. A natural and important extension would be to model migration between the daughter populations, which would presumably involve altering the nature of the missing data. However it was not obvious how to implement in computer code the isolation without gene flow model in a way that would allow such an alteration to the model without substantial alterations to the code. The PAC likelihood approach thus shares with full likelihood methods based on MCMC or importance sampling a certain lack of flexibility and extensibility. It can be argued that this is a serious drawback, in comparison perhaps with methods based on summary statistics and Monte Carlo simulations (e.g. Beaumont *et al.* 2002, Voight 2006, Becquet & Przeworski in prep.).

The prior probability distribution on S is given by the probability $p(A)$ that a $(k+1)^{\text{th}}$ line coalesces in the ancestral population, given the amount of drift in the daughter population, which can be calculated exactly (appendix 3.10). Of course, since under the isolation without gene flow model the coalescence must occur in either the ancestral or daughter population, $p(D) = 1 - p(A)$. Conditional on S, the prior probability distribution on X is

$$p(x|S=D) = \begin{cases} \frac{1}{k_{z*}} & \text{if } z_x = z_* \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

$$p(x|S=A) = \mathbb{E}\left(\frac{a_{z_x}}{a_1+a_2}\right)\frac{1}{k_{z_x}}, \tag{3.6}$$

where z_* and z_x are the population labels of h_* and h_x respectively, and notation like k_{z_x} refers to the number of lines sampled so far from population z_x. a_i is the unknown number of the distinct ancestral lines that enter the ancestral population

from daughter population i. The expectation can be calculated using the transition probabilities in Tavaré (1984), in a similar way to the calculation in appendix 3.10 equation 3.32.

In words, conditional on $S = D$, the prior on X has zero weight on lines from the other population, and is uniform over those lines from the same population. Conditional on $S = A$, the prior probability distribution on X is uniform in the case of equal amounts of drift in the two daughter populations. In the case of unequal amounts of drift, the prior is uniform over lines from the same population but has more weight on those coming from the population with less drift. Therefore if the new line was sampled in a daughter population that has experienced considerable drift, the prior on S is weighted heavily towards $S = D$, in which case an allele from the same population is necessarily copied. Thus in this situation h_* has, as required, a higher prior probability of copying an allele from the same population, and the precise form of this prior derives from a standard population genetic model of isolation without gene flow.

A key feature of the unstructured coalescent is that the overall rate of coalescence is high when there are many uncoalesced lines. In the usual description of the coalescent this results in the much shorter intervals between the early coalescent events than between the last few coalescent events leading to the most recent common ancestor. The product of conditionals representation (3.4.2) implies thinking about the coalescent instead as a process by which a random tree with n tips is generated sequentially by starting with a single ancestral line, and repeatedly sampling a 'new' tip and allowing it to coalesce into the existing tree until there are n tips in total (this is called 'sampling through space' by Ewens (1990)). From this perspective, the

high rate of coalescence when there are many uncoalesced lines has the consequence that the last lines to be allowed to coalesce into the tree mostly do so very rapidly, while the earlier-sampled lines have a much higher probability of remaining distinct for a long time, prior to their eventual coalescence. Very informally, one can think of the late-sampled lines having little chance of making it through the crowded early periods without coalescing. One consequence is that a late-sampled line will often coalesce into the existing tree before a mutation occurs on it and so the allele at its tip will be identical to a previously sampled allele. Another consequence, anticipating section 3.4.4, is that the branch joining a late-sampled line to the rest of the tree tends to persist for a considerable distance along the chromosome on either side of the site that we are focusing on, because the short branch represents little time for a recombination event to occur in.

In order to model this property of the coalescent (from the point of view of mutations; see section 3.4.4 for the way in which recombination is modelled), LS made their emission probability dependent on k, so that for large k (i.e. when many lines have been sampled already), there is only a very small probability of observing an allele that differs from the allele it copied. Under the isolation without gene flow model, conditional on $S = D$, this should also be true. However, since $S = A$ implies a more ancient coalescence time than $S = D$ whatever the value of k, the emission probability in the new model should reflect this in the form of its dependence on both S and k. The approach taken here is to proceed as if the time t of the copying (coalescence) events is always equal to its expectation $\tilde{t}_s = \mathbb{E}(t|S = s; k, \phi)$. These expectations can be computed exactly using results from Tavaré (1984) and Fu & Li (1993) (appendix 3.10). The emission probabilities used are based on the assumption that either 0 or 1 mutations occurred on the branch joining the new line to the

existing tree and are

$$u(h_*|h_i, s) = p(h_*|h_i, s; k, \phi) = \begin{cases} 1 - \exp(-\tilde{\theta}\tilde{t}_s) & \text{if } h_* \neq h_i \\ \exp(-\tilde{\theta}\tilde{t}_s) & \text{if } h_* = h_i, \end{cases} \tag{3.7}$$

An exception occurs when $k = 1$ in which case $2\tilde{t}_A$ is substituted for $\tilde{t}_A$ in order to account for the time on the branch ancestral to the first line ($S = D$ is impossible for the second line sampled, as the random orderings are chosen such that haplotypes are sampled alternately from each population). Note that although quantities such as $p(A)$, $p(x|A)$ and $u(h_*|x, A)$ depend on k, this dependence is generally left implicit.

The value of $\tilde{\theta}$ depends on the way in which the sampled loci were ascertained. If the loci were selected randomly, without regard to whether or not they show polymorphism, then $\tilde{\theta}$ is treated as a parameter of the model to be estimated. Alternatively, if loci were only included if they were polymorphic in the sample then an arbitrary small value of $\tilde{\theta}$ is used. For the results presented in this chapter I used $\tilde{\theta} = 1/\mathbb{E}(\mathrm{T}_{\mathrm{total}})$, where $\mathrm{T}_{\mathrm{total}}$ is the expected length of the genealogy given ϕ, which can be calculated approximately. Note that $\tilde{t}$ is used in an analogous way when modelling recombination, as half the amount of time available for recombination events on the ancestral lines connecting the $(k+1)^{\mathrm{th}}$ haplotype with the haplotype that it copies (see section 3.4.4).

As in the unlinked case (3.4) in LS, the AC probability under the isolation model is obtained by averaging the emission probability over the prior probability distribution

on the unknown missing data:

$$\hat{p}(h_*|h_1,\ldots,h_k) = \sum_{s\in\{D,A\}} \sum_{x=1}^{k} u(h_*|h_x,s)p(s,x;k)$$

Let $\bar{u}(h_*|s) = \sum_{x=1}^{k} u(h_*|h_x,s)p(x|S=s)$ denote the emission probability averaged over which allele is copied, given that $S=s$. Then the AC probability under isolation without gene flow is

$$\hat{p}(h_*|h_1,\ldots,h_k) = p(A)\bar{u}(h_*|A) + (1-p(A))\bar{u}(h_*|D). \tag{3.8}$$

Equations 3.3, 3.7 and 3.8, together with the procedure for averaging over orders and the expressions for $p(A)$ and $\tilde{t}_s$ given in appendix 3.10, specify the new PAC likelihood for unlinked data under the isolation without gene flow model.

3.4.4 A PAC likelihood for loosely linked data under isolation

In loosely linked data, correlation is anticipated between the patterns of polymorphism at nearby sites as a result of linkage. In LS this is modelled by introducing correlation into the prior on the missing data at nearby sites. Recall that in LS, for each new haplotype h_*, there is a sequence of unobserved quantities $X_1, X_2, \ldots, X_L$ specifying which of the previous k haplotypes is copied at each of the L sites. LS assume that these random variables form a Markov chain with the following transition probabilities between the hidden states x and x' at sites l and $l+1$ (their equation A1):

$$p(x \to x') = \left[1 - e^{-\rho_l/k}\right]\frac{1}{k} + I(x'=x)e^{-\rho_l/k}, \tag{3.9}$$

where $I(\texttt{argument})$ is a notational device that is equal to 1 if `argument` is true and equal to zero otherwise. Here ρ_l is equal to $2N_e$ times the distance between the two sites in Morgans; this is sometimes referred to as the 'population genetic map' distance between the sites. The idea behind (3.9) is that either there is a 'switch' in which haplotype is copied or there isn't. If there isn't (probability $\exp(-\rho/k)$), then the same haplotype is necessarily copied at the next site. If there is, then the haplotype that is copied at the next site is a draw from the uniform prior on the k previously-sampled haplotypes (including the same one that was being copied). The switching events represent the effects of recombination, although again there is no formal correspondence, and indeed LS found that the estimator of ρ based on (3.9) tends to underestimate the true value. Note that the rate ρ/k of switching between two sites ρ units apart decreases with k, in keeping with the idea outlined in section 3.4.3 that late-sampled lines coalesce rapidly into the existing tree, presenting little opportunity for recombination.

The approach taken in this chapter to modelling loosely linked data follows that of LS: prior correlation between the hidden states at nearby sites is introduced, parameterised by ρ, F_1 and F_2, in a way that attempts to capture some features of the genealogical process with recombination under the isolation model. The prior correlation is introduced by assuming that the sequence of hidden states along the chromosome $(S_1, X_1), (S_2, X_2), \ldots, (S_L, X_L)$ is a Markov chain, and so the model of loosely linked data is determined by specifying transition probabilities in this two-dimensional state space. In accordance with the considerations in section 3.3.2, the model should have the feature that, conditional on $S = A$ at site l, for large values of $2rT = \rho_{z*} F_{z*}$ there is rather little correlation between the haplotype that is copied at sites l and $l + 1$. The transition probabilities used here are, as with the prior dis-

tribution on (S, X) based on consideration of the standard population genetic model of isolation, and on an attempt to equate fairly explicitly the switching events in the copying process with recombination events in that model.

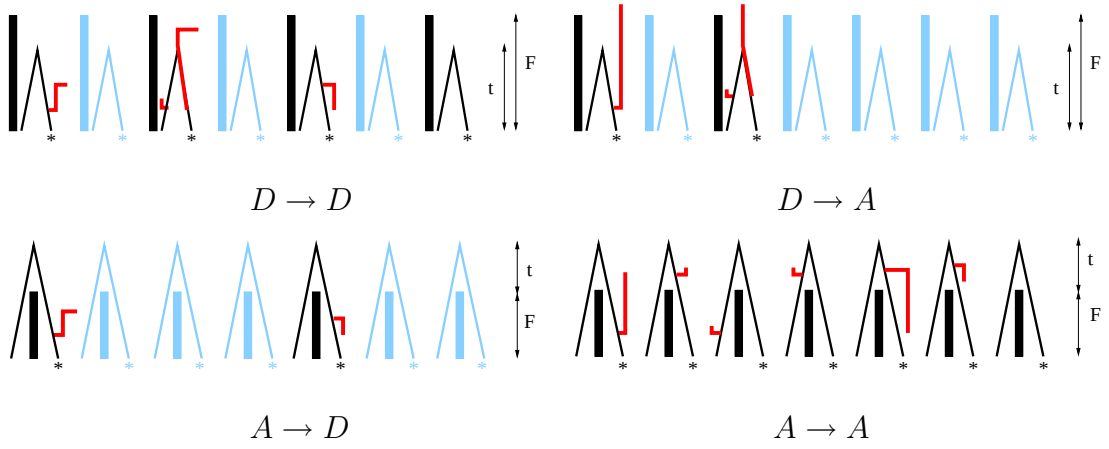


Figure 3.3: **Transitions between daughter and ancestral hidden states.** Asterisks indicate the $(k+1)^{\text{th}}$ haplotype (the 'new line'), which at the current locus is copying the other line drawn (the 'other line'). Each panel illustrates the different ways in which the indicated transition between hidden states can occur between the current and next loci. Trees in light blue represent types of event that cannot result in that panel's transition. Red indicates events whose probability must be included in the transition probability, except the short rising red lines which indicate that the transition depends only on the occurrence of the recombination, and not otherwise on the fate of the recombinant line. The 7 types of event are, from left to right, recombination on the new line at daughter level, recombination on the new line at ancestral level, recombination on the other line at daughter level, recombination on the other line at ancestral level, hit by recombinant line from daughter level, hit by recombinant line from ancestral level, no interrupting event.

Figure 3.3 illustrates the way in which transitions between the two copying levels can result from various sorts of events in the genealogy at the two sites. The following discussion is perhaps least questionable if it is assumed that the new line coalesces at site l with a line ancestral to just one of the k haplotypes sampled so far. I will refer to this haplotype as the 'other' haplotype, and to the two lines leading from

the sampled haplotypes to their coalescence as the 'new line' and 'other line'. In this case it seems reasonable to equate the state of affairs in the genealogical process with that in the copying process in which the 'new' haplotype is copying the 'other' haplotype. Of course it will frequently be the case that the new line coalesces with a line ancestral to a clade of haplotypes at site l. In that case it is less clear which haplotype the new haplotype should be said to be copying at site l. This lack of exact correspondence between the genealogical process and the process under which the approximate likelihood is computed is a feature of the 'copying' approximation in general; it is not specific to the model of population history considered here.

Figure 3.3 is divided into four panels, each representing one of the four possible transitions among the states $\{A, D\}$ (including 'transitions' to the same state). Each panel is divided into 7 columns (schematic genealogies) each representing a particular type of genealogical event. The types of genealogical events represented by the 7 columns are the same in each of the 4 panels, and types of events that cannot result in that panel's transition are greyed out. For example, the first column in each panel represents a recombination event occurring in the daughter population on the line joining the new haplotype to the existing tree; such an event can result in all four transitions. The fifth and sixth columns represent events in which a recombination event occurs in the part of the genealogy which is ancestral to neither the new nor the other line. The line at site $l + 1$ ancestral to that recombination event then coalesces into the part of the tree ancestral to the new haplotype, thus changing the identity of the haplotype most closely related to the new haplotype. Some or much of the evidence of recombination in real data presumably does result from these and other more complicated sorts of events, and versions of the transition probabilities were investigated in which some attempt was made to model them. However this resulted

in a considerable increase in the complexity of the expressions without obviously commensurate improvement in the statistical properties of the estimators (results not shown), and the results presented in this chapter are based on the transition probabilities given below that do not attempt to model such events. The last (seventh) column represents the situation in which no relevant recombination occurs, and hence the same haplotype is copied at the same level at the next site.

Transitions from the daughter population

I consider two possibilities involving recombination: a recombination occurs on either the new line or the other line, prior to their coalescence at time t (event R_t; columns 1 and 3 of the top left and top right panels in figure 3.3). The coalescence time of these lines is assumed to be the conditional expectation $\tilde{t}_D$, measured in units of twice the effective size N_{e*} of the new haplotype's population. On this time scale the recombination events occur at rate $\rho_{l*} = 2N_{e*}r_l$. Conditional on such a recombination, the new line subsequently coalesces into the tree at the next site either prior to T ($S' = D$) or after T ($S' = A$). Conditional on the recombination and the level at which the subsequent coalescence occurs, the haplotype that is copied at the next site is drawn from the prior on X. Additionally, if $S' = D$ and the 'transition' is to the same haplotype, the probability that neither sort of recombination event occurs contributes to the transition probability. Thus the corresponding entries in the matrix of transition probabilities are

$$p(x, D \to x', S') = p(R_t)p(S'|R_t)p(x'|S') + (1 - p(R_t))I(S' = D, x' = x), \quad (3.10)$$

where

$$p(R_t) = 1 - \exp(-2\rho_{l*}\tilde{t}_D). \tag{3.11}$$

In principle, evaluating $p(S'|R_t)$ requires averaging the probability of surviving back to the ancestral population over the unknown number of uncoalesced lines at the unknown time of the recombination event; in practice I substitute the marginal (prior) probability $p(S')$, which is larger than $p(S'|R_t)$ in the case $S' = D$ and smaller in the case $S' = A$.

Transitions from the ancestral population

In this case, it is necessary to classify the recombination events according to whether they occur on the new line in the daughter population (event R_D^n), on the other line in the daughter population (event R_D^o) or on either line in the ancestral population (event R_A). The case $A \rightarrow D$ is straightforward, as only a recombination event on the new line in the daughter population, followed by recoalescence in the daughter population, can effect this transition. Thus

$$p(x, A \rightarrow x', D) = p(R_D^n)p(D|R_D^n)p(x'|D), \tag{3.12}$$

where $p(R_D^n) = 1 - \exp(-\rho_{l*}F_*)$. Again, I substitute $p(D)$ for $p(D|R_D^n)$.

The case $A \rightarrow A$ is more complex, as this transition can result from any of the following (mutually exclusive) combinations of events:

- a recombination on the new line in the daughter population followed by ancestral recoalescence;

- no recombination on the new line in the daughter population, but recombination on the other line in the daughter population;
- recombination on neither the new or other line in the daughter population but an ancestral recombination on one or other line.

Conditional on the occurrence of one of these events, the haplotype copied at the ancestral level is drawn from the prior. Additionally, if the transition is to the same haplotype, then the probability that none of these recombination events occurs contributes to the transition probability. Thus

$$p(x, A \to x', A) = \left(p(R_D^n)p(A) + (1 - p(R_D^n))p(R_D^o) + (1 - p(R_D^n \cup R_D^o))p(R_A) \right) p(x'|A) + I(x' = x)(1 - p(R_D^n \cup R_D^o \cup R_A)), \tag{3.13}$$

where the probabilities of the various combinations of recombination events are

$$\begin{aligned} p(R_D^n) = p(R_D^o) &= 1 - \exp(-\rho_* F_*) \\ p(R_D^n \cup R_D^o) &= 1 - \exp(-2\rho_* F_*) \\ p(R_D^n \cup R_D^o \cup R_A) &= 1 - \exp(-2(\rho_* F_* + \rho_A \tilde{t}_A)). \end{aligned} \tag{3.14}$$

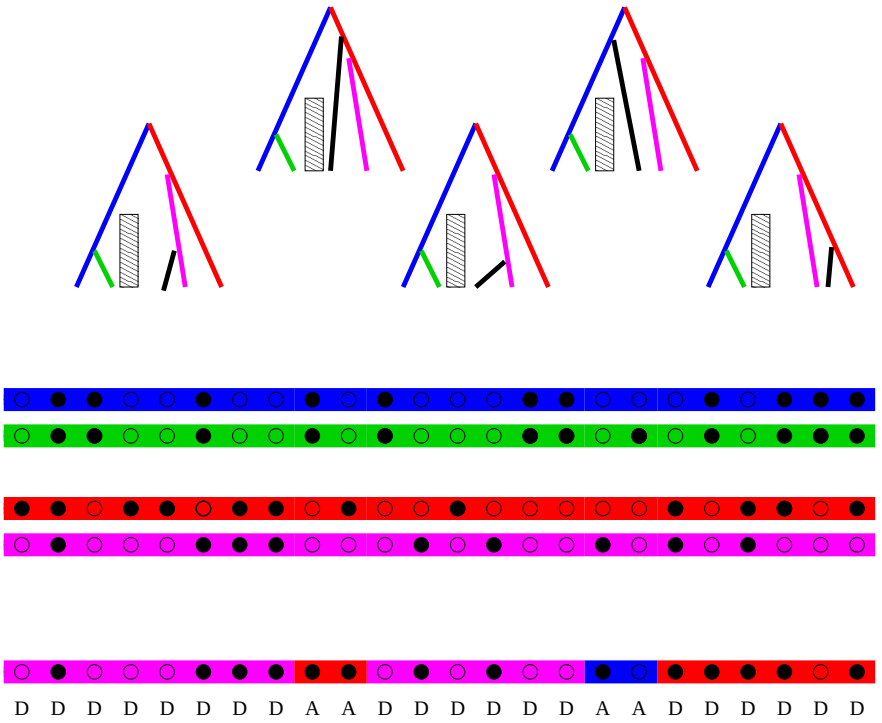


Figure 3.4: **The copying process in the new PAC model for loosely linked data**
Two haplotypes have been sampled from each population so far (i.e. $k = 4$), and these are depicted as coloured horizontal lines grouped into two groups of two. The 'new' haplotype is depicted below; it was sampled from the same population as the lower two of the four. At each site along the haplotypes, small circles represent which of the two alleles is present (filled or open). Each of the 4 haplotypes has its own colour. The new haplotype at the bottom is made up as a mosaic of these colours, indicating which of the four is copied at each site. Letters below the new haplotype indicate whether the copying occurred in the daughter (D) or ancestral (A) population at each site. This particular colouring of the new haplotype and sequence of letters represents one possible path through the missing data. The path that is illustrated is one that might have high prior probability when F is relatively large because (i) the sections that are copied at the daughter level are much longer than those copied ancestrally, and (ii) the ancestral-level copying has made more errors (mutations) than the daughter-level copying. For each of the 5 copied sections, a schematic genealogy is drawn above that might correspond to the state of the copying process below. In the trees, the new line is depicted in black. Although the relationships of the coloured lines in the genealogies are depicted as remaining the same, note that this is not an assumption of the model. The computation of the AC probability corresponds to summing the probability of the data over every possible colouring and sequence of letters (path) along the new haplotype, weighting each term by the prior probability of the path given the model parameters.

As a result of the discrete nature of the missing data and the Markov assumption about the sequence of hidden states along the chromosome, the approximate conditional probabilities $p(h_{k+1}|h_1, \ldots, h_k)$ can be computed using the forward algorithm for hidden Markov models (HMMs; e.g. Rabiner 1989), as in LS (see appendix 3.8).

Informally, the new PAC likelihood corresponds to considering the $(k+1)^{\text{th}}$ haplotype to be made up as a mosaic of sections of haplotype copied from one of the previous k haplotypes (figure 3.4). Points where there is a switch in the haplotype being copied correspond to recombination events. When the likelihood is evaluated for large F, higher probability is associated with copying short sections in the ancestral population and copying these less faithfully, in line with the above considerations. Conversely, when the likelihood is evaluated for small F, higher probability is associated with copying longer sections in the ancestral population and copying these more faithfully. A large set of 'paths' $(S_1, X_1), (S_2, X_2), \ldots, (S_L, X_L)$ through the missing data are defined by the set of all possible choices of which haplotype is copied, and when, at every position along the chromosome. The prior probability of such a path will depend on the value of the parameters F_1, F_2 and ρ. The PAC likelihood corresponds to the probability of the data averaged over this probability distribution on possible paths, and the forward algorithm for HMMs is used to compute this average efficiently. The HMM computations are described in more detail in appendices 3.8 and 3.9.

The algorithms described above and in the appendices were implemented in R (R Development Core Team 2006), and all code is available from the author. As an indication of the computing time, it takes approximately 13 seconds to compute the PAC likelihood at a single point in parameter space for a single ordering of an

alignment of 32 haplotypes (16 from each population) at 60 segregating sites on a computer with a 3.0 GHz 'Intel Pentium 4' processor running Linux.

3.5 Results

In this chapter I focus on the case of SNP data. I do not discuss the possible effects of SNP ascertainment procedures nor ways in which they might be modelled. Thus I effectively assume that the probability of sampling a particular site is, conditional on it being polymorphic in one or other or both populations, independent of the frequencies of the alleles in those populations. The model for resequenced data is therefore essentially identical, with the difference that θ is a parameter of the model rather than taking an arbitrary value. Alignments of resequenced data may contain large blocks without polymorphism and a computationally efficient method for computing the AC probability in that case is described in appendix 3.9.

In order to simulate such SNP data I selected a mutation rate such that many more segregating sites resulted than were desired, and then sampled the desired number from these, uniformly with respect to their chromosomal location and sample frequencies. All simulated data sets comprise 16 haplotypes sampled from each population, and 60 SNPs. These values were chosen in order to resemble the real data analysed in section 3.5.2. PAC likelihoods were evaluated using 5 random orderings of the haplotypes, subject to the restriction that the population labels alternate in the ordering.

3.5.1 Simulated data

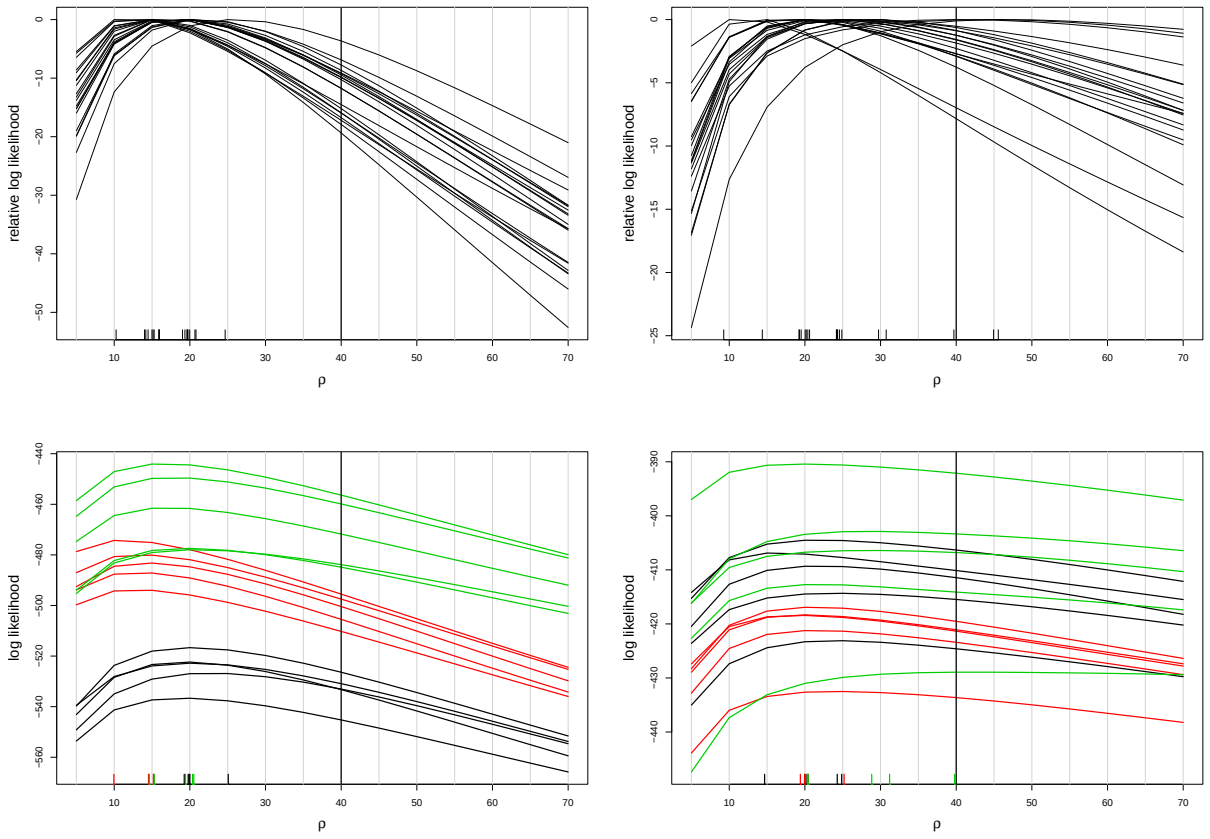
Under the frequentist statistical paradigm, the success of a new estimator is evaluated by considering it to be a random variable whose value is determined by random data sets drawn from the model. Thus for particular values of the parameters of the model, one studies the distribution of the estimator, typically focusing on some measure of its central location (e.g. its mean or median) and spread (e.g. variance, or inter-percentile ranges). In this section I simulate data sets from a standard model of population genetics using the computer program `ms` (Hudson 2002) and study these properties of the new estimators. Although the term bias strictly refers to the mean difference between the random value of the estimator and its true value in the limit as the amount of data tends to infinity, I use the term to refer to any apparent discrepancy between the central tendency of our estimators and the truth. The bias *sensu strictu* of parameters such as ρ (and $2Nm$; see chapter 2) may be of limited practical relevance because it is strongly affected by the presence of occasional very high estimates (see Fearnhead & Donnelly 2001, Hudson 2001, Li & Stephens 2003).

Estimating the rate of recombination relative to the rate of drift

A simple application of the new PAC likelihood is to model haplotypes at ascertained SNPs, assuming that recombination occurs homogeneously throughout the genomic region sampled, and assuming that $F_1 = F_2 = 0$, i.e. the standard constant-sized panmictic equilibrium neutral model. In this case there is a single parameter $\rho = 2N_e r$ which measures the rate of recombination relative to drift. Although this inference problem is not the motivation for the new method, and although several other methods have been designed specifically for this problem (e.g. Kuhner *et al.* 2000, Fearnhead & Donnelly 2001, Hudson 2001, Nielsen 2000, Li & Stephens 2003,

McVean *et al.* 2004), it is important to understand the behaviour of the new method in such simple cases.

The top left panels of figure 3.5 illustrate log likelihood surfaces for ρ for 20 data sets simulated and analysed assuming this model. These are relative surfaces, i.e. each surface has been shifted vertically so that its maximum value is zero. Thus one can compare the extent to which the data support different values of the parameter, but one cannot compare the actual probabilities of the different simulated data sets. A substantial downward bias in the maximum PAC likelihood estimator is evident: for these simulations, the model fits the data best when ρ is approximately half its true value. The top right panels of figure 3.5 illustrate log likelihood surfaces for data sets simulated and analysed under a structured model with relatively little drift in the larger population since the split ($F = 0.1$), and in which the smaller population has had an effective size since the split one tenth that of the larger (i.e. $\alpha = 0.1$). These parameter values correspond approximately to the estimates of F and α in the analysis of data from South American and Siberian human populations 3.5.2. In this case ρ is equal to $2N_{e1}r$, i.e. the rate of recombination relative to the rate of drift in the larger/ancestral population. The downward bias is present in this case also, although it is slightly less pronounced, and the difference between the likelihoods at $\hat{\rho}_{\text{pac}}$ and the true value of ρ is less than under the panmictic model. In both cases (panmixia, and recent split with unequal population sizes) the bias does not seem to depend heavily on the number of haplotypes sampled. The lower panels of the figure illustrate, for three of the 20 data sets, log likelihood surfaces for each of the 5 orders used when calculating the PAC likelihood.



$F_0 = F = 0$ $F_0 = F = 0.1$, $\alpha_0 = \alpha = 0.1$

Figure 3.5: **Simulated data: likelihood surfaces for ρ**
The upper panels show log likelihood surfaces for ρ for 20 different data sets of 60 SNPs, with 16 haplotypes sampled from each population. The vertical position of each surface has been shifted so that its maximum value is 0. The true value of ρ is indicated by the vertical line at $\rho = 40$. Light grey vertical lines show the values of ρ at which the likelihood was evaluated, and the positions of the MLEs are shown by the vertical tick marks above the x axis. Each of the log likelihood curves was obtained by averaging the curves for 5 different orderings of the haplotypes. The lower panels show the log likelihood curves obtained for each ordering, for 3 of the 20 data sets. The data in the left panels were simulated under panmixia, and panmixia was assumed when evaluating the likelihood by fixing $F = 0$. In the right panels the parameter values used in the simulation and when evaluating the likelihood were $F = 0.1, \alpha = 0.1$.

It is in general unsurprising that $\hat{\rho}_{\text{pac}}$ is biased, since the PAC likelihood depends on ρ in a way that only crudely resembles the way in which an exact likelihood would depend on ρ. Indeed Li & Stephens (2003) observe a comparable downward bias when making maximum PAC likelihood estimates of ρ, as do Fearnhead & Donnelly (2001) whose estimator is based on the exact likelihood. When $F = 0$ the new PAC likelihood differs from that of LS only in the way in which the probability of copying a different haplotype at the next site decreases with increasing k: in place of the factor $1/k$ used in equation 3.9 (their equation A1), I use $2/(k+1)$, which is the expected time of coalescence of a $(k+1)^{\text{th}}$ line (Fu & Li 1993) and exceeds $1/k$ for $k > 1$. Thus, since in the new PAC likelihood switching occurs at a higher rate for fixed ρ, one would expect the degree of underestimation to be worse than that in LS.

Figure 3.6 illustrates the dependence of $\hat{\rho}_{\text{pac}}$ on the true value ρ_0 for 3 different models of population history, with the values of F and α fixed at their true values in the simulations. For each model, the relationship between $\hat{\rho}_{\text{pac}}$ and ρ_0 appears roughly linear over the range of ρ_0 investigated, and the figures also show the result of fitting a linear regression with an intercept of zero to the estimates. These results suggest that the bias in $\hat{\rho}_{\text{pac}}$ could be remedied by a treatment similar to that applied by LS when designing their $\hat{\rho}_{\text{PAC-B}}$ estimator. That is, for the likelihood at some value ρ, substitute the likelihood at a different value $\tilde{\rho} = \beta\rho$, where a suitable value of β is estimated from simulations.

Estimating the amount of drift in each population since their separation

For the isolation model the procedure suggested above for dealing with bias in $\hat{\rho}_{\text{pac}}$ could be justified only if it results in improved estimation of the parameters F and α, which is the main concern here. Since it is not obvious that this will be the case,

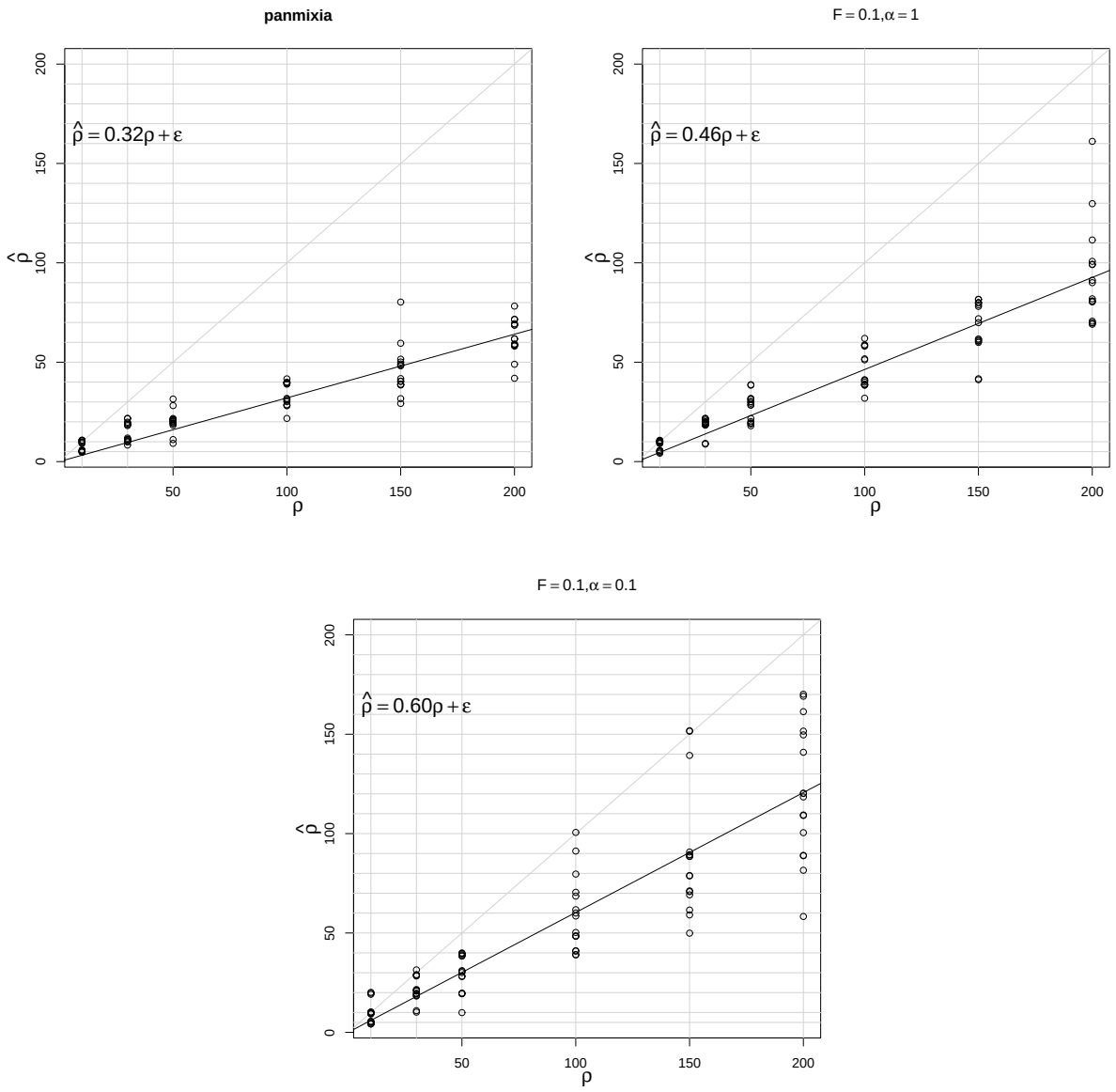


Figure 3.6: **Dependence of $\hat{\rho}_{\mathbf{pac}}$ on ρ_0, when F and α are fixed at their true values.**

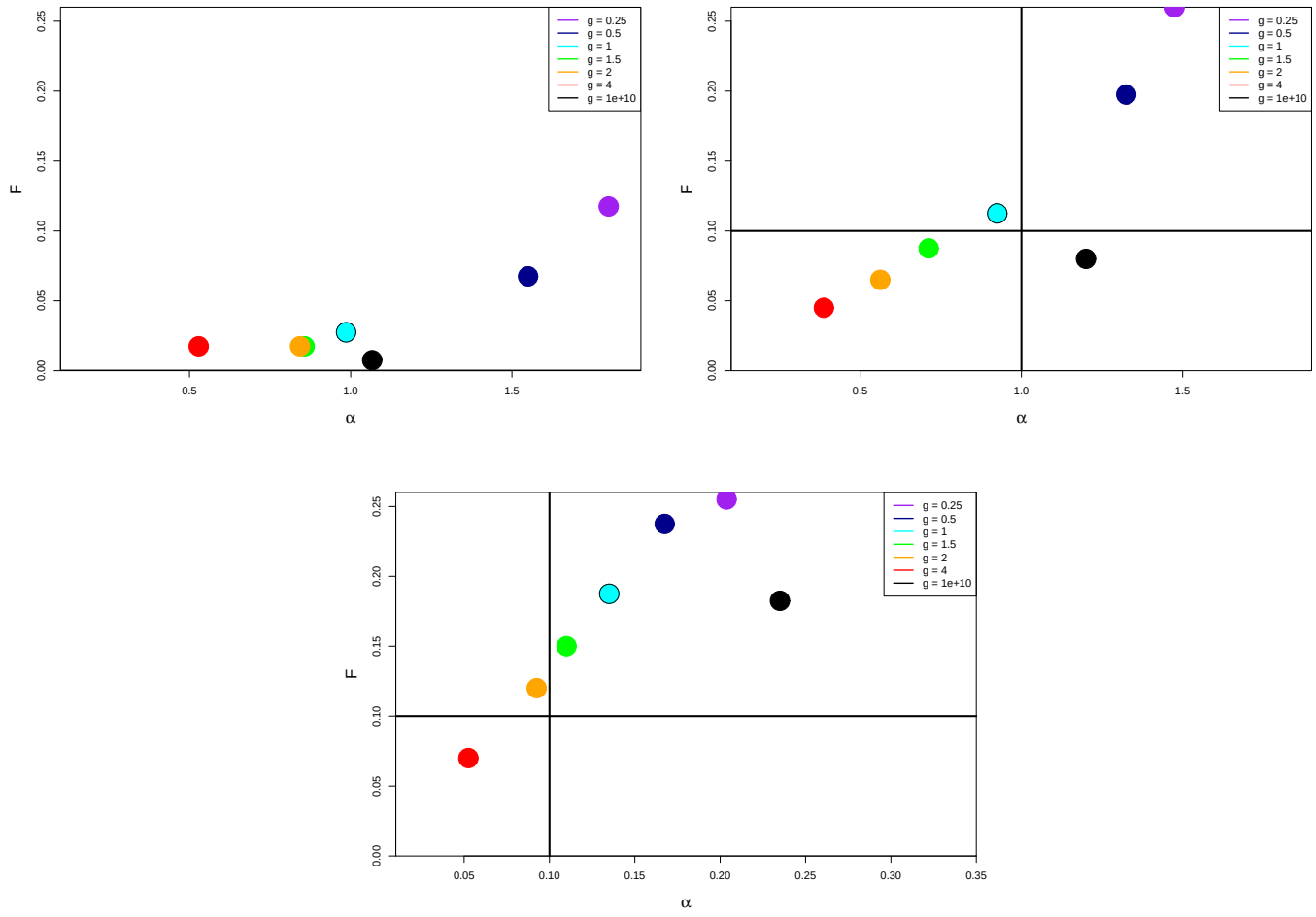


Figure 3.7: **Effect of $\tilde{\rho}$ on the joint estimation of F and α.** Circles show the mean of the 16 estimates of (F, α) resulting from 16 different data sets used in figure 3.6. The colour of the circle indicates the value of g, which determines the value of $\tilde{\rho}$ used when evaluating the likelihood (see text). The same 16 data sets were analysed for each value of $\tilde{\rho}$. The true values of F and α are indicated by horizontal and vertical lines respectively (in the top left panel the true value is $F = 0$) and correspond to the three models examined in figure 3.6.

I investigated the effect of the value $\tilde{\rho}$ used in the model on the distribution of the joint estimates $(\hat{F}_{\text{pac}}, \hat{\alpha}_{\text{pac}})$. For each of the 16 data sets simulated with $\rho_0 = 50$ used in figure 3.6, and for each of a series of values of $\tilde{\rho}$, I evaluated the PAC likelihood at a grid of points (F, α). For each of seven values of $\tilde{\rho}$, figure 3.7 plots the mean value of the 16 maximum PAC likelihood estimates. The three panels correspond to the same models of population history as the three panels in figure 3.6. Note that as F approaches zero, α becomes undefined and the likelihood surface becomes flat in the direction parallel to the x-axis. The positioning along the x-axis therefore contains increasingly little meaningful information for those points with small mean value of $\hat{F}_{\text{pac}}$; the marginal mean of $\hat{\alpha}_{\text{pac}}$ was computed excluding any joint estimates for which $\hat{F}_{\text{pac}} = 0$.

For each data set the values of $\tilde{\rho}$ used were $g\hat{\rho}^0_{\text{pac}}$, where $\hat{\rho}^0_{\text{pac}}$ is the estimate of ρ for that data set from figure 3.6, and $g \in \{0.25, 0.5, 1.0, 1.5, 2.0, 4.0, \infty\}$, the hope being that if $g = 1.0$ resulted in an estimator of (F, α) with reasonable properties, then the following would describe a possible procedure for estimating F and α.

1. Use the data from the larger population to obtain an estimate $\hat{\rho}^0$ assuming panmixia.

2. If $\hat{\rho}^0$ was obtained from some other procedure believed to have relatively little bias, then set $\tilde{\rho} \leftarrow \beta\hat{\rho}^0$, where β is a regression coefficient like that estimated in figure 3.6, perhaps reestimated using simulations matching the particular data set in hand. Otherwise set $\tilde{\rho} \leftarrow \hat{\rho}^0$.

3. Hold ρ in the new PAC model fixed at $\tilde{\rho}$ while searching for a maximum PAC likelihood over a grid of points (F, α).

From the results in figure 3.7, it is clear that if $\tilde{\rho}$ is made larger, the new PAC likelihood responds by fitting a model in which both the amount of drift in the larger population, and the amount of drift in the smaller population relative to that in the larger population, are smaller. It seems natural that the first of these responses should occur, in order that the model fit the same perceived patterns of LD in the data, but the reason for the asymmetric response does not seem obvious. For the model with equal amounts of drift in each population (middle panel), the estimator of (F, α) for $g = 1$ (i.e. $\tilde{\rho} = \hat{\rho}^0$) does seem to be the least biased, providing some grounds for optimism regarding an estimation procedure such as that outlined above.

However, with ten times as much drift in the second population (right panel), F is seriously overestimated, bringing into question the utility of the procedure as currently implemented for analysis of such data sets. This is unfortunate, since many real examples of population divergence involve considerable disparity in the amount of drift experienced in each daughter population. A familiar example is the colonisation of the planet by humans originating in Africa; as is typical in such situations, the colonising populations have experienced much more drift than those that remained *in situ*. Indeed, Mayr (e.g. 1963) suggested that drift (and selection) associated with colonisation of new areas by relatively small populations ('founder effects') might be a general phenomenon of central importance to the process of biological diversification.

Estimating the amount of drift when it is the same in both populations

Researchers studying isolated populations are likely to be most concerned with dating the separation, and therefore the performance of the new model at estimating F is a key question when assessing its utility. Figure 3.8 illustrates the distribution of $\hat{F}_{\text{pac}}$ for a range of true values F_0. For each of the ~ 220 data sets simulated at each

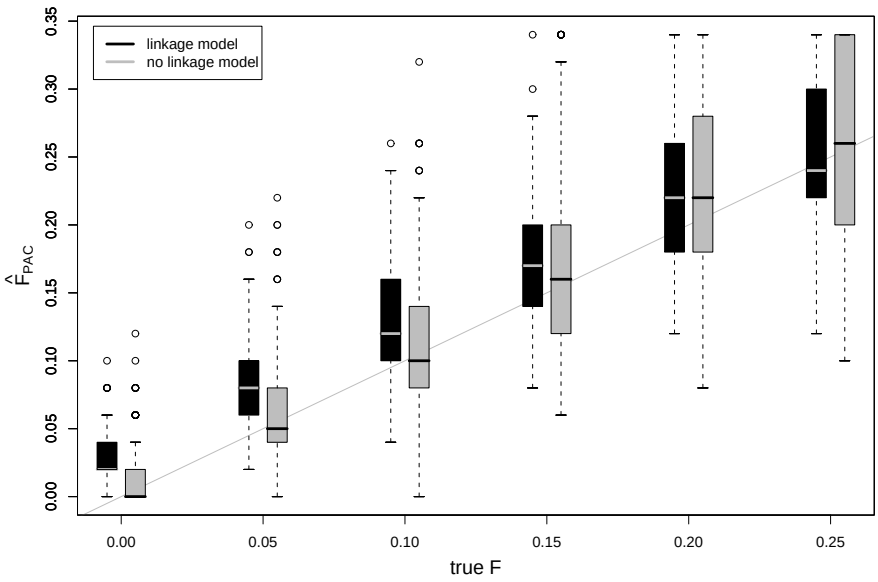


Figure 3.8: **Estimation of F when the daughter populations have drifted by equal amounts**
The figure shows the dependence of $\hat{F}_{\text{pac}}$ on F_0, with α fixed at its true value of $\alpha_0 = 1$, and ρ fixed at 0.46 times its true value of $\rho_0 = 50$.

value of F_0 I found the maximum PAC likelihood estimate of F under two different assumptions about the data. The first estimator $\hat{F}^{\mathrm{u}}_{\mathrm{pac}}$ assumes (incorrectly) that the sites in the data are unlinked by setting $\tilde{\rho}$ to a very large value when evaluating the PAC likelihood. The second estimator $\hat{F}^{\mathrm{l}}_{\mathrm{pac}}$ uses $\tilde{\rho} = \beta\rho_0$, where $\beta = 0.46$ is the estimate of the slope of the regression in the middle panel of figure 3.6. Both estimators assume (correctly) that the amount of drift is equal in the two daughter populations by holding α fixed at 1.0.

The incorrect assumption of no linkage should have two consequences for the estimator $\hat{F}^{\mathrm{u}}_{\mathrm{pac}}$. Firstly, since the no linkage model in effect assumes that the data contain more independent pieces of information than is actually the case, the likelihood surface should be more tightly curved around its maximum than that under the linkage model – this implies that the model exaggerates the degree to which the data support values of F in the vicinity of the maximum relative to other values. Secondly, since all the information in the data deriving from haplotype structure is discarded, the variance of the estimator $\hat{F}^{\mathrm{u}}_{\mathrm{pac}}$ should exceed that of $\hat{F}^{\mathrm{l}}_{\mathrm{pac}}$. Note that the bias of an estimator is not necessarily affected by discarding information.

Three features of figure 3.8 deserve comment. Firstly, the no-linkage model seems to result in an estimator with little bias. Secondly, it is evident that the linkage model does reduce the variance of the estimate of F, although for these simulations only when the true amount of drift since the split is quite large (roughly, $F_0 > 0.1$). Thirdly, the linkage model seems to result in an upward bias, the magnitude of which decreases as the amount of drift in the data increases. In particular, while the no-linkage model tends to conclude that the data have been sampled from an unstructured model when that is the case ($F_0 = 0$), the linkage model tends incorrectly

to result in a small but non-zero estimate of F. This is undesirable, as whether or not the data have been sampled from a structured population is a question of great importance.

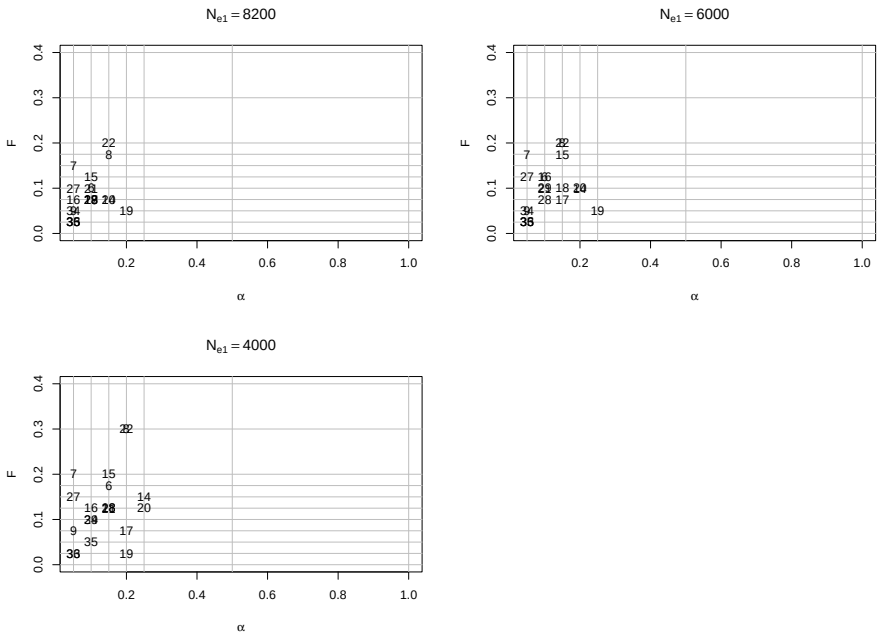
Although it is clearly inefficient when there is linkage, the estimator which ignores linkage is fairly unbiased and in contrast to the linkage model, this is true when in fact there is no structure. The no-linkage model could therefore be used for parameter estimation, and as the basis for a model-based hypothesis test of panmixia against an alternative under which the population labels are specified. I have compared the likelihood surfaces based on the bivariate frequency spectrum under the no-linkage PAC model to the exact likelihood surface estimated by a simple importance sampling scheme and found them to be very similar (results not shown). Although the main concern here is with extracting information from haplotype structure, it should be noted that an efficient means of computing an approximate likelihood for unlinked data with good statistical properties, under a model of isolation which includes recent mutation, is itself of value. In theory, analysing loosely linked data under the assumption of no linkage discards valuable information but does not introduce bias, although in order to form confidence intervals for the resulting estimates some assessment of the degree of linkage in the data (i.e. correlation between sites) must be made.

3.5.2 SNP data from American and Asian human populations

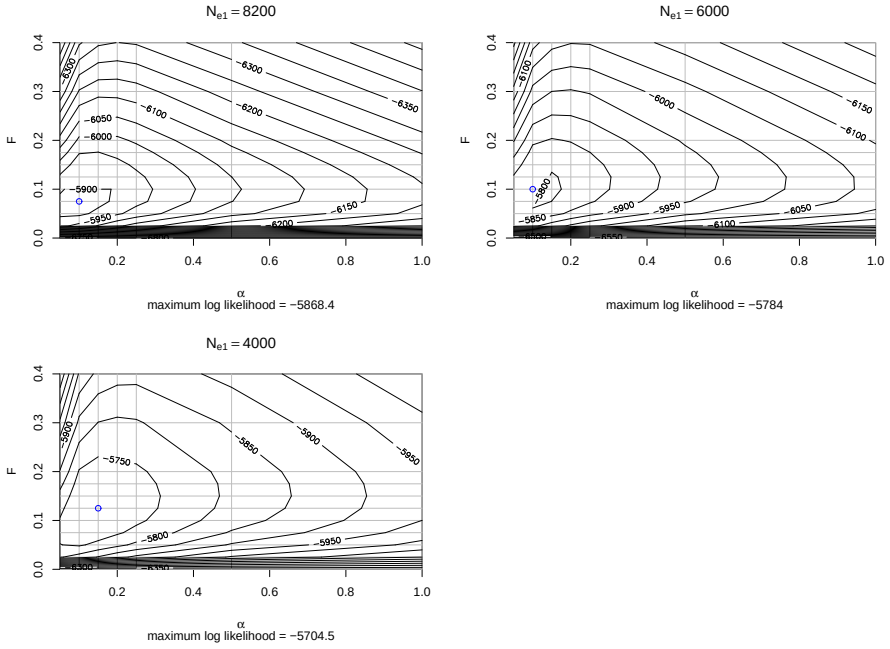
There is considerable evidence that the original human occupants of the Americas are descended from Asian populations. They are thought to have arrived during the late Pleistocene, but the details of this colonisation process remain uncertain. Questions of interest include the the number of independent colonisations, the time(s) of the colonisation(s), and the origins of the colonists within Asia. As an example of

the use of the new method, I apply it to data from some indigenous American and Asian populations represented in the data set of Conrad *et al.* (2006). The data used here are 18 genomic regions of average length ~ 90 kb, in each of which there are 60 SNPs (their 'core regions'). Conrad *et al.* (2006) used `fastPHASE` (Scheet & Stephens 2006) to infer haplotypic phase within each genomic region, and the resulting inferred haplotypes are used as the data here. See Conrad *et al.* (2006) for further details of the data set.

Figure 3.9 illustrates some results of fitting the new model to data from the Yakut of Siberia, and the Karitiana of Amazonian Brazil. These data were selected merely for illustration of the new method; many pairs of American and Asian populations could be selected from the data set of Conrad *et al.* (2006), and the analysis presented here is intended only as an initial exploration of the information in the full data set about the colonisation of the Americas. For each population, I selected 16 haplotypes at random from among those available. Despite the evidence presented by Conrad *et al.* (2006) for heterogeneity of recombination rates within many of the genomic regions (including several 'recombination hotspots'), I assumed that recombination occurred homogeneously within each region. One of the analyses of Conrad *et al.* (2006) (making use of the map of recombination rates in the human genome estimated by Kong *et al.* (2002)) resulted in estimates of the per-generation recombination probability r_i within each genomic region i, under the assumption that these do not differ among populations. I was therefore able to investigate the effect of assuming various effective populations sizes N_{e1} for the Yakut, allowing $\rho = 2N_{e1}r_i$ for region i to be determined by the corresponding estimate of r_i. The estimate of the effective size of the Yakut in Conrad *et al.* (2006) is 8200. Since it is apparent from section 3.5.1 that $\tilde{\rho}$ in the new model corresponds to a considerably larger actual value of



(a) MLEs for individual regions



(b) Likelihood surface for all 15 regions

Figure 3.9: **Drift since human colonisation of the Americas**. The figure shows results of analysing SNP data from 15 chromosomal regions from the Karitiana (Amazonian Brazil) and the Yakut (Siberia). Results are shown for three different assumptions about the effective size of the ancestral population, which is assumed to be the same as that of the Yakut.

$4N_{e1}r$, I evaluated the PAC likelihood at points of a grid in the $F - \alpha$ plane for $N_{e1} = 8200, 6000$ and 4000.

The three panels of figure 3.9a show the location of the joint estimates $(\hat{F}_{\text{pac}}, \hat{\alpha}_{\text{pac}})$ resulting from independent analysis of each genomic region, for each value of N_{e1}. Recall that, under this parameterisation, $\hat{F}_{\text{pac}}$ is the 'amount of drift since the split' in the Yakut population (i.e. $\hat{F}_{\text{pac}} = T/2N_{e1} = T/2N_{e\text{Yakut}}$), and $\hat{\alpha}_{\text{pac}}$ is the ratio of the effective size of the Karitiana to that of the Yakut (i.e. the ratio of the 'amount of drift since the split' in the Yakut to that in the Karitiana). The three panels of figure 3.9b plot estimated contours in the PAC likelihood surface based on all the data (i.e. that resulting from multiplication of the 18 PAC likelihood surfaces for the independent genomic regions). As expected from section 3.5.1, the highest overall likelihood is obtained under the model with $N_{e1} = 4000$, approximately half the size estimated by Conrad *et al.* (2006). The phenomenon of reduction in the estimate of both F and α in response to larger assumed effective size is also apparent, as it is in the results of section 3.5.1.

The data strongly suggest that, after the separation of the ancestral (Asian) population, substantially more drift occurred in the ancestry of the Amazonian population than in the ancestry of the Siberian population, in accordance with strong prior information suggesting the same. However, this suggests that suitable models for the data (ones in which α takes a fairly small value) are ones under which the estimators resulting from the new method have undesirable statistical properties (section 3.5.1). In particular, it seems that the estimate of $\hat{F}_{\text{pac}} \approx 0.1$ is likely to be an overestimate. Assuming the estimated effective size of 8200 for the Yakut and a simple model of constant effective sizes since the separation, $F = 0.1$ corresponds to 1640 generations

since the separation, or $(24,600, 41,000)$ years assuming generation times of $(15, 25)$ years respectively. This is considerably older than estimates based on archaeological evidence, which tend to favour a date of 15,000 years or earlier. If some of the drift in the histories of the two populations is the result of population bottlenecks, then the estimated time of the separation would be more recent. Occasional population bottlenecks seem especially plausible in the history of the Amazonian population, perhaps as a result of colonisation events at various stages of the presumed dispersal across Beringia and southward.

3.6 Discussion

In this chapter I have presented a new statistical model for haplotypes sampled from two populations. The model supposes that the haplotypes are descendents of a panmictic population which existed T generations ago, at equilibrium with respect to mutation, recombination and drift. Since that time, the haplotypes are assumed to have evolved independently under these three processes in two isolated populations. The long term effective size of those populations may have been unequal, resulting in different expected amounts of drift F_1 and F_2 in the two isolates. If it is assumed that the effective sizes of the two populations, although perhaps different, have remained constant since their separation then the expected amount of drift in daughter population i is $F_i = T/2N_{ei}$.

The motivation, outlined in sections 3.3.2 and 3.4.1, was that

1. because of recombination, a sample of genomes contains a lot of independent information about the history of the sampled population;

2. much of this information is contained in the 'haplotype structure' (*sensu* section 3.3.2);

3. it is not obvious how to design summary statistics that efficiently capture this information;

4. and yet no likelihood or approximate likelihood methods exist for models of population history other than that of an unstructured population of constant size.

The approach taken was to extend the approximate likelihood framework of Li & Stephens (2003) to the more complex model of population history considered here, and this chapter can be viewed as an investigation of the potential utility of that general approach.

3.6.1 Biases in parameter estimation

I have focused on the estimation of the parameters $\rho = 2N_{e1}r$ (the rate of recombination relative to drift in population 1) and F_1 and F_2 (or, equivalently, $\rho, F = F_1$ and $\alpha = F_1/F_2$). In section 3.5.1 I presented results illustrating some aspects of the frequentist behaviour of estimators based on the new approximate likelihood. In accordance with the results of Li & Stephens (2003), the method tended to underestimate ρ. To restate the observation, apparently a given rate of 'switching' in the copying process corresponds to a higher rate of recombination in the genealogical process. Although one would obviously not expect an exact equivalence, the reason for the direction is unclear. Since the exact likelihood estimator of Fearnhead & Donnelly (2001) also seems to underestimate ρ, perhaps the reason does not lie in the approximations made by the PAC likelihoods.

The magnitude of this bias depends on the model under which the data were simulated (figure 3.6) and on the assumed values of the other model parameters F and α. It is natural that there should be this latter dependency, because some of the information in the data about F and α derives from the chromosomal scale at which similarity in patterns of variation due to linkage decays, which is determined by ρ. I made a preliminary investigation of whether a simple alteration to the value of ρ used in the approximate likelihood calculation resulted in reasonably unbiased estimators of F and α (see section 3.5.1 and figure 3.7). The alteration was based on the observed bias when estimating ρ under restricted versions of the model. I found that while this seemed promising when the true amount of drift in the two populations is the same, it resulted in quite severe overestimation of F when drift differed by a factor of ten between the populations, as it may do in reality. Therefore important tasks that remain are to develop reasonable procedures for coping with bias in such higher-dimensional parameter spaces, and to improve the approximate likelihood such that the bias is reduced. In figure 3.7 the estimator based on the assumption that the sites are unlinked also performed reasonably well when drift was equal, but poorly when unequal. As always, it is possible that some of the behaviour of the model of unequal drift is the result of an error in implementation (a bug).

When the daughter populations have drifted by equal amounts, the new model performs reasonably well at estimating F, and it is apparent that it succeeds in capturing relevant information contained in haplotype structure (figure 3.8). There is an upward bias in $\hat{F}_{\mathrm{pac}}$, even when the data are drawn from an unstructured model, which is absent when no attempt is made to model linkage. Naively, one would expect that the introduction of structure into a model of data from an unstructured population would result in a decrease in likelihood because of penalties associated with

copying haplotypes in the 'other population' — a population which is entirely artificial and haplotypes which might very well be closely related. A possible explanation (suggested by G. Coop) is that this cost is overcome by an increase in likelihood resulting from the better ability of the structured model to fit variation in coalescence times around their expectations. It is interesting to note that, when modelling a panmictic population with the objective of inferring haplotypes from genotype data, Stephens & Scheet (2005) found it advantageous to increase the dimension of the hidden state space in a way that can be viewed as allowing two different 'copying times' as opposed to the single 'copying time' of Li & Stephens (2003). Since the unlinked model does not appear to benefit in this way, perhaps it is the time for recombination, rather than the time for mutation, which is being better fit.

3.6.2 Approximations made in the model

Many of the problems encountered when using PAC models for inference must lie in the way that the PAC likelihood approximates the true likelihood under a standard coalescent model. The various approximations made in this chapter include the following: the PAC likelihood depends unnaturally on the ordering of haplotypes; the copying times correspond to a model in which, conditional on S, k and ϕ, there is no variance in coalescence time; the prior on the missing data is not conditioned on the data observed so far; there is no formal justification for the mutation model (equation 3.7); the transition probabilities are inexact and the process generating genealogies along a chromosome is in any case not Markov (Wiuf & Hein 1999).

3.6.3 Recombination rate variation & model-based statistics

An important conclusion of recent studies in human genetics is that recombination does not occur homogeneously along the chromosome, but instead that there are 'recombination hotspots' at a density of approximately 1 per 50 kb — regions of chromosome under 2 kb in length in which the rate of recombination exceeds the background rate by 10 times or more (see e.g. Myers *et al.* 2005). Data sets suited to making this inference in other taxa are rare, but the phenomenon has also been observed in chimpanzees *Pan troglodytes* and mice *Mus domesticus* and in the yeast *Saccharomyces cerevisiae*. In order to use haplotype structure information to reach accurate conclusions about population history, this rate variation should be taken into account. Data analysis problems of this sort, in which it is deemed necessary to account for more complex features of the data-generation process, provide one argument for the use of methods based on the likelihood under a formal model for the data. For example, in the inference problem considered in this chapter, while it may be possible to design summary statistics that capture information about the parameters of the model contained in haplotype structure, it is surely harder to do so while using information about variation in recombination rates along the chromosome in a reasonable fashion. In contrast it would be straightforward to incorporate such information into a formal model such as that presented here, in which the recombination rate between each marker locus features explicitly.

3.6.4 Prospects

It seems probable that the quantity of available data on genomic variation in populations will continue to increase rapidly. Initially, the richest sources will probably be organisms that are the subject of genome-sequencing projects (see e.g. JGI 2006,

NHGRI 2006), and their close relatives. The inference problem addressed in this chapter lies at the heart of evolutionary biology, and in order that evolutionary biology responds effectively to these data further progress on the statistical problem is necessary. Here I have described the problem and one possible statistical framework for its solution, and investigated one particular implementation. The details described here will be modified if this general approach is pursued, but the objectives of this work will have been satisfied if it contributes helpfully to the challenge of revealing the considerable information that loosely linked genomic variation must contain about recent evolutionary history.

3.7 References

Bahlo, M. & Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor Popul Biol* **57**, 79–95.

Beaumont, M. (2001). Conservation genetics. In *Handbook of Statistical Genetics* (edited by D. Balding, M. Bishop & C. Cannings), chapter 29, pages 779–809. Wiley.

Beaumont, M. A., Zhang, W. & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–35.

Becquet, C. & Przeworski, M. (in prep.) .

Beerli, P. & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763–73.

Conrad, D., Coop, G., Jakobsson, M., Wen, X., Wall, J. D., Rosenberg, N. A. & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *submitted* .

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

Ewens, W. J. (1990). Population genetics theory - the past and the future. In *Mathematical and Statistical Problems in Evolution* (edited by S. Lessard), pages 177–227. Kluwer.

Falush, D., Stephens, M. & Pritchard, J. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–87.

Fearnhead, P. & Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–318.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet. Res.* **60**, 209–220.

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

Gaggiotti, O. E. & Excoffier, L. (2000). A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proc Biol Sci* **267**, 81–7.

Gay, J. (in prep.) .

Griffiths, R. C. & Tavare, S. (1994). Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46**, 131–159.

Hellenthal, G. & Stephens, M. (2005). A new method for estimating rates of gene conversion from population data. *http://www.stat.washington.edu/garretth/WNARsubmission.pdf* .

Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–60.

Hudson, R. & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics* **159**, 1805–17.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–8.

Jennings, W. B. & Edwards, S. V. (2005). Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. *Evolution Int J Org Evolution* **59**, 2033–47.

JGI (2006). US Department of Energy Joint Genome Institute `http://www.jgi.doe.gov/sequencing/DOEprojseqplans.html`. URL.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. & Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics* **31**, 241–247.

Kuhner, M., Yamato, J. & Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421–30.

Kuhner, M. K., Yamato, J. & Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–401.

Li, N. & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–33.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–4.

Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324.

Myers, S. R. & Griffiths, R. C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**, 375–94.

NHGRI (2006). National Human Genome Research Institute `http://www.genome.gov/10002154`. URL.

Nicholson, G., Smith, A., Jónsson, F., Gústafsson, O., Stefánsson, K. & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society B* **64**, 695–715.

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–42.

Nielsen, R., Mountain, J. L., Huelsenbeck, J. P. & Slatkin, M. (1998). Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* **52**, 669–677.

Nielsen, R. & Slatkin, M. (2000). Likelihood analysis of ongoing geneflow and historical association. *Evolution* **54**, 44–50.

Nielsen, R. & Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–96.

Nordborg, M. (2001). *Handbook of Statistical Genetics*, chapter Coalescent theory. Wiley.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286.

Scheet, P. & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing geneotypes and haplotypic phase. *American Journal of Human Genetics* **78**, 629–644.

Stephens, M. (2001). Inference under the coalescent. In *Handbook of Statistical Genetics*, pages 213–238. Wiley.

Stephens, M. & Donnelly, P. (2000). Inference in molecular population genetics. *Philosophical Transactions of the Royal Society Series B* **354**, 1–31.

Stephens, M. & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449–62.

Stephens, M., Smith, N. J. & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–89.

Tavaré, S. (1984). Line of descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26**, 119–164.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Voight, B. F. (2006). *Approaching Human Genetics from a Population-based Paradigm*. Ph.D. thesis, University of Chicago.

Wakeley, J. (2006). *Coalescent theory*. Roberts and Co.

Watterson, G. A. (1985). The genetic divergence of two populations. *Theoretical Population Biology* **27**, 298–317.

Wilson, D. J. & McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**, 1411–25.

Wiuf, C. & Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology* **55**, 248–259.

3.8 Appendix A: The 'forward' and 'backward' algorithms

When modelling loosely linked data, the approximate conditional sampling distributions of haplotypes have the form of a 'hidden Markov model' and so evaluation of the probability of the observation sequence (the haplotype), and evaluation of the posterior probability distribution on hidden states at each site, are standard procedures, effected according to the 'forward' and 'backward' algorithms respectively (see e.g. Rabiner 1989). However, since it is important to avoid certain computational inefficiencies, I describe the computations here as they apply to the PAC model.

Define the 'forward probability' for each hidden state at site l to be the joint probability of the hidden state and the data up to and including site l, conditional on the haplotypes sampled so far and the model parameters:

$$\alpha_l(x, s) = p(h_{*1}, \ldots, h_{*l}, X_l = x, S_l = s | h_1, \ldots, h_k, \phi). \tag{3.15}$$

As in section 3.2, I will leave the dependence on ϕ implicit hereafter. The approximate conditional probability $p(h_*|h_1, \ldots, h_k)$ is obtained as usual by summing the forward probabilities at the last site,

$$p(h_*|h_1, \ldots, h_k) = \sum_{x,s} \alpha_L(x, s), \tag{3.16}$$

and the PAC likelihood based on a single ordering of all the haplotypes is computed using (3.3).

The forward algorithm is initialised by setting the forward probabilities at the first (say leftmost) locus to

$$\alpha_1(s,x) = \tilde{p}(s)p(x|s)u(h_{*1}|h_{x1},s) \tag{3.17}$$

for each hidden state pair (s,x).

For resequenced data the physical spacing of consecutive pairs of marker loci is equal (1 bp) and therefore, under the assumption of homogeneous recombination rates along the chromosome, so is the rate of recombination between them. In this case, the transition probabilities between the hidden states are the same for all consecutive pairs of loci and the Markov chain specified by the transition probabilities (3.10), (3.12) and (3.13) has a stationary distribution $\pi(s,x)$. For resequenced data, in order that the prior distribution on the hidden states does not depend on the chromosomal location, I evaluate $\pi(s,x)$ numerically and use $\tilde{p}(s) = \sum_x \pi(s,x)$ in the initialisation. For irregularly-spaced SNPs however, the rates of recombination between consecutive marker loci vary, and therefore so do the transition probabilities, and the Markov chain on hidden states has no stationary distribution. In this case I use $\tilde{p}(s) = p(s)$ and the prior therefore differs along the chromosome.

The forward probabilities at sites to the right are computed recursively, using the values at the adjacent site to the left, according to

$$\alpha_l(x',s') = u(h_{*l}|h_{x'l},s') \sum_{x,s} \alpha_{l-1}(x,s)p(x,s \to x',s'). \tag{3.18}$$

For computational efficiency it is important to avoid an unnecessary extra loop over haplotypes by storing the quantities $f_{l-1}^{(D)} = \sum_x \alpha_{l-1}(x,D)$ and $f_{l-1}^{(A)} =$

$\sum_x \alpha_{l-1}(x, A)$ and performing the computation in (3.18) instead as

$$\alpha_l(x', s') = u(h_{*l}|h_{x'l}, s')\Big[f_{l-1}^{(D)} r(D \to s')p(x'|s') + f_{l-1}^{(A)} r(A \to s')p(x'|s') + \alpha_{l-1}(x', s')q(s')\Big]. \tag{3.19}$$

In this expression $r(s \to x', s')$ is the probability with which at the next site haplotype x' is copied at level s', conditional on copying at level s at the current site, *as a result of a switch.* $q(s')$ is the probability that no switch occurs (and therefore the same haplotype is copied at the next site), conditional on copying at level s', and is given by

$$q(s') = \begin{cases} 1 - p(R_t) & \text{if } s' = D; \text{ see (3.11)} \\ 1 - p(R_D^n \cup R_D^o \cup R_A) & \text{if } s' = A; \text{ see (3.14),} \end{cases} \tag{3.20}$$

The probabilities of transitions as a result of a switch can be defined as

$$r(s \to s') = \frac{p(x, s \to x, s') - I(s' = s)q(s)}{p(x|s')}. \tag{3.21}$$

Define the 'backward probability' for each hidden state at site l to be the joint probability of all the data to the right of l, conditional on the hidden state, the haplotypes observed so far and the model parameters:

$$\beta_l(x, s) = p(h_{*,l+1}, \ldots, h_{*L}|X_l = x, S_l = s, h_1, h_2, \ldots, h_k, \phi). \tag{3.22}$$

The posterior probability that site l is in hidden state (x, s) is proportional to the product of the forward and backward probabilities at that site

$$p(X_l = x, S_l = s|h_1, \ldots, h_{k+1}) = \frac{\alpha_l(x, s)\beta_l(x, s)}{\sum_{x',s'} \alpha_l(x', s')\beta_l(x', s')}. \tag{3.23}$$

The backward algorithm is initialised by setting these probabilities to 1 for all hidden states at the last locus (since there is no data to the right of the last locus). The backward probabilities at loci to the left are computed recursively, using the values at the adjacent site to the right, according to

$$\beta_l(x, s) = \sum_{x',s'} p(x, s \to x', s')u(h_{*,l+1}|h_{x',l+1}, s')\beta_{l+1}(x', s'). \tag{3.24}$$

The analogous efficiency measure in the backward algorithm to that described above for the forward algorithm is to store the quantities $b_{l+1}^{(D)} = \sum_x u(h_{l+1}^*|h_{x,l+1}, D)\beta_{l+1}(x, D)$ and
$b_{l+1}^{(A)} = \sum_x u(h_{l+1}^*|h_{x,l+1}, A)\beta_{l+1}(x, A)$ and to perform the computation in (3.24) instead as

$$\beta_l(x, s) = r(s \to D)b_{l+1}^{(D)} + r(s \to A)b_{l+1}^{(A)} + q(s)u(h_{*,l+1}|h_{x,l+1}, s)\beta_{l+1}(x, s). \tag{3.25}$$

3.9 Appendix B: Computing the PAC likelihood efficiently for resequenced data

Resequenced data may feature blocks of sites in which the k haplotypes sampled so far all have the same allele as the $(k+1)^{\text{th}}$ haplotype. It is unnecessary to compute the forward and backward probabilities explicitly at each such site because the emission

probabilities remain constant, and if the blocks of monomorphic sites are large it may be computationally inefficient to do so. Let the emission probability of observing the same allele i as that on the copied haplotype, conditional on $S = s$, be

$$u_0(s) = u(i|i, s).$$

Let P be an $m \times m$ matrix, where $m = k_{z_*} + k$, containing the probabilities of all the possible transitions multiplied by the corresponding emission probability, with 'daughter' events preceding 'ancestral' events along each margin. That is,

$$P_{xx'} = \begin{cases} p(D, x \to D, x)u_0(D) & \text{if } x \leq k_{z_*} \text{ and } x' \leq k_{z_*} \\ p(D, x \to A, x)u_0(A) & \text{if } x \leq k_{z_*} \text{ and } x' > k_{z_*} \\ p(A, x \to D, x)u_0(D) & \text{if } x > k_{z_*} \text{ and } x' \leq k_{z_*} \\ p(A, x \to A, x)u_0(A) & \text{if } x > k_{z_*} \text{ and } x' > k_{z_*}. \end{cases}$$

In the forward case, suppose that site l is the leftmost of a block of B monomorphic sites. The forward probabilities at site $l + B - 1$ are required so that those at the polymorphic site $l + B$ can be computed. They are

$$\alpha_{l+B-1} = \alpha_l P^B, \tag{3.26}$$

where α_l is a row vector containing the forward probabilities in the order corresponding to the margins of P. That is

$$\alpha_l = [\alpha_l(D, 1), \ldots, \alpha_l(D, k_{z_*}), \alpha_l(A, 1), \ldots, \alpha_l(A, k)]. \tag{3.27}$$

In the backward case, suppose that l is the site to the left of the rightmost site in a block of B monomorphic sites. The backward probabilities at polymorphic site $l - B + 1$ are required so that the backward probabilities at the site to the left can be computed. They are

$$\beta_{l-B+1} = P^B \beta_l, \tag{3.28}$$

where β_l is a column vector arrayed in the same way as α_l. The matrix P^B can be computed as usual via an eigenvector decomposition. I.e. let V be a matrix containing the eigenvectors of P in its columns, and Λ be an $m \times m$ matrix containing the eigenvalues of P along its diagonal and zeroes elsewhere. Then

$$P^B = V \Lambda V^{-1}.$$

3.10 Appendix C: Results from coalescent theory used in the PAC likelihood

The prior on the missing data features the quantity $p(A) = p(A; k)$, which is the marginal (single locus) probability that a newly sampled chromosome coalesces into the existing tree in the ancestral population, when k chromosomes have already been sampled from the same population. Let $g_{ij}(t)$ be the probability that i lines coalesce to $j \leq i$ over scaled time t. Tavaré (1984) gives exact expressions for $g_{ij}(t)$ (see also Wakeley 2006, ch. 3 p. 66). Let $H_k(t)$ be the probability that a newly sampled $(k+1)^{\text{th}}$ line has not coalesced by scaled time t. Suppose that a lines remain distinct at scaled time F_*. The probability that a particular one of the $k + 1$ lines was not

involved in any of the coalescences is

$$\prod_{j=k+1}^{a+1} \frac{\binom{j-1}{2}}{\binom{j}{2}} = \prod_{j=k+1}^{a+1} \frac{j-2}{j} = \frac{a(a-1)}{k(k+1)}.$$

$H_k(t)$ can be obtained by averaging this quantity over the distribution on the number a of distinct lines at scaled time t:

$$H_k(t) = \sum_{a=2}^{k+1} g_{k+1,a}(t) \frac{a(a-1)}{k(k+1)}. \tag{3.29}$$

Let F_* be the drift parameter for the population from which the new line was sampled. Then

$$p(A) = H_k(F_*). \tag{3.30}$$

$\tilde{t}_D$ is the expected coalescence time of the new line, measured on the same time scale as F_*, conditional on coalescence before F_*. Since $1 - H_k(t)$ is the cumulative density function (cdf) of the coalescence time, $\tilde{t}_D$ can be obtained as

$$\tilde{t}_D = \int_0^{F_*} H_k(t) dt. \tag{3.31}$$

I evaluate the integral numerically in R.

$\tilde{t}_A$ is the expected coalescence time, conditional on coalescence after T, measured in units of $2N_{eA}$. It is equal to $F_1 = T/2N_{eA}$ plus the expected amount of scaled time between T and the coalescence of a $(k+1)^{\text{th}}$ line, given that coalescence occurs after T (i.e. in the ancestral population). Conditional on the numbers a_1 and a_2

of distinct ancestral lines entering the ancestral population from the two daughter populations, this is simply the expected time to coalescence of the $(a_1 + a_2)^{\text{th}}$ line under panmixia, which is $2/(a_1 + a_2)$ (Fu & Li 1993). $\tilde{t}_A$ can therefore be calculated by averaging this quantity over the joint distribution on (a_1, a_2). Since a_1 and a_2 are independent, for the case in which the new line was sampled in population 1,

$$\tilde{t}_A = F_1 + \frac{1}{p(A)} \sum_{a_1,a_2} g_{k_1+1,a_1}(F_1) \frac{a_1(a_1-1)}{k_1(k_1+1)} g_{k_2,a_2}(F_2) \frac{2}{a_1+a_2}, \tag{3.32}$$

where k_1 and k_2 are the numbers of haplotypes sampled so far from each daughter population ($k_1 + k_2 = k$).

CONCLUSION

This dissertation has focused on the problem of making inferences about population history from genetic variation in natural populations. I have argued that, in addition to the intensive study of genomic variation and genomic function in tractable model systems, and continued progress towards a molecular biological understanding of phenotypic variation (which has a medical dimension in humans), a fundamentally important task in evolutionary biology is achieving genuine integration of the fields of systematic biology and biogeography with population genetics. This will involve (i) improving the ability to collect informative data sets from a phylogenetically diverse range of study organisms, and (ii) providing a general statistical framework for making inferences about the most recent portion of evolutionary history.

Genomic and phenotypic differentiation, and evolution of intrinsic reproductive incompatibilities, require reductions in levels of gene flow, and by far the most frequent way in which this occurs is spatial separation of populations (e.g. Mayr 1942, Coyne & Orr 2004). The most recent portion of evolutionary history is therefore characterised by spatial replacement of populations, and research at the intersection of population genetics, biogeography and systematics involves characterisation of spatial variation in genomes, of intrinsic and extrinsic reproductive isolation between spatially replacing populations, of the population genetic basis for spatial variation in phenotype, and of the inferences about evolutionary history that may be drawn from these spatial data.

Much important work in these areas has already been carried out, frequently labelling itself as 'phylogeography'. However, this field has been dominated by detailed reconstructions of genealogies of mitochondrial genomes (e.g. chapter 1), some of the problems of which were discussed in chapters 2 and 3. The phylogeographic literature is addressing some of the most complex and ambitious inference problems in population genetics, and it is therefore worrying that it has been so widely criticised from a statistical point of view (e.g. Barton & Wilson 1996, Knowles & Maddison 2002, Irwin 2002, Hey & Machado 2003). There is a tendency towards pronounced understatement of the uncertainties that remain after the data have been analysed, and this problem is exacerbated by the positive spin that is *de rigueur* in scientific publication.

The immediate future in population genetics is particularly exciting because of the potential for new and very large genomewide data sets in many areas. The high information content of large data sets dictates that substantial resources should be allocated to development and implementation of careful statistical analyses. However such genomewide data sets will initially be associated with genome sequencing projects, and in the immediate future it will continue to require considerable innovation to generate large polymorphism data sets affordably in organisms of arbitrary phylogenetic provenance. Furthermore, the inherently spatial nature of the data at the intersection of population genetics, biogeography and systematic biology means that the spatial distribution of samples is critical in determining the questions that may be answered, and neither fancy genotyping technologies nor fancy computer programs will obviate the often logistically-challenging requirement for field collection of sufficient numbers of samples at sufficient spatial density throughout large areas. Continuation of intensive field surveys, and collection and preservation of suitable

material for phenotypic and genetic analyses, is clearly essential.

Issues of experimental design in spatial population genetics have been somewhat neglected. There has been work on the theoretical and statistical aspects in the panmictic setting (e.g. Pluzhnikov & Donnelly 1996, Felsenstein 2006), but relatively little work in structured settings (e.g. chapter 2). The spatial distribution of samples is an important aspect, although one frequently has much less control over this than would be ideal. There is also a need for clearer discussion of the procedure for deciding on a genotyping strategy that is suited to the aims of a particular study; currently such decisions are heavily influenced by historical allegiance to particular genetic markers. An argument can also be made for more pragmatic choice of study organisms. Although the ideal in systematics is to describe biological diversity with little regard for tractability of study systems, if current priorities in low-level systematics include establishing a general framework for the analysis of large spatial phenotypic-population genetic data sets, then it may be appropriate to start with organisms for which obtaining large amounts of polymorphism data has been facilitated by previous work. On the other hand, vast areas of the south-east corner of Amazonia studied in chapter 1 now comprise small patches of forest in a matrix of land cleared for agriculture, and to ensure that we know anything about the distribution of genetic and phenotypic diversity among forest organisms in these areas, collection of study material should be done now.

A key challenge in data analysis is to improve statistical methods for the analysis of spatial genetic data. In particular, there is a need for better methods to identify zones across which gene flow is unusually low when there is isolation-by-distance. Recent work in this area has focused on unlinked genetic markers. An example is the

work of Guillot *et al.* (2005), who incorporate the information on sample locations into the prior on cluster membership in a model-based clustering scheme similar to that of Pritchard *et al.* (2000). It is also possible to adapt methods from the spatial statistics literature to model spatial variation in allele frequencies at unlinked loci (e.g. Vounatsou *et al.* 2000, Wasser *et al.* 2004), and these approaches could be extended to make inferences about spatial heterogeneities in rates of gene flow. In some organisms, rates of recombination relative to mutation may be so high that there is little linkage between physically nearby polymorphisms. When this is not the case, making full use of the information on haplotype structure in loosely-linked spatial genetic data is an important challenge. Chapter 3 made a start on this problem in a very simple model of population history. Much future work is required on the problem of making inferences about the evolutionary history of geographically structured populations, but there are grounds for optimism as genomes are large, genomic variation is potentially highly informative and its study is the best available opportunity for learning about the evolutionary past.

References

Barton, N. & Wilson, I. (1996). Genealogies and geography. In *New uses for new phylogenies* (edited by P. H. Harvey, A. J. L. Brown & J. M. Smith). Oxford University Press.

Coyne, J. A. & Orr, H. A. (2004). *Speciation*. Sinauer.

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol* **23**, 691–700.

Guillot, G., Estoup, A., Mortier, F. & Cosson, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170**, 1261–1280.

Hey, J. & Machado, C. (2003). The study of subdivided populations — new hope for a difficult and divided science. *Nat. Rev. Genet.* **4**, 535–543.

Irwin, D. E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution* **56**, 2383–94.

Knowles, L. & Maddison, W. (2002). Statistical phylogeography. *Mol. Ecol.* **11**, 2623–2635.

Mayr, E. (1942). *Systematics and the origin of species*. Columbia University Press, New York, U.S.A.

Pluzhnikov, A. & Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* **144**, 1247–62.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–59.

Vounatsou, P., Smith, T. & Gelfand, A. E. (2000). Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics* **1**, 177–189.

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B. & Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proc Natl Acad Sci U S A* **101**, 14847–14852.