Genetic Differentiation in Spatially Structured Populations with Particular Regard to the Atlantic salmon (*Salmo salar* L.)

Ian Joseph Wilson

B.Sc. (University of Durham)
M.Sc. (University of Sheffield)

A thesis presented for the degree of Doctor of Philosophy at the University of Edinburgh, Scotland.

In collaboration with SOAEFD Marine Laboratory, Aberdeen.

January 1996
Abstract

The interpretation of genetic diversity within subdivided populations is a difficult problem. This thesis develops new theory for its interpretation, and applies it to the problem of within river variation in Atlantic salmon (*Salmo salar* L.) populations which have been the subject of many surveys of genetic diversity throughout their range. These surveys have found lower levels of genetic variation than many fish species, even those occupying similar environments such as the Brown Trout (*Salmo Trutta* L.). Genetic differentiation has been observed both between samples from different rivers, and from samples from different tributaries of the same river. The levels of differentiation are of considerable interest from both scientific and commercial viewpoints. Despite this intense study the reasons behind the levels of differentiation between populations are unclear, and it is uncertain whether there are distinct stocks of fish within rivers. As an attempt to understand this differentiation and to guide any future studies new mathematical and statistical models for the riverine environment are developed.

The riverine environment within which salmon breed is unlike any of the previously produced models for genetic differentiation. This thesis shows that the branching pattern within rivers can lead to a distinct pattern of genetic variation which is unlike that seen in linear habitats. These habitats may lead to increased differentiation between populations. A stochastic approach to genetic variation, the coalescent process, allows us to separate the processes involved in the production of genetic data, into the genealogical and mutational processes. This approach allows us more flexibility in the modelling of genetic data. The various processes which have been invoked as causes for the observed genetic variation are analysed using these new tools.

Statistical analysis of genetic variation in populations has traditionally proceeded using either Wright's $F_{ST}$ or Slatkin's rare alleles approach. Here I introduce a new maximum likelihood approach to the analysis of data from subdivided populations which is more independent of the number of alleles in the population than either of the other two approaches. This technique also allows us to obtain confidence limits for the dispersal rate between populations.

When animals are sampled from the wild the boundaries between populations are not always distinct. This can result in sampling a mixture of several populations. Here I look at the power of tests for detecting such population admixture, and to examine the power of tests for mixed populations. Studies on the Atlantic Salmon are reanalysed with regard to this new theory, and the future direction of the study of population differentiation in salmonids is reappraised.
Acknowledgements

Many people have helped me over these past years, and have been instrumental in the production of this thesis. Thanks must go to my supervisors, Drs. Eric Verspoor, and Rob Fryer at the Marine Laboratory in Aberdeen, and Professor Nick Barton at Edinburgh University for sharing their knowledge and guiding me through my postgraduate work, and more recently for reading through countless chapter drafts.

The knowledge of salmon biology by members of the Stock Structure section of SOAEFD in the Marine Laboratory in Aberdeen and the Freshwater Fisheries Laboratory in Pitlochry is unsurpassed; and I appreciated discussions on salmon biology, ecology, and life history with Eric Verspoor, Alan Youngson, Anne McLay, and John Webb from Aberdeen; and David Hay, Ian McLaren, and Mike Donaghy from Pitlochry. Special thanks must go to everyone who took me on their fieldwork trips, for their patience with a novice biologist (although I now consider myself proficient in freshwater carpentry).

All credit for my knowledge of the arcane workings of enzyme electrophoresis, PCR, and single locus probes is due wholly to Jo Murphy, Caroline Thompson, and David Knox.

Specific help was given by Eric Verpoor, who supplied me with unpublished data on genetic variation on the river Dee, and maps of the rivers Dee and Teno, and by Stuart Baird who taught me the power of recursive programming in Pascal; his demonstration program is the ‘root’ of all simulation programs in Chapter 3.

The offices in Aberdeen, and Edinburgh have been friendly throughout, and I am especially grateful for the kindness and encouragement of those with whom I shared offices while writing up, and to everyone else; all support was greatly appreciated.

Thanks to my family for their support through my long student career, and to everyone in Edinburgh and Aberdeen who have made my time enjoyable.

Special thanks to Selina, for all her help and support for the last 3 years.

Financial support for this work was provided by a SOAEFD Marine Laboratory studentship.
# Contents

## CHAPTER 1

### INTRODUCTION

1.1 BACKGROUND ................................................................. 1  
1.2 POPULATION DIFFERENTIATION AND MANAGEMENT ............. 3  
1.2.1 Genetic Variation ......................................................... 3  
1.2.2 Population Structure .................................................... 5  
1.2.3 Local Adaptation ........................................................... 7  
1.2.4 Homing, Straying and Dispersal ....................................... 8  
1.2.5 Population Differentiation and Fisheries Management ............ 10  
1.3 THE ATLANTIC SALMON (*Salmo salar* L.) ............................. 11  
1.3.1 Life History ................................................................. 11  
1.3.1.1 Commercial Importance ........................................... 12  
1.3.2 Genetic Variation in the Atlantic Salmon ............................. 13  
1.3.3 Direct Measures of Homing .......................................... 14  
1.3.4 Stocking and Farmed Escapes ....................................... 14  
1.3.5 Local Adaptation in Atlantic Salmon ............................... 15  
1.3.5.1 Evidence of Selection at the MEP-2 Locus .................... 18  
1.4 SUMMARY .................................................................. 19

## CHAPTER 2

### CLASSICAL MODELLING OF GENE FLOW WITHIN RIVER SYSTEMS .... 21  
2.1 INTRODUCTION .............................................................. 21  
2.2 PRELIMINARIES - CLASSICAL POPULATION GENETICS .......... 22  
2.2.1 The Wright-Fisher Model .............................................. 22  
2.2.2 The Effective Size of a Population .................................. 23  
2.2.3 Mutation Models .......................................................... 24  
2.2.4 The Kolmogorov Forward Equation for the Distribution of Gene Frequencies .......... 25  
2.2.4.1 Wright's Formula for the Distribution of Gene Frequencies ........................................ 26
## Contents

2.2.4.2 The Infinite Island Model ................................................................. 27
2.3 MEASURES OF GENETIC VARIABILITY AND DISTANCE ....................... 28
  2.3.1 Within Population Measures .............................................................. 28
    2.3.1.1 The probability of identity in state ............................................... 29
    2.3.1.2 The probability of identity by descent .......................................... 29
    2.3.1.3 Wright’s Inbreeding coefficient .................................................... 30
  2.3.2 Fixation Indices .................................................................................. 30
  2.3.3 Genetic Variation Between Populations ................................................ 31
    2.3.3.1 F-Statistics .................................................................................... 31
    2.3.3.2 Identity Coefficients Between Populations ...................................... 33
  2.3.4 Measures of Distance between Populations ........................................ 33
    2.3.5 Pairwise Measures ............................................................................ 34
      2.3.5.1 Pairwise \(F_{ST}\) ................................................................. 34
      2.3.5.2 Nei’s Distance, \(D\) .................................................................. 34
  2.4 THE INFLUENCE OF SELECTION ............................................................. 35
  2.5 THE FINITE ISLAND MODEL ................................................................. 37
    2.5.1 Finite Island Model with Selection .................................................... 39
  2.6 ISOLATION BY DISTANCE ..................................................................... 40
    2.6.1 Clines and Isolation by Distance ...................................................... 42
  2.7 THE EFFECT OF BRANCHING ON MEASURES OF POPULATION DIFFERENTIATION ................................................................................... 43
    2.7.1 Introduction ....................................................................................... 43
    2.7.2 Model .............................................................................................. 44
    2.7.3 Comparison of linear habitats and trees of order \(k\) ............................... 50
    2.7.4 Results for Small Trees ................................................................. 50
    2.7.5 Simulation ....................................................................................... 52
    2.7.6 \(F_{ST}\) within a tree system ............................................................. 57
  2.8 MORE REALISTIC RIVER SYSTEMS ....................................................... 60
    2.8.1 Single Branch Point ........................................................................ 61
    2.8.2 More Branch Points ........................................................................ 63
  2.9 DISCUSSION .......................................................................................... 64

CHAPTER 3

THE COALESCENT IN SUBDIVIDED POPULATIONS ........................................ 66

3.1 INTRODUCTION ...................................................................................... 66
3.2 THE COALESCENT PROCESS .................................................................69
  3.2.1 The Coalescent process in a single population .........................72
    3.2.1.1 Pairwise Coalescence Times ...........................................72
    3.2.1.2 Sample of r genes .......................................................74
  3.2.2 The Structured Coalescent .......................................................75
  3.2.3 Simulating the Coalescent Process .........................................77
3.3 MEASURES OF DIVERSITY FROM THE COALESCENT .............79
  3.3.1 Within Population Measures .................................................79
  3.3.2 Between Population Measures ..............................................81
3.4 THE COALESCENT AND EFFECTIVE POPULATION SIZES ..........83
3.5 THE FINITE ISLAND MODEL .............................................................85
  3.5.1 Moment Generating Functions ..............................................85
    3.5.1.1 Identity and Distance Measures .................................86
    3.5.1.2 Effective Population Size .........................................89
  3.5.2 The Distribution of Pairwise Coalescence Times in an Island Model ....90
3.6 HIERARCHICAL ISLAND MODEL .....................................................92
3.7 POPULATION FISSION .................................................................100
  3.7.1 Distance and Diversity .........................................................102
3.8 EXTINCTION AND RECOLONISATION .........................................106
  3.8.1 Island model with no migration ...........................................106
    3.8.1.2 Identity and Distance Measures ................................108
    3.8.1.3 Mean Coalescence Times ..........................................108
    3.8.1.4 Effective Population Size .........................................109
    3.8.1.5 Examples ...............................................................109
    3.8.1.6 Testing the models forward in time .........................110
  3.8.2 Extinction/Recolonisation Model with Migration ....................113
3.9 THE ISLAND-FISSION MODEL .......................................................114
3.10 ISOLATION BY DISTANCE .........................................................117
  3.10.1 One Dimensional Habitats .................................................118
    3.10.1.1 The Laplace Transform of Malecot’s equation ..............119
    3.10.1.2 Wright’s Method of Neighbourhood Size ..................119
  3.10.2 Finite Linear Habitats .......................................................122
3.11 RANGE EXPANSION .................................................................123
  3.11.1 Simple Model .................................................................124
  3.11.2 Range expansion with subsequent dispersal ......................127
CHAPTER 4

MAXIMUM LIKELIHOOD ESTIMATION OF GENE FLOW

4.1 INTRODUCTION

4.2 MODEL

4.2.1 Limitations in Using the Beta Distribution

4.2.2 Test of Beta Distribution

4.2.3 Limitations of Binomial Sampling

4.3 EXTENSION TO K ALLELES PER LOCUS

4.4 PRIVATE ALLELES AND THE BETA BINOMIAL

4.5 ESTIMATION USING MAXIMUM LIKELIHOOD

4.5.1 Estimating $F_{ST}$ using Maximum Likelihood

4.5.2 Other Maximum Likelihood Estimators

4.5.3 Estimation over Several Loci

4.6 THE MAXIMUM LIKELIHOOD ESTIMATOR OF M

4.6.1 Tests of the Estimators

4.7 TESTS ON SIMULATED BETA-BINOMIAL DATA

4.8 OTHER ESTIMATORS OF GENE FLOW

4.8.1 Gene Flow Estimates from Coalescence Times

4.8.1.1 Estimating Mean and Variance of Coalescence Times

4.8.2 Gene Flow Estimators from Frequency Data

4.8.2.1 Nei's $G_{ST}$

4.8.2.2 Weir and Cockerham's $\theta$

4.9 SIMULATION STUDIES

4.9.1 Details of Programs

4.9.2 Coalescence Times

4.9.3 The Effect of Mutation

4.9.4 Maximum Likelihood Estimates

4.9.5 Estimation of $F_{ST}$ using Maximum Likelihood

4.9.6 More than One Locus

4.10 DISCUSSION
## CHAPTER 5

**THE POWER OF STATISTICAL TESTS OF POPULATION DIFFERENTIATION**

5.1 **INTRODUCTION** .................................................. 175

5.2 **TESTS FOR HARDY-WEINBERG DISEQUILIBRIUM** ............. 177

5.2.1 Power for 2 Allele Tests ............................................. 179

5.2.2 Power for Simulated Data ............................................. 182

5.3 **POWER OF TESTS OF GENETIC DIFFERENTIATION** ......... 184

5.3.1 Test Statistics ......................................................... 185

5.3.1.1 Definitions ......................................................... 185

5.3.1.2 Chi-Squared Tests ................................................ 185

5.3.1.3 G-Test ............................................................... 185

5.3.1.4 Fisher's Exact Test ................................................ 186

5.3.1.5 Maximum Likelihood Estimate for Gene Flow .............. 186

5.3.1.6 Weir and Cockerham's θ ........................................ 186

5.3.2 Power of Tests for Simulated Data ............................... 187

5.4 **DISCUSSION** ......................................................... 190

## CHAPTER 6

**THE STRUCTURE OF ATLANTIC SALMON POPULATIONS**

6.1 **INTRODUCTION** .................................................. 193

6.2 **F_{ST} WITH OVERLAPPING GENERATIONS** .................. 195

6.3 **THE ABERDEENSHIRE DEE** ..................................... 201

6.3.1 Description ......................................................... 201

6.3.2 Tagging Experiments .............................................. 203

6.3.3 Genetic Data ......................................................... 208

6.4 **OTHER STUDIES OF GENETIC VARIATION** ................. 210

6.5 **DISCUSSION** ......................................................... 217
CHAPTER 7

DISCUSSION ........................................................................................................................................... 220

7.1 THEORETICAL RESULTS ........................................................................................................... 220

7.2 MANAGEMENT IMPLICATIONS ............................................................................................ 223

REFERENCES ....................................................................................................................................... 227

APPENDIX 1: ALGORITHMS FOR SIMULATIONS ON TREE STRUCTURES .................. 243

APPENDIX 2: ALGORITHMS FOR QUICK RANDOM DRIFT CALCULATIONS .......... 244

APPENDIX 3: ALGORITHMS FOR GENERATION OF TREES .............................................. 246

APPENDIX 4: RANGE EXPANSION SIMULATION ........................................................................... 249

APPENDIX 5: MAXIMUM LIKELIHOOD ESTIMATION USING SYSTAT ...................... 253

APPENDIX 6 MUTATION ALGORITHMS ....................................................................................... 256

APPENDIX 7: PERMUTATION ALGORITHMS ........................................................................... 258

APPENDIX 8: GENEALOGIES AND GEOGRAPHY ................................................................. 260

APPENDIX 9: GENEALOGIES AND GEOGRAPHY (CONDENSED VERSION) .......... 295
Chapter 1

INTRODUCTION

1.1 Background

Knowledge of the extent and distribution of genetic diversity in the Atlantic salmon (*Salmo salar* L.) is important to its management, and more generally, is crucial to the understanding of evolutionary processes such as adaptive differentiation. The Atlantic salmon has been the subject of a considerable number of surveys of genetic variation and differentiation, yet these surveys, while informative, have not provided the required information on the subdivision of the species (in the wild) into genetic populations. Yet, effective management of wild salmon, in relation to its conservation, exploitation, and enhancement, requires knowledge of the biology of populations. A major part of this is information about the geographic scale spanned by populations.

Populations, and stocks in fisheries management, can be defined as geographic groupings of animals which interbreed randomly, but among which interbreeding is restricted. If populations are more or less reproductively isolated then they are best managed as separate units. Thus knowledge of levels of interbreeding between geographic locations is required to define management units.

Most Atlantic salmon live in rivers for part of their lives before migrating to sea and returning to freshwater to spawn. It is known (through tagging studies) that some degree of homing to natal rivers, and to tributaries within larger river systems, occurs, and juvenile Atlantic salmon are restricted in their movement within rivers. Thus population subdivision is likely.
One way of assessing the levels of interbreeding among different geographical groups would be to look at the reproductive relationships of individuals, by means of physical tagging, and observation. However, the riverine environment is alien to our experience and direct observation is difficult. In addition most Atlantic salmon spend a large part of their life at sea, so observation over the entire life cycle of a fish is extremely difficult.

An alternative approach is to use spatial patterns of the distribution of heritable genetic variation to try to infer dispersal rates and distances. If there is variability at a genetic locus, there is information about population history: hence we can potentially calculate which individuals are most closely related. Two individuals who have a recent common ancestor are likely to share more DNA variants than two unrelated individuals. The patterns of genetic variation observed in a sample from a wild population then represent the degree to which members of the sample share a common ancestry. Thus the patterns will reflect the history of interbreeding which are influenced by selection, gene flow, extinction of local populations, and other population processes.

How much information is contained in spatial patterns of differentiation, and what may be inferred from them, are current topics of research. Genetic variation in natural populations of Atlantic salmon has been surveyed for the last 30 years, mainly in the form of allozyme variation. The levels of genetic variation seen are low in relation to other species (Johnson 1984). New techniques, developed in the last 10 years, reveal greater levels of genetic variability (Verspoor et al. in prep.), and may allow us greater accuracy in determining the relative amounts of inbreeding. However, whether this extra information is worth the additional work and expense is unclear.

Genetic differentiation among wild populations is widespread. To relate differentiation to ecologically important factors requires population genetic models. Previously, most models produced were for populations in continuous habitats; and are inappropriate for salmon which live in rivers which are branching and linear. The effects of this structure on population differentiation are unknown. Another potential complication is the adaptation of life history traits to local habitats, which is observed for wild Atlantic salmon within the same river. Whether this differentiation constitutes genetic adaptation of local populations, or whether the adaptations are
due to phenotype plasticity is unclear. If we use selectively neutral genetic variation to assess population structure and gene flow this is not an issue. There is, however, evidence that variation in at least some allozyme loci in Atlantic salmon is non-neutral. Non-neutrality may be of particular importance for Atlantic salmon if farmed escapees, which have been shown to reproduce in the wild, are significantly genetically differentiated from natural populations, as it may affect the nature and extent of any induced genetic change.

This thesis examines the various processes which can cause observable patterns in genetic variation in the Atlantic salmon, looks at their effects, and patterns, and develops methods which may be used to analyse them. By way of introduction the analysis of genetic variation in relation to population structuring, local adaptation, dispersal, and its application in the context of fisheries management will be reviewed, and the evidence obtained so far for local adaptation and population differentiation in the Atlantic salmon (Salmo salar L.) will be examined.

1.2 Population Differentiation and Management

1.2.1 Genetic Variation

The actual levels of within species genetic variation in natural populations have only come to be appreciated in the last 40 years. Previously, levels of phenotypic variability for meristic characters gave an indication that there was variation, and even geographic subdivision (e.g. Thompson 1931). Dobzhansky and Wright (e.g. 1941, studies reviewed in Wright 1978) surveyed populations of Drosophila pseudoobscura in California recording the frequency of recessive lethal alleles in different populations, which showed geographic variation, and hence a degree of reproductive isolation. Since then considerable levels of genetic variation have been seen in the nuclear DNA of most species of animals and plants, observable through the range of techniques: blood group proteins, allozymes, micro-, and mini-satellites, and substantial genetic variation has also been observed in cytoplasmically transmitted elements (mitochondrial DNA in animals, chloroplasts in plants (Hartl & Clark 1989)). It is arguable whether any of these techniques produce estimates of genetic variation typical of the genome as a whole. The variation may be less (scoring allozymes does not necessarily show all the genetic variation), or more (the
allozymes chosen for investigation in a species may be those shown to be variable in other species biasing the levels of variation seen). These arguments are explored more fully in Gillespie (1991).

The levels of genetic variation seen in these first studies were alarming to some people due to the perceived genetic loads they would place on the population. As an answer to this “the neutral theory of molecular evolution” due to Kimura and co-workers (Kimura 1983) held that most genetic variation in population is neutral, or at most weakly selected against, so that this load theory did not apply. However, it has been argued (e.g. Gillespie 1991) that genetic load theory is misleading, and that it is the distribution of observed genetic differences which are important. Gillespie (1991) provides a synthesis of an alternative view to the neutral theory. See also McDonald (1983) for a review of studies of selective local differentiation in allozymes.

Models of genetic variation for selective neutral loci are more workable than for selected loci. Hence most studies of genetic variation from natural populations assume, at least in the first instance, that the variation is neutral (Kimura 1983), an assumption made, for the most part, in this thesis. This assumption is questionable, and the influence of selected alleles on levels of genetic differentiation will be considered for data from Atlantic salmon populations.

For this study the absolute levels of genetic variation are not important in themselves; instead the patterns of shared ancestry between different populations which can, in theory, reveal patterns of population structure are the topic of study.

---

1 The genetic load is defined by Crow (1958) to be the difference in mean population fitness and that of a 'perfect' genotype, as a ratio to the 'perfect' genotype. The implication of Kimura (1983) is that, considering the segregational load, with 2000 loci with a heterozygote advantage of 0.01, and selection eliminating unfit genotypes prior to reproduction, that 22,000 offspring must be produced by each individual. Gillespie (1991) argues that the load theory only applies to the range of heterozygosity seen in natural populations.
1.2.2 Population Structure

The importance of population structure in evolution has been controversial for most of the twentieth century. Before that time, workers such as Wallace believed that most new species would evolve at the limits of a species range by allopatric speciation (Endler 1977). In contrast Darwin believed that sympatric speciation was also possible. In the 1930’s the mathematical theory of natural selection was being developed by Fisher (1930), Wright (1931), and others. Fisher believed that the phenomenon of local races which had been observed previously (Thompson 1931) was due to differential selection in distinct subpopulations, and that large population sizes seen in natural populations would make the influence of random drift in gene frequencies inconsequential compared with these selective processes. An opposing view, championed by Sewall Wright (1931, 1943), held that the subdivision of large populations into smaller subpopulations drove evolution because random drift in gene frequencies could mean that small populations could overcome the epistatic interactions between genes which he thought would otherwise stop adaptation. This is the ‘Shifting Balance Theory of Evolution’ (Wright 1978). In this theory there are three phases of evolution in a subdivided population. First, random drift, due to small local population sizes, results in the crossing of threshold values in the surface of fitness values (epistatic interactions are assumed to cause multiple local maxima); then in the second phase, selection within the local population leads to a new fitness peak superior to the previous one; and finally the fitter individuals migrate from the local population to swamp the whole population. The first two phases have been shown to be possible theoretically (Barton & Whitlock 1996 and references therein) but there is still considerable debate about whether Wright’s shifting balance is a force in natural populations and whether local subdivision plays any part in evolution (Endler 1977). New models of population subdivision such as the metapopulation model, where individual populations become extinct and are recolonised from the other subpopulations (Gilpin & Hanski 1989; Hastings & Harrison 1994), have also been developed.

There are many possible processes resulting in population structure: limited gene flow causing isolation by distance (Felsenstein 1976, Slatkin 1987a); extinction/recolonisation (Gilpin & Hanski 1989); the effects of population fission/fusion (Neel 1978); and range expansion (Nichols & Hewitt 1994, Sokal et al. 1991). These have all been used as explanations of the patterns of genetic
differentiation seen within populations when the genes concerned are neutral. When we have non-neutrality, adaptation to local conditions is taken to be a possible reason for the subdivision (Endler 1977). In some ways all will affect the levels of differentiation to some extent, but the interactions between these processes are hard to investigate. Selection in particular can occur in so many different ways that to incorporate it into models is very difficult. However, theoretical results show that, in some circumstances, selection will produce gene frequency distributions within populations which are indistinguishable from neutrality (Gillespie 1991).

Comparative studies such as those by Waples (1987) give us an idea of the differences in population structuring and levels of gene flow seen in animals from similar environments. The use of statistics such as $F_{ST}$ allow us to compare the levels of variation seen in different species (Waples 1987), and from the same species in different locations, and over varying spatial scales (Jordan 1992). Wright (1931) showed that in the absence of selection, levels of gene flow greater than one per generation meant that substantial differentiation could not occur between subpopulations, and levels higher than four meant that little differentiation would be seen. This topic has been reviewed by Felsenstein (1976) and Slatkin (1985a, 1987a).

Models of genetic variation with population structure have, until recently, been based on following gene frequencies, or gene frequency distributions (e.g. Crow & Kimura 1970). While these models give important information, their structure masks processes which ultimately determine levels and patterns of genetic variation. Two genes are identical when they share a common ancestor in the past and no mutations have occurred in either line-of-descent since then, or, more rarely, when back mutations have occurred to produce the same allele. Modelling using a coalescent approach brings the genealogical aspect of genetic variation to the fore; concentrating on the times at which genes in a sample share common ancestors. This approach is of use when considering strategies for sampling from natural populations; the effect of different types of mutation can be considered with the same patterns of relationship between genes.

The coalescent approach is of particular use when considering small samples relative to the total population size, as the mathematics are more tractable. This is usually the case when in surveys of natural populations; a coalescent approach is thus
valuable when modelling surveys from natural populations. Effective sampling strategies require knowledge of the precision of estimates, and the power of tests for genetic differentiation or various types of genetic data. A coalescent approach is extremely useful for investigating the properties of estimators for a variety of population genetic models.

1.2.3 Local Adaptation

A complicating factor for population differentiation models is local genetic adaptation. Numerous studies have claimed local adaptation at the ecological, morphological, and behavioural levels, where it may be easier to show differentiation (Carvalho 1993) than at a genetic level, but it is unclear whether the adaptations are genetic, or just due to phenotypic plasticity, particularly in fish.

Local adaptation of genotypes is dependent on gene flow being low so that any alleles which are selected under local environmental or ecological conditions are not swamped (Slatkin 1973, Nagylaki 1975). In a one dimensional habitat there are reasonably simple conditions on the ratio of the strength of selection to the amount of dispersal for an allele favoured in one portion of the habitat to prevail. However, the amount of dispersal which is required to swamp local adaptation on small spatial scales in one-dimension is fairly low, and when the relative fitnesses of alleles are of the order of the proportion of dispersers, local adaptation is unlikely to occur. Where there are more than a single locus, and we have epistatic interactions then the relationships become much more complicated (e.g. Emlen 1991), and it is difficult to create general, and robust models. In order to create workable models it is necessary to know how local adaptation acts within populations.

To demonstrate local adaptation of a feature, and that this is under genetic control, we must show (McDonald 1983):

- that the feature has a genetic basis;
- that differential expression of the feature leads to differential reproductive fitness;
- a mechanism of selection responsible for maintenance of the trait in a population.
Chapter 1: INTRODUCTION

These three conditions serve to rule out alternative explanations such as phenotypic plasticity, or panmixia with differential survival of genotypes in local environments. Nevertheless, these are strict conditions, and to unequivocally show the action of selection in natural populations is very difficult (Christiansen 1989). Few studies on enzyme polymorphism have convincingly shown the role of selection on allozymes in natural populations (but see Gillespie 1991 for examples of selection on allozymes). One exception in fish is that of the killifish (*Fundulus heteroclitus*). The kinetics, and other functional differences, between alternate forms of various enzymes, in particular lactase dehydrogenase enzyme (*Ldh*), have been extensively studied for many years (Powers et al. 1991 and references therein). Functional differences have been found in several enzymes which show clinal variation up the Eastern seaboard of the North America where there is a large temperature gradient. This suggests that many enzymes, at least for this fish, may be under selection. However, the dearth of examples indicates that this may not be representative of enzymes as a whole. The killifish may not be typical as the clines are believed to have formed from secondary contact between two forms of the fish separated for the last ice age (Powers et al. 1991). Nevertheless there are a number of studies on fish which indicate that selection on enzymes, or on closely linked genes, may be important to fitness (Torrissen et al. 1993, Allendorf et al. 1983, Danzmann et al. 1988). The large amounts of work involved in such studies preclude experiments which could answer this question conclusively.

At what scale local adaptation may occur in fish species is unclear. There is greater difficulty for looking at local adaptation in fish than other species because of their phenotypic plasticity. Heritabilities of traits are generally much lower in fish than in other vertebrates (Allendorf et al. 1987) and age, and size at sexual maturity, are interrelated in fish in a way that permits great flexibility in life history strategy without loss of reproductive success (Myers 1986). These factors make the study of local adaptation in fish populations extremely difficult.

1.2.4 Homing, Straying and Dispersal

In Atlantic salmon, and other animals where migration to breeding grounds occurs, the rates of dispersal are not usually thought of as being adaptive, but are regarded as straying (e.g. Mills 1989). However, when we consider the amount of dispersal between populations, then there is the possibility that the rates of straying are
adaptive. There has been considerable theoretical work on the evolution of dispersal. Early work assumed that dispersal was a strategy in variable environments; was a group selective strategy (Johnson & Gaines 1990); or was inbreeding avoidance (e.g. Greenwood 1987). However Hamilton and May (1977) have shown that dispersal could be adaptive even in stable habitats with no inbreeding depression. Work on dispersal, reviewed by Johnson and Gaines (1990), indicates that the levels of dispersal can be a strategy under selection, and that for most situations there is an optimal dispersal rate, which is greater than zero. The evolution of dispersal, when there is variation in the time until germination for plant seeds (which is in some ways analogous to the semelparous breeding with variable times in freshwater and in the marine environment which is the natural state of Atlantic salmon in Scotland), has been studied (Cohen & Levin 1991), and has indicated that levels of dispersal are lower when there is variation in the germination time (analogous to variable maturation times in Atlantic salmon). The patterns of dispersal in space and time, under this model, depend on the patterns of spatial and temporal correlation in fitness. Viewed in this way dispersal can be seen as an ‘averaging’ effect over uncertainty in nature (Cohen & Levin 1991), however it is unclear whether this can be considered advantageous.

Dispersal rates cannot be considered to be completely adaptive. Homing in Atlantic salmon has been thought to occur for some time, since at least the 16th century (Stabel 1984, Mills 1989). Many authors have claimed that this is a strategy so that local adaptation may occur (Ryman 1983). However, other animals which travel long distances also home with considerable accuracy, for example Laysan albatrosses return to an isolated island to breed on average 19 metres from the nest in which they fledged (Greenwood 1987), but it is unlikely that local adaptation is happening on this scale. An alternative explanation is that the only strategy for successful breeding that an Atlantic salmon may know is to spawn where it was spawned, or, using an optimality argument, it may be ‘cheaper’ in some sense to equip a salmon to home than to equip it to find a successful spawning site. Homing, however, may provide the opportunity for populations to become locally adapted. There is evidence of genetic variation in homing ability in Pacific salmon (Reisenbichler 1988) and circumstantial evidence of heritable variation (a prerequisite for natural selection) from return rates of ranched Atlantic salmon to the vicinity of hatcheries (Kallio-Nyberg & Ikonen 1992). However measurement of the heritability of
dispersal distances has proved to be an extremely difficult technical problem (Johnson & Gaines 1990).

1.2.5 Population Differentiation and Fisheries Management

Fish are a major food resource. No other major resource is so predominantly gathered from wild populations. To exploit this resource optimally it is necessary to understand the biology of the fish (Ståhl 1987), a difficult problem for marine fishes, as the marine environment is a difficult habitat for us to study. Yet, in order to exploit fish resources efficiently it is necessary to understand the size and distribution of the exploited populations, or stocks. The stock concept has a long history in fisheries management, but can be defined in exactly the same terms as a population in ecology, namely geographic groupings of plants or animals which interbreed randomly but among which interbreeding is restricted (Pitcher & Hart 1982). Also, for certain species temporal stocks can also be seen (Aspinwall 1974). If stocks are viewed from an evolutionary viewpoint then we may define stocks as locally adapted sub-populations or demes (Ståhl 1987, Carvalho 1993), however local adaptation is an unnecessary condition. An alternative viewpoint regards stocks as economic resources (e.g. Gulland 1983).

Genetic variability allows us to get information on the size of local stocks by using genetic markers as tags. If there is little gene flow between stocks then it may be possible to identify stocks by one, or a combination of genetic markers. The earliest attempts to use genetic variation in the Atlantic salmon were of this type. Physical tagging had previously indicated that the stocks caught in the Greenland fishery were from a variety of natal rivers along the East coast of North America and in Europe, while the Faroese fishery caught fish of European origin (Mills 1989). Surveys of genetic data were an attempt to estimate the relative proportions of salmon of North American and European salmon in the commercial fisheries (Ryman 1983). The Greenland fishery was surveyed, along with rivers in North America and Europe. This allowed estimates of the relative contributions of the two continents to the Ocean fishery to be made (Bermingham et al. 1991). This data together with evidence from tagging studies made clear that salmon from different rivers, even in nearby localities were divided into more or less reproductively isolated
subpopulations. As more detailed studies were performed, it became clear that there was further substructuring within river systems (Verspoor et al. 1991).

This leads to a second problem for anadromous salmonid fisheries management, namely over what scale should stocks be managed in the freshwater environment. This is important for the management of local fisheries because the tributaries can be taken as stocks, and the conventional view is that separate stocks of fish should be managed individually. The presence of difference in gene frequencies between tributaries within rivers is not, in itself, evidence for separate stocks within each tributary (differential selection in different localities could account for this); there should be reproductive isolation in order to define separate stocks.

A third potential interaction between fisheries management, and genetics, is the possible problem of differential harvesting. For example, over the past 60 years the average size of Chinook salmon caught along the north coast of North America has declined by more than 50% and the average age at maturity has declined by approximately 2 years (refs. in Porter et al. 1986). One explanation for this effect is that later maturing fish are differentially harvested and thus the genetic makeup of the population is altered. Models for maturity in the Atlantic Salmon by Myers (1986) predict the same effects if older fish are differentially harvested. This topic will not be considered in this thesis.

1.3 The Atlantic Salmon (Salmo salar L.)

1.3.1 Life History

The Atlantic salmon (Salmo salar L.) has a range, apart from introductions in Australasia and South America, through the North Atlantic from Greenland and Iceland in the north to Massachusetts and the Bay of Biscay in the south (Mills 1989). It is predominantly anadromous and always spawns in rivers; spends the first 1-5 years of its life in rivers; then, in most cases, smolts and migrates to sea where it spends between 1-3 further years (Mills 1989). Fish returning after 1 winter at sea are grilse, the remainder are multi-sea-winter (2SW, 3SW etc.). Precocious maturation of males in fresh water occurs at a variable frequency throughout the
salmon’s range. Tagging and genetic studies have shown that the salmon returns to its natal river, and even to tributaries within its natal river (Youngson et al. 1994).

Our knowledge of the Atlantic salmon’s life at sea is extremely limited, and most of this comes from commercial fisheries. There have been commercial fisheries, presently not active, off the coasts of Greenland and the Faroe Islands which catch salmon during their time at sea, and returning Atlantic salmon are caught along the coastlines of Ireland and Britain, yet much of the rest of their migration and life at sea is a mystery.

An important feature of the Atlantic salmon when we discuss its evolution and local adaptation is the phenotypic plasticity which is observed, which as a consequence makes the study of life history traits extremely difficult. Salmon from the same river system may spend vastly different times in freshwater, may spend varying amounts of time at sea, and may return at different times of the year to their natal river. The estimates of heritability for Atlantic salmon are generally low (e.g. Thorpe et al. 1983), and estimates from cultured populations may be misleading.

There are a number of resident races of Atlantic salmon, both in North America, and Europe (see Mills 1989). A large number of alternative names have been suggested for these forms. However, Verspoor (1994) has shown that where there are sympatric anadromous and non-anadromous Atlantic salmon then the genetic differentiation between these two types is of the same order as that between different continents. However it is not consistent for all comparisons and thus is insufficient for the life history forms to be labelled as separate species. Some authors (e.g. Mills 1989) suggest that the North American and European forms should be designated as separate species, or sub-species.

1.3.1.1 Commercial Importance

There are a number of different aspects to the commercial use of the Atlantic Salmon:

- sports fisheries throughout Europe and North America;

- commercial fisheries both offshore: Greenland and the Faroe Islands; and inshore fisheries along the coast, and in river estuaries; and
commercial salmon farming.

As a consequence the scientific effort expended on the salmon has been considerable for the last hundred years.

1.3.2 Genetic Variation in the Atlantic Salmon

There have been many studies on genetic variation in salmonids over the last 30 years. Most of these have confirmed the presence of genetic variation between rivers in the juvenile stage, and also for returning adults. This has resulted in a huge store of potentially useful data, but its literature, particularly for the Atlantic salmon, is confused and no adequate review exists to date.

The first studies on Atlantic salmon considered transferrin alleles (Payne et al. 1971). These first attempts to examine genetic variation in the Atlantic salmon attempted to use genetic markers as tags for discriminating the natal location of salmon caught by the Greenland fishery (Ståhl 1987). Since that time 260+ (Wilson & Verspoor submitted) allozyme studies have been completed on Atlantic salmon in North America and Europe. Mitochondrial DNA variation has also been assayed (Bermingham et al. 1991, Hovey et al. 1989, O'Connell et al. 1995).

One major finding of these surveys is that the levels of genetic variation in the Atlantic salmon are low compared to other fish species, and relative to other salmonids (Ryman 1983, Johnson 1984). Relatively large levels of population differentiation are found between fish from the Baltic, the rest of Europe, and North America. However, the levels of genetic differentiation for fish within these major geographic groupings are small. The MEP-2* locus is clinally distributed in Europe and North America (Verspoor & Jordan 1989). This is discussed in a later section.

There are differences in allele frequencies between rivers, and between tributaries within rivers, but these are generally small (e.g. Jordan et al. 1992, Heggberget et al. 1986). There are potential problems in comparing surveys reported by different authors because of the difference in nomenclature used for variant alleles. In this thesis the main concern is variation within river systems. Due to the large number of surveys which have been performed, and problems associated with allele
nomenclature making comparisons between studies hazardous, only surveys done on single river systems are considered. A list of studies used is given in Chapter 6.

Most surveys on Atlantic salmon are done on juvenile fish from tributaries of the main rivers as it is easier to electro-fish on small tributaries, and the owners are more agreeable to the loss of a few juveniles. This adds somewhat to the problems of assessing population structure as the structure of the main river is unknown.

1.3.3 Direct Measures of Homing

Direct measures of homing from tagging studies have been done since 1903 (Mills 1989). Accurate measures for the amount of dispersal are difficult to obtain as we can only infer from the proportion of marked fish the probability that a returning fish is natal to the river. Added to this problem is the tendency of Atlantic salmon to enter non-native rivers and swim upstream, before returning to the sea and entering another river reported by some authors (Laughton 1991, Clarke & Purvis 1990). Another problem is that early tags were large, and clumsy, and may produce differential mortality in tagged fish. Hence, these studies do not tell us about dispersal distances but just the frequencies of returning fish from natal, and non-natal rivers. This gives an upper limit to the amount of gene flow, as the non-native fish may not breed as successfully as natal fish. Recent studies with microtags on the Girnock burn in Aberdeenshire have indicated that the homing shows levels of site fidelity of greater than 90% for females, and greater than 50% for males to a particular small tributary (Youngson et al. 1994, D. Hay pers. comm.). Additionally, these studies show that the vast majority of male spawners in the earlier half of the season are natal to the river, but later on more straying may occur, which suggests that the later season entrants are opportunists, who may have already spawned elsewhere, and whose contribution of genetic material will be minimal.

1.3.4 Stocking and Farmed Escapes

One of the most controversial aspects of wild fisheries management are the consequences of escapes of farmed salmon. A related concept is stocking of rivers with non-native fish with ‘desirable’ traits. In some ways these two processes are the same, but while stocking has long been considered beneficial to the local salmon stocks, farmed escapees are usually considered detrimental (e.g. Wilkins 1984). This is due to the potential spread of disease, and the consequences of poorly adapted
fish on the local populations (Hansen & Bakke 1989). Many farmed salmon in Scotland are of Norwegian origin (Verspoor et al. 1991), and the possible use of farmed salmon of Canadian origin may be of even more concern. Artificial propagation of fish in farmed and hatchery environments may result in a loss of genetic variation (Allendorf et al. 1987). A gene bank of threatened stocks of salmon in Norway has recently been initiated (Hansen & Bakke 1989).

Webb and co-authors have used the flesh colourant canthaxanthin to track the breeding success of escaped, farmed salmon. This chemical was used widely in the fish farming industry until mid-1990 (Webb et al. 1991, Webb et al. 1993a). This pigment is detectable in eggs, alevins, and first feeding fry (Webb et al. 1993b). Fry of farmed origin have been found in most rivers along the West coast of Scotland (Webb et al. 1993b), at variable frequency with a mean of 5.1%. In a separate set of studies (Webb et al. 1991, Webb et al. 1993a) on the river Polla, close to a major farm escape, farmed fish were shown to contribute about 50%, and 5% of the parr in successive years after the escape. Hybridisation has been shown to occur more frequently between salmon and wild Brown trout (Salmo trutta) with the hybridisation being between female farmed salmon and male trout (Youngson et al. 1993).

The influence of escaped salmon on natural stocks is of concern mainly on the West Coast of Scotland, as the salmon farms are more common there. There is evidence from other species in Scotland that non-native introduced species can invade, and displace the native genetic races. For example, Sika deer are in the process of replacing Red deer in the Kintyre Peninsular (Abernethy 1995).

Guidelines are needed for acceptable levels of introgression of farmed salmon into Atlantic salmon stocks. There is some controversy on this point with Bentson (1994) arguing that little threat is posed by farmed salmon.

1.3.5 Local Adaptation in Atlantic Salmon

There is little direct evidence for local adaptation in the Atlantic salmon. Taylor (1991) reviewed the evidence for local adaptation in salmonids. Pacific salmon
show considerable variation between populations in a number of traits, which may be adaptive including:

- orientation (Quinn 1982),
- homing (Reisinbichler 1988), and
- swimming capabilities (Taylor and McPhail 1985).

Several published studies have shown direct evidence of genetic differences in the Atlantic salmon, which are likely to be adaptive. Hansen and Jonsson (1991) reared eggs from two natal rivers in Norway in a hatchery. These authors showed that the return times from these hatchery reared fish were different: which indicates the genetic basis behind life history variation; however, no evidence of additive genetic variance, which is a prerequisite for natural or artificial selection, was presented. Ranching of salmon has indicated that probability of return may be selected, and is thus heritable (Kallio-Nyberg & Ikonen 1992).

Riddell et al. (1981) tested the hypothesis that the differences in body morphology and time of downstream migration for juvenile Atlantic salmon were adaptive, and under genetic control. Controlled breeding experiments were performed to investigate the genetic basis of phenotypic differences observed between natural populations of salmon parr from tributaries of the main river. These experiments were performed under hatchery conditions. The amounts of genetic variation in life history characters were then analysed to find what proportion are due to within population variation.

Body morphology was significantly different between populations, and Riddell et al. (1991) argued that body shape was important for swimming efficiency and hence was important under differing stream conditions. The heritability estimates on these characters are extremely small.

These studies do not provide incontrovertible evidence for local adaptation. The study by Hansen and Jonssen (1991) did not adequately show heritability for the traits, and Riddell et al. (1981) did not show the adaptive basis to the studied traits. There are many studies on ecological, morphological, and behavioural traits, each of which give some weak evidence for adaptation but the extreme plasticity of the
salmonids in general (Taylor 1991) mean that genetic differences cannot be
automatically invoked.

More dramatic evidence comes from the ectoparasite Gyrodactylus salaris
Malmberg 1957, endemic to the Baltic, which was accidentally introduced to
Norwegian salmon rivers in the 1970's. This parasite reduced the density of salmon
parr within some rivers in Norway to close to zero (Johnsen & Jensen 1986).
Hatchery reared stocks of Baltic salmon show resistance to this parasite, whereas
salmon of Norwegian origin show no resistance (Bakke et al. 1990).

Fish within a large river system such as the Aberdeenshire Dee in Scotland show
large variation between populations in a variety of life history traits (Webb &
McClay unpubl., J. Webb, A. McLay, E. Verspoor, A. Youngson pers. comm.) such
as:

- spawning time,
- time of return from the sea,
- egg size,
- smoltification timing and
- proportion of grilse, and 2 sea-winter fish.

All these properties have selective effects, as the difference in conditions between the
lower tributaries of the Dee, on an agricultural flood plain, and the upper tributaries,
in highland streams are large. Even if these traits are not adaptive they provide
barriers to gene flow between fish from opposite ends of the river, as the fish breed
up to 2 months apart.

Atlantic salmon show considerable variation in the percentage of males that mature
precociously in fresh water. Myers (1986) has carried out mathematical modelling of
life history parameters using an evolutionary stable strategy (ESS) approach
(Maynard Smith 1982) which indicates that increased levels of male precocity may
be related to increased mortality in the smolt or spawning migrations. Thorpe et al.
(1983) have shown that the incidence of precocious maturation in salmon was to
some extent genetically determined. Some authors (e.g. Porter et al. 1986) have
taken the decline in the run of spring salmon as evidence for local adaptation because the spring fish are differentially captured prior to spawning.

There is evidence to support adaptation over large scales from transplantations of wild salmon from within Scotland (Verspoor pers. comm.), and strong suggestions from unsuccessful restocking of Spanish rivers (Garcia de Leaniz et al. 1989). However, adaptation over such long ranges is not local when the water conditions are so different between Northern Europe and Spain.

The alternative viewpoint is that salmon are a supremely plastic species, and that the adaptive traits are simply phenotypic responses to the environment in which they are found. Naturalisations of salmon to the great lakes and to New Zealand lakes have taken place (Mills 1989), though these were extremely difficult to achieve.

All of these studies ignore the marine phase of the salmon’s life history during which there may be considerable selection (Mommsen et al. 1980).

1.3.5.1 Evidence of Selection at the MEP-2 Locus

Genetic variation in the diallelic malic enzyme (MEP-2*) (Cross & Ward 1980) is clinally distributed both in North America and Europe (Verspoor & Jordan 1989), with the 100 allele predominating in southern populations. The allele has been found to be associated with high summer temperatures within river systems (Verspoor & Jordan 1989, Verspoor et al. 1991). This pattern of gene frequencies is consistent with selection acting on this, or a closely linked locus, with respect to temperature. Jordan and Youngson (1991) showed that, in culture, there were some growth differences between the different MEP-2* genotypes. Pringle (1996) expanded this study and investigated the growth performance of the 3 genotypes in tank conditions in three different temperature regimes. In this study he found association between increased growth and the *125 allele in various temperature groups. However, there was no obvious association between growth and temperature. While this study only investigates growth for a short period of life, and is not a measure of fitness, the growth rate in freshwater may be correlated with fitness, particularly as growth rates in freshwater has been shown to affect the frequency of precocious maturation (Thorpe 1986). How natural selection affects this allele frequency is unclear; however, the evidence for some sort of selective effect is incontrovertible, particularly because of the association with temperature within single river systems.
Chapter 1: INTRODUCTION

The major effects of this locus may be in the marine environment, which is extremely difficult to investigate. However, Mommsen et al. (1980) have shown that MEP-2* plays an important role in the spawning migration of Sockeye salmon (Oncorhynchus nerka).

1.4 Summary

There have been a considerable number of studies of genetic variation in natural populations of Atlantic salmon (Salmo salar). Data from these investigations indicate unequivocally that there is population subdivision between continents, between regions, between drainages within regions, and even within individual river systems. Local adaptation has been observed in some studies between fish from different rivers, and there is a large amount of inconclusive evidence that there are adaptations within rivers. However, work on this problem is still in its infancy. The levels of variation in Atlantic salmon are low compared to most species of fish, and even compared to closely related species such as the brown trout (Salmo trutta). The levels of population differentiation seen within river systems are small, as are the levels seen from neighbouring river systems. This form of differentiation is expected from results on homing in the Atlantic salmon which indicate that fish home to their natal river, and even to streams within their natal river. The adaptive significance of this has yet to be unequivocally demonstrated although there is considerable circumstantial evidence that this is the case.

Strong evidence that variation at MEP-2*, or at a closely linked locus, is selected has been obtained but it remains unclear whether this is indicative of the genome as a whole or is an isolated case. The nature of the selective differences which can be attributed to this locus are unknown.

The degree of importance that local adaptation of Atlantic salmon stocks have to sports fisheries is a large, difficult, and to a great extent unanswered question. There is little concrete evidence for widespread local adaptation, though this sort of evidence is notoriously difficult to obtain (Taylor 1991). There is a considerable amount of inconclusive evidence both for enzyme loci and for life history traits. Attempting to answer the above question is beyond the scope of this thesis and assumptions of selective neutrality are made, almost entirely, in this thesis. Instead,
the view is taken that evaluation of the levels of genetic variation between populations is one of the preliminary steps by which we may infer the degree of local adaptation. To understand the processes involved in local populations, and hence to predict the consequences of the escapes of farmed salmon we need to have information both on local structuring, and on the degree of local adaptation.

Chapter 2 reviews the literature of classical population genetics, and produces a new model for neutral genetic variation on a tree structure. Chapter 3 moves from the genetics of whole populations to the interpretation of samples of genes from natural populations using the coalescent. This model is of use when considering sampling strategies from natural populations, as the effects of mutation are modelled explicitly, and separately from the genealogy of genes. The coalescent process is utilised to investigate some classical models, and also new models which have relevance to some of the hypothetical explanations, short of natural selection, which have been postulated for the patterns of genetic variation seen in populations.

Chapter 4 describes an improved maximum likelihood technique for calculation of levels of gene flow from natural populations. Chapter 5 is an investigation into the power of exact tests for Hardy-Weinberg disequilibrium, and population subdivision. The application, and consequence of the power of tests on genetic data from wild populations of Atlantic salmon is discussed.

Chapter 6 applies methods from the previous chapters to data for Atlantic salmon genetic variation and tagging returns. Particular use is made of unpublished data on the River Dee in Aberdeenshire, Scotland.

Chapter 7 discusses results presented in this thesis and attempts to ascertain where our level of knowledge of the Atlantic salmon currently stands. Recommendations on the future direction of studies are made in light of the results presented in this thesis.
2.1 Introduction

The previous chapter reviewed work on the differentiation of local populations (demes) in natural populations of Atlantic salmon. This chapter first reviews the theory relating genetic differentiation to limited dispersal. A new model for genetic differentiation in branching structures such as river systems, where dispersal is constrained to a tree like structure, is then developed.

Population structure has implications for the management of populations of Atlantic salmon, as well as having implications for evolution. There are two separate considerations from a management perspective: with populations where there is limited or no dispersal then the populations must be considered as separate units. Parallel to this, if there is limited gene flow then local adaptation may be an important consideration.

In both of the above cases it is important to know the levels of dispersal between populations. In order to gauge dispersal, we need to understand how population structure affects the patterns of genetic diversity observed in natural populations. The levels of local adaptation possible in one and two dimensional models are dependent on the gene flow between populations, the selective advantage of an allele
and the area over which an allele is selected (Nagylaki 1975, Slatkin 1973). Therefore, in order to investigate the possible importance of local adaptation it is important first to understand the levels of gene flow.

Parallel to the production of models for genetic differentiation has been the development of statistics to describe genetic variation seen within populations. There is considerable confusion in the literature between parameters used in models and statistics that can be measured from natural populations (Chakraborty & Danke-Hopfe 1991).

Atlantic salmon breed in rivers, which are superficially linear habitats. However, the branching structure is likely to have an effect on the levels of population differentiation seen. Few or no models for genetic differentiation have been developed on tree structures (but see Sawyer 1978). The final part of this chapter develops a model for gene flow on a simple tree structure, using the classical models reviewed in this chapter. These results are extended to examine the consequences of isolation by distance between branch points using simulation studies.

### 2.2 Preliminaries - Classical Population Genetics

#### 2.2.1 The Wright-Fisher Model

The Wright-Fisher model, produced independently by Wright (1931) and Fisher (1930), is the framework that has been and is most commonly used for building models incorporating random genetic drift in non-overlapping generations. The model deals with the changes in frequencies of alleles A and a in a randomly mating diploid population of size N, assuming no mutation and no selection. Under the Wright-Fisher model, the probability of j alleles of type A, given i alleles in the previous generation is

\[
P_j = \binom{2N}{j} \left( \frac{i}{2N} \right)^j \left( \frac{2N-i}{2N} \right)^{2N-j}.
\]  

(2.1)

The mean gene frequency in the next generation for replicate populations starting with a frequency of \( p = \frac{i}{2N} \) is then \( p \) and the variance in gene frequencies is \( p(1-p)/2N \) from known properties of the binomial distribution. The justification
for this model is random sampling with replacement from the underlying gene-
frequencies in the previous generation, which is equivalent to binomial sampling from
an infinite pool of gametes. Equation 2.1 defines a Markov chain with absorbing
states at 0 and 2N. Any population under this model will eventually go to fixation
for either allele. The time until fixation, and the probability of fixation for the
alternate alleles is a function of the population size and the number of alleles of each
type at the start.

The above model for the changes in the number of alleles of the two types is
equivalent to a multinomial distribution for the number of descendants of each of the
genes in the population. For this model the probability of the vector $x \in \mathbb{N}$
is given by

$$
\Pr(x) = \frac{(2N)!}{x_1! \cdots x_{2N}!} \left( \frac{1}{2N} \right)^{x_1} \cdots \left( \frac{1}{2N} \right)^{x_{2N}},
$$

where $\sum x_i = 2N$. Considering the model in this way is of more importance in the
next chapter.

### 2.2.2 The Effective Size of a Population

Most populations in nature do not follow the Wright-Fisher model. There may be
two sexes, overlapping generations and other factors which affect the number of
individuals contributing genetic material to the next generation.

The effective size of a population, discussed in Crow and Kimura (1970), is a way of
relating other mating structures to the 'ideal' Wright-Fisher model. There are 2
commonly used effective population size measures: inbreeding effective population
size (here $N_{E}^{I}$); and variance effective population size ($N_{E}^{V}$). Ewens (1982)
discusses the difference between them, and introduces another measure, the
eigenvalue effective population size ($N_{E}^{E}$). Other measures of population size based
on properties of the Wright-Fisher model are possible, one of which is considered in
the next chapter.
All the above measures are based on comparison with a Wright-Fisher model. The effective population size is that size of population in a Wright-Fisher model which would produce the same \((N_{E}^{I})\) rate of increase of inbreeding, or \((N_{E}^{V})\) variance in gene frequency as a Wright-Fisher model, when the assumptions on random mating are relaxed. \(N_{E}^{E}\) is defined by comparing the leading eigenvalue of the transition matrix given by (2.1) to the leading eigenvalue of the model under consideration. Crow and Kimura (1970) derive effective population size for a variety of models.

### 2.2.3 Mutation Models

Mutation is needed to provide new sources of variation in models with finite population size at steady state, otherwise the genetic variation will decay over time. We thus need models which approximate the process of mutation in nature. Several mutation models exist, which correspond to some idealised form of the processes in nature. Some of the most commonly used mutation models are:

- **Infinite alleles**: every mutation gives a new allele.

- **Infinite sites**: under this model an allele is assumed to consist of an infinite array of independent sites. Every mutation then changes a unique site on the array so that, given the array of sites for every member of a sample, the mutational history of the sample can be reconstructed. This is a closer approximation to the structure of DNA than the infinite alleles model.

- **K-alleles**: In this, each mutation switches the allele to one of the \(k-1\) other allelic types with some probability - usually equal but a matrix of transition probabilities \(p_{ij}\) between alleles can also be studied. This is useful when thinking about allozymes or microsatellites where there is some natural ordering of the possible alleles. Examples of these models are the stepwise mutation models for allozymes (Ohta & Kimura 1973) and for microsatellites (Slatkin 1995, Valdes et al. 1993).

A k-allele model corresponds to dispersal if we assume that dispersers come from a large population with mean gene frequency \(\bar{p}\). In this case migration is equivalent to mutation with rates \(A\) to \(a\) of \(m(1-\bar{p})\) and from \(a\) to \(A\) at a rate \(m\bar{p}\).
2.2.4 The Kolmogorov Forward Equation for the Distribution of Gene Frequencies

Under some models we can calculate the distribution of gene frequencies, by taking a continuous diffusion approximation for the gene frequency. Using a method based on that in Crow and Kimura (1970, p372), let \( f(x,t) \) denote the distribution of gene frequencies at time \( t \). Conditioning on the distribution at time \( t \) we have

\[
f(x,t+\partial t) = \int f(x-\varepsilon,t)g(x-\varepsilon,x,t,t+\partial t)d\varepsilon,
\]

(2.3)

where \( g(y,y+\varepsilon,s,s+\partial t) \) is the probability that the gene frequency changes from \( y \) to \( y+\varepsilon \) in the time period \( (s,s+\partial t) \). Taking a Taylor expansion of the integrand about \( x \) and ignoring terms of order \( \varepsilon^3 \) we get

\[
f(x,t+\partial t) = \int f(x,t)g(x,x-\varepsilon,t,t+\partial t)d\varepsilon - \int \varepsilon \frac{\partial f(x,t)}{\partial x}g(x,x-\varepsilon,t,t+\partial t)d\varepsilon \\
+ \int \varepsilon^2 \frac{\partial^2 f(x,t)}{\partial x^2}g(x,x-\varepsilon,t,t+\partial t)d\varepsilon.
\]

(2.4)

Dividing through by \( \partial t \) and taking the limit as \( \partial t \to 0 \) we have the equation

\[
\frac{\partial f(x,t)}{\partial t} = \frac{V(x,t)}{2} \frac{\partial^2 f(x,t)}{\partial x^2} - M(x,t) \frac{\partial f(x,t)}{\partial x},
\]

(2.5)

where

\[
M(x,t) = \lim_{\partial t \to 0} \frac{1}{\partial t} \int \varepsilon g(x,x-\varepsilon,t,t+\partial t)d\varepsilon \quad \text{and}
\]

\[
V(x,t) = \lim_{\partial t \to 0} \frac{1}{\partial t} \int \varepsilon^2 g(x,x-\varepsilon,t,t+\partial t)d\varepsilon.
\]

(2.6)

Thus \( M(x,t) \) and \( V(x,t) \) can be thought of as the mean and second moment of the change in gene frequency over a single generation. This equation can be used to derive the distribution of gene frequencies under the Wright-Fisher single population model of the previous section, and of other models. Crow and Kimura (1970, Chapter 8) give examples of its uses.
2.2.4.1 Wright's Formula for the Distribution of Gene Frequencies

The approximate distribution of gene frequencies for a pair of alleles was calculated by Wright (1931, 1943, 1951). The distribution of gene frequencies in this case is taken to be the distribution of frequencies over time (such that successive time points are independent) or over replicate populations.

The reasoning behind this is that the changes in gene frequency over time are due to two forces, the directional forces, which are mutation and selection, and random genetic drift. In Section 2.2.3 we showed how the effects of migration from a large population with gene frequency $p$ could be shown to be equivalent to mutation. The directional force gives an equilibrium expected gene frequency and the random drift tends to lessen the genetic variation. Thus we can, by using a diffusion approximation, obtain an expression for the distribution of gene frequencies under these opposing forces.

More rigour may be obtained by first deriving the Kolmogorov forward equation (2.5) for the distribution of gene frequencies over time (Crow and Kimura 1970). At steady state there is no change in the distribution over time so that $\partial f / \partial t = 0$. The equation can then be solved to get

$$f(p) = \frac{C}{\sigma_v^2} e^{-\frac{\Delta p}{\sigma_v^2}}, \quad (2.7)$$

where $\Delta p$ is the sum of directional changes in gene frequency, produced by selection, mutation, or gene flow (if you consider gene flow from a central gene pool) which is equal to $M$ of (2.6) in steady state; $\sigma_v^2$ are the random genetic processes, such as random genetic drift, with a mean change in gene frequency of 0 and a variance of $\sigma_v^2$; which is equal to $V$ of (2.6) at steady state. $C$ is a normalising constant such that $\int_0^1 f(p) dp = 1$.

This can be used to get a variety of distributions for gene frequencies if there is some systematic pressure on gene frequencies which is constant over time.
For example, if we have a mutation rate of \( \lambda \) from A to a and \( \mu \) from a to A then taking \( p \) as the frequency of A and considering a Wright-Fisher population of \( N \) diploids with random mating:

\[
\Delta p = (1 - p)\mu - p\lambda,
\]

\[
\sigma^2 = \frac{p(1-p)}{2N},
\]

\[
f(p) = \frac{Bp(1-p)}{2N} e^{\frac{4N}{\lambda}(\frac{(\mu(1-p)-\lambda p)dp}{p(1-p)})} = C p^{4N\mu^{-1}}(1-p)^{4N\lambda^{-1}},
\]

and \( C \) is a normalising constant.

### 2.2.4.2 The Infinite Island Model

The original model for the combination of gene flow and random genetic drift was derived by Wright (1931). This assumed a Wright-Fisher population on a single island with gene flow from a mainland population with gene frequency \( \bar{p} \), at a rate \( m \), i.e. a proportion \( m \) of the gametes came from the mainland population rather than being native to the island. This is equivalent to an infinite number of islands with overall mean gene frequency \( \bar{p} \) contributing a proportion \( m \) of the gametes in any one island each year. Gene flow then causes a directional change in gene frequency towards \( \bar{p} \). The gene frequency on the island can then be shown to be beta distributed using Equation 2.7. Thus we have:

\[
f(p) = \frac{p^{4Nm\bar{p} - 1}(1-p)^{4Nm\bar{q} - 1}}{\beta(4Nm\bar{p}, 4Nm\bar{q})},
\]

where \( \beta \) is the beta function (Abramowitz & Stegun 1965)

\[
\beta(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.
\]

The assumptions behind this model will be examined further in chapter 5.

From the above equations it can be seen that the migration rate and the effective population size contribute to the distribution of gene frequencies only as a product, so that it is not the absolute migration rate which matters but the number of migrants per generation \( Nm \).
2.3 Measures of Genetic Variability and Distance

In order to quantify the differences in genetic variability seen in models we need measures of genetic variation within populations, and measures of differences between populations. There have been many attempts to construct measures for between and within population measures, many of which are reviewed by Nei (1987). There are two types of between population measures, the distance measures where the interest is in the differences between populations, and diversity measures where the interest is in partitioning of variation into that within and between populations.

2.3.1 Within Population Measures

The main measures of variation within a population, when we are considering genetic data which can be described as alleles (rather than sequences) are the:

- number of alleles,
- proportion of polymorphic loci, and the
- gene diversity.

For sequence data there are measures analogous to the above three.

The first two measures are descriptive. The expected number of alleles seen varies with the sample size taken and the proportion of polymorphic loci is dependent on the loci studied (Kimura 1983). There is a natural bias in this measure to examine loci which have proved to be polymorphic in other species which will upwardly bias the proportion of polymorphic loci (as discussed in Chapter 1).

Gene diversity is equal to the expected heterozygosity under complete random mating, as described by the expectation under Hardy-Weinberg equilibrium. This definition gives an analogous measure for haploid organisms or genes:

\[ h = 1 - \sum_{\text{alleles}} p_i^2, \]  

(2.10)
where $p_i$ is the frequency of the $i$-th allele. This is the measure which is most commonly used to describe the variation seen in natural populations. To compare its value to that from mathematical or computer models we need analogous quantities for these models. The theoretical measures are the probabilities of identity, whose expectations depend on the mutation models used. We also need measures, for diploid populations, of the difference between this expected gene diversity, and the actual gene diversity, which is the observed level of homozygosity in the population.

2.3.1.1 The probability of identity in state

This is equal to the probability that two alleles selected at random from a population are identical. The relationship of this measure to the gene diversity can be seen easily by noting that the probability of identity in state is simply $\sum p_i^2 = 1 - h$, where $h$ is the gene diversity.

2.3.1.2 The probability of identity by descent

This measure is due to Malecot (1948). It was originally defined to be the probability that both alleles are copies of the same single allele in some reference ancestral population. This is identical to Wright's inbreeding coefficient (1922) defined in the next section (although that uses correlations rather than probabilities). However, there is some confusion in the definition of identity by descent in the literature, as with some authors (e.g. Slatkin 1985, 1991, 1993, Herbots 1995) it is implicit that there is also no mutation in either line-of-descent back to the ancestral allele.

In this thesis the identity by descent will be defined to be: the probability that two alleles are copies of the same allele in some reference ancestral population, and that there have been no mutations in either line-of-descent. This definition is identical to the probability of identity in state for mutation models where there is no back mutation. With small mutation rates then all identity measures are very close to each other.

When we simulate genetic variation in populations, then it is usual, and easier, to use 2 allele models with mutation between the two allelic states. In this case there is a difference between the two measures of identity.
2.3.1.3 Wright’s Inbreeding coefficient

The original definition for the inbreeding coefficient (here written $F_t$) was due to Wright (1922) who defined it to be “the correlation between homologous genes of uniting gametes under a given mating pattern, relative to the total array of these in random derivatives of the foundation stock”. The inbreeding coefficient has its foundation in animal breeding where the amount of inbreeding seen within a stock after artificial selection is an important consideration, and so has been used for prescribed mating systems. The relation between this measure and the probability of identity in state is shown by Crow and Kimura (1970). This measure is more fully explored in the next section.

For data from simulations with two alleles the probability of identity by descent in a single population can be inferred from the probability of identity in state over replicate populations (irrespective of whether the simulation is haploid or diploid) provided the mean gene frequency in the population is known. If a pair of genes in the population are identical in state then they can be identical because of back mutations or they can be identical by descent. Hence, denoting the probability of identity by descent by $f_d$, the probability of identity in state can be written as

$$f_d + \left( E(p^2) + E(1-p)^2 \right)(1-f_d).$$

Hence, by taking the expectation we get the following expression for $f_d$:

$$f_d = \frac{E(p^2) - E[p]^2}{E[p](1-E[p])} = \frac{Var(p)}{\bar{p}(1-\bar{p})},$$

where $\bar{p}$ is the expected gene frequency. This allows the identity by descent to be obtained from simulations with only 2 alleles.

2.3.2 Fixation Indices

The fixation index (Wright 1951) is a measure of deviation from Hardy-Weinberg proportions for diploid individuals. It is a measure of the difference between the observed gene diversity at a locus for a pair of alleles and the expected gene diversity at that locus under Hardy-Weinberg. We can describe the genotype frequencies as
(1 - F)x_1^2 + Fx_1; 2(1 - F)x_1x_2, and (1 - F)x_2^2 + Fx_2 when the alleles frequencies are \( x_1 \) and \( x_2 \). The fixation indices can be positive, zero, or negative. When there are \( k \) alleles at a locus then \( k(k-1)/2 \) fixation indices can be defined. This deviation from Hardy-Weinberg can be due to many factors including:

- selection,
- assortative mating and
- population subdivision.

When the deviation from Hardy-Weinberg is due to population subdivision or assortative mating, rather than selection on the differing alleles then one fixation index may be used as the deviation from Hardy-Weinberg is expected to be the same for all loci, and all alleles.

The fixation index is a statistic. It can be calculated from a population, or estimated from a sample from a population, but it can also be used as a parameter in population genetic models (see Chapter 5 for an example).

### 2.3.3 Genetic Variation Between Populations

The above measures concern a single population. When there are more than one population further questions of interest arise:

- How much of the total genetic variation is accounted for by the subdivision of the population into subpopulations?
- What is the ‘distance’ between the populations (where by distance we mean a measure of the difference between populations)?

The starting point for the investigation of the partitioning of genetic variation between populations is the fixation index described in Section 2.3.2.

#### 2.3.3.1 F-Statistics

When there is population subdivision a deficit of heterozygotes in the subpopulations relative to the gene-diversity of the population as a whole is observed, known as the Wahlund effect (Crow and Kimura 1970). Fixation indices can be used to describe
Chapter 2: CLASSICAL MODELS OF GENE FLOW

this deficit but there are now two levels of the population; the total population, and the individual populations. Wright's F statistics were again originally defined for the purposes of animal breeding but were later refined by Wright (1943) for describing the properties of hierarchically subdivided natural populations. These were defined in terms of inbreeding coefficients, which were defined as the correlation between uniting gametes relative to hierarchies in the population.

The original definition for this is in terms of the total population T, the subdivisions S, and the individuals I. Then $F_{IT}$ is the correlation between gametes which unite within individuals relative to the gametes in the total population; $F_{ST}$ is the correlation between gametes within populations relative to the gametes within the whole population; and $F_{IS}$ is the correlation between uniting gametes within a subdivision relative to the gametes of that subdivision, averaged over all subdivisions.

This is a clumsy definition, but has the benefit of being clear that the F measures are statistics, and are properties measured from populations. The statistics can thus have a theoretical expectation under models for the differentiation of the populations under consideration. A simpler definition was presented by Nei (1973) who used the expected heterozygosity within and between populations, this measure is commonly known as $G_{ST}$. These hierarchical statistics can be extended to further levels in the population in a natural way. Weir and Cockerham (1984) provided a further definition for $F_{ST}$ as a function of variance components from a nested analysis of variance.

If there is no heterozygote deficit within subpopulations then the expected value of $F_{IS}$ will be zero, and we only need to consider $F_{ST}$. By using a similar argument to that in Section 2.3.1.3 it is clear that we have

$$E[F_{ST}] = \frac{Var(p)}{p(1-p)},$$

where the variance is measured over subpopulations. Here $E[ ]$ is written to denote an expectation over the evolutionary process.
The infinite island model considered in Section 2.3.1 had a Beta distribution of gene frequencies. The variance in alleles frequencies over replicate populations is thus
\[ \frac{\bar{pq}}{4Nm + 1} \] and hence
\[ E[F_{ST}] = \frac{1}{1 + 4Nm}. \] (2.14)

This is a familiar result for the probability of identity in an island model. In the next chapter this will be derived in a different way, which is more informative about the processes generating the values of $F_{ST}$ seen. The difficulty underlying these values is that while the F statistics are defined in terms of whole populations (and are themselves stochastic realisations of the evolutionary process) we attempt to estimate them from samples from populations. This will be discussed in Chapter 4.

### 2.3.3.2 Identity Coefficients Between Populations

The identity coefficients introduced in Section 2.3.2 can be used equally well for between population measures. If we assume that all populations have the same breeding patterns and structure then we can define $f_0$ and $f_1$ to be the probabilities of identity for pairs of genes sampled at random within and between populations. A further definition for $F_{ST}$ was provided by Nei (1973), and used explicitly by Crow and Aoki (1984):
\[ E[F_{ST}] = \frac{f_0 - \hat{f}}{1 - \hat{f}}, \] (2.15)

where the expectation is over the evolutionary process. This provides us with a method for calculating $F_{ST}$ which is more amenable to mathematical modelling.

### 2.3.4 Measures of Distance between Populations

The above F-statistics allow the partitioning of genetic variation between different levels of the population structure. One simple measure for 'distance' between populations could thus be $F_{ST}$. This statistic does not include any information about the geographic positions of the different populations.
Using the identity coefficients of the previous section when there is no geographical information in the samples, such as in an island model, then there are two measures of genetic variation, $f_0$, the probability of identity within a population and $f_1$, the probability of identity for two genes sampled from different populations. Thus it is possible to devise several measures, using the condition that when $f_0 = f_1$, the distance is zero and the distance is positive when $f_1 > f_0$. The case when $f_0 < f_1$ is possible but unlikely. Most measures of distance reviewed in Nei (1987) can be written as functions of these probabilities of identity within and between populations.

These probabilities can be estimated from surveys of natural populations. Consider a locus with $k$ alleles and for which $r$ subpopulations are surveyed completely (to ignore sampling problems) then if the frequency of allele $j$ in population $i$ is denoted by $x_{ij}$ then $f_0 = \frac{1}{r} \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij}^2 / r$ and $f_1 = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{k} \sum_{l=1}^{k} x_{ij} x_{il}$. In this thesis measures which have some evolutionary interpretation, rather than measures such as Roger's Distance, or Pearson's coefficient of racial likeness will be considered.

### 2.3.5 Pairwise Measures

#### 2.3.5.1 Pairwise $F_{ST}$

When there are additional measurements on the observed populations then it may be informative to get distances between pairs of populations. One measure is derived from Wright's $F_{ST}$ for groups of subpopulations. This is a pairwise measure of $F_{ST}$ defined to be

$$F_{ST} = \frac{f_0 - f_1}{2 - f_0 - f_1}. \quad (2.16)$$

Slatkin (1993) used this measure to compare geographic distance to genetic distance.

#### 2.3.5.2 Nei's Distance, $D$

Nei's $D$ (Nei 1987) can be defined in terms of probabilities of identity as

$$D = -\log \left( \frac{f_1}{f_0} \right). \quad (2.17)$$
The mean distance can be calculated for a set of subpopulations which gives us another estimate for the partitioning of genetic variation within and between populations.

2.4 The Influence of Selection

Most studies of genetic variation in natural populations cite selection as a possible reason for any patterns in the genetic variation seen. The interaction between selection and gene flow has been reviewed by Felsenstein (1976). The easiest type of selection to include in an analysis like this is balancing selection with an equal degree of over dominance in each population (if replicate populations are considered) or over time.

Using the formula for the distribution of gene frequencies within a single population (Equation 2.7) the effect of selection on the distribution can be approximated. Consider again a 1 locus 2 allele system but now with differential survival to the next generation of the three genotypes. Assume that the three genotypes AA, Aa and aa have fitnesses 1-r, 1, and 1-s respectively and that a proportion m of the alleles are from long range migration from a population with mean gene frequency \( \bar{p} \); and that the population has an effective population size of N diploids. Also assume weak selection so that Hardy-Weinberg equilibrium is maintained.

Now the directional change in gene frequency \( \Delta p \) is the sum of the effects due to selection and gene flow. The effect of selection is then:

\[
\Delta p_s = \frac{(1-r)p^2 + p(1-p)}{(1-r)p^2 + 2p(1-p) + (1-s)(1-p)^2} - p = \frac{pq((1-r)p + p(1-p))}{bp^2 + 2hp(1-p) + c(1-p)^2}.
\]

(2.18)

Now the distribution of gene frequencies \( f(p) \) can be calculated using (2.7) to give

\[
f(p) = C(\bar{\omega})^{2N} p^{4Nm\bar{\omega}^{-1}} (1-p)^{4Nm(1-\bar{\omega})^{-1}},
\]

(2.19)

where C is a normalising constant, and \( \bar{\omega} \) is the mean fitness of the population (which is equal to \( (1-r)p^2 + 2p(1-p) + (1-s)(1-p)^2 \)). This is a very general equation (assuming weak selection). To look at the properties in a more simple
situation, symmetric overdominance, let $r=s$, and let $r$ vary. The influence of selection on the distribution of gene frequencies can be seen in Figure 2.1.

The effect of underdominance is to widen the distribution function, increasing the variance. Overdominance steepens the distribution function, decreasing the variance. The effects of this in the values of $F_{ST}$ can be seen in Figure 2.2. When the migration rate is high small selection coefficients make little difference to $F_{ST}$, but for lower migration rates the effect of selection can be large. This sort of equal balancing selection with equal pressure in each population is almost certainly extremely rare in natural populations, but it gives an idea of the sorts of effects seen when there is a degree of overdominance.
When the degree of overdominance changes between different locations, this process becomes more difficult to model. If we assume that selection coefficients for the genotypes AA:Aa:aa, are $1-r:1:1-s$, then the expected gene frequency of A is $r/(r+s)$. If we ignore genetic drift we can then get an expression for $F_{ST}$. For example if $r$ is uniformly distributed between 0 and 0.1, and $s = 1-r$ then the expected value of $F_{ST}$ is $1/3$ (the variance of the distribution of gene frequencies is $1/12$ and the expected mean gene frequency is $0.5$).

### 2.5 The Finite Island Model

All of the previous models are for single islands with gene flow from a mainland population (or alternatively for an infinite number of islands with a fixed mean gene frequency). When there are only a finite number of islands then there is drift in the mean gene frequency among the islands as well as within each individual island. This original model was extended to incorporate a finite number of islands and the general solutions were provided by Latter (1973).
Simple linear mathematics can be used to solve the finite island model. The only coefficients of interest are the probability of identity within a single population \( f_o \), and the probability of identity between populations \( f_i \). Consider a population made up of \( k \) islands each with an effective population size of \( N \) diploids and a proportion \( m \) of the gametes in any subpopulation are from parents from outside that subpopulation. The gametes mutate at a rate \( \mu \) per generation, and assuming that in each generation the events take place in the following order:

Adults → Random Mating within Demes → Production of Infinite number of Gametes → Mutation → Dispersal → Population Regulation → Adults.

The probabilities of identity by descent can be calculated by conditioning on the events in the previous generation. In this case the probability of identity within a deme is the sum of the probability that the 2 genes are both descendants of the same gene in the previous generation \( \frac{i}{2N} \) and the probability that the 2 genes in the previous generation were identical by descent, and there was no mutation. Similar reasoning is used for genes in different demes to get the following simultaneous equations:

\[
\begin{align*}
\frac{n_{i_0}}{2N} + \frac{2N-1}{2N} (1-\mu)^2 \left[ (1-m)^2 f_o + 2m(1-m) f_i + m^2 \left( \frac{f_o}{k-1} + \frac{(k-2)f_i}{k-1} \right) \right], \\
\frac{n_{i_1}}{2N} + \frac{2N-1}{2N} (1-\mu)^2 \left[ \left(1-\frac{m}{k-1}\right)^2 f_i + 2m(1-m) \left( \frac{f_o}{k-1} + \frac{(k-2)f_i}{k-1} \right) + m^2 \left( \frac{f_o}{k-2} + \frac{(k-3)f_i}{k-2} \right) \right]
\end{align*}
\]

Solving these equations and using Equation 2.15 gives us a messy expression for the expected value of \( F_{ST} \). Approximating by ignoring second order terms in \( m \) and \( \mu \) we get the result of Nei (1973),

\[
E[F_{ST}] = \frac{1}{4Nm_0^2 + \frac{4N\mu k}{1+\frac{M}{k-1}}}
\]

This is similar to the result for the infinite island model but now the number of populations and the mutation rate have an effect. Slatkin (1989, 1991) defined the effective number of migrants per population to be \( M \) such that \( E[F_{ST}] = \frac{1}{1+M} \).
2.5.1 Finite Island Model with Selection

Results for selection in subdivided populations have been reviewed by Felsenstein (1976). The interaction between selection, random drift and migration is difficult to model, because of the various possible combinations of selection coefficients between populations which are possible. One way of modelling heterozygote advantage which varies between populations is to use the model of Section 2.4 extended to deal with a finite number of populations. Table 2.1 gives the degree of population subdivision, as measured by $F_{ST}$, for 100 populations which have selection coefficients $1-r : 1: 1-s$ for the three genotypes, where $r$ is uniformly distributed between 0, and $h$, and $s = h-r$. The migration and mutation rates are denoted by $m$ and $\mu$ respectively.

These values are based on simulations. These are forward in time with discrete generations. The number of A alleles in a deme is calculated from the frequency of A in that deme ($p_D$) and in the rest of the population ($p_o$) in the previous generation by drawing from a binomial distribution $Bn(2N, p + \Delta p_s)$ where

$$p = (1-m)(1-\mu)p_D + m(1-\mu)p_o + (1-m)\mu(1-p_D) + (1-m)(1-\mu)(1-p_o),$$

$N$ is the diploid population size, and $\Delta p_s$ is given by Equation 2.18. Estimated $F_{ST}$ values are calculated by taking replicate samples every 500 generations after 2000 generations (to allow the population to reach steady state) to get estimates of the variance in gene frequencies between demes and then using Equation 2.13.

When the migration rate is 0 the values of $F_{ST}$ are reduced by increasing $h$, because in this case less populations are fixed for the alternate alleles. In the other cases increasing $h$ reduces the amount of subdivision seen. This indicates that the effects of spatially varying selection are extremely difficult to model, and the effects on expected values of $F_{ST}$ are not quantifiable without detailed knowledge.
Chapter 2: CLASSICAL MODELS OF GENE FLOW

<table>
<thead>
<tr>
<th>Migration rate</th>
<th>Expected $F_{ST}$ $h=0$ (m=0 N large)</th>
<th>$h$</th>
<th>Estimated $F_{ST}$</th>
<th>Standard error $F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>0.01</td>
<td>0.914</td>
<td>2.1x10^{-3}</td>
</tr>
<tr>
<td></td>
<td>0.333</td>
<td>0.05</td>
<td>0.845</td>
<td>1.9x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.716</td>
<td>2.1x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.20</td>
<td>0.569</td>
<td>1.4x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.40</td>
<td>0.493</td>
<td>1.0x10^{-3}</td>
</tr>
<tr>
<td>0.001</td>
<td>0.690</td>
<td>0.0001</td>
<td>0.687</td>
<td>2.5x10^{-3}</td>
</tr>
<tr>
<td></td>
<td>0.333</td>
<td>0.0005</td>
<td>0.675</td>
<td>2.7x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.001</td>
<td>0.668</td>
<td>2.6x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.005</td>
<td>0.649</td>
<td>2.9x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.01</td>
<td>0.638</td>
<td>2.8x10^{-3}</td>
</tr>
<tr>
<td>0.01</td>
<td>0.195</td>
<td>0.01</td>
<td>0.189</td>
<td>1.5Ex10^{-3}</td>
</tr>
<tr>
<td></td>
<td>0.333</td>
<td>0.05</td>
<td>0.197</td>
<td>1.5x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.227</td>
<td>1.3x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.20</td>
<td>0.259</td>
<td>1.1x10^{-3}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.40</td>
<td>0.290</td>
<td>1.0x10^{-3}</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0240</td>
<td>0.01</td>
<td>0.026</td>
<td>2.5x10^{-4}</td>
</tr>
<tr>
<td></td>
<td>0.3333</td>
<td>0.05</td>
<td>0.027</td>
<td>2.6x10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.10</td>
<td>0.034</td>
<td>2.9x10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.20</td>
<td>0.056</td>
<td>3.7x10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.40</td>
<td>0.100</td>
<td>4.4x10^{-4}</td>
</tr>
</tbody>
</table>

Table 2.1 Values of $F_{ST}$ seen for a finite island model with varying overdominance between populations. Population size = 100 diploids per deme with 100 demes and mutation rate = 10^{-4}.

2.6 Isolation By Distance

The island model has no geographic content. Any population is equally likely to exchange genetic material with any other. For natural populations this is an unlikely scenario. Populations are more likely to exchange genetic material with neighbours than with widely separated populations.

The next level of complexity in the structure of a subdivided population is to assume that the population is divided up into an infinite array of local subpopulations which
only exchange genetic material with their neighbours rather than with the whole population. The term 'stepping stone' was coined by Kimura as an analogy with stepping stones over a river. One can only move to a stone one step forward or backwards in the chain.

The main features of the infinite stepping stone models were described in the 1960's (Kimura and Weiss 1964; Weiss and Kimura 1965). Malecot (1948) gave solutions for continuum models of gene flow; these will be discussed in a later chapter.

The essential feature of stepping stone models of population genetic variation are discrete randomly mating subpopulations (demes) which exchange a proportion of their genetic material with neighbouring demes in each generation.

The details of their model will not be considered here as a variant of their 1-dimensional model is constructed in a later section. However, the main results can be summarised as:

\[
\begin{align*}
1-d & \quad F(d) \propto \exp \left( -d \sqrt{\frac{2\mu}{m}} \right), \\
2-d & \quad F(d) \propto K_0 \left( -d \sqrt{\frac{8\mu}{m}} \right) \approx \begin{cases} \\
\frac{\exp \left( -d \sqrt{\frac{8\mu}{m}} \right)}{\sqrt{d}} & \text{for } d \text{ large, and} \\
-\log(4d) & \text{over small scales} \end{cases} \\
3-d & \quad F(d) \propto \frac{\exp \left( -d \sqrt{\frac{6\mu}{m}} \right)}{d},
\end{align*}
\]

where \(d\) is absolute distance (in deme spacing) between demes, \(\mu\) is the mutation rate, \(m\) is the total amount of genetic material exchanged in each generation (over all directions), and \(K_0\) is the Bessel function of the third kind of order 0 (Abramowitz & Stegun 1965). The mutation rate can either be interpreted as mutations or as long range migration from a population depending on how we interpret \(F\).

For populations of Atlantic Salmon in rivers one dimensional habitats are most likely to be of interest. In this case we expect an exponential decline in identity over space for salmon sampled over a linear stretch of river. These equations indicate that the
correlation decays much faster over three dimensions than over two dimensions, and much faster over two dimensions than over one dimension. The case of three dimensions would be very rare anyway, restricted, perhaps, to some species of free floating marine invertebrates.

Many other kinds of stepping stone models have been studied. Endler (1977) simulated a stepping stone model in 2 dimensions where the demes were arranged in a hexagonal lattice. Sawyer (1976) provides a more technical set of results with generalisations of the more simple models. Maruyama (1969, 1970, 1972) gives solutions for finite stepping stone models. Carmelli and Cavalli-Sforza (1976) deal with a number of more specialised models of gene flow.

2.6.1 Clines and Isolation by Distance

The amount of gene flow is of interest if we want to investigate the amount of local adaptation. Nagylaki (1975) showed that conditions for the existence of an allele which is advantageous in a region of length \( b \) (where the allele is selected against outside that region) in a finite region is given by the dimensionless parameter \( b \sqrt{2s/\sigma^2} \), where \( \sigma^2 \) is the standard deviation in dispersal distance. Clines do not form when this value is less than \( \pi/4 \). This is not including the effects of random genetic drift. When random drift is added to models of clines the general shape is unaltered (Slatkin & Maruyama 1975, Felsenstein 1975b), but there is noise around the ‘average’ cline.
2.7 The Effect of Branching on Measures of Population Differentiation

2.7.1 Introduction

The natural world only rarely provides the completely homogeneous environment required by the usual stepping stone and island models of population structure. Geographic heterogeneity is a common feature of the habitats of a wide variety of creatures, due to physical barriers and obstacles such as rivers, seas, and mountains. These barriers have large effects on the movement of animals. The majority of models of population subdivision, however, assume that there are no barriers or aids to dispersal, they all exchange dispersers with each other at the same rate, or at least with the same probability distribution. The basis of most models of genetic variability, in terms of geography are either a series of subpopulations (demes) each of which exchanges migrants with each other at the same rate (Section 2.2.4.2) or a series of demes arranged in a lattice structures which exchange migrants with their 2 neighbours in a 1 dimensional structure (Kimura and Weiss 1964); their 4 neighbours in a square lattice (Kimura and Weiss 1964) or 6 (Endler 1977) neighbours in a hexagonal pattern of demes in 2 dimensions (Section 2.6). These models give us predictions of the patterns of genetic variation seen in continuous habitats in 1 or 2 dimensions but not for more complicated conditions. Nagylaki et al. (Nagylaki 1988, Nagylaki and Barchilon 1988) have investigated the results of a barrier to gene flow in 1 dimensional habitats, which showed that a barrier produced a sharp step in the covariance of gene frequencies.

Here, one very specialised type of impediment to dispersal, namely gene flow on a branching structure, will be investigated. Many animals and plants live in habitats which are to some extent branching: riverine fish and animals; and plants which are restricted in range to the proximity of rivers. Alpine plants and animals live on valley floors which may restrict their direction of movement to these valley floors.

Figure 2.3 shows the difference in branching structures seen in Scottish rivers: the Spey is essentially a long single river with smaller tributaries whereas the Kyles are more branched with no main river. There is evidence in many fish species for population differentiation within a single river system (see e.g. Carvalho 1993), but little evidence that this can be accounted for by physical distance within the river.
system (Verspoor et al. 1991). The branching structure of the system is thus an important area for investigating the causes and patterns of this variation.

In Chapter 1 some of the evidence for local adaptation in Atlantic salmon was reviewed. The detailed information required for modelling such local adaptation is not available, and therefore neutral genetic variation is assumed. It is possible to argue how selection, and local adaptation would affect this neutral variation by analogy to the processes in 1 and 2- dimensions (Slatkin 1977, Nagylaki 1975).

![Figure 2.3 A schematic diagram of the River Spey and the Kyles of Sutherland.](image)

### 2.7.2 Model

Consider a neutral genetic model in which discrete randomly mating demes with non-overlapping generations of effective population size N (so that there are 2N genes) are at the nodes of an infinite tree structure of order k (i.e. every deme is connected to k others and there is a unique path between any two demes) with dispersal occurring between neighbouring demes only. This is equivalent to there being randomly mating populations between the nodes. For k=2 this reduces to a linear
stepping stone model. Figure 2.4 illustrates a small part of such an infinite tree for k = 3.

Figure 2.4. Part of an infinite tree of order 3.

Assume an infinite alleles model with mutation, random drift and migration occurring in the order

Adults → Random Mating within Demes → Production of Infinite number of Gametes → Mutation → Dispersal → Population Regulation → Adults.

Define $F_d$ to be the probability of identity by descent of a pair of alleles sampled from demes a distance d apart ($F_0$ is the probability of identity of 2 genes sampled with replacement from within a population) and assume complete random mating so that this is the same as the probability of homozygosity (we are treating the diploid
case as identical to the haploid). This is sufficient specification as the path between any two demes is unique.

The model assumes random union of gametes within each population to produce an infinite number of gametes of which a proportion $m$ disperse to adjacent demes. A proportion $\mu$ of the genes in the gametes mutate to produce completely novel alleles and then population regulation reduces the population size to $N$ individuals ($2N$ genes) by selecting at random from the infinite pool of gametes present.

Then for $d \geq 2$, $k \geq 2$, using similar reasoning to that of Section 2.5, we obtain

$$F_d = (1-\mu)^2 \left( \frac{m}{k} \right)^2 \left[ (k-1)^2 F_{d+2} + 2(k-1)F_d + F_{d-2} \right]$$

$$+ (1-\mu)^2 \left[ \left( \frac{2m}{k} \right)(1-m) \left( (k-1)F_{d+1} + F_{d-1} \right) + (1-m)^2 F_d \right],$$

(2.23)

where $F_d$ is the probability of identity in the previous generation. For brevity let $\alpha = (1-\mu)(1-m)$ and $\beta = (1-\mu)m/k$ (c.f. Kimura and Weiss 1964).

Now if we assume that at equilibrium $F_d = F_e$ then this is a difference equation of order 4 and so, provided all the roots are different, any solution is of the form

$$F_d = a_1 \lambda_1^d + a_2 \lambda_2^d + a_3 \lambda_3^d + a_4 \lambda_4^d,$$

(2.24)

where the $\lambda_i$'s are determined by replacing $F_d$ by $x^d$ and solving Equation 2.23. The roots of this equation are:

$$\lambda_1 = \frac{1-\alpha + \sqrt{(1-\alpha)^2 - 4(k-1)\beta^2}}{2(k-1)\beta}, \quad \lambda_2 = \frac{1-\alpha - \sqrt{(1-\alpha)^2 - 4(k-1)\beta^2}}{2(k-1)\beta},$$

$$\lambda_3 = \frac{-(1+\alpha) + \sqrt{(1-\alpha)^2 - 4(k-1)\beta^2}}{2(k-1)\beta}, \quad \lambda_4 = \frac{-(1+\alpha) - \sqrt{(1-\alpha)^2 - 4(k-1)\beta^2}}{2(k-1)\beta},$$

(2.25)

Since $F_d \leq 1$ for all $d$, then because $|\lambda_1| \geq 1$ and $|\lambda_4| \geq 1$ we get $a_1 = a_4 = 0$.

In order to calculate $a_2$ and $a_3$ we need boundary conditions. We have
\[ F_1 = \beta^2 \left( (k-1)^2 F_3 + (2k-1)F_1 \right) + 2\alpha\beta \left( (k-1)F_2 + F_0 \right) + \alpha^2 F_1, \]  
(2.26)

and after random drift \( F_0 = \frac{1}{2N} + \frac{(2N-1)F_0}{2N}, \) so that

\[ F_0 = \frac{1}{2N} + \frac{2N-1}{2N} \left[ \beta^2 \left( k(k-1)F_2 + kF_0 \right) + 2k\alpha\beta F_1 + \alpha^2 F_0 \right]. \]  
(2.27)

After some manipulation we get

\[
\begin{align*}
\alpha_2 &= \frac{(k-1)U}{(k-1)U + (k-1)V + (2N-1)UV}, \text{ and } \\
\alpha_3 &= \frac{(k-1)U}{(k-1)U + (k-1)V + (2N-1)UV} \\
\text{where} \\
U &= (1+\alpha)(k-2) + k\sqrt{(1+\alpha)^2 - 4(k-1)\beta^2}, \text{ and} \\
V &= (1-\alpha)(k-2) + k\sqrt{(1-\alpha)^2 - 4(k-1)\beta^2}.
\end{align*}
\]  
(2.28)

This gives a complete solution. Figures 2.5 and 2.6 indicate the difference between trees of order 2 (i.e. a linear habitat) and trees of higher orders for a variety of test statistics. A branching structure, rather than the more conventional one dimensional model (here represented by a tree of order 2) decreases the expected values of identity by descent considerably. The expected values seen for pairwise differences are both decreased (for pairwise FST statistics) and increased (for pairwise Nei’s D statistics). This represents the different philosophies for the two statistics, which are discussed more thoroughly in the next chapter. Global FST values are not possible for these infinite branching structures, but values based on demes at a distance are possible, and seen later in this section.
Figure 2.5 The identity by descent plotted against distance between samples. $m = 0.01$, $\mu = 10^{-6}$ and population size =100. Circles indicate trees of order 2, triangles and squares indicate trees of orders 3, and 4 respectively. All values from Equation 2.24, using values from Equations 2.25 and 2.28.
Figure 2.6 a-b. Pairwise $F_{ST}$ (a) and Nei's D (b) plotted against distance between samples for $M_{ig} = 0.01$, mutation rate $= 10^{-4}$ and population size $= 10^4$. Circles indicate trees of order 2, triangles and squares give trees of orders 3 and 4 respectively.
2.7.3 Comparison of linear habitats and trees of order k

The leading root in the expression for $F_d$ is $\lambda_2$ so the qualitative behaviour is dominated by this term. Now

$$
\lambda_2 = \frac{(1 - \mu)m + \mu - \sqrt{\mu^2 + 2m\mu(1 - \mu) + m^2(1 - \mu)^2(1 - 4(k - 1)/k^2)}}{2(k - 1)(1 - \mu)m / k}
$$

$$
= \frac{k}{2(k - 1)} \left[ 1 - \frac{\mu}{(1 - \mu)m} - \frac{(k - 2)^2}{k^2} + \frac{2\mu}{m(1 - \mu)} + \frac{\mu^2}{m^2(1 - \mu)^2} \right]. 
$$

(2.29)

As $\mu \to 0$: $\lambda_2 \to \frac{1}{k - 1}$ and the probability of identity is influenced most by the branching structure of the population in the limit. This explains the major differences between the usual 1-dimensional model and these models on branching structures.

2.7.4 Results for Small Trees

The analytic results presented above are for trees with infinite numbers of branching points. While this is a reasonable approximation for large finite tree structures this remains to be tested. Edge effects may be important for smaller trees. For trees with very small numbers of nodes it is, however possible to get numerical results by solving the simultaneous equations.

Consider a tree with the structure seen in Figure 2.7, i.e. with 4 nodes, which is the simplest possible tree structure. There are 4 different identity probabilities to keep track of here: the probability of identity of two genes sampled from a 'leaf' node $F_1$; the identity for 2 genes from the centre node $F_c$; the probability of genes selected from different 'leaves' $F_{11}$; and the probability of identity of genes selected from a 'leaf' and the centre $F_c$. 
More assumptions are needed for a tree of this type, namely what happens at a 'leaf'. Here it is assumed that a proportion $m/k$ of the gametes in a 'leaf' deme are from the central deme in any generation, and that the migration rate from a 'leaf' deme to a central deme is the same. The equations to be solved are then:

\[
F_t = \frac{1}{2N} + \frac{2N - 1}{2N} (1 - \mu)^2 \left[ \left(1 - \frac{m}{k}\right)^2 F_t + \left(\frac{m}{k}\right)^2 F_c + \frac{2m}{k} \left(1 - \frac{m}{k}\right) F_{tc} \right],
\]

\[
F_c = \frac{1}{2N} + \frac{2N - 1}{2N} (1 - \mu)^2 \left[ (1 - m)^2 F_c + \frac{m^2}{k} ((k-1)F_{tt} + F_t) + 2m(1-m)F_{tc} \right],
\]

\[
F_{tc} = (1 - \mu)^2 \left[ (1 - \frac{m}{k}) (1 - m) F_{tc} + (1 - m) \left(\frac{m}{k}\right)^2 F_c + \left(1 - \frac{m}{k}\right) mF_t + \frac{m^2}{k} F_{tc} \right], \text{ and}
\]

\[
F_{tt} = (1 - \mu)^2 \left[ (1 - \frac{m}{k})^2 F_{tt} + \frac{2m}{k} \left(1 - \frac{m}{k}\right) F_{tc} + \left(\frac{m}{k}\right)^2 F_c \right]. \tag{2.30}
\]

The four simultaneous equations can be easily solved using a mathematics package such as Mathematica (Wolfram 1991). Values of $F_{ST}$ for this finite model can be calculated by using Equation 2.15. In this case,
\[ f_0 = \frac{F_c + kF_i}{k+1} \quad \text{and} \quad f = \frac{F_c + kF_i + k(2F_{le} + (k-1)F_{li})}{(k+1)^2}. \] (2.31)

Table 2.2 compares the amount of differentiation in this model compared to an island model (from Equation 2.21).

<table>
<thead>
<tr>
<th>migration rate (m)</th>
<th>( F_{ST} ) for tree of height 2</th>
<th>( F_{ST} ) for equivalent finite island model (k+1 demes; migration rate m/k)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k=2</td>
<td>k=5</td>
</tr>
<tr>
<td>0.1</td>
<td>0.030</td>
<td>0.083</td>
</tr>
<tr>
<td>0.01</td>
<td>0.172</td>
<td>0.417</td>
</tr>
<tr>
<td>0.001</td>
<td>0.665</td>
<td>0.870</td>
</tr>
</tbody>
</table>

Table 2.2. A comparison of \( F_{ST} \) for a finite island model and a tree of height 2. In all comparisons \( N=100 \) diploids, and the mutation rate = \( 10^{-5} \).

This model shows similar levels of differentiation as island models with similar migration rates out of the 'edge' demes, and the same number of populations. The tree of height 2 shows greater levels of differentiation for high levels of gene flow and less differentiation for low levels of gene flow. This indicates that for extremely small trees then the island model of population differentiation is a suitable model for the patterns of genetic variation seen. With a tree of height 4 there are 10 identity by descent values to calculate, producing a system of 10 simultaneous equations to solve similar to those of Equations 2.30. Increasing the height of the tree again would require the solution of 20 simultaneous equations. Instead of attempting these it is better to use simulation techniques.

2.7.5 Simulation

The analytic results presented above are for trees with infinite numbers of nodes, and very small numbers of nodes. Simulation studies were performed to test the theory
and to see how large a tree is needed to reproduce the results seen for infinite tree structures, which are not biologically realistic. Exact results are known for trees of order 2 (Kimura and Weiss 1964, Weiss and Kimura 1965, Maruyama 1969, 1970, 1972) so the results for trees of this order can be compared with the analytic theory. Simulation studies were done on trees of order 2 and 3, 4, 5 and 6.

The simulation studies were performed on trees shaped as in Figure 2.4 but varying the height (defined here to be the maximum distance from an internal node to a 'leaf'). A tree of height \( i \) of order \( k \) has a total of \( 1 + k((k - 1)^{i-1} + 1) \) demes. A serendipitous feature of computer programming is that using tree structures is a well established technique (Carrano et al. 1993 pp.480) which allows easy recursion. The original program (in Pascal) was based on an algorithm by Stuart J. E. Baird. Trees of this shape were chosen to facilitate the understanding of edge effects, as the distance from the edge can then be given by a single number.

To simplify the simulation process a 2 allele model was used with reversible mutation between the 2 alleles at the same rate \( \mu \) and discrete generations. In each generation, for each deme, the frequency of alleles of type A in the deme (denoted \( p_a \)), and in its \( k \) neighbours (\( p_k \)) are calculated. The number of alleles of type A in the next generation is calculated using a Binomial with parameters \( 2N \) and

\[
p = (1 - m)(1 - \mu)p_a + m(1 - \mu)p_k + (1 - m)\mu(1 - p_a) + (1 - m)(1 - \mu)(1 - p_k).
\]

Simulations were written in C using Borland C++ Version 4. Algorithms are given in Appendix 1.

In order to increase the speed of simulation the transition probabilities for each deme given the numbers of allele A in the deme and in its neighbours were stored in a matrix. Thus the transition required only a single random number. This, however, restricted the deme size (because of memory constraints) to 15 diploid individuals (or 30 haploids). Algorithms for the transition matrix approach are given in Appendix 2.

The reversible mutation means that the mutation rate is doubled in these calculations compared with the theory (because the theory is for an infinite alleles model where every mutation is to a new allele). Equations 2.23-2.28 are as before but now \( F_d \) is not the identity by descent but is defined as \( r_d = E(\hat{p}_i\hat{p}_j)/(\bar{p}(1 - \bar{p})) \) where \( \hat{p} = p - \bar{p} \) and \( \bar{p} \) is the expected gene frequency over time and \( i \) and \( j \) are demes a
distance $d$ apart. This relationship between infinite allele and finite allele models was discussed in Section 2.3.1. This is the measure used by Kimura and Weiss (1964), which they called the correlation in gene frequencies, but it is not the true correlation (measuring the gene frequency from its expected value over all time and space rather than its expectation in the two populations given the frequencies in the previous generation). Replacing $F_d$ by $r_d$ in Equations 2.23-2.28 does not affect the equations so we can perform simulations and then estimate $F_d$ by the formula

$$r_d = \frac{1}{n_{obs}} \sum_{i,j} \sum_{k} \frac{(x_j - \bar{p})(x_k - \bar{p})}{\bar{p}(1 - \bar{p})},$$

(2.32)

where $j$ and $k$ are demes separated by a distance $d$ and with equal mutation rates in both directions so that $\bar{p} = 0.5$. The observations were at time intervals such that successive observations are uncorrelated. For all sets of parameters, runs tests (Rice 1988) were performed to check the independence of successive observations.

Figures 2.8 and 2.9 show the estimated identities by descent from trees of order 2 and 3 compared to the theoretical results. These identities were calculated for distances away from the central node. The results shown in Figure 2.9 indicate that the theory fits the data well when the trees have large numbers of nodes, a low migration rate, and a comparatively large mutation rate. Figure 2.8 illustrates the identity by descent within the central deme for trees of different heights, compared to the theory for infinite trees. The theory fits well for small migration rates even for small trees, and for large mutation rates (which were necessary to prevent fixation of alleles). For larger migration rates the fit is not good for small trees. These large mutation rates are equivalent to long distance migration at a low rate which may be seen in natural populations.
Figure 2.8 The identity by descent within the central deme for small trees of order 3 compared to the theoretical identities (lines). Migration rate =0.01 (squares, theory given by solid line) and m=0.1 (triangles theory given by dotted line). N=12, and μ=0.001; results based on 1000 replicates: Error bars represent ±2 s.e.
Figure 2.9a-b Theoretical i.b.d. compared to simulation results for 1000 replicates for trees of order 2 (a) and 3 (b). Squares m=0.2; circles m=0.1; triangles m=0.05; stars m=0.01. Simulations based on trees of height 11 with N=12 and μ=0.001. Error bars represent ±2 s.e.
2.7.6 $F_{ST}$ within a tree system

The exact global values of $F_{ST}$ seen within a finite tree system can in principle be calculated exactly, as in Section 2.7.4, but this is only feasible for small trees. Values for $F_{ST}$ can be estimated from simulated trees by using Equation 2.15 so that a value may be calculated by estimating the variance in gene frequencies among all the populations in a single generation using the usual formula for the estimation of variance. For simulated data we can either estimate $\bar{p}$ from the data and use this to calculate the variance or use $E(p)$ which we know for equal mutation rates from A to a as from a to A to be 0.5. The formula for estimating the variance is then correspondingly changed. The values calculated for these large trees were very similar for the two methods. Figure 2.10 illustrates the estimated values for $F_{ST}$ using both methods of estimation for a range of sizes of trees of order 3.

Figure 2.10 $F_{ST}$ from simulations on trees using calculated mean gene frequencies (open symbols), and expected gene frequencies (solid symbols). Triangles for migration rate of 0.01, and squares for migration rate of 0.1. Simulations based on tree of order 3 with $\mu=0.001$. The values are based on 1000 replicates.
An estimate for the expected value of $F_{ST}$ can also be calculated by using the expression in terms of identity coefficients (Wright 1951; Slatkin 1985b, 1991, Equation 2.15). $f_0$ can be calculated from the expectation from an infinite tree, and $\bar{f}$ can be calculated by taking a weighted average of $f_i$ for all distances $i$. Table 2.3 illustrates the calculation of an approximation for $\bar{f}$ for a tree of order 3 of height 4 (so that there are a total of 46 nodes and so 2116 distances) from the theoretical results from an infinite tree.

<table>
<thead>
<tr>
<th>distance</th>
<th>probability of distance ($\times 2116$)</th>
<th>i.b.d. for infinite number of populations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>m=0.1</td>
</tr>
<tr>
<td>0</td>
<td>46</td>
<td>0.0489</td>
</tr>
<tr>
<td>1</td>
<td>90</td>
<td>0.0237</td>
</tr>
<tr>
<td>2</td>
<td>132</td>
<td>0.0118</td>
</tr>
<tr>
<td>3</td>
<td>168</td>
<td>0.0059</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>0.0029</td>
</tr>
<tr>
<td>5</td>
<td>288</td>
<td>0.0015</td>
</tr>
<tr>
<td>6</td>
<td>384</td>
<td>0.0007</td>
</tr>
<tr>
<td>7</td>
<td>384</td>
<td>0.0004</td>
</tr>
<tr>
<td>8</td>
<td>384</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

Table 2.3 Probability of identity and probability of sampling genes from demes for a tree of height 4.

Figure 2.12 shows the expected values for $F_{ST}$ for demes within varying distances from a central deme. The pattern is similar to that for pairwise $F_{ST}$ values in Figure 2.6. Figure 2.13 shows the fit of estimated $F_{ST}$ values to the theory. The fit is not good. The reason for this is that the correlation in gene frequencies between neighbouring demes will increase the variance in allele frequencies when measured away from the expected value over all space and time, and reduce the estimated variance using the mean gene frequency.
Figure 2.12. Overall $F_{ST}$ values for the area covered by a distance for migration rates of 0.01 (open symbols) and 0.1 (solid symbols) with mutation rate $= 10^{-6}$ and population size = 100. Circles tree of order 2, triangles trees of order 3.

Figure 2.13. Overall $F_{ST}$ values for the area covered by a distance compared to simulation results for 1000 replicates for trees of order 3. Squares $m=0.2$; circles $m=0.1$; triangles $m=0.05$; stars $m=0.01$. Simulations based on trees of height 11 with N=12 and $\mu=0.001$. Error bars represent $\pm 2$ s.e.
2.8 More Realistic River Systems

The tree structures analysed above give very different answers to the expected amount of variation seen a given distance apart from those of a one-dimensional habitat. Real river systems do not have such a regular branching structure as seen in Figure 2.3. How this affects the levels of variation is of interest. A more realistic model for gene flow on a real river system would be to assume that it is a series of one dimensional habitats joined together as in Figure 2.14. The relative importance of branching and the linear habitats can then be investigated.

It would be possible to approximate this type of continuous habitat by using a stepping stone model with branch points, and in principle it is possible to solve the equations for a discrete infinite branching structure with $r$ linear habitats between each branching point. However, the number of variables becomes very large. Simulations are used to investigate the properties of this model.

![Figure 2.14 Branching structure with isolation by distance between the branch points.](image-url)
2.8.1 Single Branch Point

As a first approximation to a real river system consider a number of linear stepping stones joined at one end as in Figure 2.16. The simulation program is based on that used for simulations on a tree structure.

Figure 2.16 System of semi-infinite habitats.

Figure 2.17 indicates the probability of identity for demes close to the barrier for a typical simulation run. The behaviour of this model is qualitatively different than that for a 1 dimensional habitat. Figure 2.18 illustrates the difference for demes close to and further away from the branch point. The identity by descent is reduced close to the branch point.
Figure 2.17 The probability of identity for pairs of demes on the same side of the branch point measured from deme 1 for k=2 (circles), 3 (triangles), 4 (squares) and 5 (stars). The line is the theory for a linear habitat. In all cases m=0.1, N=15, and \( \mu = 0.0005 \). Error bars give ± s.e. Simulations are based on 2000 replicates on semi-linear habitats of length 80 demes.

Figure 2.18 The probability of identity for pairs of genes with 1 gene at positions 0 (squares), 1 (triangles), and 4 (circles). In all cases m=0.1, N=15, k=3, and \( \mu = 0.0005 \). The error bars represent ± s.e. The simulations are based on 2000 replicates on semi-linear habitats of length 80 demes.
2.8.2 More Branch Points

Simulation studies were performed with nodes of order 3 to evaluate the relative contributions of the length of the branches and the number of branch points. Figure 2.19 below shows the estimated probability of identity for simulated trees with 1, 2, and 5 demes between branch points. The identity by descent in these structures is intermediate between that of a linear habitat and a branching structure, but as in the branching structure the probability of identity by descent is reduced considerably compared to a linear habitat.

![Figure 2.19](image)

Figure 2.19 The estimated probability of identity by descent in a tree structure with linear stepping stones between nodes. Upper solid line is expected probability of identity for a linear habitat, and lower line for a tree structure of order 3. Squares represent a single deme between branch points, circles 2 demes, and triangles 5 demes. Error bars represent ±s.e. Simulations based on m = 0.1, μ = 0.001, and N=12. Values based on 500 replicate values. All values measured from the central branch point.
2.9 Discussion

In this chapter classical models for population differentiation caused by limited dispersal, and relevant statistics and definitions for this and subsequent chapters were reviewed. A new model for dispersal when dispersal was limited to a tree structure was also developed. The amount of differentiation between populations in a branching structure is far more than that seen in linear habitats. Even a single branch point can have a considerable effect on the probability of identity seen within a simulated population. The levels of gene flow needed to produce significant genetic differentiation in rivers are therefore less than that needed in linear habitats. However, when the number of populations is small the system was shown to be approximated by an island model.

This is a similar result to those of Nagylaki (1988, Nagylaki & Barchilon 1988) for a barrier to gene flow in a one dimensional habitat, where the barrier causes a sharp drop in the correlation of gene frequencies. Sharp drops as the result of branching points were not observed in the simulations presented here. These results stress the importance of the geography of an area to the patterns of genetic differentiation seen. Most models of genetic differentiation consider highly homogeneous habitats, without the influence of the many impediments to gene flow which characterise natural populations, whether or not these impediments are visible to us.

Rivers may look reasonably uniform to us, barring deep pools and riffles, but to migrating salmon, whose navigation within rivers is thought to depend on olfaction (Stabell 1984), it may be completely different. In this case tributaries are thought to be recognisable by the chemical content of river water. Thus it is likely that a population structure, somewhat like the one presented in this chapter, will be dependent on the confluences of tributaries, which will define boundaries in the chemical content of water, although it has been shown that salmon have some idea of distance within the main stem of rivers (Heggberget et al. 1986). Nevertheless this could still be due to differences in water trace element quantities due to the chemical influence of tributaries. In mountainous areas, or where dispersal is constrained along riverbanks, this form of branching structure is likely to be of importance.

These models give predictions of the amount of genetic differentiation seen in river systems of differing shapes, namely that for highly branched river systems the
amount of genetic differentiation seen would be higher. The effect of branching is seen even when there is only a single branch point, and isolation by distance, although the effect is greatest when there are many branch points.

If the populations of river systems are branched in this way then we would predict that there should be a reduction in the probability of identity with increasing distance, and also that populations on more branching systems should show greater differentiation for comparable levels of gene flow between populations. This leads us to hypothesise that the main causes in genetic variation within river systems are not the distances between populations but are rather the shape of the river systems. In light of this the lack of correlation between genetic and geographic distance in salmon populations within river systems reported by some authors (Verspoor et al. 1991, Jordan et al. 1992) become comprehensible when it is known that salmon home within river systems.

There are problems in that the values of $F_{ST}$ calculated from these branching structures do not agree with the theory for values calculated for a single generation. Correlations between neighbouring demes overestimate $F_{ST}$ compared to the expectation calculated when the variance is calculated using the expected gene frequencies over both time and space, underestimate $F_{ST}$ when the variance is calculated using observed gene frequencies. Pairwise $F_{ST}$ values agree with the theory.

This lack of agreement suggests that a better approach to looking at patterns of genetic variation observed from natural populations may be to consider the properties of samples from the populations rather than whole populations. This approach is taken in the next chapter.
Chapter 3

THE COALESCENT IN SUBDIVIDED POPULATIONS

3.1 Introduction

Over the past 25 years there have been a number of studies investigating levels of genetic variation in salmonids (reviews Elo 1993, Altukov & Salmenkova 1994). The first surveys were based on blood group data, but as technology has advanced this technique has been successively replaced by other methods. Allozyme electrophoresis, and modern techniques such as mini- and micro-satellite DNA sequences, RAPDs, mitochondrial DNA, and genetic fingerprinting have all been used to estimate the levels of genetic variation in natural populations. These methods should produce ever more detailed data on the levels of variation seen within and between populations, and a comparison of the data obtained by the various methods should give us additional information about the evolutionary history of the Atlantic Salmon. The various studies have allowed statistics such as Wright's $F_{ST}$ and Nei's D to be calculated with ever increasing precision. However, the behaviour of these statistics under competing hypotheses for the maintenance of genetic variation and with the mutation rates of the various types of loci differing by many orders of magnitude, is still unclear. The differences in mutational processes which lead to variation at these loci are qualitatively different and this leads to yet further complication.

The classical theory of identity by descent (Wright 1931, 1943, 1951; Malecot 1948; reviewed in Chapter 2 ) provides models for the maintenance of genetic variation in
natural populations and gives predictions for the amount and patterns of genetic diversity for a variety of population structures as shown in the previous chapter. A natural extension of this theory allows a different perspective on the patterns of genetic variation by focusing on the relatedness of the genes within a sample and making explicit the processes by which the genetic variation is produced, namely extremely rare mutations. This separation of the allelic states from other population genetic processes enables an investigation into the effects of varying mutation rates for the different types of data seen.

Surveys of genetic diversity within and between populations were, until recently, based on allozyme data, for which the number of alleles observed per locus is usually small. With mini- and micro-satellite data the number of alleles observed is very high but no phylogeny of alleles can usually be constructed. For sequence data a phylogeny of alleles can, in principle, be constructed, and hence there is more information in a sample of alleles than just allele frequencies; a genealogy of the sampled genes can be inferred. This gives potentially far more informative data than previously. Using inbreeding coefficients would seem to be wasting this extra information. Felstenstein (1992a, b) has shown that pairwise measures are indeed less efficient at estimating population parameters for a single population than maximum likelihood measures based on the full genealogy.

The distribution of coalescence times under different models of population structure are important for an understanding of what genetic surveys may achieve. The techniques for interpreting genealogies in terms of historic events and continuing gene flow are in their infancy. Slatkin and co-authors (Slatkin and Maddison 1989; Hudson et al. 1992) and Templeton et al. (1994) have produced schemes for attempting to infer information about population structure from genealogies. Slatkin's method uses parsimony to attempt to infer the minimum amount of gene flow consistent with the observed genealogy. This showed good results for low levels of gene flow, as would be expected for any parsimony method with extremely rare events. The method of Templeton et al. (1994) uses the (presumed) dates of the earliest splits in the gene genealogies to separate processes which happened in the distant past to those closer to the present. This is a more empirical technique but lacks any real theoretical justification. The behaviour of pairwise measures of genetic diversity for the simple island and stepping stone models are well
understood, as shown in the previous chapter. However, under more complicated models the behaviour is not well understood.

There are inherent difficulties, aside from the practical difficulties in phylogeny reconstruction, in attempting to use the extra information contained in genealogies. Neigel and Avise (1993) used genealogies to attempt to get a measure of the dispersal rate of individual animals but this technique is flawed (Barton and Wilson 1995, 1996) as ancestral genes do not perform a random walk when there is local population regulation. Parsimony techniques have been used to attempt to both date and locate 'mitochondrial Eve'. Barton and Goldman (1992) and Templeton (1993, 1994) illustrated the difficulties of using parsimony techniques in getting geographic information from this mitochondrial data from humans. There are problems in allocating a series of events to a sample based only on a single gene where the statistical power contained in a single loci is likely to be low in most cases (Barton and Wilson 1995, 1996). The combination of data over many loci has not been examined, mainly due to most studies being done on mtDNA. This may change in the future.

The information in pairwise measures has been used to postulate possible population histories in Salmonids (e.g. Kazakov & Titov 1991). The behaviour of pairwise measures under these alternative population histories is thus an area for investigation, as is the possibility of using other, more informative, techniques to investigate alternative population histories for a species. Two of the most common explanations for observed patterns of genetic variation are that of gene flow between populations and that the populations are derived from a single ancestral population which split in the past. Complicating these two models are the effects of extinction and recolonisation of populations.

Using the coalescent in subdivided populations allows the use of, in the most part, simple linear models to describe the distribution of coalescence times both within and between populations. Computer simulations are simple to use when the sample sizes are larger. Models for island populations have been produced previously (Slatkin 1991; Strobeck 1987; Herbots 1995), as has a general migration matrix model (Herbots 1995). These models will be reviewed and the effects of different events in population history in terms of pairwise measures of genetic variation, and in terms of the distribution of coalescence times will be investigated.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

3.2 The Coalescent Process

The coalescent process is a model for the genealogy of a population, which is related to the genetic composition of that population. Most work has concentrated on models for a small sample in the absence of selection, but selection may be included (Kaplan et al. 1991, Hey 1991). In this chapter only neutral loci will be considered. A population of $N$ diploids has a total of $2N$ genes at a given locus. Considering the history of these genes, then in the previous generation, since each individual has 2 parents there are $2N$ parents, going back another generation there are $4N$ grandparents. The total population size is always finite so that some of the ancestors of different individuals must be shared as we go into the past. This ‘pedigree’ is different to the genealogies for any individual genes and each loci may have a different genealogy due to separate chromosomes or recombination. All genes in a population will share a common ancestor with probability one as we go into the past. This fact provides us with the basis for the coalescent process, where we do not keep track of ancestors but of lines-of-descent which are the paths by which genes ‘flow’ through the population. For haploids the number of ancestors stays constant as we go back in time but again at some point in the past all the genes in the population will have descended from a single copy. When further conditions are added then properties of the history of the sample (or the genealogy) can be calculated. The distribution of alleles in the population can then be calculated by considering mutations along the lines of descent.

This model for patterns of genetic variation is implicit in the work on identity by descent by Wright (1943, 1951) and by Malecot (1948) but the ancestry of the individuals was not considered explicitly until the work by Watterson (1975). Kingman (1982a, b) formalised this approach to the study of genetic variation in

---

1 In principle a diploid population could have a prescribed mating system in which no two individuals shared both parents in any generation, or a haploid population could consist of a series of isolated subpopulations. This would mean that no coalescence events would occur. This is only possible for artificial examples, and if we go back in time ancestors must be shared (without independent divine creations!).
populations. Reviews on this field are by Hudson (1990) and, more mathematically orientated, by Tavaré (1984).

In his series of papers Kingman (1980, 1982a, b) proved that the Kingman coalescent (defined below) gave a good description for the relationships between genes in a population for a large class of the general exchangeable models (Cannings 1974). The Wright-Fisher model is a special case of this class of models for which the variance in the number of descendants per individual is finite as the total population size tends to infinity.

The Kingman coalescent for a sample of \( r \) genes was defined in terms of equivalence relations which can be written as groupings of the integers from 1 to \( r \). These groupings correspond to the genes which share the same common ancestor at some time \( \tau \) in the past. The equivalence classes of a set of \( r \) genes at the moment of sampling, prior to any coalescence events, is

\[
(1),(2),\ldots,(r-1),(r).
\]  

(3.1)

After all the genes have coalesced then all the genes in the sample are in a single equivalence class

\[
(1,2,3,\ldots,r-1,r).
\]  

(3.2)

For example, consider the genealogy with 5 genes in Figure 3.1. Labelling the genes 1 to 5 from the left then the equivalence classes at time 0 is \((1),(2),(3),(4),(5)\). At time \( \tau_4 \) there is just the single equivalence class \((1,2,3,4,5)\). The equivalence classes at times \( \tau_1 \) and \( \tau_2 \) are \((1),(2),(3),(4,5)\) and \((1,2),(3),(4,5)\) respectively.

Kingman's coalescent is the stochastic process by which the equivalence classes change from groupings of type (3.1) to those of type (3.2). This is defined in terms of a continuous time Markov-Chain for the Kingman's coalescent, but alternative discrete time processes can be used. For the Kingman's coalescent, time is usually expressed in units of \( 2N \) generations, the diffusion limit. For reviews of Kingman's coalescent process see Hudson (1990) and Watterson (1984).

The coalescent process for a whole population is a fairly complicated expression for the Wright-Fisher model; but if we just consider a small sample compared to the population size then it is considerably simplified.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

Figure 3.1 The coalescence times for a sample of 5 alleles from a single population.

A simplified coalescent process on a small sample compared to the population size is an advantageous method to use for modelling the genetic variation of samples from natural populations because:
• The process deals with a sample rather than the whole population; hence it is amenable to efficient simulation, and simulations are guaranteed to be in steady state (Hudson 1990).

• The inbreeding and dispersal events by which the patterns of genetic variation are formed is brought to the fore in the model.

• The influence of mutation is separated from the genealogy.

• It allows us to use the idea of perfect data; that is, data from which the entire genealogy of the population is known, and the times of all the coalescent events.

For these reasons the coalescent will be used to investigate the effects of various kinds of population structuring, both the standard island; and one-dimensional models, and also the effects of population fission, extinction/recolonisation and range expansion.

3.2.1 The Coalescent process in a single population

3.2.1.1 Pairwise Coalescence Times

The idea behind the coalescent is to trace back in time the history of a sample of alleles. Any pair of genes sampled from a single population will (with probability one) be descended from a single ancestor at some point in the past for a finite population size.

Consider a Wright-Fisher population with constant population size 2N haploids per generation and non-overlapping generations so that the numbers of offspring from any individual are multinomially distributed. Label the generations \( t = 0, 1, 2, \ldots \) with \( t=0 \) being the time of sampling and \( t=i \) being \( i \) generations in the past. Now consider a single neutral locus. Taking a pair of genes at random without replacement from

\[ 2N \text{ haploids are considered so that the population size is equivalent to that for a diploid population; in this case the coalescent process is exactly markovian, rather than the approximation used for diploid populations.} \]
this population the probability of coalescence in the first generation in the past (t=1) is simply the probability that the two alleles share the same parental allele. For haploid genes this is then just $1/2N$.

If no coalescence occurs in generation 1 then the probability of coalescence in generation 2 is the same. This gives a geometric distribution for the time until coalescence. So we have

$$\Pr(\text{coalescence in generation } i) = p_i = \left(1 - \frac{1}{2N}\right)^{i-1} \left(\frac{1}{2N}\right).$$

(3.3)

For a population of N diploids we have to take account of sex. The two genes in any generation can now be in the same individual (in which case no coalescence can take place in the next generation), or in different individuals. Consider a sample of 2 genes at generation 0, assuming that the genes are in different individuals. The probability of the genes coming from the same individual is $1/N$, and the probability of the alleles being the same is $1/2$ so that the probability of identity in the previous generation is $1/2N$. The probability that the two genes are in the same individual, but do not coalesce is also $1/2N$. In subsequent generations we need to keep track of the location of the two genes.

Denote the probability of a pair of genes being in the same individual in generation $i$ by $s_i$, and the probability of being in a different individual by $d_i$. The rate at which these two probabilities decay is the rate of coalescence. The state at time $i$ is a Markov chain, and the transition equations are:

$$d_{i+1} = \left(1 - \frac{1}{N}\right)d_i + s_i, \text{ and}$$

$$s_{i+1} = \frac{d_i}{2N}.$$ 

(3.4)

Thus we obtain the difference equation

$$d_i = a_1 \lambda_1^i + a_2 \lambda_2^i,$$

(3.5)

and using the initial conditions $d_0 = 1$ and $d_1 = \frac{N-1}{N}$.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[ \lambda_1 = \frac{N - 1 - \sqrt{N^2 + 1}}{2N}, \quad \lambda_2 = \frac{N - 1 + \sqrt{N^2 + 1}}{2N}, \]

\[ a_2 = \frac{N - 1 + \sqrt{1 + N^2}}{2\sqrt{1 + N^2}} \text{ and } a_1 = 1 - a_2. \]

The leading eigenvalue of this is \( \lambda_2 \), which is approximately equal to \( \frac{2N - 1}{2N} \) for large \( N \), and \( a_2 \) is approximately \( \frac{2N - 1}{2N} \). The probability that the coalescence is in the \( i \)-th generation is \( \frac{d_{i-1}}{2N} \) and thus the probability of coalescence is in the \( i \)-th generation is approximately geometric with mean \( 2N \), the same as for haploid populations with \( 2N \) individuals. This approximation will be used in this chapter when we consider samples from diploid populations.

### 3.2.1.2 Sample of \( r \) genes

Consider a sample of \( r \) genes from a single locus from a haploid population of \( 2N \) individuals with non-overlapping generations. Since we are assuming finite total population size, at some point in the past there is a common ancestor to all these alleles (see Figure 3.1 for a diagram of one possible history of a sample). We can work back to this original time by considering in turn each of the individual coalescent events. If we assume that the population from which the sample (of size \( r \)) is taken follows a Wright-Fisher model then the time until coalescence for any pair of genes is distributed as a geometric distribution with mean \( 2N \) as shown in the previous section. Here we assume that \( r \) is small compared to the total population size \( 2N \) so that the probability of more than one coalescent event in any one time period is small, as is the probability of 3 genes coalescing into a single one.

Now, this discrete distribution can be approximated by an exponential distribution with mean \( 2N \). There are a total of \( r(r-1)/2 \) distinct pairs of genes in the population at time 0, and the first coalescence can occur between any of these pairs of genes. Until the first coalescence event, the times of these events are independent. As we have already assumed that the probability of more than one event in a time period is small, the time until the minimum of these events defined as \( t_1 \) (i.e. the first coalescent event) can be calculated as:
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[ \Pr(t_1 < t) = 1 - \Pr(t_1 > t) \]
\[ = 1 - \left( e^{-\frac{t}{4N}} \right)^{r(r-1)/2} \]
\[ = 1 - e^{-\frac{r(r-1)t}{4N}}. \]  

(3.6)

Hence defining \( f_i(s) \) as the distribution of time from the \( i \)-th to the \( i+1 \)-th coalescence event \( f_i(t) = \frac{r(r-1)}{4N} \text{Exp} \left( \frac{-r(r-1)t}{4N} \right) \), which is an exponential distribution with parameter \( r(r-1)/4N \). The next coalescent event is now the minimum of \( (r-1)(r-2)/2 \) distinct events, which is distributed as exponential with parameter \( (r-1)(r-2)/4N \). We continue this until all the genes have coalesced.

Then the distribution of time between the \( i \)-th and \( i+1 \)-th coalescence event is given by

\[ f_i(t) = \frac{(r+1-i)(r-i)}{4N} \text{Exp} \left( \frac{(r+1-i)(r-i)t}{4N} \right). \]  

(3.7)

The time until all the genes have coalesced is thus distributed as the sum of \( r-1 \) independent exponential distributions with parameters \( 1/(2N), 3/(2N), \ldots, r(r-1)/(4N) \).

The probability for any shape of a tree for this neutral model in a single population can be calculated using the method described in Harding (1971). The distribution of the shape of the tree is easy to calculate in this case because any of the lines of descent may coalesce with equal probability.

### 3.2.2 The Structured Coalescent

The structured coalescent is a coalescent process where there are a number of distinct populations which exchange migrants. This idea was mathematically formalised by Herbots (1995). The additional information contained in the location of the lines-of-descent complicates the model. The further assumptions made for the structured coalescent are that any migration events between demes occur as a Poisson process so that the time between migration events has an exponential distribution (Feller 1950). This assumption is equivalent to the classical models discussed in Chapter 2 in continuous time. This type of model has been used by many authors to attempt to model the variation seen within and between subdivided populations.

Strobeck (1987) showed that the expected number of sites differing (under an infinite sites model of mutation) between genes sampled at random from within the same population for a subdivided population is the same as would be expected for a single panmictic population under some conditions. This is not explicitly a coalescent model but the number of sites differing under this model is a synonym for time. Slatkin (1987b) obtained similar results again using the infinite sites model rather than time.

The result due to Strobeck (1987) is that with a population of haploids which is subdivided into $n$ demes of sizes $2N_i$ ($i=1,\ldots,n$) with a backward migration matrix $M$ (the probability of an individual in deme $i$ having been in deme $j$ a generation before is $m_{ij}$) then if the migration between demes is conservative (that is the expected number of individuals migrating into a deme is the same as the number migrating out or $N_i \sum_{i \neq j} m_{ij} = \sum_{i \neq j} N_j m_{ij}$) then for a randomly chosen deme (weighted by the size of the deme) the expected coalescence time for genes sampled from the same deme is given by:

$$E[T_0] = 2 \sum_{i=1}^{n} N_i,$$

where $T_0$ is the time until coalescence for two genes sampled from a single deme (under weighting by size of deme). Thus the mean time until coalescence is independent of the migration rates.

Tightening the assumptions so that all demes have the same population size ($N_i = N \quad \forall i$) and assuming that all demes have the same migration pattern (the formalised conditions are given in Strobeck (1987)) then the mean coalescence time for genes sampled from any deme is $2Nn$.

These results are extended and formalised by Herbots (1995) who gives a more general result under the conditions of conservative migration and where sampling is weighted by size of deme. This can be done by considering the moment generating
functions of the time until coalescence within and between demes. Theorem 5.1 of Herbots (1995) gives us

\[ f_0 + N_T \mu \tilde{f} = 1, \]  

where \( f_0 \) is the probability of identity within a single deme (under weighting by size of deme), \( \tilde{f} \) is the probability of identity for genes sampled at random from the whole population, \( \mu \) is the mutation rate, and \( N_T = \sum_i N_i \). The definitions for the identity by descent for a coalescent and discussions about their approximation are given in Section 2.3.

By differentiating (3.9) with respect to \( \mu \) then we obtain

\[ E[T_0^\nu] = 2\nu N_T E[T^{\nu-1}], \]  

where \( T_0 \) and \( T \) are the times until coalescence for genes sampled at random from within the same subpopulation (weighting by size) and from the population at random. With \( \nu = 1 \) this gives us (3.8).

### 3.2.3 Simulating the Coalescent Process

Simulating genealogies using the coalescent is a very efficient way of generating genetic data under a population genetic model. For example, to simulate the sampling of \( r \) genes from a single population then a simulation can proceed by generating the times of coalescence for the \( r-1 \) independent coalescent events by generating them from the appropriate exponential distributions given in equation 3.7. The shape of the genealogical tree for this model can be simulated by noting that any two lines of descent may coalesce with equal probability so that any labelled shape of the tree is equally probable. The shape of a tree may then be simulated by coalescing any of the extant lines at random.

Gene frequencies can be obtained from the coalescence times and the tree shape by assigning mutations to the branches proportional to the branch lengths according to the mutation models used (some mutation models were described in Section 2.3).
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

The method given for generating genealogical data from a single population can be extended to the structured coalescent, however the position of all the lines-of-descent must be tracked as the lines will not join at random except within single populations.

Two algorithms for generating a tree by this method for the island model are given in Appendix 3.

Algorithm *make_tree_island* is a discrete generation time algorithm which goes back one generation at a time. At each time point the program loops through all the remaining lines-of-descent and a uniform random number between 0 and 1 is generated for each line-of-descent. If the random number is less than \( m \), the migration rate, then the line is moved to another deme at random. Next, the program loops through all pairs of lines-of-descent. If the lines are in the same deme then another uniform(0,1) random number is generated, and if this is less than \( \frac{1}{2N} \) then the two lines are joined. The algorithm then adds one to the time counter and repeats the whole process until only a single line-of-descent remains.

Algorithm *make_tree_island2* is a more complex, and faster, algorithm, in continuous time. This algorithm is event driven. The algorithm proceeds by generating an exponential variable which is the time (going into the past) of the next event. The time until the next event (if there are \( n \) lines-of-descent), is the minimum of \( n \) exponentials each with rate \( m \) (the migration events) and a number of exponentials of rate \( \frac{1}{2N} \) (the number is the sum of the possible coalescent events). Which event occurs is generated using a uniform random number with probability of an event determined by the ratio of the rates. If there is a coalescence lines-of-descent are joined at random within demes (the deme being determined relative to the rate of coalescences within it). Again this algorithm proceeds until there is only a single line-of-descent remaining.

It is comparatively simple to extend these algorithms to work with generalised migration structures, with different migration rates and population sizes. The second, continuous time, algorithm is used in subsequent chapters. The first was used as a check of the theory, and of the second, considerably faster, algorithm.
3.3 Measures of Diversity from the Coalescent

3.3.1 Within Population Measures

Pairwise measures of genetic diversity and distance are readily amenable to investigation using coalescence times as they are based on the moments of coalescence times, or rather of the number of mutations, between individual alleles. Other measures such as the number of alleles and the gene diversity are more complicated since they depend on the complete distribution of coalescence times. Felsenstein (1992a, b) showed that using statistics based on the full distribution of times gave more accurate measures than those based on pairwise times. The full distribution of times for a subdivided population must condition on the positions of the lines in the past and while it may in principle be written down (Barton and Wilson 1995) it is difficult to infer any information from the distribution. Simulation is the best way to investigate the properties. Pairwise measures are focused on first.

The probability of identity for a pair of alleles sampled from a single population can be thought of in terms of coalescence times by noting that it is simply the probability that no mutation events have occurred in either of the lines of descent since the coalescence of the pair of alleles.

Define $c_i$ to be the probability of coalescence $i$ generations before the present for a discrete time model, and $c(t)$ the distribution function for the coalescence time distribution in continuous time. Then, assuming that each mutation produces an allele which is new to the population (infinite sites, or infinite alleles), the probability of identity is given by:

$$P = \sum_{i=0}^{\infty} c_i (1-\mu)^{2i} \quad \text{if we use discrete time steps or}$$

$$P = \int_0^\infty c(t) e^{-2\mu t} dt \quad \text{for a continuous time approximation.} \quad (3.11)$$

Now this is the probability distribution function in $(1-\mu)^2$ for the discrete time version. The moment generating function of a random variable with probability distribution function $f(x)$ is $E[e^{ax}] = \int_{-\infty}^{\infty} e^{ax} f(x) \, dx$, if this exists, so the probability of
identity can be calculated from the moment generating function of coalescence times by putting $s = -2\mu$. The moment generating function of an exponential distribution with parameter $\lambda$ is $\frac{\lambda}{\lambda - s}$, and the time until coalescence for the continuous model is exponential with mean $1/2N$. This has a moment generating function $1/(1-2Ns)$ so, the probability of identity is $1/(1+4N\mu)$ as shown by other methods (Nei 1987).

Slatkin (1991) approximated the above results by noting that for small $\mu$, $f$ could be approximated, in both cases, by $1 - 2\mu \bar{t}$ where $\bar{t}$ is the mean time until coalescence so that pairwise measures could be expressed in terms of mean coalescence times. For the single population $\bar{t} = 2N$ giving probability of identity $1-4N\mu$. Taylor expansions of higher orders may also be used to get more accurate approximations for the identity by descent. These approximations are not very good (and not necessary for this simple model) for $N\mu > 10^{-2}$ and will not converge for $N\mu \geq 1/4$.

For substitution events the mutation rate is of the order of $10^{-6}$ so for small populations the approximation may be suitable. With STR (single tandem repeat) loci, mutations do not correspond to simple substitution events of this kind so the mutation rates are correspondingly higher, and the approximation is not suitable.

Figure 3.2 illustrates the exact value with first and second order approximations obtained by Taylor expansion. For this single population the first order approximation is not suitable for $N\mu < 0.01$. 

Figure 3.2. The exact \( f \) value \( 1/(1+4U) \) (solid line) compared with Slatkin’s approximation (dashed line) and second order approximation (dotted line) for \( U (= N\mu) \) from 0.0 to 0.2.

3.3.2 Between Population Measures

Wright’s \( F_{ST} \) (Wright 1943, 1951) is defined in terms of identity coefficients within subpopulations, and for pairs of alleles drawn at random from the whole population as given in Equation 2.15.

\[
F_{ST} = \frac{f_0 - \bar{f}}{1 - \bar{f}}
\]

where \( f_0 \) is the probability of identity within a population and \( \bar{f} \) is the probability of identity for a pair of genes sampled from the population at random. As seen before these can be calculated by using the moment generating functions of the time until coalescence for pairs of genes sampled at random from within a population or between populations.

Consider a single random mating deme of size \( 2N \) haploids, with non-overlapping generations, within a subdivided population. The probability of identity by descent
for 2 genes sampled at random from this population \((= E[e^{-2\mu}]\) within this population can be calculated conditioning on the first event. Possible events are coalescence of the 2 genes, mutation in either of the lines of descent, or a migration of 1 of the lines of descent from another deme. If migration or mutation events are rare then the time between successive migration or mutation events is approximately exponentially distributed. The time until the first of the three events is then the minimum of three exponentials and hence is exponential with the sum of the rates (as in Equation 3.7), and the relative probabilities of coalescence and migration are given by their relative rates (Feller 1950). Thus we get:

\[
f_o = \Pr[i.b.d./first \ event \ coal] \Pr[coal] + \Pr[i.b.d./first \ event \ mig] \Pr[mig] + \Pr[i.b.d./first \ event \ mutation] \Pr[mutation] \\
= 1 \times \frac{1}{2N} + \frac{2m}{2N + 2m + 2\mu} \Pr[i.b.d./first \ event \ migration] \\
= \frac{1}{1 + 4Nm + 4N\mu} \cdot \frac{4Nm}{1 + 4Nm + 4N\mu} f_i, \tag{3.12}
\]

where \(f_i\) is the probability of identity for genes sampled at random from different demes. Assuming that the whole population is very large so that the probability of identity once the line of descent has left the deme is small \((f_i << 1)\) then the expected value of \(F_{ST}\) (over the evolutionary process) can be calculated:

\[
E[F_{ST}] \approx f_o \approx \frac{1}{1 + 4Nm + 4N\mu}. \tag{3.13}
\]

If the mutation rate is small

\[
E[F_{ST}] \approx f_o \approx \frac{1}{1 + 4Nm}. \tag{3.14}
\]

Note that for simplicity in the remainder of this chapter, \(F_{ST}\) will be used for the expected value of \(F_{ST}\) over the evolutionary process. Equation 3.14 is identical to
the value for the infinite island model (Wright 1951) which was derived in the previous chapter. This explains the observation that the values of $F_{ST}$ observed for other models of population structure are approximated by the island model (Slatkin and Barton 1989), because events within populations predominantly determine expected values of $F_{ST}$ over the evolutionary process, rather than the geographic position of populations. This is a justification of the use of the island model to try to get estimates of gene flow. The relation does not hold in all population structures: if we consider a 1-dimensional population then the migration event can only be to the 2 adjacent populations, and so this will not hold, in 2 dimensions it is closer.

Nei’s $D$ can be defined (Nei 1987) as

$$D = -\log_{10} \left( \frac{f_1}{f_0} \right).$$  \hspace{1cm} (3.15)

Proceeding as for $F_{ST}$ if we assume that the distribution of time until the pair of genes are in the same deme is $w(t)$ then:

$$f_1 = f_0 E_w \left[ e^{-2\mu t} \right],$$

so that

$$D = -\log_{10} \left( E_w \left[ e^{-2\mu t} \right] \right).$$

$D$ depends only on the moment generating function of the time spent until the pair of genes are in the same deme. Hence assuming that the amount of time until the genes are in the same deme is distributed as an exponential with mean $\tau$ then

$$D = -\log \left( \frac{1}{1 + 2\tau \mu} \right).$$  \hspace{1cm} (3.16)

This illustrates the differences between these two pairwise statistics. $F_{ST}$ only depends (to a first approximation) on what happens within any one deme whereas Nei’s $D$ only depends on what happens when the two genes are in separate demes. The interpretation of these two statistics is therefore completely different.

### 3.4 The Coalescent and Effective Population Sizes

The coalescent process can also be used to get a measure of effective population size. The inbreeding effective population size (Ewens 1982) can also be calculated
from these mean coalescence times (Nei and Takahata 1993). These results did not explicitly define effective population size but used the mean time until coalescence for a pair of alleles sampled at random from the population.

It is not immediately obvious from the definitions of effective population size given in Chapter 2 what this new measure represents. Effective population sizes are measured for the whole population, whereas pairwise coalescence times are for a pair of genes sampled from the population. This measure is dependent on the population structure remaining the same until the pair of alleles has coalesced, so that there is no dependence on historic events.

This method does give familiar results to those derived using other methods. For example consider a single population in which the population size cycles, i.e. the population size is $N_0$ for even numbered generations and $N_1$ for odd numbered generations. Then the classical approach gives $N_E$ (for inbreeding, variance and eigenvalue effective population sizes) to be the harmonic mean of $N_1$ and $N_0$.

The probability of coalescence in generation $j$, $p_j$ is:

$$p_j = \left(1 - \frac{1}{N_1}\right)^{\frac{j-1}{2}} \left(1 - \frac{1}{N_0}\right)^{\frac{j-1}{2}} \frac{1}{N_1} \quad \text{for } j \text{ odd}$$

$$p_j = \left(1 - \frac{1}{N_1}\right)^{\frac{j}{2}} \left(1 - \frac{1}{N_0}\right)^{\frac{j}{2}} \frac{1}{N_0} \quad \text{for } j \text{ even}.$$

Hence the expected time until coalescence is

$$\sum_{j=1}^{\infty} 2j(1 - 1/N_1)^j(1 - 1/N_0)^{j-1}1/N_0 + \sum_{j=0}^{\infty} (2j + 1)(1 - 1/N_1)^j(1 - 1/N_0)^j1/N_1$$

$$\approx \frac{2}{1/N_0 + 1/N_1} \text{ ignoring terms of order } 1/N^2.$$

The coalescence effective population sizes for island models have been explored previously (Nei and Takahata 1993).
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

3.5 The Finite Island Model

The finite island model has been studied many times previously; Latter (1973) and Nei (1973) have used either the correlation between uniting gametes or the probability of identity by descent to investigate the model. Various authors (Strobeck 1987; Slatkin 1991, Nei and Takahata 1993, Herbots 1995) have used the coalescent as a tool. In this section a coalescent approach is used to derive the usual identities associated with this model, and investigate the distribution of coalescence times within the model. As shown in the previous chapter the infinite island model can be a good approximation to the patterns of variation seen in natural populations. All natural populations are finite, and to do any further analysis of the island model using the coalescent a finite total population size is needed. Using the finite island model the geographic or temporal (Aspinwall 1974, Jordan 1992) separation of the samples is abstracted into a finite number of discrete populations each of which exchange migrants at the same rate.

3.5.1 Moment Generating Functions

Consider a finite island model with k randomly mating subpopulations of constant size N diploids and non-overlapping generations with a migration rate of m per generation. This is equivalent to saying that a gene has a probability m of being the descendant of a gene in another population in the previous generation.

The probability of identity by descent for genes sampled from within and between populations can then be calculated using a similar method to that of Equation 3.12, however, by calculating the moment generating functions we can get the probability of identity directly from Equation 3.11, and we can also obtain the moments of the time until coalescence.

To do this we use the result (Rice 1988) that if X has moment generating function $m_X$ and Y has moment generating function $m_Y$ then if X and Y are independent then the sum $X+Y$ has moment generating function $m_X m_Y$. Let $T_o$ and $T_i$ denote the times until coalescence for genes in the same and different demes respectively.

Consider first a pair of genes in different demes. Then the time until coalescence is the sum of the time until the pair of genes are in the same deme for the first time,
and the time until coalescence within a single deme. Now the time until the genes are in the same deme is exponential with rate $2m/(k-1)$, and thus has moment generating function $2m/(2m-s)$. Hence $E[e^{sT_0}] = E[e^{sT_0}] = \frac{2m/k}{2m/k-s}$.

For a pair of genes within the same deme we condition on whether the first event is a migration or a coalescence. The time until this event is the minimum of two exponentials and so is exponential with the sum of the rates giving a moment generating function $(1/(2N) + 2m)/(1/(2N) + 2m-s)$ (here we assume that the probabilities of coalescence and migration are small so that the probability of simultaneous events is ignored). Thus we can write down

$$E[e^{sT_0}] = E[e^{sT/\text{first event coal}}] \times E[e^{sT/\text{first event mig}}] \times \Pr[\text{coal}] + E[e^{sT/\text{first event mig}}] \times \Pr[\text{mig}] \times E[e^{sT_0}]$$

Hence we get the following simultaneous equations:

$$E[e^{sT_0}] = \frac{1}{4mN+1-2sN} + \frac{4mN E[e^{sT_0}]}{4mN+1-2sN} \quad \text{and}$$

$$E[e^{sT_1}] = \frac{2m E[e^{sT_0}]}{2m-s(k-1)}.$$  

which give moment generating functions for the time until coalescence within and between demes:

$$E[e^{sT_0}] = \frac{2m-s(k-1)}{-8m^2N+(2m-s(k-1))(1+4mN-2Ns)} \quad \text{and}$$

$$E[e^{sT_1}] = \frac{2m}{-8m^2N+(2m-s(k-1))(1+4mN-2Ns)}.$$  

The mean time until coalescence for samples taken within and between demes can be calculated by differentiating with respect to $s$ and putting $s=0$: this gives $2Nk$ and $2Nk+(k-1)/2m$ as expected (Slatkin 1991, Strobeck 1987).

### 3.5.1.1 Identity and Distance Measures

We can now calculate $f_0$ and $f_1$ by putting $s=-2\mu$ in Equations 3.18. Using these $f$'s and Equation 2.15 gives the expected values of $F_{ST}$ over the evolutionary process:
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[
F_{ST} = \frac{1}{1 + 4Nmk^2 / (k-1)^2 + 4N\mu k / (k-1)^2},
\]

(3.19)

as obtained by Nei (1975).

Using the mean coalescence time approximation for identity by descent discussed in Section 3.3; Slatkin (1991) calculated \(F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}} = \frac{1}{1 + 4Nmk^2 / (k-1)^2}\). Using this approximation the mutation rate cancels to a first order, and so \(F_{ST}\) is “relatively” unaffected by the mutation rate as compared to the identity by descent shown in Figure 3.3. Table 3.1 gives the percentage error in Slatkin’s approximation for \(F_{ST}\) compared to the exact solution given by Equation 3.19. This approximation is reasonably accurate when the mutation rate is an order of magnitude lower than the migration rate.

Other measures of genetic variation between populations such as Nei’s genetic distance \(D\) (Nei 1973, 1987) can also be written in terms of identity coefficients within and between populations. We then have \(D = -\log(f_1 / f_0)\) and so

\[
D = -\log\left(\frac{m}{m + \mu(k-1)}\right) \approx \frac{\mu(k-1)}{m} \text{ ignoring terms of order } \left(\frac{\mu}{m}\right)^2,
\]

In this case there is a major dependence on \(\mu\), and no dependence on \(N\), as shown in Figure 3.4.
Table 3.1 Percentage error in the expected values of $F_{ST}$ seen by Slatkin’s method and by Equation 3.19. Values calculated from island model with 100 populations.

<table>
<thead>
<tr>
<th>Nm</th>
<th>Nμ</th>
<th>1</th>
<th>0.1</th>
<th>0.01</th>
<th>$10^{-3}$</th>
<th>$10^{-4}$</th>
<th>$10^{-5}$</th>
<th>$10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td>9.66</td>
<td>0.966</td>
<td>0.0966</td>
<td>0.009</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>79.6</td>
<td>7.96</td>
<td>0.796</td>
<td>0.080</td>
<td>0.008</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>286.9</td>
<td>28.7</td>
<td>2.87</td>
<td>0.287</td>
<td>0.029</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>402.4</td>
<td>40.24</td>
<td>4.240</td>
<td>0.240</td>
<td>0.024</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>$10^{-3}$</td>
<td></td>
<td>403.9</td>
<td>40.38</td>
<td>4.038</td>
<td>0.0404</td>
<td>0.0040</td>
<td>0.0004</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 3.3. $F_{ST}$ as a function of Nm for the finite island model for $Nμ = 0.01$ (solid line); 0.1 (long dashed line); and 1 (short dashed line) with 10 islands.
3.5.1.2 Effective Population Size

The effective population size of the island population is simply the mean time until coalescence for a pair of genes sampled at random from the population, as defined in Section 3.4. For the finite island model we have:

\[ N_E = 2Nk + \frac{(k-1)^2}{2mk}. \]  (3.20)

The effective size of a population is thus increased by the subdivision. A single population, with the same number of diploid individuals (Nk), would have effective population size 2Nk.
3.5.2 The Distribution of Pairwise Coalescence Times in an Island Model

All the measures of genetic variation examined in this chapter use the moment generating function of the pairwise coalescence times. Though the moments uniquely determine the distribution, the distribution of coalescence times can give us useful information. This distribution of coalescence times has been shown to provide a better estimate for population size than pairwise measures (Felsenstein 1992a,b). The distribution of a set of coalescence times and the null probability of a genealogy can be calculated for a set of observations from a single Wright-Fisher population as described in Section 3.2. The distribution of shapes for the genealogy is then the same as for any tree produced by a simple branching process (Harding 1971) so the null probability of any tree shape can be calculated. When there is population subdivision the picture is more complicated. The times of coalescence are affected by the position, and the shape of any tree is no longer the same as that produced under a simple branching process.

Consider a population of k randomly mating demes with discrete non-overlapping generations. Assume that the migration rate per generation $m$ is small, and the population size per deme $N$ is large, so that the probability of two events in the same time step is small. Using a continuous approximation, let $b(t)$ and $w(t)$ denote the distribution function of coalescence times for a pair of alleles sampled from different demes, and the same deme respectively. Then we can write down equations for the probability that the coalescent event is $t+1$ generations in terms of the distribution of the location of the 2 genes $t$ generations in the past.

$$b(t+1) = \left(1 - \frac{2m}{k-1}\right)b(t) + \frac{2m}{k-1}w(t),$$
giving

$$b(t+1) - b(t) = \frac{2m}{k-1}(w(t) - b(t)).$$

$$w(t+1) = \left(1 - 2m - \frac{1}{2N}\right)w(t) + 2mb(t),$$
so

$$w(t+1) - w(t) = 2mb(t) - \left(\frac{1}{2N} + 2m\right)w(t).$$
Hence the differential equations for the distribution of coalescence times within and between populations can be written down (where prime denotes the differential with respect to t)

\[ b'(t) = \frac{2m}{k-1}(w(t) - b(t)) \quad \text{and} \]

\[ w'(t) = 2mb(t) - \left( \frac{1}{2N} + 2m \right)w(t), \]

with the boundary conditions \( b(0) = 0 \) and \( w(0) = 1/2N \).

These give the solutions

\[ b(t) = \frac{m}{\lambda N(k-1)} \left( e^{\frac{(a_2 - \lambda)t}{\lambda}} - e^{\frac{(a_2 + \lambda)t}{\lambda}} \right) \quad \text{and} \]

\[ w(t) = \frac{1}{4N\lambda} \left( (\lambda - a_1)e^{\frac{-(a_2 + \lambda)t}{\lambda}} + (\lambda + a_1)e^{\frac{-(a_2 - \lambda)t}{\lambda}} \right), \]

where

\[ a_1 = \frac{2m}{k-1} - 2m - \frac{1}{2N}, \quad a_2 = \frac{2m}{k-1} + 2m + \frac{1}{2N}, \quad \text{and} \]

\[ \lambda = \sqrt{a_1^2 + \frac{16m^2}{k-1}}. \]

Herbots (1995) derived the scaled distribution of coalescence times for this model using the moment generating functions. Figure 3.5 illustrates the difference between the distribution of times within and between demes, compared to that for a single population of the same total population size. The subdivision does not alter the mean time until coalescence within a region but increases the probability of coalescence in the first few generations, and also a long time in the past.
Figure 3.5 The distribution of coalescence times for a pair of alleles sampled from within the same deme (dashed line) and from different demes (dotted line) for an island model with 10 demes, a population size of 100 diploids per deme, and a migration probability of 0.001 per gamete per generation. The solid line is the pairwise time for genes from a single panmictic population of size 1000 diploids.

3.6 Hierarchical Island Model

Wright (1951) showed how the levels of structure within a population could be analysed in term of hierarchical \( F \) statistics. Carmelli and Cavalli-Sforza (1976) investigated this model and Slatkin and Voelm (1991) used a mean coalescence time approach to investigate this model assuming that the rate of gene flow between regions was much less than that within regions. In this section the partitioning of genetic variation into between region and within region components will be analysed. This approach to the partitioning of genetic variation was pioneered by Dobzhansky and Wright (1941, see also Wright 1978) who used the frequency of a recessive lethal gene in populations to get an estimate of the structuring within populations.
This model of population structure has been used extensively as an approximation for the real population structure which will be considerably more complicated. Jordan (1992) and Elo (1993) have used this model to obtain estimates of the population structure within Atlantic salmon populations, based on within river systems, between regions, and between continents. The interpretation of these F statistics in terms of population parameters are unknown in general. Slatkin and Voelm (1991) concentrated on the cases where gene flow within regions was very much greater than that between regions. While this may be true for most species, it is not at all clear if it is true for the Atlantic salmon; so the consequences of other patterns of gene flow would be useful.

Define the model as follows:

- r regions each with k Wright-Fisher populations with non-overlapping generations, each with N diploid individuals, and a migration rate of m of which a proportion p is between populations within regions and a proportion 1 − p between regions.

Let

- \( t_0 \), \( t_1 \), and \( t_2 \) be the time of coalescence for individuals sampled from within a population; between populations within a region and between regions respectively.
- \( \tilde{f}_R \) to be the probability of identity for 2 alleles sampled at random from within a region. As before let \( f_0 \) and \( \tilde{f} \) be the probability of identity for a pair of alleles sampled at random from within a population, and from the whole population respectively.

In order to get the moment generating functions for these times we follow the method of Section 3.5.1. We want to calculate the moment generating functions of times of coalescence within demes; between demes within regions and between regions. Ignoring mutation there are 3 possible events for a pair of genes within the same deme:

- a coalescence which occurs at a rate \( 1/(2N) \);
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

- a migration between demes within a region, which happens at rate pm;
- or a migration between regions which happens at a rate \((1-p)m\).

To do this we condition on the first event to get three simultaneous equations; using the same arguments as in Section 3.5.1 we obtain:

\[
E[e^{sT_0}] = \frac{1}{4mN + 1 - 2sN} + \frac{4Nm(1-p)E[e^{sT_1}]}{4mN + 1 - 2sN},
\]

\[
E[e^{sT_1}] = \frac{2mE[e^{sT_0}]}{2m(p + (k-1)(1-p)) - s(k-1)} + \frac{2m(1-p)(k-1)E[e^{sT_2}]}{2m(p + (k-1)(1-p)) - s(k-1)}
\]

and

\[
E[e^{sT_2}] = \frac{2m(1-p)E[e^{sT_0}]}{2m(1-p) - s(r-1)k} + \frac{2m(1-p)(k-1)E[e^{sT_1}]}{2m(1-p) - s(r-1)k}.
\]

Following Nei (1973) we define:

\[
F_{ST} = \frac{f_o - \tilde{f}}{1 - \tilde{f}}, \quad F_{SR} = \frac{f_o - \tilde{f}_R}{1 - \tilde{f}_R} \quad \text{and} \quad F_{RT} = \frac{\tilde{f}_R - \tilde{f}}{1 - \tilde{f}}.
\]

Note that as expected 

\[(1 - F_{ST}) = (1 - F_{SR})(1 - F_{RT}).\]

In a similar way Nei’s \(D\) can be partitioned into

\[
D = -\log\left(\frac{f_2}{f_0}\right), \quad D_{RT} = -\log\left(\frac{f_2}{f_1}\right) \quad \text{and} \quad D_{SR} = -\log\left(\frac{f_1}{f_0}\right),
\]

with \(D = D_{RT} + D_{SR}\).

To calculate the hierarchical F statistics for this model, the moment generating functions for pairs of alleles sampled at random from within a region and from within the whole population are required.

Now

\[
\tilde{f} = \left[\left(E[e^{2\mu T_0}] + (k-1)E[e^{2\mu T_1}]\right) + (r-1)kE[e^{2\mu T_1}] \right] / kr
\]

and

\[
\tilde{f}_R = \left(\frac{E[e^{2\mu T_0}] + (k-1)E[e^{2\mu T_1}]}{k}\right),
\]

(3.26)
which gives

\[ F_{ST} = \left(1 + \frac{4Nkr(m(k-1) + mp + \mu(k-1))(mr(1-p) + \mu(r-1))}{m((1-p)(kr^2(k-2) - (2r-1)) + k(r-1)^2) + \mu(r-1)(k-1)(r(k-1))}\right)^{-1}, \]

\[ F_{SR} = \left(1 + \frac{4Nm(k(k-1) + 4kNmp + 4k\mu(k-1))}{(k-1)^2}\right)^{-1} \]

and

\[ F_{RT} = \left(1 + \frac{r((k-1)^2 + 4Nm(k(k-1) + 4kNmp + 4\mu(k(k-1))(mr(1-p) + \mu(r-1)))}{(r-1)^2(m(k-1)+mp+\mu(k-1))}\right)^{-1}. \]

These are very complicated expressions. As shown in Section 3.3.2 the influence of the mutation rate on the values of \( F_{ST} \) statistics is minimal as long as \( t \) is small compared to \( m \). Taking the limits of the above equations as \( \mu \to 0 \):

\[ F_{ST} = \left(1 + \frac{4Nmkr^2(1-p)(k-1+p)}{(1-p)(kr^2(k-2) - (2r-1)) + k(r-1)^2}\right)^{-1}, \]

\[ F_{SR} = \left(1 + \frac{4Nm{k}{(k-1)} + 4Nm{k}{p}}{(k-1)^2}\right)^{-1} \] and

\[ F_{RT} = \left(1 + \frac{r^2(1-p)((k-1)^2 + 4Nm(k(k-1) + 4Nm{k}{p})}{(r-1)^2(k-1+p)}\right)^{-1}. \]

If both \( k \) and \( r \) are large then \( F_{ST} \approx (1+4Nm)^{-1} \) which indicates that the effective migration rate is unchanged. This is to be expected after the results of Section 4.3.2, as the probability of identity once a line-of-descent has left the deme will be small.

Figure 3.6 compares the values (based on Equation 3.27) of the three \( F \) statistics against the proportion of migrants within a region. The proportion of heterozygote deficit within regions (\( F_{SR} \)) is relatively constant as the proportion migrating within and between regions varies, while the values of \( F_{RT} \) and \( F_{ST} \) increase as the proportion of migrants within regions increases. Slatkin and Voelm (1991) used approximations based on the mean times of coalescence which give approximations (using my notation):
$F_{SR} = \left(1 + \frac{4k^2Nmp}{(k-1)^2}\right)^{-1}$ and 

$F_{RT} = \left(1 + \frac{4Nm(1-p)kr^2}{(r-1)^2}\right)^{-1}$ \hspace{1cm} (3.29)

Equations 3.29 are not consistent with Equations 3.28 except for $p$ close to 1 (which are the conditions under which Slatkin and Voelmin's (1991) equations were derived) as shown in Figures 3.7, and 3.8.

For a standard island model with the same number of populations the expected value of $F_{ST}$ is:

$F_{ST} = \left(1 + \frac{4Nm(kr)^2}{(kr-1)^2} + \frac{4N\mu kr}{(kr-1)}\right)^{-1}$ \hspace{1cm} (3.30)

and when the migration rate between all the populations is the same (when $p = (k-1)/(kr-1)$) then we obtain this value. Figure 3.7 shows the overall $F_{ST}$ value compared to that for an island model with the same total number of populations and the same migration rate. This illustrates that the influence of the extra level in the hierarchy is to increase the value of $F_{ST}$ relative to that expected with a single level of population substructuring.

Slatkin's approximation for $F_{ST}$ (from the identity $(1 - F_{ST}) = (1 - F_{SR})(1 - F_{RT})$ and Equation 3.28) is good when most of the migration occurs between demes within a region. The variation shown between regions can still be large even when a large proportion of the total migration is between regions as in Figure 3.6 In this case the inbreeding is mostly within demes for $p<0.75$, and within regions for $p>0.75$. This is particularly pronounced when the number of regions is large and the number of populations within regions are small. Slatkin’s approximation for $F_{SR}$ is also reasonable when the proportion migrating within a region is close to 1 (Figure 3.8) but for less than 1 Slatkin’s approximation is an overestimate, as the value of $F_{SR}$ is almost constant for this large number of populations per region. Figure 3.9 indicates the dependence of $F_{SR}$ on the number of populations within each region. When the number of populations per region is large then $F_{SR}$ is almost constant but for small numbers of populations then $F_{SR}$ decreases as the proportion of migrants within regions increases.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

Figure 3.6 F statistics for a hierarchical island model with parameters $N=100$, $m=0.05$, $k=5$, $r=40$, and $\mu=10^{-7}$. $F_{RT}$ (line with long dashes); $F_{SR}$ (dotted line); $F_{ST}$ (solid line).

Figure 3.7 $F_{ST}$ vs. the proportion of migration within a region for $N=100$, $m=0.05$, $k=20$, $r=20$, and $\mu=10^{-7}$. Solid line $F_{ST}$ for the Island model; dotted line, exact value; long dashed line, Slatkin's approximation.
Figure 3.8  Exact value of $F_{sr}$ and Slatkin's approximation for varying proportion of migrants within a region for $N=100$, $k=20$, $r=20$, $\mu = 10^{-7}$. Exact: $m=0.1$ (solid line) and $m=0.01$ (dotted line). Slatkin's approx.: $m=0.1$ (short dashes) and $m=0.01$ (long dashes).

Figure 3.9 $F_{sr}$ within a region for the parameters $N=100$ $m=0.05$ $r=20$, and $\mu = 10^{-7}$. $K=20$ (solid line), $k=10$ (long dashes), $k=5$ (short dashes), $k=2$ (dotted line).
Using Equations 3.23 and 3.25, and using a Taylor expansion we get expressions for Nei’s D statistics:

\[ D \approx \frac{((k - 1)(kr - 1) - p(k^2r - 2kr + 1))\mu}{(k - 1 + 2p - kp - p^2)m}, \]

\[ D_{RT} \approx \frac{(k - 1)kr\mu}{(k - 1 + p)m} \quad \text{and} \]

\[ D_{SR} \approx \frac{(k - 1 + p(kr - 1))\mu}{(k - 1 + p)m}, \quad (3.31) \]

ignoring terms of higher order in \( \mu \) and \( m \).

Figure 3.10 gives the 3 statistics compared with the Nei’s D statistic for an island model with the same number of populations. There is no local differentiation without there being a reasonable proportion of the migrants within regions, as compared with the F statistics. This is to be expected because Nei’s D statistics only look at the amounts of time spent between demes, so if a large proportion of migration is outside the region then the time spent within regions and between regions will be approximately the same.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

Figure 3.10 Nei's D as a function of the proportion of migrants within each region. Overall D (solid line), $D_{SR}$ (dotted line), $D_{RT}$ (long dashes). Parameters $k=10$, $r=10$, $\mu=10^{-4}$, and $m=0.01$. The line with short dashes is the value for Nei's D found for a single population of 100 sites.

3.7 Population Fission

The previous models in this chapter have assumed that the population has had the same structure over all time. This is a very strong assumption when we know that Atlantic salmon, for example, have recolonised most of their present range over the last 10,000 years after the last ice age (Behnke 1972). The coalescent process was originally developed for these stationary population structures but the process has been used to attempt to model the patterns of variation seen in expanding and contracting populations (Harvey 1994; Nee et al. 1995). An alternative to the island model for the genetic variation seen within and between populations is to assume that there is no continuous migration seen within and between populations but that the populations are all derived from the split of an ancestral population at some point in the past.
Variation in natural populations of animals and humans has been proposed to stem from fissions and fusions of distinct subpopulations. The Yanomama, an Amerindian tribe from Venezuela and Northern Brazil have been studied in detail by Neel, Smouse and others (reviewed by Neel 1978). The primary cause of the genetic similarities between villages has been conjectured to be the fission of ancestral tribal villages to form new distinct villages. This fission model may also be appropriate for some species, where a rise in sea level, or some other catastrophic event, causes the complete reproductive isolation of a population. It is a good null model for the variation seen in natural populations, and a natural alternative to the island model for continuous migration between populations. This approach, when applied to cladistic measures between species and groups of species, is called "vicariance biogeography" (Brundin 1988).

Consider k populations each of N diploids, which all originated by sampling gametes from an ancestral population of \( N_A \) diploids R generations in the past. With allele frequencies there is no way of dating the fission events if there is no information on the relative population sizes; however, with genealogical data this may become possible.

The distribution of coalescence times within and between populations can be written down. Let the distribution within and between populations be \( f_w(t) \) and \( f_b(t) \) respectively. Then

\[
f_w(t) = \frac{1}{2N} e^{\frac{t}{2N}} \quad \text{for } t \leq R \\
= \frac{1}{2N_A} e^{\frac{t}{2N_A}} R \left( \frac{1}{2N_A} - \frac{1}{2N} \right) \quad \text{for } t > R \text{ and}
\]

\[
f_b(t) = 0 \quad \text{for } t \leq R \\
= \frac{1}{2N_A} e^{\frac{(t-R)}{2N_A}} \quad \text{for } t > R.
\]

The moment generating functions for these distributions can be easily calculated to be:
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[ E[e^{Rt_s}] = \frac{1 - e^{-R \left( \frac{1}{2N} - s \right)}}{1 - 2Ns} + \frac{e^{-R \left( \frac{1}{2N} - s \right)}}{1 - 2N_A s} \text{ and} \]

\[ E[e^{Rt_b}] = \frac{e^{Rs}}{1 - 2N_A s}. \]

(3.33)

3.7.1 Distance and Diversity

The mean times until coalescence for alleles sampled from the same and different populations are thus

\[ \bar{t}_w = 2N + (2N_A - 2N)e^{R/2N} \text{ and} \]

\[ \bar{t}_b = R + 2N_A. \]

(3.34)

If we assume that \( N_A = N \) then for a sample from the same population the fission event has no effect on the mean time until coalescence and for alleles sampled from different populations the mean time until coalescence is the sum of the time until fission and the mean time until coalescence in a single population.

We can again calculate \( F_{ST} \) for this model using the method described in Section 3.3.2. With different population sizes before and after fission the expression for \( F_{ST} \) is very complicated. \( F_{ST} \) increases very quickly as we increase the time since fission. Figure 3.11 illustrates the influence of the time since population fission on the values of \( F_{ST} \) seen for two different mutation rates, when \( N = N_A \). With the fission event 200 generations in the past, \( F_{ST} \) is approximately 0.5, which indicates extreme differentiation. This corresponds to the time since fission being equal to twice the effective population size of any subpopulation. The values of \( F_{ST} \) are clearly influenced minimally by the increase in two orders of magnitude of the mutation rate.

Figure 3.12 shows the influence of the population sizes before and after the fission event on \( F_{ST} \). This illustrates that the values of \( F_{ST} \) are increased by small population sizes after the fission event and reduced by small population sizes prior to the fission event (i.e. \( F_{ST} \) increases as \( N \) decreases and as \( N_A \) increases). However, the influence of population sizes prior to the fission event is minimal compared to the large effect of population sizes after the fission event.
Figure 3.11 $F_{ST}$ against the time since fission: solid line $\mu = 10^{-4}$ dotted line $u=10^{-4}$, $N=100$ and $k=10$.

An approximation for $F_{ST}$ can be obtained by letting $N_A = N$ and taking the limit as $\mu \to 0$. This gives

$$F_{ST} = \frac{1}{1 + \frac{2kN}{(k-1)R}} \approx \frac{1}{1 + \frac{2N}{R}} \quad \text{for } k \gg 1. \quad (3.35)$$

Equating this to the value seen for the finite island model with $k$ islands we see that for $k$ large the migration rate in an island model is equivalent to $1/2R$. Thus a migration rate of 0.01 per gamete per generation gives the same heterozygote deficit as a population fission 50 generations ago, independent of the local population size. There is thus a lack of resolution in $F_{ST}$ statistics for the detection of different types of population structuring. Only some combination of statistics might allow us to distinguish between opposing population histories. Further implications of this problem are discussed at the end of this chapter.

When the population sizes are the same before and after the fission event then Nei’s $D$ is simply $2R\mu$, but otherwise
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[ D = \log \left( \exp \left( \frac{-R}{2N} \right) + \exp \left( 2R \mu \right) \frac{1 + 4N_A \mu}{1 + 4N \mu} \left( 1 - \exp \left( -R - 2R \mu \right) \right) \right). \] (3.36)

The influence of the population sizes before and after the population fission is shown in Figure 3.13. The centre line shows a straight line relationship with the time since fission, but when the population sizes differ then there is a large deviation from a straight line relationship. The dependence of Nei’s D on the differences in population sizes before and after fission are shown in Figure 3.14. Increasing population size after fission leads to decreases in Nei’s D, and increases in population size after fission lead to increases in D. This statistic is far more sensitive to changes in these parameters than is \( F_{ST} \). In this case it should be possible to detect the effects of large changes in population size by the combination of \( F_{ST} \) and Nei’s D, given information about current population sizes, and an estimate of mutation rates as \( F_{ST} \) should enable the estimation of the population size after fission, and Nei’s D can then determine the relative size of the population before the fission event.

![Figure 3.12 FST as a function of population sizes after fission for a model with R = 200 k=10 u=10^-6 and for N_A = 10, 100, 1000, and 10,000. The solid line is N_A =10, and line with short dashes N_A =10,000.](image)
Figure 3.13 Nei's D against the time since fission. Mutation rate $10^{-4}$. $N=10$ and $N_A=1000$ (solid line); $N=N_A=1000$ (long dashes); and $N=1000$ and $N_A=10$ (dotted line).

Figure 3.14 Nei's D against the population size after fission. The time since fission $R = 200$, $\mu=10^{-4}$. Solid line $N_A=10$, lines above this $N_A=100$, 1000, 10000, respectively.
3.8 Extinction and Recolonisation

3.8.1 Island model with no migration

The meta-population model of population structure (Gilpin & Hanski, 1989, Hastings & Harrison, 1994, Hanski & Gilpin, 1995) contends that the main defining events in the history of a “meta-population” (a population in the usage of the rest of the chapter) are the extinction of local populations followed by recolonisation from other subpopulations. The structuring of populations in this way is thought to have a major effect on their evolution by some (Wright, 1931), and the evidence for this is reviewed in Barton and Whitlock (1996). This section shows how the genetic variation of a sample from such a population may be modelled using a coalescent approach. Section 3.4 of this chapter showed how variation in population size could be modelled using the geometric mean in population sizes, or alternatively by varying the time scaling (Hudson, 1990). An extinction event when viewed backwards in time is equivalent to the forced migration of a set of individuals out of a subpopulation, with, or without, an extra possibility of coalescence due to the bottleneck. The maintenance of genetic variability, and effective population size of this type of population structure have been considered previously (Maruyama & Kimura, 1980; Slatkin, 1977; Whitlock & Barton, 1996), but in this chapter the coalescent approach will be used to examine the properties of a sample from such a population.

Consider a group of $k$ subpopulations each of $N$ diploid individuals and with a probability $\gamma$ of becoming extinct in each generation. If extinction occurs then the population is recolonised from one or more of the other $k-1$ populations. The rate of population growth after the colonisation event and the distribution of subpopulations from which the colonisers arrive determine what happens to extant lines. If the individuals come from the other $k-1$ populations at random without replacement then this is equivalent to distributing the extant lines at random to the other subpopulations. Alternatively, if the population growth is small, or the recolonisation occurs from only one individual then the extant lines coalesce, and this line 'migrates' to one other population. For intermediate growth, and recolonisation events there is some intermediate distribution of instantaneous coalescence and distribution of the extant lines. These processes can be modelled by assuming that proportions $\rho$ and $1-\rho$ of the colonisers come from the same and
from different populations respectively, and that the probability of a coalescent event caused by inbreeding after the extinction event is given by $\lambda$. For example if after a coalescent event the population is recolonised by 2 diploid individuals, which are always from different populations, and the population grows to its carrying capacity $N$ in a single generation, then $\rho = 0$ and $\lambda = 1/4$.

Whitlock and Barton (1996) use $\phi$ as a measure of the relatedness of colonisers for a model of the entire meta-population which is analogous to $\rho$ here. In this section we are concerned with developing a coalescent model for genetic variation in small samples from a population rather than for the entire population. Here it is assumed that extinction events are independent between generations and between populations, so that the distribution of time until the next extinction event in any population is exponential.

The moment generating functions for the times until coalescence for genes sampled from the same and different populations can be calculated (taking events as extinctions or coalescences) in a similar way to those of Section 3.5.1:

$$E[e^{sr}] = E[e^{sr}/\text{first event extinction}]Pr(\text{extinction})$$
$$+ E[e^{sr}/\text{first event coalescence}]Pr(\text{coalescence}).$$ (3.37)

The rates of coalescence, movement to different demes, and movement to the same deme for genes in the same deme are $\gamma(1-\lambda)\rho$ and $\gamma(1-\lambda)(1-\rho)$ respectively, so that the time until the first event is exponential with rate $1/2N + \gamma$. For genes in different demes the time until coalescence is the sum of time until genes are in the same deme, which is exponential with rate $\gamma/(k-1)$, and the time until coalescence once in the same deme. So, we need to solve

$$E[e^{sr}] = \frac{2N\gamma(1-\lambda)\rho E[e^{sr}] + (1-\rho)E[e^{sr}]}{2N\gamma + 1 - 2Ns} + 1$$
and

$$E[e^{sr}] = \frac{2\gamma E[e^{sr}]}{2\gamma - (k-1)s}. $$ These give (3.38)
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

\[ E[e^{st_2}] = \frac{(1 + 2\gamma \lambda N)(-2\gamma - s + ks)}{-2\gamma - 4\gamma^2 \lambda N - s + ks + 2\gamma Ns + 2\gamma kN \rho Ns - 2\gamma kN \rho s} \quad \text{and} \]

\[ E[e^{st_3}] = \frac{2\gamma(-1 - 2\gamma \lambda N)}{-2\gamma - 4\gamma^2 \lambda N - s + ks + 2\gamma Ns + 2\gamma kN \rho Ns + 2\gamma kN \rho s} \quad (3.39) \]

Now, the moment generating function for times between populations is of the form \(1/(1+as)\), therefore the distribution of pairwise times between populations is exponential. The distribution of times for genes sampled from within a single population is more complicated.

3.8.1.2 Identity and Distance Measures

\( F_{ST} \) and Nei’s D are calculated as in Section 3.3.2. This gives

\[ F_{ST} = \left(1 + \frac{2kN\gamma[1+k+(k-1)(\lambda \rho - \lambda - \rho)]}{(k-1)(1+2\gamma \lambda N)} + \frac{4kN\mu}{(k-1)(1+2\gamma \lambda N)} \right)^{-1} \quad \text{and} \quad (3.40) \]

\[ D = -\log \left(1 + \frac{\mu (k-1)}{\gamma} \right). \quad (3.41) \]

As before Nei’s D is influenced solely by the time spent before the genes are in the same deme whereas the F statistic depends on the behaviour within each subpopulation.

If, as for the island model, we assume that the mutation rate is small, and that the number of populations large then \( F_{ST} \approx \left(1 + \frac{2N\gamma (1-\lambda)(1-\rho)}{1+2N\gamma \lambda} \right)^{-1}. \)

3.8.1.3 Mean Coalescence Times

The mean times until coalescence are calculated as before, by differentiating the moment generating function for times within and between by s, and putting s=0. Thus we get:

\[ \bar{t}_0 = \frac{N[2+(k-1)(1-\lambda)(1-\rho)]}{1+2\gamma \lambda N} \quad \text{and} \quad \bar{t}_i = \bar{t}_0 + \frac{(k-1)}{\gamma}. \]
3.8.1.4 Effective Population Size

The effective population size is calculated as before, by calculating the mean time until coalescence for a pair of alleles sampled at random from the population.

\[ N_E = \frac{N \left[ 2 + (k - 1)(1 - \lambda)(1 - \rho) \right]}{1 + 2\gamma AN} + \frac{(k - 1)^2}{k\gamma}. \] (3.42)

It is clear from this that increasing \( \gamma, \lambda, \) or \( \rho \) decreases the coalescent effective population size. Maruyama and Kimura (1980) considered the effective population size of haploid lines subject to random extinction. In this case the factor of 2 in Equation 3.42 should be replaced by unity, and \( \rho \) and \( \lambda \) are equal to 1 giving

\[ N_E = N + \frac{(k - 1)^2}{k\gamma} \]

which is approximately the afore-mentioned author’s result.

3.8.1.5 Examples

The equations above are complicated expressions. As examples consider three extreme kinds of recolonisation.

**Type 1:** When a population becomes extinct all the individuals in that population in the next generation are drawn at random with replacement from the individuals present in the generation at which the population became extinct. This is equivalent to \( \lambda = 0 \) and \( \rho = 1/k \).

**Type 2:** After an extinction all the individuals are drawn at random with replacement from the alleles present in a single population. In this case \( \lambda = 0 \) and \( \rho = 1 \).

**Type 3:** After an extinction the population is recolonised from a single individual drawn at random from the population for which \( \lambda = 1 \) and \( \rho = 1 \). This is only possible for haploid organisms.

Obviously all these recolonisation scenarios are simplifications of what happens in a real population where there is sampling from a subset of the population, and limited inbreeding occurs during population growth. However, they give an idea of the range of behaviours that can occur. The mean times until coalescence between and within demes are given for these 3 types in Table 3.2, along with both the exact
value of $F_{ST}$, and Slatkin's approximation using the mean coalescence times. Figure 3.15 shows the effect of the recolonisation method on the degree of inbreeding seen.

### 3.8.1.6 Testing the models forward in time

In order to test these models, simple discrete generation time simulations were run forward in time for the three different models. These simulations were based on binomial sampling from the gene frequency of the previous generation (after the expected change due to mutation), unless there was an extinction event. In this case the frequency was taken from: type 1) the rest of the population; Type 2) the frequency in a single population drawn at random; or type 3) 0, or 1 at random proportion to the population allele frequencies. The fit to the model was good for the three models here. Values for $F_{ST}$ were calculated using Equation 2.15. Figure 3.16 indicates the fit of the coalescent data to a classical model for extinction/recolonisation model of type 1 forward in time.
Chapter 3: THE COALESCENT IN SUBDIVIDED POPULATIONS

Table 3.2: The mean times until coalescence, and F statistics for 3 models of extinction/recolonisation.

<table>
<thead>
<tr>
<th>E/ type</th>
<th>$\bar{t}_w$</th>
<th>$\bar{t}_b$</th>
<th>$F_{ST}$ from m.g.f.'s</th>
<th>$F_{ST}$ from mean coalescence times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$kN$</td>
<td>$kN + \frac{k-1}{2\gamma}$</td>
<td>$\frac{1}{1 + \frac{2N\gamma k^2}{(k-1)^2} + \frac{4N\mu k}{(k-1)}}$</td>
<td>$\frac{1}{1 + \frac{2N\gamma k^2}{(k-1)^2}}$</td>
</tr>
<tr>
<td>2</td>
<td>$2N$</td>
<td>$2N + \frac{k-1}{2\gamma}$</td>
<td>$\frac{1}{1 + \frac{4N\gamma k}{(k-1)^2} + \frac{4N\mu k}{(k-1)}}$</td>
<td>$\frac{1}{1 + \frac{4N\gamma k}{(k-1)^2}}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{2N}{2N\gamma + 1}$</td>
<td>$\frac{k-1 + 2N\gamma (1+k)}{2\gamma (1+2N\gamma)}$</td>
<td>$\frac{1}{1 + \frac{4N\gamma k + 4N\mu k(k-1)}{(1+2N\gamma)(k-1)^2}}$</td>
<td>$\frac{1}{1 + \frac{4N\gamma k}{(1+2N\gamma)(k-1)^2}}$</td>
</tr>
</tbody>
</table>
Figure 3.15 $F_{ST}$ for three extinction recolonisation models: solid line type 1; dashed line type 2; and dotted line type 3. All models for $N=100$, $k=100$, and $\mu=10^{-7}$.

Figure 3.16 Fit of coalescent models for $F_{ST}$ compared to results for a classical model forward in time. Line indicates theory, simulations indicated ± 2s.e. Simulations for 400 replicates for 100 populations with 40 diploids and $\mu = 10^{-4}$. 
3.8.2 Extinction/Recolonisation Model with Migration

Extinctions and recolonisations may be only one process in the population: migration between extant populations may also be important. To investigate this a model incorporating both extinction/recolonisation and migration is briefly considered.

Again, consider k demes each with N diploid individuals with a probability of extinction of the populations of \( \gamma \) per generation for each deme; proportions \( \rho \) and \( 1 - \rho \) of the colonisers come from the same and from different populations respectively; the probability of a coalescent event caused by inbreeding after the extinction event given by \( \lambda \); and a dispersal probability of \( m \) per gamete per generation.

To model this behaviour consider the possible events: an extinction; a coalescent event; and a migration event which happen with relative probabilities \( \gamma \), \( 1/2N \), and \( 2m \) for pairs of genes in the same population, and with relative probabilities \( 2\gamma \) and \( 2m \) for genes in different populations. Conditioning on the time of the first event leads to equations for the moment generating functions:

\[
E[e^{\lambda T}] = \frac{2N\gamma (1-\lambda)\rho E[e^{\lambda T}] + [4Nm + 2N\gamma (1-\rho)]E[e^{\lambda T}] + 2N\gamma \lambda + 1}{4Nm + 2N\gamma + 1 - 2Ns} \quad \text{and} \\
E[e^{\gamma T}] = \frac{2(m+\gamma)E[e^{\gamma T}]}{2(m+\gamma) - (k-1)s}.
\]

Solving the above equations and using the results in Section 3.32 we obtain the following expressions for \( F_{ST} \) and Nei’s D:

\[
F_{ST} = \left(1 + \frac{2Nk\gamma (1 + k + (k-1)(\lambda \rho - \lambda - \rho)}{G} + \frac{4k^2Nm}{G} + \frac{4Nk(k-1)\mu}{G}\right)^{-1},
\]  

where

\[
G = (k-1)^2 + 2N\gamma k(k+1) - 2N\gamma \lambda (k-1) - 2N\gamma \rho k(k-1) + 2N\gamma \lambda \rho k(k-1) + 4N\mu k(k-1)
\]

and

\[
D = -\log \left( \frac{\gamma + m}{\gamma + m + \mu(k-1)} \right).
\]
For $F_{ST}$ all the evolutionary parameters enter the equation in the same way and there is no interaction except between the extinction/recolonisation parameters. The combination of parameters combines to give $M$, the effective migration rate (Slatkin 1991). The number of parameters which affect Nei’s $D$ are small: just the extinction rate, the migration rate, the mutation rate, and the number of populations. There is no way, using these pairwise measures, of separating the effects of the different evolutionary forces which could be the cause of the population subdivision. This is shown further in the next section where we investigate the interaction between a fission event and migration.

The mean times until coalescence between and within populations and the effective population size for this type of model can be calculated in the same way as in the previous section. This gives:

$$\bar{t}_0 = \frac{2Nkrn + Ny(2 + (k - 1)(1 - \lambda)(1 - \rho))}{(\gamma + m)(1 + 2\lambda N)}; \quad \bar{t}_1 = \bar{t}_0 + \frac{k - 1}{2(\gamma + m)}$$

and

$$N_E = \frac{2Nkrm + Ny(2 + (k - 1)(1 - \lambda)(1 - \rho))}{(\gamma + m)(1 + 2\lambda N)} + \frac{(k - 1)^2}{2k(\gamma + m)}.$$

When $\gamma = 0$, these give the same value as for the island model, and when $m = 0$, we get the answers of the previous section.

### 3.9 The Island-Fission Model

As shown in Section 3.7 the fission model produces a great deal of genetic variation if there is no migration between populations. No natural populations have lived in their present locations for an infinite amount of time; for example the Atlantic salmon has recolonised Europe since the last ice age, about 10,000 years ago. This means that the degree of genetic divergence and its distribution are due to several factors. Using the distribution of gene frequencies within an island model we can investigate the relative effects of population fission/recolonisation and subsequent dispersal on the patterns of genetic divergence between populations.

The distribution of times until coalescence for an island model was calculated in Section 3.5.2. This allows us to calculate the distribution of coalescence time when
there is a population fission event in the past by conditioning on whether the pairs of alleles have coalesced prior to that time. Assume an island model with \( k \) populations of \( N \) diploids and a migration rate of \( m \) which is produced by the fission of a single population of \( N_A \) diploids \( R \) generations in the past. Denoting the cumulative distribution functions for the time of coalescence for pairs of genes sampled within and between populations for an island model by \( F_W(t) \) and \( F_B(t) \) (the probability density functions are given in Equation 4.22), then the distribution of coalescence times within \( f_{FW}(t) \), and between \( f_{FB}(t) \), can be written down:

\[
f_{FW}(t) = \begin{cases} 
F_W(t) & \text{for } t \leq R \\
\frac{(1-F_W(R))}{2N_A} e^{-\frac{t-R}{2N_A}} & \text{for } t > R
\end{cases}
\]

\[
f_{FB}(t) = \begin{cases} 
F_B(t) & \text{for } t \leq R \\
\frac{(1-F_B(R))}{2N_A} e^{-\frac{t-R}{2N_A}} & \text{for } t > R
\end{cases}
\]

where prime denotes taking the differential with respect to \( t \). Figure 3.17 illustrates the distributions of coalescence time. The peak in probability is the point at which the fission event occurs, so producing a peak in the probability of coalescence. The mean times until coalescence, and the moment generating functions for samples from within and between populations can be calculated (but are extremely complicated expressions) Figure 3.18 illustrates the influence of population fission on the degree of differentiation seen in populations. If the migration rate is high then there is little effect of the fission event on expected values of \( F_{ST} \). However, with lower migration rates the effect is significant. This is a similar result to that of Latter (1977) who used the probability of identity directly.
Figure 3.17 The distribution of coalescence times for the island fission model. Solid lines within populations, dotted line between populations. $N=100, m=0.005, k=10, R=4000$, and $N_A = 100$.

Figure 3.18 $F_{ST}$ for an island model after a population fission event. The model is for $N=100; N_a=100; k=20; and m = 0.1, (solid line) m=0.01, (long dashed line) and m=0.001 (dotted line). A pure fission model with $m=0$ is given by the line with short dashes.
3.10 Isolation by Distance

The island and hierarchical island models previously derived give expressions for the amount of inbreeding within populations when there is no geographical information. Slatkin and Barton (1989), using simulations, showed that the values of $F_{ST}$ observed from two-dimensional habitats gave good estimates for the effective migration rate ($Nm$). Section 3.3.2 showed why this can be true when the probability of identity for two genes which are not in the same population is much less than that within a population. However, the relative geographical distances between populations may give us additional information about the populations especially as modern computers allow us to perform far more involved statistical calculations.

A brief description of the major results for the identity by descent was given in the last chapter. Wright (1951) used the method of path coefficients to get expressions for the amount of genetic differentiation seen in continuous two dimensional populations. Malecot (1948) used the probability of identity by descent to analyse similar models. There are some theoretical difficulties with these models, because without local density regulation the equations lead to clumping of individuals (Felsenstein 1975a). Barton and Wilson (1995, 1996) extended Wright’s work on identity by descent in two dimensions to include the distribution of coalescence times for this type of model.

Stepping stone models for population differentiation give fewer theoretical problems for identity by descent. Kimura and Weiss (1964), and Weiss and Kimura (1965) obtain expressions in one, two and three dimensions. Slatkin (1991) obtained expressions for the mean time until coalescence for genes sampled from a circular habitat. Herbots (1995) obtained expressions for the moment generating functions of coalescence times for a variety of isolation by distance models for stepping stones. When there are infinite habitats, or habitats in a circular arrangement, symmetry allows the removal of large numbers of parameters. Unfortunately, natural habitats are rarely so simple. While solutions for the moment generating functions of the pairwise coalescence times are possible for any set of migration rates and population sizes (see, for example, Herbots (1995)) these are seldom simple to calculate explicitly, and while numerical solutions are easier, as shown in a later section for a special case, simulation is often the best way forward. In this section some
expressions for the identity by descent for both infinite and finite one-dimensional habitats will be derived, and simulation techniques will be used to test these results. Similar results for two-dimensional habitats have been derived elsewhere (Barton and Wilson 1995, 1996). The habitats of Atlantic salmon can be described as one-dimensional, or at least as the union of one-dimensional habitats at different scales (linear tributaries joined to create rivers, and rivers along the coast). A one dimensional habitat would appear to be a better approximation for these structures.

3.10.1 One Dimensional Habitats

In Chapter 3 the decrease of probability of identity with distance was discussed for one dimensional habitats, and a model for a branching structure was developed. Using equations for the probability of identity we can get an expression for the moment generating function of the distribution of coalescence times (since the probability of identity is simply the moment generating function of the coalescence times), and hence expressions for the mean time until coalescence in one-dimensional habitats. There are two ways to get approximate distributions for coalescence times in a one-dimensional population:

1. Use Wright's (1951) neighbourhood size argument to get an approximate distribution for the times; as shown in Barton and Wilson (1995). This is a more heuristic method.

2. Or use the continuous approximation for the probability of identity by descent for a one dimensional population and use an inverse Laplace transform to obtain the distribution. This is not really feasible for a two dimensional population but for one dimension it is possible. A reference book containing extensive Laplace transforms such as Gradshteyn and Ryzhik (1980), or a symbolic mathematical computer program such as Mathematica (Wolfram 1991) will give this.

Either method should give a reasonable approximation for the distribution of times for small time, but the Laplace transform is probably more robust because it is based on an equation rather than a summation. These distributions are approximate. From random walk results it can be guaranteed than any two genes on the line will coalesce with probability 1, but that the mean time until coalescence will be infinite.
We can show the latter result by differentiating (Equation 2.40) with respect to $-m/2$ and letting $\mu \to 0$.

### 3.10.1.1 The Laplace Transform of Malecot's equation

Using Equation 2.40, which was discussed at length in Chapter 2, and taking the inverse Laplace transform (Gradshteyn & Ryzhik 1980), we get the distribution

$$f(t) = \frac{\exp\left(-\frac{d^2}{4mt}\right)}{2(2N-1)\sqrt{m\pi t}} - \frac{\exp\left(d\left(\frac{t}{2(2N-1)m} + \frac{t}{4(2N-1)m}\right)\right) \text{erfc}\left(\frac{\sqrt{t}}{2(2N-1)\sqrt{m} + \frac{d}{2\sqrt{mt}}}\right)}{4(2N-1)^2 m}$$

(3.47)

where \text{erfc} is the complementary error function (Abramowitz and Stegun 1965).

### 3.10.1.2 Wright's Method of Neighbourhood Size

For the method, originally due to Wright (1943), we consider Equation 12b in Barton and Wilson (1995). The argument for this is based on Wright's (1943) derivation for the correlation in gene frequencies in different locations. The idea behind this is that as we go into the past the size of the pool of ancestors from which the gene may be descended increases. This means that as we go back 1 generation the probability that two genes in the same location are identical by descent is $1/(2N_b)$ (by definition, Wright's neighbourhood size $N_b$ in a one-dimensional habitat is $N_b = 2N\sqrt{\pi m}$, and $4\pi Nm$ in a two dimensional habitat). A normal approximation to the migration distance (rather than just being -1 and +1 with probability $m/2$) will be used, which is justified by diffusion arguments.

For a linear habitat the pool of ancestors is drawn from the real line. As we go back in time the standard deviation of the distance from the start point increases with the square root of time and so the size of the pool of ancestors is $2N_b\sqrt{t}$. We can use this argument to get a distribution for the times of coalescence. First consider a pair of genes at the same location. Denote by $c_j$ the probability of coalescence in the $j$-th generation. The probability of coalescence in the first generation is then just $\frac{1}{2N_b}$. The probability of coalescence in the second generation can then be calculated by using:
\[
\frac{1}{2N_b \sqrt{2}} = \frac{1}{2N_b} c_0(1) + c_0(2). \tag{3.48}
\]

Extending this over more generations we then get:

\[
\frac{1}{2N_b \sqrt{i}} = c_0(i) + \sum_{j=1}^{i-1} \frac{c_0(i-j)}{2N_b \sqrt{j}}. \tag{3.49}
\]

Hence we can get the result

\[
c_0(i) = \frac{1}{2N_b} \left( \frac{1}{\sqrt{i}} - \sum_{j=1}^{i-1} \frac{c_0(i-j)}{\sqrt{j}} \right). \tag{3.50}
\]

This can then be extended to genes a distance \(d\) by noting that the probability that a pair of genes a distance \(d\) apart initially are in the same region after \(t\) generations is simply \(\exp\left(-\frac{d^2}{4t}\right)\) so that we may extend the original equation simply to get

\[
c_d(i) = \frac{1}{2N_b} \left[ \frac{1}{\sqrt{i}} \exp\left(-\frac{d^2}{4i}\right) - \sum_{j=1}^{i-1} \frac{c_0(i-j)}{\sqrt{i-j}} \exp\left(-\frac{d^2}{4(i-j)}\right) \right]. \tag{3.51}
\]

Calculating the distribution of coalescence times using the above equations is then possible numerically for several hundred generations on computer, after which rounding errors become significant.

Figure 3.19 gives the cumulative distribution of coalescence times (using both Equations 3.47, and 3.51). The two expressions are almost identical when \(d > 0\) but when \(d = 0\) there is some difference. A simulation study was performed to test the two methods against simulated data, the simulation program is described in Appendix 4. Figure 3.20 illustrates the results from the simulation against the results due to Wright’s method, which gives a very good approximation. From Figures 3.19 and 3.20 it is clear that Wright’s method gives a better approximation for \(d=0\). Figure 3.21 illustrates the distribution function using method 2 for a variety of initial distances apart.
Figure 3.19 Cumulative distribution of coalescence times using method 1 solid lines; method 2 - shapes for pairs of genes separated by distances 0 - squares; 2- circles; 5- triangles; and 10- diamonds for a linear habitat with population size 10 diploids, and a migration rate of 0.1 per gamete per generation. The solid lines are eclipsed for distances >0.

Figure 3.20 The cumulative distribution of coalescence times (using method 2) solid lines with estimated cumulative frequency from simulations with 5000 replicates for pairs of genes separated by distances 0 - squares; 2- circles; 5- triangles; and 10- diamonds for a linear habitat (1000 demes) with population size 10 diploids, and a migration rate of 0.1.
0.01

0.006

0

Figure 3.21 The distribution of coalescence times (using method 2) for pairs of genes separated by distances 0 - solid line; 2- long dashed line; 5- short dashed line; and 10- dotted line units for a linear habitat with population size 10 diploids, and a migration rate of 0.1 per gamete per generation.

3.10.2 Finite Linear Habitats

The results presented above are for infinite linear habitats. When we have a finite number of habitats a theory is needed for the finite line for comparative purposes. The times until coalescence in circular linear habitats have been previously investigated (Slatkin 1991), and the identity by descent in finite linear habitats, and circular habitats have been examined previously (Maruyama 1969, 1970, 1972).

Consider a finite linear habitat as in Figure 3.22, with k demes each with N diploids and a migration rate of $\frac{m}{2}$ between neighbouring demes. For a habitat of length $k$ there are a total of $2\sum_{i=1}^{k/2} i$ simultaneous equations to solve when $k$ is even. Letting $g_{ij} = E[e^{\lambda_{ij}}]$; the following simultaneous equations need to be solved:
\[ g_{1,1} = \frac{1}{1 + 2N_m - 2N_S} + \frac{N_m}{1 + 2N_m - 2N_S} \{2g_{2,1}\}, \]
\[ g_{i,i} = \frac{m}{3m - 2s} \{g_{i,2} + g_{i-1,1} + g_{i+1,1}\} \text{ for } 1 < i < k, \]
\[ g_{i,k} = \frac{m}{2(m - s)} \{g_{i,2} + g_{k-1,k}\}, \]
\[ g_{i,d} = \frac{1}{1 + 4N_m - 2N_S} + \frac{N_m}{1 + 4N_m - 2N_S} \{2g_{i+1,i} + 2g_{i-1,i}\} \text{ for } 1 < i < k, \]
\[ g_{i,j} = \frac{m}{2(2m - s)} \{g_{i-1,j} + g_{i+1,j} + g_{i,j-1} + g_{i,j+1}\} \text{ for } 1 < i, j < k, \]
\[ g_{i,j} = g_{j,i} \text{ and } \]
\[ g_{i,j} = g_{k+1-j,k+1-i} \text{ for } j > i, j > k + 1 - i. \]

Solving these equations gives expressions for:

1. The identity by descent by letting \( s = -2\mu \).

2. The mean and second moment, by differentiating by \( s \) and setting \( s = 0 \) and by differentiating by \( s \) twice and setting \( s = 0 \) respectively.

This allows us to compare the behaviour for this finite line to that expected for an infinite linear habitat and gives a check on simulations which are needed for more complicated models. The standard simulation in one dimension is based on Equations 3.52 for pairwise models. The rate until the next event (coalescence or migration) is calculated (this is simply the sum of all the individual rates for all possible coalescences or migrations), and then the time of the event is generated from an exponential distribution. This is carried on, moving lines of descent between demes as necessary, until all the lines have coalesced. Any mutations can then be generated. Details of the simulation program are given in Appendix 4.

### 3.11 Range Expansion

A feature of most species which now live in temperate climates is that at some point in the last 20,000 years, and indeed several times in the last million years, they have recolonised at least part of their range because of the retreat of the ice caps after the last Ice age. For Atlantic salmon which have a northern limit to their range in hostile conditions in Iceland, Norway, Russia and Greenland then all of these
territories must have been recolonised since the retreat of the ice approximately 10,000 years ago as they were under a vast ice cap (as was most of Scotland). Glacial refugia are thought to have been in Spain for the European salmon and also in North America (Behnke 1972). Long distance dispersal of invading organisms has been modelled previously (Nichols & Hewitt 1994), which indicated that a leptokurtic distribution of migration distances gives different patterns of genetic variation to that of Gaussian dispersal. In this section a different situation is examined; where the recolonisation is constrained by the climatic conditions which change over an (ecologically) long time so that the times of recolonisation may be regarded as predictable. This is more applicable to species such as the Atlantic salmon which live in hostile conditions (under river ice) and so would presumably have recolonised as the ice retreated.

This is similar to a model (forward in time) for the spread of alleles due to demic diffusion, a process of local expansion of populations, followed by dispersal (Sokal & Menozzi 1982). This model is equivalent to a spatial version of stages 2 and 3 of Wright's shifting balance (Wright 1978). This has been used to explain the patterns of genetic variation of European, and Near-Eastern humans (Sokal & Menozzi 1981, Sokal et al. 1991). The patterns of genetic variation have been shown to be concordant with the spread of agriculture, the timing of which is well known from archaeology.

Section 3.9 on population fission shows how this might affect pairwise measures of genetic differentiation, but the patterns when there is a spatial element to this fission are not so clear. In this section a simple model for this sort of recolonisation will be developed, and the differentiation seen in this model compared to models just involving isolation by distance, which have been developed in the previous section.

### 3.11.1 Simple Model

The Atlantic salmon breeds in rivers along linear coastlines so a one-dimensional model will be sufficient to specify position. Consider a linear array of $k$ demes which are colonised in series as shown in Figure 3.23. Assume that the colonising individuals are selected at random from the previous deme.
Figure 3.23 Model for recolonisation.

The simplest possible model of this sort is to assume that by time \(-R\) in the past all demes have been recolonised, and at time \(-2R\) in the past there were \(k-1\) demes and so on until at time \(-(k-1)R\) there was only a single deme. Also assume that there have been no other dispersal events other than the range expansion, and that every deme has an effective population size of \(N\) randomly mating diploids with non-overlapping generations. Number demes from 1 (the original deme), to \(k\) (the most recently colonised deme).

The distribution for the time until coalescence for genes sampled from the \(i\)-th and \(j\)-th demes can then be calculated by conditioning on the time when both genes are in the same deme to give:

\[
\begin{align*}
  f_{ii}(t) &= \frac{1}{2N} e^{-\frac{t}{2N}} \quad \text{for } i = j \text{ and} \\
  f_{ij}(t) &= \begin{cases} 
    0 & \text{for } t < (k - \min(i, j))R \\
    \frac{1}{2N} e^{-\frac{t-(k-\min(i, j))R}{2N}} & \text{otherwise}
  \end{cases} \quad \text{when } i \neq j. 
\end{align*}
\]  

Hence one can derive the moment generating functions for the time until coalescence for genes taken from demes \(i\) and \(j\):

\[
\begin{align*}
  E[e^{st}] &= \frac{1}{1 - 2Ns} \quad \text{and} \\
  E[e^{st}] &= e^{(k-\min(i, j))R s} \quad \text{when } i \neq j. 
\end{align*}
\]
The give pairwise $F_{ST}$ values and Nei’s D measures shown in Figure 3.24. This indicates that most differentiation between populations should occur at the end of the species range which was closest to the deme from which the recolonisation started. For Nei’s D there is a linear relationship with distance for distances measured from the most recently colonised population, whereas with pairwise $F_{ST}$ values the relationship is not linear. This gives us something to look for when we believe that the species has undergone a range expansion event. Pairwise $F_{ST}$ values, or Nei’s D values should increase with distance for pairwise measures measured from the most recently colonised populations. Alternatively $F_{ST}$ values for several populations should increase with greater geographic range from the most recently colonised populations, but this will not happen in the same way for the oldest populations.

![Figure 3.24 Nei’s D and pairwise $F_{ST}$ values for the range-expansion model. 20 populations, each of 100 diploid individuals, with $\mu=10^{-4}$. $R=100$: solid line distance from the most recent, dashed line distance from the oldest population. $R=1000$: long dashed line distance from the most recent, dotted line distance from the oldest population.](image)

This is obviously a very crude model. The simplest development is to assume that the recolonisation events occurred at some point in the past and since then there has been no dispersal between populations. Assume that the first recolonisation event occurred now at $V+R$ generations in the past, and the subsequent events at $V+2R$, $V+3R$, ..., $V+(k-1)R$. Then the moment generating functions of the time until coalescence for genes taken from demes $i$ and $j$ are given by:
\[ E[e^{\sigma_T}] = \frac{1}{1-2Ns} \quad \text{and} \]
\[ E[e^{\sigma_D}] = e^{\frac{(k-\min(I,J))(R+\nu)s}{1-2Ns}}. \quad (3.55) \]

This alters the properties of Nei’s D and pairwise F\(_{ST}\) values. Figure 3.25 illustrates the effect of the increase on pairwise F\(_{ST}\) and D values. Again the greatest differentiation is seen between populations where the glacial refugia is closest but the differences between the populations are now all much greater which erodes the differences seen for the first model. The gap after the last recolonisation event causes a jump in the pairwise distance between the last two demes in the array, which masks the differences between the distances measured from the two ends.

![Figure 3.25 Nei's D and pairwise F\(_{ST}\) values for a recolonisation model with R=100, k=20, \(\mu = 10^{-4}\), V=100: Solid line distances from the oldest population, dashed line distances from the most recent population. V=1000: line with long dashes distance from the oldest population, dotted line distance from the youngest population.](image)

### 3.11.2 Range expansion with subsequent dispersal

When there is dispersal after a range expansion, which is a more likely evolutionary scenario, then the distributions of time until coalescence become much more
complicated. While it is possible to get a distribution for times until coalescence for a one dimensional habitat it is not trivial except when the number of demes is infinite. As an approximation, if the rates of dispersal are high and the time until the recolonisation is long so that the actual geographic structure is unimportant, the distribution of times for an island fission model will be reasonable.

If this is not the case then the best way to proceed here is to use simulation. In all subsequent simulations pairwise F_{ST} values will be used based on the mean coalescence times (Slatkin 1991) given in Section 3.5.1.1 as this gives more accurate values for F_{ST} with the same number of replicate simulations. The details of the model are as in the last section. When a deme has been recolonised (looking forward in time) then the simulation moves any lines of descent in that deme to the neighbouring deme. The simulation was checked by using the mean coalescence times when M=0.0, and when S is very large (so that all the genes have coalesced due to migration) which are known analytically (from Equations 3.52 and 3.55).

Equations 3.52 were solved numerically using a modified version of routine gaussj from Press et al. (1992) (a numerical Gauss-Jordan elimination routine), and the expected values for pairwise F_{ST} calculated using 
\[
\frac{(f_0 - f_d)}{2 - f_0 - f_d}.
\]

Figure 3.26 illustrates the pairwise F_{ST} values for an increasing time until range expansion with the theoretical expressions for no migration, and a simple 1-dimensional model. The isolation by distance effect caused by migration is depressed for all the simulations, except when the time since range expansion is very large. Only when the time since range expansion is small are we able to see the effect caused by the range expansion, shown in the last section. By examination, the crucial time scale determining the relative effects of migration and range expansion seems to be \( S \approx 1/m \). This indicates that unless gene flow is very low then patterns due to gene flow will dominate the effect of range expansion pairwise in measures of pairwise F_{ST}. 
Figure 3.26 The relationship between pairwise $F_{ST}$ values and the distance between samples for a recolonisation model with subsequent gene flow. All Figures $k=6$, $\mu=10^{-6}$, $m=0.001$, $N=1000$, and $R=100$ generations. Long dashed line the expected pairwise $F_{ST}$ values from the first population, solid line expected pairwise $F_{ST}$ values from the last population, short dashed line the expected isolation by distance given no recolonisation. The simulated pairwise values (squares from the last triangles from the first) are based on 10,000 replicates. Figure a) $S=0$, b) $S=1000$, c) $S=10,000$, and d) $S=100,000$. 
3.12 Discussion

The coalescent process is an extremely useful technique for investigating patterns of genetic variation under many models of population structure. Using an approximation based on a small sample size compared to the size of the population also allows a different perspective on observations of patterns of variation, as being the results of a few coalescent events (the total sample size-1). In this chapter previous work on the coalescent has been reviewed, and the distribution of pairwise coalescence times in an island model has been calculated. This chapter has also developed novel techniques for using the coalescent to investigate other processes: the hierarchical island model; extinction/recolonisation; fission; and range expansion for pairwise samples. All of the above models are suitable for larger sample sizes.

The chapter has dealt mostly with pairwise measures and probability distributions, which have shown a lack of discrimination between different kinds of population subdivision (for example the difference between migration and extinction/recolonisation island models). The use of these pairwise measures is necessary for the analysis of allozyme data particularly when the amount of variation is low, as in the Atlantic salmon. While the use of a combination of measures may allow us to draw inferences for population fission the major problem is that pairwise measures based on identity, with no additional geographic information, are determined by the relative rates of occurrence of only four events in the history of the sample. These events are mutation, coalescence between a pair of genes, and genes migrating in and out of the same population. These events may be influenced by a number of population processes, some of which have been described in this chapter. However the small number of genealogical events makes it very difficult to draw any conclusions from the events.

Additional information such as the geographic distribution of the samples may allow us to disentangle some of the effects (such as recolonisation). Nevertheless, when the samples are close geographically (such as in the Atlantic salmon) competing hypotheses are indistinguishable from purely pairwise genetic data for one locus. Direct measures of dispersal may give one source of such information, or alternatively if the samples come from islands where the timing of island creation is known then this information can be used (Thorpe et al. 1995). Otherwise, the long periods of time in which the patterns are produced makes disentangling the processes
underlying the four genealogical events very difficult. Wright’s $F_{ST}$ is predominantly influenced by the probability of coalescing within a deme before one of a pair of alleles has left the deme, the process by which this leaving takes place is more difficult to quantify. Nei’s $D$ gives an idea of the amount of time before genes are in the same deme. The private alleles method (Slatkin 1985a), although not portrayed as such, is an extension of this pairwise measure, but is based on the probabilities of a whole set of alleles coalescing in a single population before any of them have left the deme. Other methods of analysis may prove to be more useful.

Examining the allele frequency spectrum may be informative. Braverman et al. (1995) used this to examine the effects of deleterious mutations on sequence variation, and its use for geographic variation may also be informative. Nee et al. (1995) use the distribution of sequence differences to attempt to get information about the rates of population growth, or decay. Single genes are unlikely to give us much information by themselves, and it is in the combination of genes over different loci (analogous to linkage disequilibrium studied in hybrid zones) that is likely to give the most information. Parsimony has a mixed reception; while it may be useful in certain situations (Hudson et al. 1992), the use of this method to determine both the time and location of ‘Mitochondrial Eve’ has been widely criticised (Barton and Goldman 1992, Templeton 1993, 1994). A danger with the use of phylogenies is that “just so” stories for the observed patterns of variation are tempting. The small number of events shows us that to base whole population histories on small samples for a single loci is a very dangerous thing to do as the number of events is so small.

A maximum likelihood approach to finding the relative likelihoods of the different evolutionary scenarios is possible but the computational power required for this is huge. There are attempts at present (Griffiths & Tavaré 1994, Kuhner et al. 1995) to produce general maximum likelihood approaches to calculating the effective migration rates for a 2 island model. This approach must be a way forward though, especially if we use loci for which an accurate phylogeny of the extant alleles in the population can be ascertained.

Barton and Wilson (1995, 1996) showed that isolation by distance is not directly related to mean coalescence times. In this chapter the results have been extended to other models of population structure, and the conclusion is the same: that measures of coalescence for single loci, and pairwise measures are not reliable for
reconstructing the history of a population without other information. This chapter has shown that for a large class of models the effective migration rate (Slatkin 1991) gives a good measure of aspects of the combination of inbreeding and dispersal within any one population. In subsequent chapters a maximum likelihood approach to estimating this parameter is developed, and with additional information on dispersal conclusions may be drawn about within deme events.
Chapter 4

MAXIMUM LIKELIHOOD ESTIMATION OF GENE FLOW

4.1 Introduction

Traditional statistics can be thought of as performing two functions; which are often complementary: hypothesis testing, and parameter estimation. In this thesis these will be kept, as much as possible, separate. This chapter will consider the problem of estimating levels of gene flow between subpopulations, and in the next chapter the power of tests of population subdivision will be addressed. A variety of models for population subdivision were developed in Chapter 3. These models emphasised that data on genetic variation sampled from natural populations are the result of very few events: the total sample size less one coalescences, the rates of which are determined by migration, extinction/recolonisation, fission, or other processes. The mutational process is superimposed over this genealogy. It would appear from this that the best way to measure the extent of genetic variation caused by the population subdivision is to reconstruct the mutational and coalescent events which have produced the observed genetic data and to make inferences on the set of likely reconstructions. The theory in this area of research is in its infancy, as discussed in Chapter 3. With many types of data, e.g. allozyme electrophoresis and micro-satellite loci, the construction of a genealogy is impossible, so pairwise measures are most commonly used for making inferences about population processes (in particular overall $F_{ST}$, and pairwise $F_{ST}$ values). Also frequently used is Slatkin’s private alleles method (Slatkin 1985a), which relies on those alleles present in only a single population. This has
been shown to give the same information as pairwise measures (Barton & Slatkin 1986).

The estimation of statistics such as $F_{ST}$ has been considered by several authors including Weir and Cockerham (1984), Slatkin and Barton (1989), and Slatkin (1989). Cockerham and Weir (1993), and Chakraborty and Danker-Hopfe (1991) discuss the differences between the various estimators of gene flow based on $F$ statistics. Hudson et al. (1992) considered the calculation of $F_{ST}$ and a cladistic method for DNA sequence data. This chapter will use the island model for genetic differentiation and consider allele frequencies rather than sequence data, as found for allozyme, micro-satellite data, or mtDNA data without knowledge of the genealogy. Sequencing is an expensive and time consuming method for collecting data. Minisatellites are an easier way to collect large quantities of information but the interpretation of this data, which generally comprises a large number of alleles and loci, is more difficult. These loci have high mutation rates and the effects of this on measures of gene flow are largely unquantified; however Slatkin (1995) has considered specific mutation models for STR (single tandem repeat) loci. The use of a coalescent approach to simulating data allows the mutation rate to be varied. Then it is possible to investigate the performance of alternate estimators of gene flow, and tests of population differentiation over a range of parameters for a well understood system.

One measure of the amount of gene flow between populations is $M$ ($\hat{M}$ in Slatkin 1989, 1991) which is defined as the number of migrant genes per generation which would produce the same expected level of differentiation (as measured by $F_{ST}$) as if the populations came from an infinite island model. This has intuitive appeal because the parameter $M$ is defined in terms of a very well understood system, and, as shown in Chapter 3, most expressions for the expectation of $F_{ST}$ under the evolutionary process, for a variety of population genetic models, are most easily expressed as $\frac{1}{1 + \alpha}$. As discussed in Chapter 3, when the probability of identity for genes from different populations is small, then $E[F_{ST}]$ is equivalent to the probability of a pair of genes coalescing prior to leaving a deme. This implies that it has the same meaning in many different arrangements of populations. Slatkin and Barton (1989) showed that estimates of $F_{ST}$ from 2 dimensional populations, both for stepping stone models and
models of isolation by distance, gave similar values for $F_{ST}$ to those from an island model.

Chapter 3 has described the properties of pairwise coalescence times under a variety of models. A sample of larger size can be made and the mean and variance of the simulated set of coalescence times measured. This allows another estimate, the 'best' pairwise measure based on location of the samples and the complete genealogy. Hence comparisons can be made between estimators available from the 'real' simulated gene frequencies, and the coalescence times underlying the gene frequencies. Previous studies on estimating $F_{ST}$ have used simulations based on classical population genetic models (Slatkin & Barton 1989). Simulations based on the coalescent process can be performed with lower mutation rates, as the mutational process can be repeated until we have a variable sample, and thus the performance of estimators can be examined over a larger range of parameter values.

In this chapter, the use of maximum likelihood estimation of $M$ and $F_{ST}$, assuming a beta distribution for allele frequencies, will be investigated. This method for estimating levels of gene flow has been considered before by Wehrhahn and Powell (1987), and Slatkin and Barton (1989) but here a natural extension to their approaches to deal with more than 2 alleles per locus, and sampling uncertainty, is developed. This estimator of $M$ has the advantage that it should be able to combine both rare allele and frequency data into a single estimator. The maximum likelihood estimator will be presented, and its performance evaluated on simulated data, compared to other indirect estimators of the amount of gene flow between populations, and to pairwise approaches based on coalescence times.

### 4.2 Model

Consider the infinite island model of population sub-division, as described in Section 2.2.4.2; with an infinite number of islands each with $N$ diploid individuals with random mating within each population; and two alleles $A$ and $a$. Now, under a diffusion approximation, the frequency of allele $A$ in the local populations will follow a beta distribution with parameters $4M\bar{p}$ and $4M\bar{q}$ (Section 2.2.4.2) where $\bar{p}$ is the average frequency of the allele over all populations and $\bar{q} = 1 - \bar{p}$. $M$ is defined to equal $N_em$ where $N_e$ is the effective population size in each of the local
subpopulations and \( m \) is the proportion of migrant gametes per generation. The gene frequencies in local populations follow a distribution

\[
f(p) = \frac{p^{4M_0^0}(1-p)^{4M_0^0-1}}{\beta(4M_0^0,4M_0^q)}, \quad (4.1a)
\]

where \( \beta \) is the beta function (Abramowitz and Stegun 1965)

\[
\beta(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \quad (4.1b)
\]

Then it can be shown that for a sample of size \( n \), with replacement, from a local population, the number \( x \) of allele A will follow a Beta-Binomial distribution:

\[
P(X = x) = \binom{n}{x} \frac{\beta(4M_0^0 + x, 4M_0^q + n - x)}{\beta(4M_0^0, 4M_0^q)}. \quad (4.2)
\]

This distribution is associated with over-dispersion in binomial models (Liang and McCullagh 1993). The two sampling forces at work are therefore separated into the genetic sampling which gives a beta distribution to the underlying gene frequencies and binomial sampling from these underlying gene frequencies. As \( M \) becomes large this approaches the binomial distribution with parameters \( n \) and \( \bar{p} \) as expected for a sample from a single panmictic population.

We can see the connection between \( F_{ST} \) and the beta-binomial model by using the usual definition of \( F_{ST} \) in an infinite island model namely:

\[
E[F_{ST}] = \frac{\text{Var}(p)}{\bar{p}(1-\bar{p})}. \quad (4.3)
\]

where, as in Chapter 3, \( E[\ ] \) denotes the expectation over the evolutionary process. Now the variance of the beta distribution with parameters \( a \) and \( b \) is \( \frac{ab}{(a+b)^2(a+b+1)} \) so that if a beta distribution for the gene frequencies between demes is assumed then this distribution gives the expectation of \( F_{ST} \) to be exactly \( 1/(1+4M) \). If \( n \) diploids are sampled with replacement from a subpopulation, letting \( x \) be the number of alleles of type A, then:
Chapter 4: MAXIMUM LIKELIHOOD ESTIMATION

\[ ar(x) = \frac{n\bar{p}(1 - \bar{p})(4M + n)}{(4M + 1)} \]

so that

\[ E[Var(p)] = \frac{\bar{p}(1 - \bar{p})(4M + n)}{n(4M + 1)} \]

Hence using (4.3) an estimator of \( F_{ST} \):

\[ \hat{F}_{ST} = \frac{4M}{n} + 1 \]

Thus an estimator based purely on the variance in observed gene frequencies will overestimate \( F_{ST} \). Weir and Cockerham (1984) use correction factors for the sample size to account for this but here a maximum likelihood approach is used instead. This simplifies the complications of varying sample sizes and allows the separation of the two sampling processes in the population.

4.2.1 Limitations in Using the Beta Distribution

As shown in Chapters 2 and 3, the expectation of \( F_{ST} \) is most commonly described by a formula of the form \( \frac{1}{1 + a} \) where \( a \) is some function of the evolutionary parameters. In the infinite island model \( a = 4Nm \). When mutations, and/or a finite total population size are added to the model then the form of \( a \) becomes more complicated. Using a parameter such as the effective migration rate \( M \) to describe these inseparable parameters is a way of ignoring the complexity of different factors. In most cases, therefore, the use of this single parameter is more useful than guesses of unknowable parameters. Exceptions to this are when we know the number of natural populations where corrections can be made to get more accurate expressions for the migration rate. This definition for \( M \) is really just a way of rewriting \( E[F_{ST}] \), but it is a more easily conceptualised quantity when it is believed that the population differentiation is caused by limited gene flow, and allows the distinction between the statistic \( F_{ST} \), and a parameter used for estimation purposes to be made clear.

The beta distribution is known to fit the distribution of gene frequencies for an island model when there are an infinite number of populations. When the number of populations is finite, drift in the population mean gene frequency will occur and so
mutation between the two alleles A and a is needed to prevent fixation. In this case the beta distribution with parameters $4M\bar{p}$, and $4M(1 - \bar{p})$ is not exact and a quasi-equilibrium distribution, as described in Barton and Slatkin (1986), is required, which depends on the mutation rate. If the total population size is large then this will make very little difference to the distribution as the change in an individual population mean per generation will be much greater than the change in the mean gene frequency over all populations.

Other models for population subdivision were described in Chapter 3. For the simplest extinction/recolonisation model of Section 3.8.1.5. (Type 1) the expected change in gene frequency in a single generation ($\Delta p$) is $\gamma \bar{p} - \gamma (1 - \bar{p})$ so the distribution of gene frequencies is again beta distributed with parameters $2N\gamma \bar{p}$ and $2N\gamma (1 - \bar{p})$ (see Section 2.2 for the derivation). When there is correlation in the location of founding individuals, or inbreeding before the population reaches the carrying capacity, then this distribution will not hold exactly.

The distribution of gene frequencies after population fission can also be calculated. Assuming a gene frequency of $\bar{p}$ prior to the fission event then Crow & Kimura (1970) give the distribution of gene frequency after time $t$:

$$f(x | \bar{p}, t) = \sum_{i=1}^{m} \bar{p}^i(1 - \bar{p})^{(i+1)(2i+1)} F(1-i, i+2, 2, \bar{p}) F(1-i, i+2, 2, x)e^{-\frac{i(i-1)p}{4N}} , \tag{4.5}$$

where $F$ is the hypergeometric function (Abramowitz and Stegun 1965). This is an extremely complicated function, and estimation for fission events will not be considered here.

### 4.2.2 Test of Beta Distribution

The quasi-equilibrium distribution gives a beta distribution to the underlying frequency of A in the sub populations with parameters $4Nm\bar{p} + 4N\mu$ and $4Nm(1 - \bar{p}) + 4N\lambda$, where $\mu$ and $\lambda$ are the mutation rates from a to A, and A to a, respectively. In this case the expected gene frequency of A, $\bar{p}$, will be $\frac{\mu}{\lambda + \mu}$.

Barton and Slatkin (1986) used simulations to test the efficacy of this approximation for rare allele data. Further simulations were carried out to test these
approximations over the range of model parameters used later in this chapter. These simulation results are from a simple island model simulation forward in time.

A Kolmogorov-Smirnov test (D'Agostino and Stephens 1986) was performed comparing the observed empirical distribution function to that calculated from a beta distribution with the parameters $4Nmp$ and $4Nm(1 - p)$ where $p$ is the mean gene frequency calculated among all sub-populations (the mutation rates used in the simulations were small compared to the migration rates so the mutation terms were ignored). Figure 4.1 illustrates the correspondence between the empirical distribution functions and the calculated cumulative distribution function. These results indicate that the beta distribution is a good approximation for the gene frequencies for a finite island model.

Figure 4.1 Plots of the empirical distribution functions for simulated island models compared to the cumulative distribution functions for a beta distribution. 40 populations with 40 diploids. Mutation rate $10^{-4}$. Triangles $m=0.01$; squares $m=0.05$; circles $m=0.1$. Cumulative distribution functions based on observed mean gene frequencies, x. D gives the K-S statistic.
4.2.3 Limitations of Binomial Sampling

Under the beta-binomial distribution sampling takes place with replacement from the underlying population. For genetic data the samples are taken without replacement so that a binomial distribution is no longer applicable. The exact sampling distribution is the hypergeometric. Assuming a total population size of $2N$ genes of which $r$ are of type $A$ then if one takes a sample of size $n$ genes the distribution for the number of type $A$ is given by

$$P(X = x) = \binom{r}{x} \binom{2N-r}{n-x} \frac{1}{\binom{2N}{n}}.$$  \hfill (4.6)

The mean and variance of the number of alleles of type $A$, $X$, is then:

$$E(X) = \frac{nr}{2N}, \quad \text{and} \quad \text{Var}(X) = \frac{mr(2N-r)}{4N^2} \left(1 - \frac{n-1}{2N-1}\right).$$  \hfill (4.7)

If we take $r/2N$ to be $p$ we can see that the variance of the sample is reduced relative to binomial sampling by a fraction $(n-1)/(2N-1)$. If this fraction is appreciable then alternative methods to calculate the amount of gene flow must be used as the beta-binomial is not applicable. However if the fraction sampled is small then this is a reasonable approximation. When sampling takes place at a juvenile phase before population regulation in a species with a large number of juveniles relative to adults, sampling can be assumed to be taken with replacement from the adult population. This is a reasonable way of sampling species which have large numbers ($>10^3$) of offspring. For Atlantic salmon this is the usual way in which data are collected.

4.3 Extension to $k$ Alleles per Locus

For samples from natural populations there are often more than 2 alleles per locus, in particular for micro or mini-satellite data. There has been some disagreement over the best way to combine over alleles (Weir & Cockerham 1984, Chakraborty &
Danke-Hopfe 1991, Cockerham & Weir 1993). In many surveys the least common alleles are combined.

For a locus with \( k \) alleles, a generalisation of the beta distribution is the Dirichlet distribution so that the joint distribution of the alleles in a sub population, given that the frequencies of the \( k \) alleles over all populations are \( \overline{p}_1, \ldots, \overline{p}_k \), is:

\[
f(p_1, \ldots, p_k) = \frac{p_1^{4M_p-1} \cdots p_k^{4M_p-1}}{D(4M_p_1, \ldots, 4M_p_k)},
\]

where
\[
D(\alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}.
\]

Taking a sample of size \( n \) from a subpopulation, the probability of obtaining \( x \) is:

\[
P(X_1 = x_1, \ldots, X_k = x_k / \sum_{i=1}^{k} x_i = n) = \frac{n!}{x_1! \cdots x_k!} \frac{D(4M_p_1 + x_1, \ldots, 4M_p_k + x_k)}{D(4M_p_1, \ldots, 4M_p_k)}.
\]

For \( k = 2 \) this is simply the beta-binomial distribution.

The multinomial-Dirichlet has the property that if any allele is removed from the analysis then the remainder of the alleles follow the same marginal distribution. This means that if the extant alleles in the sample are grouped together in some way, then the frequencies follow the same marginal distribution whatever form the grouping takes. This is a desirable property as with some genetic data (e.g. allozyme data) the alleles seen are grouped according to weight and to charge which is an arbitrary grouping. Microsatellite loci are comprised of repeats of a short sequence of DNA, which can presumably be lost or gained. This means that a microsatellite allele which gives the same pattern on a gel may have come from a set of convergent mutations. This method has the advantage that these convergent mutations should not effect the estimates for gene flow because of this exchangeable property.
4.4 Private Alleles and the Beta Binomial

The use of alleles present in samples from a single sub-population (private alleles) to estimate levels of gene flow was suggested by Slatkin (1985a). The presence of alleles in a single sub-population at high frequency is an indication of low levels of gene flow because the allele has not spread to other populations. Tables were produced using simulations to estimate the levels of gene flow using private alleles. Figure 4.2 shows the estimated frequency spectra for differing migration rates produced by simulations (details of the simulation program are given in Section 4.9.1). The relative frequency of single copies of alleles decreases with the migration rate. Thus even single alleles, or the lack of them, give information about the rates of gene flow between subpopulations.

![Frequency spectrum of alleles found in only a single deme. All bars based on 2000 replicates for 20 populations of 100 diploids with 10 diploids sampled from each. Mutation rate = 10^{-5}; migration rate = 0.01 (obliquely hatched bar), 0.005 (open bar), 0.001 (black bar).]
It is common in studies of population subdivision to define alleles to be polymorphic if the frequency of the most common allele does not exceed 0.95 (however, lack of resolution in gel electrophoresis may make this necessary); and only to include these alleles in calculations of $F_{ST}$.

Barton and Slatkin (1986) showed that a beta distribution could model rare alleles. Balding and Nichols (1995) showed that the distribution of alleles in a single population (for a $k$-allele, rather than an infinite allele model), would follow a Dirichlet distribution. A Dirichlet distribution, therefore will fit the distribution of a number of rare alleles in the population (if we consider a single rare allele and group all the rest of the alleles together then we get a beta distribution from the Dirichlet).

Consider an allele which is at frequency $\bar{p}$ in the population. The probability of it not being present in a sample of $n_i$ genes from a single population, which is denoted $g(\bar{p}, n_i)$ is:

$$g(\bar{p}, n_i) = \frac{\beta(4\bar{p}, 4(1-\bar{p})+n_i)}{\beta(4\bar{p}, 4(1-\bar{p}))}.$$  \hspace{1cm} (4.10)

Thus if we have $r$ samples of size $n_i$, $i=1,..,r$ the probability of a private allele which is denoted $h(\bar{p}, n)$ where $n = n_i$, $i=1,..,r$ is:

$$h(\bar{p}, n) = \prod_{j=1}^{r}(1 - g(\bar{p}, n_j)) \prod_{i=1, i\neq j}^{k} g(\bar{p}, n_i).$$  \hspace{1cm} (4.11)

Therefore, with $r$ samples each of equal size, so that $n_i = n \forall i$, the probability of a private allele is simply the binomial probability $rg(\bar{p}, n)^{r-1}(1 - g(\bar{p}, n))$. In this situation the distribution of the frequency of private alleles (which is denoted $f_p(x)$) can be calculated by conditioning on the allele being private:

$$f_p(x) = \frac{r\beta(4N\bar{p} + x, 4M(1-\bar{p})+n-x)}{h(\bar{p}, n)\beta(4N\bar{p}, 4M(1-\bar{p}))}.$$  \hspace{1cm} (4.12)

Figure 4.3 shows $f_p(x)$ compared with simulated data (the simulation program used is described in Section 4.9.1). In this case $\bar{p}$ was estimated from the mean private allele frequency over all replicates. The beta-binomial distribution gives a reasonable approximation to the observed frequencies. The predominant class of private alleles are those with only a single copy. These are commonly ignored in studies.
Figure 4.3 Frequency spectrum of alleles found in only a single deme. First four bars based on 2000 replicates for 20 populations of 100 diploids with 10 diploids sampled from each. Migration rate = 0.01; mutation rate = $10^{-6}$ (obliquely hatched bar), $10^{-5}$ (black bar), $10^{-4}$ (horizontal lined bar), $10^{-3}$ (open bar). Grey bar is beta-binomial probability of observing this number of private alleles assuming the observed mean private allele frequency for all 2000 replicates.

Figure 4.3 also shows that the frequency spectrum of private alleles is unaffected by changes of 4 orders of magnitude in the mutation rate. This is somewhat counterintuitive, as one would expect that increasing the mutation rate would increase the number of private alleles at low frequencies.

### 4.5 Estimation using Maximum Likelihood

Using a beta-binomial, or multinomial-Dirichlet distribution has been shown, in the previous section, to provide a good description of number of alleles of each allelic
type in a subdivided population, where the differences are caused by limited dispersal. Maximum likelihood estimators of $M$, or $F_{ST}$ can thus make more use of the information provided by the samples than can conventional pairwise estimators.

Consider a set of genetic data for $r$ subpopulations consisting of samples of size $n_i$ with $k$ alleles with the number of allele $j$ in sample $i$ written as $x_{ij}$. The likelihood of the parameters $M$, and $\bar{p}$ is then

$$L(M, \bar{p} / X, n) = \prod_{i=1}^{r} \frac{n_i}{x_{i1} \cdots x_{ik}} \left( \frac{D(4M\bar{p}_1 + x_{i1}, \ldots, 4M\bar{p}_k + x_{ik})}{D(4M\bar{p}_1, \ldots, 4M\bar{p}_k)} \right).$$ (4.13)

with the added constraint $\sum_{i=1}^{r} \bar{p}_i = 1$.

The maximum likelihood estimates of this distribution cannot be obtained exactly but using a suitable maximisation technique the maximum likelihood estimates of $M$ and the $\bar{p}_i$'s can be calculated by numerically maximising $\log(L)$. Many commercially available packages contain routines for numerical maximisation. An example of performing maximum likelihood estimation of gene flow using SYSTAT (Wilkinson 1992) is given in Appendix 5.

A support surface can be calculated from the data which can be used to give approximate confidence intervals for $M$, or alternatively parametric bootstrapping (Efron & Tibshirani 1993) from the distribution with parameters given by the maximum likelihood estimators will also give approximate confidence intervals.

### 4.5.1 Estimating $F_{ST}$ using Maximum Likelihood

The previous section has concentrated on estimating $M$ using maximum likelihood. However, it is equally feasible to use an estimate of $M$ to construct an estimator for $E[F_{ST}]$ using the multinomial-Dirichlet distribution. In this case $E[F_{ST}] = 1/(1+4M)$ and so to get a maximum likelihood estimator we simply maximise Equation 4.13 replacing $M$ by $(1/E[F_{ST}] -1)/4$. This will give the same point estimate as maximising over $M$; the support curve will be different.
4.5.2 Other Maximum Likelihood Estimators

For a single locus with 2 alleles the maximisation is quick, and without problems. When there are few (<5) alleles the maximisation is still fairly simple, and not too time intensive. When there are additional alleles, or more loci each with several alleles, the maximisation becomes extremely time consuming. An alternative estimator can be constructed by assuming that the allele frequencies are the observed gene frequencies, \( \hat{p}_i = \frac{\sum_{j=1}^{r} x_{ij}}{\sum_{j=1}^{r} n_j} \), and then the maximisation is over a single parameter, \( M \). This is much faster. In the simulation experiments presented in this chapter the reduced parameter maximum likelihood estimator is denoted by \( \hat{M}_s \).

Another estimator may be constructed when the mutation rate is so high relative to the migration rate that both processes can be considered together. Under these conditions, and when the number of subpopulations sampled is small relative to the total population, the observed gene frequencies, \( \hat{p}_i \), will be unbiased estimators of the gene frequency of the population, but their variance will be large. In this case there is little reason for the assumption that allele frequencies in the demes sampled are representative of the population as a whole (this has been shown to give a good approximation in the case of private alleles when we only look at the presence of alleles in a single population so that the probability of a migrant allele having the same allele which is identical by descent is very low). In these circumstances it may be more appropriate to assume that all extant alleles have the same frequency (it is just the case that in the populations one looks at they are perturbed away from this frequency). This is a natural way to consider the distributions of fingerprint loci (implicit in Balding & Nichols 1995) and at other highly mutable loci where it is believed (or it is a reasonable approximation) that no alleles are more prevalent just by the actions of mutation (this is an extremely strong assumption, and unlikely to be reasonable in many cases, as mutations are will not give all alleles with equal probability).

Under these conditions it has been shown by Balding and Nichols (1995) that the expected distribution of gene frequencies of these types is given by a Dirichlet distribution. In this case an estimator can be constructed which assumes that all alleles are of equal frequency (in some underlying population), or that mutation
events are so frequent that the probability of a particular allele by mutation, or migration from another deme is the same.

The estimator for M which is produced by assuming all $\bar{p}_i$'s are the same is denoted by $\bar{M}_R$.

4.5.3 Estimation over Several Loci

It is conceptually simple to estimate M over several loci adding the log-likelihood at each locus and then maximising over all loci for all of the estimation procedures described above. However, migration of individuals between populations will involve alleles at all loci, and hence estimates based on different loci will not be independent even with physically unlinked loci. The magnitude of this effect will depend on the frequency of gene flow, and the relative populations sizes. The effect of this on estimates on gene flow has not been considered here.

4.6 The Maximum Likelihood Estimator of M

4.6.1 Tests of the Estimators

Practical use of any estimator for gene flow requires several properties:

1. consistency across different types of genetic data,

2. use of all available information,

3. no sensitivity to removal of either a) alleles or b) sites, and

4. ease of use.

In this chapter the mean of the estimators over replicates will be measured, which gives an indication of the bias of the estimators. Also considered is the mean squared error (M.S.E.) which is defined as

$$M.S.E. = E[(X - x_0)^2].$$

(4.14)
Where \( x_0 \) is the true value of the quantity and \( X \) is an estimator of it. In this chapter simulated data will be used and the M.S.E. will be estimated from this by:

\[
M.S.E.(r) = \frac{1}{\text{replicates}} \sum_{\text{replicates}} (r_i - r)^2. \tag{4.15}
\]

MSE will be used to denote the estimated M.S.E. rather than \( M.\hat{S}.E. \). When estimators are unbiased the M.S.E. is equal to the variance of the estimator.

Maximum likelihood estimators are not guaranteed to be unbiased, but maximum likelihood estimators are consistent which means that as the sample size increases the expected bias will decrease as will the mean squared error.

### 4.7 Tests on Simulated Beta-Binomial Data

Tests on the efficacy of this estimation procedure will be in two stages. In this section the properties of the estimator \( \hat{M} \) based on binomial sampling from a beta distribution with parameters \( 4M\bar{p} \) and \( 4M(1 - \bar{p}) \) will be examined. In subsequent sections the estimator will be tested against simulated data from a coalescent model, and the results compared to those using alternate estimators. The following results are based on drawing \( r \) binomial samples each of size \( n \) from \( r \) independent beta variates with parameters \( 4M\bar{p} \) and \( 4M(1 - \bar{p}) \). The beta variates are generated by an algorithm based on that in RANLIB (Brown & Lovato 1995).

Figure 4.4 illustrates the properties of \( \hat{M} \) over a wide range of number of samples. The mean value of \( \hat{M} \) is too large over all numbers of samples, but disproportionately more when the number of samples is less than 10. Table 4.1 describes the situation when the total sample size is constant but the number of samples and the number per sample change.

Table 4.1 shows that the mean squared error of the estimators, and their standard deviation increases as the number of sites decreases. We should expect then that the estimator for \( \hat{M} \) should work best for genetic data when there are many samples of small size rather than a small number of samples with a large sample size, and that large sample sizes increase the precision also. This will be examined in a later section. Table 4.2 suggests that the maximum likelihood estimator is better when M
is low. The best estimates of both M and p are obtained when p is close to 0.5 as in Table 4.3.

<table>
<thead>
<tr>
<th>r = number of samples</th>
<th>n = sample size</th>
<th>̂p</th>
<th>est. s.d ̂p</th>
<th>̂M</th>
<th>est. s.d ̂M</th>
<th>MSE ̂p</th>
<th>MSE ̂M</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0.500</td>
<td>0.027</td>
<td>1.047</td>
<td>0.209</td>
<td>0.001</td>
<td>0.045</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
<td>0.499</td>
<td>0.035</td>
<td>1.085</td>
<td>0.295</td>
<td>0.001</td>
<td>0.093</td>
</tr>
<tr>
<td>40</td>
<td>50</td>
<td>0.501</td>
<td>0.039</td>
<td>1.090</td>
<td>0.324</td>
<td>0.001</td>
<td>0.112</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>0.503</td>
<td>0.051</td>
<td>1.140</td>
<td>0.434</td>
<td>0.003</td>
<td>0.206</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>0.500</td>
<td>0.065</td>
<td>1.438</td>
<td>0.742</td>
<td>0.004</td>
<td>0.737</td>
</tr>
</tbody>
</table>

Table 4.1 Properties of estimators of M and p with number of samples and total sample size fixed. Values based on 100 replicates with p = 0.5 and M = 1.0.

Figure 4.4 Estimated M against number of samples. Squares represent means, lines estimated standard deviations. Values based on 200 replicates with each sample of size n=100, M=1, and ̂p = 0.5. Solid line is the expected value.
### Chapter 4: MAXIMUM LIKELIHOOD ESTIMATION

#### Table 4.2 Properties of estimators of \( M \) and \( p \) with changing \( M \). Values based on 100 replicates for 20 populations with 100 samples per population and \( p = 0.5 \).

<table>
<thead>
<tr>
<th>( M )</th>
<th>( \hat{p} )</th>
<th>est. s.d ( \hat{p} )</th>
<th>( \hat{M} )</th>
<th>est. s.d ( \hat{M} )</th>
<th>MSE ( \hat{p} )</th>
<th>MSE ( \hat{M} )</th>
<th>MSE ( /M^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.497</td>
<td>0.089</td>
<td>0.111</td>
<td>0.034</td>
<td>0.008</td>
<td>0.001</td>
<td>0.126</td>
</tr>
<tr>
<td>0.2</td>
<td>0.487</td>
<td>0.079</td>
<td>0.226</td>
<td>0.061</td>
<td>0.006</td>
<td>0.004</td>
<td>0.109</td>
</tr>
<tr>
<td>0.4</td>
<td>0.500</td>
<td>0.065</td>
<td>0.469</td>
<td>0.143</td>
<td>0.004</td>
<td>0.025</td>
<td>0.157</td>
</tr>
<tr>
<td>0.8</td>
<td>0.507</td>
<td>0.052</td>
<td>1.053</td>
<td>0.513</td>
<td>0.003</td>
<td>0.325</td>
<td>0.507</td>
</tr>
<tr>
<td>1.6</td>
<td>0.502</td>
<td>0.039</td>
<td>1.992</td>
<td>0.749</td>
<td>0.002</td>
<td>0.709</td>
<td>0.277</td>
</tr>
<tr>
<td>3.2</td>
<td>0.501</td>
<td>0.033</td>
<td>4.112</td>
<td>2.594</td>
<td>0.001</td>
<td>7.493</td>
<td>0.732</td>
</tr>
<tr>
<td>6.4</td>
<td>0.502</td>
<td>0.030</td>
<td>7.603</td>
<td>3.731</td>
<td>0.001</td>
<td>15.232</td>
<td>0.372</td>
</tr>
</tbody>
</table>

#### Table 4.3 Properties of estimators of \( M \) and \( p \) with changing \( p \). Values based on 100 replicates for 20 populations with 100 samples per population and \( M = 1.0 \).

<table>
<thead>
<tr>
<th>( p )</th>
<th>( \hat{p} )</th>
<th>est. s.d ( \hat{p} )</th>
<th>( \hat{M} )</th>
<th>est. s.d ( \hat{M} )</th>
<th>MSE ( \hat{p} )</th>
<th>MSE ( \hat{M} )</th>
<th>MSE ( /p^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04</td>
<td>0.038</td>
<td>0.019</td>
<td>2.067</td>
<td>3.513</td>
<td>0.000</td>
<td>0.227</td>
<td>13.457</td>
</tr>
<tr>
<td>0.08</td>
<td>0.075</td>
<td>0.027</td>
<td>1.636</td>
<td>4.660</td>
<td>0.001</td>
<td>0.117</td>
<td>22.073</td>
</tr>
<tr>
<td>0.12</td>
<td>0.117</td>
<td>0.033</td>
<td>1.313</td>
<td>0.617</td>
<td>0.001</td>
<td>0.077</td>
<td>0.477</td>
</tr>
<tr>
<td>0.20</td>
<td>0.199</td>
<td>0.043</td>
<td>1.215</td>
<td>0.513</td>
<td>0.002</td>
<td>0.046</td>
<td>0.309</td>
</tr>
<tr>
<td>0.25</td>
<td>0.248</td>
<td>0.045</td>
<td>1.191</td>
<td>0.457</td>
<td>0.002</td>
<td>0.032</td>
<td>0.245</td>
</tr>
<tr>
<td>0.36</td>
<td>0.362</td>
<td>0.048</td>
<td>1.168</td>
<td>0.425</td>
<td>0.002</td>
<td>0.018</td>
<td>0.208</td>
</tr>
<tr>
<td>0.40</td>
<td>0.404</td>
<td>0.046</td>
<td>1.176</td>
<td>0.400</td>
<td>0.002</td>
<td>0.013</td>
<td>0.191</td>
</tr>
<tr>
<td>0.42</td>
<td>0.423</td>
<td>0.047</td>
<td>1.179</td>
<td>0.391</td>
<td>0.002</td>
<td>0.013</td>
<td>0.184</td>
</tr>
<tr>
<td>0.45</td>
<td>0.452</td>
<td>0.047</td>
<td>1.196</td>
<td>0.422</td>
<td>0.002</td>
<td>0.011</td>
<td>0.216</td>
</tr>
<tr>
<td>0.50</td>
<td>0.503</td>
<td>0.048</td>
<td>1.201</td>
<td>0.433</td>
<td>0.002</td>
<td>0.009</td>
<td>0.228</td>
</tr>
</tbody>
</table>

#### 4.8 Other Estimators of Gene Flow

In order to test the efficiency of the maximum likelihood approaches for estimating gene flow, their properties need to be compared with other methods. The structured coalescent simulations give us 3 pieces of information:
theoretical values for M,
• simulated times of coalescence for the sample, and
• simulated allele frequencies

Therefore, there are 3 levels of comparison to make: with the theoretical expectations; with estimates made from the coalescence times; and with alternate estimators of M from allele frequencies. The theoretical values seen under a variety of population structures have been examined in Chapter 3.

4.8.1 Gene Flow Estimates from Coalescence Times

The probabilities of identity within and between populations can be written in terms of the moment generating functions of coalescence times as shown in Chapter 3 (see also Cockerham & Weir 1993, Slatkin 1991). It was shown that using the mean coalescence time produced a simple approximation for $F_{ST}$ which was a good approximation for small mutation rates. In this chapter higher mutation rates will be used. Now (following Herbots (1995)):

$$ f = \int_0^\infty g(t)e^{-2\mu t} dt $$

where $g(t)$ is the distribution of coalescence times and $f$ is the probability of identity by descent for 2 genes sampled at random. Using a Taylor expansion of $e^{-2\mu t}$ around $\mu=0$ we have:

$$ f = \int_0^\infty g(t)[1 - 2\mu t + 2\mu^2 t^2 + o(\mu^3)]dt \approx 1 - 2\mu E[t] + 2\mu^2 E[t^2]. \quad (4.16) $$

This is extremely approximate as it depends on $\mu$ being small enough so that $\int_0^\infty g(t)\mu^3 E[t^3]dt$ is negligible. For an exponential distribution of coalescence times, this may be reasonable. In this case we can construct other estimators for $F_{ST}$ based on the mean, and assuming that we know the mutation rate, on the mean and second moment of the coalescence times, which are denoted $\hat{T}_{ST}$ and $\hat{T}_{ST}^2$ respectively:
\[ \hat{T}_{ST}^1 = \frac{i - i_0}{\hat{I}}, \quad \text{and} \]
\[ \hat{T}_{ST}^2 = \frac{i - i_0 - \mu (\hat{m}_i^2 - \hat{m}_0^2)}{\hat{I} - \mu \hat{m}_i^2}, \tag{4.17} \]

where \( \hat{m}_i^2 \) is the estimated second moment of \( t \) about 0.

### 4.8.1.1 Estimating Mean and Variance of Coalescence Times

In order to use the estimators \( \hat{T}_{ST}^1 \) and \( \hat{T}_{ST}^2 \) we require the mean and variance of coalescence times within and between demes. The data on coalescence times, for samples of size \( n_i, i = 1, \ldots, r \), consist of \( \sum n_i - 1 \) different coalescence times (assuming that events happen simultaneously). There are, however, \( \sum (n_i(n_i - 1)/2) \) pairwise coalescences within demes, and \( (\sum n_i)[(\sum n_i - 1) / 2 - \sum (n_i(n_i - 1)/2)] \) pairwise coalescences between demes so that each coalescence time is, on average, representative of \( \sum n_i / 2 \) pairwise times.

Thus there is a large correlation between the individual pairwise times (see also Felsenstein 1992a, b). Figure 4.5 illustrates this for two samples of 3 genes from demes A and B. There are then 6 pairwise coalescence times within demes and 6 between demes. The mean times within and between are calculated by weighting by the number of within and between deme coalescences.

The variance of coalescence times can be similarly calculated.
Figure 4.5 An illustration of the cause for the correlation of within and between deme pairwise measures.

This correlation between times within and between demes leads to bias in the estimation of $F_{ST}$ by either $\hat{T}^1_{ST}$ or $\hat{T}^2_{ST}$. If we have random variables $Y$ and $X$ then the approximate mean and variance of $Z = \frac{Y}{X}$ can be approximated by

$$\frac{\mu_y}{\mu_x} + \frac{1}{\sigma_x^2} \left( \frac{\sigma_x^2}{\mu_x} - \rho \sigma_x \sigma_y \right),$$

and

$$\frac{1}{\mu_x^2} \left( \sigma_x^2 \frac{\mu_y^2}{\mu_x^2} + \sigma_y^2 - 2 \rho \sigma_x \sigma_y \frac{\mu_x}{\mu_y} \right)$$

respectively. This suggests that to combine estimates over loci it would be preferable to calculate $\hat{T}^1_{ST}$ and $\hat{T}^2_{ST}$ for each locus and then to take the average (as the correlation between times means that the bias is reduced).

4.8.2 Gene Flow Estimators from Frequency Data

The estimation of $F_{ST}$ from gene frequency data is extremely confused, with some authors using $\text{Var}(p)/(\bar{p}(1-\bar{p}))$, and others using the methods due to Weir and Cockerham (1984), or Nei (1973). Some authors have attempted to ease this confusion by testing the efficiency of these different estimators, for example, Slatkin & Barton (1989). The theoretical expectation for $F_{ST}$ was calculated for a number of population structures in Chapter 3. Some of this confusion is based on the
definitions used for $F_{ST}$. The original definition was discussed in Chapter 2, however other definitions have been used in this thesis (Equation 2.15). Weir and Cockerham (1984, Cockerham & Weir 1993) constructed an estimator $\theta$ for $F_{ST}$, and argued that estimates of probability of identity for genes sampled at random from the whole population should not be made, rather estimates of the identity by descent for genes sampled at random from different, and the same populations. The reasoning behind this is that it is not possible to know the number of demes within a population, so it is better to assume that the number is so large that the differences will be negligible. An alternative method was proposed by Nei (1973) (Nei’s $G_{ST}$) who advocates estimating the amounts of inbreeding from the populations directly. The differences between these philosophies of estimation mean that there are differences in the properties of the estimators, the most obvious of which is that while $\theta$ may take negative values while $G_{ST}$ cannot.

4.8.2.1 Nei’s $G_{ST}$

Nei (1973) introduced a method ($G_{ST}$) for estimating $F_{ST}$ for gene frequency data. This method has been criticised by Weir and Cockerham (1984, 1993) because it does not adequately separate the parameter $F_{ST}$ from the statistic $G_{ST}$ used to estimate it. This is an argument about the philosophy of estimation. Weir and Cockerham assume, as in this thesis for maximum likelihood estimation, that the data are regarded as a random sample from an infinite number of replicate populations whereas Nei assumes that the data are representative of the whole population. $G_{ST}$ is defined in terms of the ‘average’ pairwise difference in the probability of identity:

$$G_{ST} = \frac{D_{ST}}{H_T},$$

where $D_{ST} = \frac{1}{r^2} \sum_{i,j} \sum_k (p_{ik} - p_{jk})^2$, $H_T = 2\sum_k \bar{p}_k (1 - \bar{p}_k)$ and $p_{ik}$ is the frequency of the $k$-th allele in the $i$-th population and $\bar{p}_k$ is the mean frequency of allele $k$ over the whole population, weighting by sample size.

Denoting the expected probability of identity of alleles sampled at random from demes $i$ and $j$ by $f_{ij}$, and the expected probability of identity for genes sampled at random from the whole population by $\bar{f}$ then the expected value of $G_{ST}$ can be calculated:
Chapter 4: MAXIMUM LIKELIHOOD ESTIMATION

\[
E(G_{ST}) = \frac{\sum (f_{ii} + f_{jj} - 2f_{ij})}{2d^2(1-f)}.
\]

If \( f_{ii} = f_0 \ \forall \ i \) and \( f_{jj} = f_i \ \forall i, j \neq i \), as in an island model then

\[
E(G_{ST}) = \frac{f_0 - \bar{f}}{1 - \bar{f}}.
\]

\( G_{ST} \) is not in this case an estimator for \( F_{ST} \) unless we sample every population, but is a statistic whose expectation is the same as \( F_{ST} \). Nei’s method when there are multiple loci is to sum \( D_{ST} \) and \( H_T \) over loci and then calculate \( G_{ST} \).

4.8.2.2 Weir and Cockerham’s \( \theta \)

Weir and Cockerham (1984, Cockerham & Weir 1993) advocate an alternative philosophy for the estimation of \( F_{ST} \). These authors regard \( \theta \) as a parameter which has the same expectation as the statistic \( F_{ST} \). Analysis of variance is used to calculate \( \theta \). Corrections are included for the number of demes sampled, and the sample size from each deme, which will be ignored here (although the estimated \( \theta \) from simulated data will use their corrections). Their equation for \( \theta \) can be shown to give the same expected value as \( G_{ST} \).

\[
\theta = \frac{\sum s_k^2}{\sum p_k (1 - p_k)} \approx \frac{1}{r} \frac{\sum p_k^2 - \sum \bar{p}_k^2}{\sum p_k (1 - \bar{p}_k)} \quad \text{so that} \quad E(\theta) = \frac{f_0 - \bar{f}}{1 - \bar{f}}.
\]

Where \( s_k^2 \) is the estimated variance in allele frequencies in allele \( k \) across populations. Cockerham and Weir (1993) argue for a slightly different method of estimation so that when there are a limited number of demes sampled \( E(\theta) = \frac{f_0 - f_i}{1 - f_i} \) where \( f_i \) is the expected probability of identity for genes sampled from different loci. To combine an estimate over loci, again these authors suggested summing both the numerator and the denominator over loci; and then calculating an overall \( \theta \).
4.9 Simulation Studies

Simulation studies were performed to test the above methods for estimating $M$ against $\theta$; against $\hat{T}_{sT}^1$, and $\hat{T}_{sT}^2$; and against the theoretical expectations which were derived in Chapter 3.

4.9.1 Details of Programs

The simulation studies were based on a coalescent approach, the advantages of which were discussed in Chapter 3. The algorithm used for generating times of coalescence and tree shapes was `make_tree_island2` described in Section 3.2.3, and listed in Appendix 3.

Gene frequencies were generated from the trees by assigning mutations to the branches of the tree proportional to their length. Mutation algorithms are shown in Appendix 6. There are two algorithms here, an infinite allele algorithm `inf_alleles`, for which every mutation produces a new allele; and a k-alleles algorithm `kalleles`, which mutates an allele to any of k-1 others at random. Both these algorithms proceed by assigning an allele to the root of the tree (the point at which all of the genes coalesce). The algorithms then recursively go through all coalescence points recursively assigning mutations according to a Poisson process of rate $\mu s$, where $s$ is time.

Mutation is thus separated from the calculation of coalescence times, and shape of the genealogy. The data were generated with the condition that the frequency of the most common allele is less than $\alpha$, where $\alpha$ is usually 0.95 so that simulated surveys with only a single, or very few mutant alleles are rejected.

When it was feasible (the total number of parameters to be estimated <6) both a full maximisation, and a maximisation over just $M$ were performed; otherwise only the single parameter estimate was determined. The maximisations were performed using Powell's method (in multidimensions), and Brent's method in one dimension (Press et al 1992). The algorithms used were based on those of Press et al (1992).

The programs for the production of data and the estimation of parameters were written in C for the IBM PC using Borland C++ Version 4.
4.9.2 Coalescence Times

As a first step towards comparing estimators, the empirical distribution of $\hat{T}_ST^1$, the estimator of $F_{ST}$ produced by considering the coalescence times, will be examined. As noted by Felsenstein (1991a,b) pairwise times are inefficient estimators so it is useful to get a baseline before further investigations into alternate estimators are considered. As a first step $\hat{T}_ST^1$ was calculated for a number of different sampling strategies which will provide the basis for the tests of estimators. Figure 4.6 illustrates the different estimated distributions for $\hat{T}_ST^1$ for 4 sampling strategies; $\hat{T}_ST^2$ performed worse than $\hat{T}_ST^1$ in all cases, and is not considered further. It is clear that using a large number of small samples is the most effective way to estimate $F_{ST}$ using mean coalescence times, for this extremely low migration rate. Table 4.4 shows that this is not the case for all migration rates, when the migration rates are higher, better estimates are obtained from sampling fewer demes with greater sample sizes in each deme.

The number of samples thus plays a crucial role in the precision of an estimator. Table 4.4 shows the mean times until coalescence within and between demes, and the corresponding estimate of $F_{ST}$ from the coalescence times. These figures show a good fit to the expected values given in Chapter 3. None of the mean times within or between demes is significantly different from the expectation. For the higher migration rate of $M=10$, good estimates could not be obtained for $M$ because in a large number of cases the observed $\hat{T}_ST^1$ values were less than 0. This emphasises the amount of noise in pairwise estimators.
Figure 4.6 Distribution of $\hat{F}^{\text{st}}$ for samples of 100 diploids from a finite island model with 100 diploids per population, and 100 populations with a migration rate of 0.001. Histograms (based on 1000 replicates) represent (clockwise from top left) 100 samples of 1 diploid; 20 samples of 10 diploids; 2 samples of 100 diploids; and 10 samples of 20 diploids. Arrows represent expected values.
Table 4.4 Statistics for samples of 100 diploids from a finite island model with 100 diploids per population, and 100 populations with a migration rate of 0.001. Means and mean squared errors based on 1000 replicates. Cells marked with an asterisk denote values where $M$ could not be estimated (because some $\hat{\theta}_{ST}^3$ are less than 0).

### 4.9.3 The Effect of Mutation

Real data are never as precise as the data presented above. There is always more uncertainty about the genealogy due to uncertainties in phylogeny reconstruction and for some types of data phylogeny reconstruction is impossible. In this section
the effects of mutation rate on $\theta$ as an estimator of $F_{ST}$ will be considered. Figure 4.7 indicates that the precision of estimators increases with increasing mutation rate.

Figure 4.7 The empirical distribution (from 200 replicates) for $\theta$ both from times (top left) and from allele frequencies $\mu = 10^{-4}$ (top right), $\mu = 10^{-5}$ bottom left, and $\mu = 10^{-6}$ bottom right. Simulations are of 100 populations each of 100 individuals with $M=0.1$. Samples of 5 diploids are taken from each of 20 populations.

With a mutation rate of $10^{-4}$, $\theta$ calculated from allele frequencies is a more precise estimator than the mean coalescence times. This can be explained in that when there is genealogical information, and a low $M$, most of the information between populations comes from the earliest split, whereas if we do not assume a genealogy then the distribution is averaged between all different alleles. These estimators give extremely accurate estimates for $E(\theta)$, but this extremely low migration rate is a rarity in natural populations (e.g. of salmon). The original aim was not to estimate $F_{ST}$ but rather $M$. The precision of estimators of $M$ using $\theta$ ($\hat{M}_\theta$) with different sampling strategies is illustrated in Figure 4.8. This Figure shows the same patterns
as observed in the estimates of $F_{ST}$. This gives us a baseline for comparison with the maximum likelihood estimators and $G_{ST}$.

![Graph showing estimates of M from θ compared to the number of populations sampled](image)

**Fig 4.8** Estimates of $M$ from $θ$ compared to the number of populations sampled (keeping total sample size = 100 diploids). Triangles represent $u = 10^{-4}$, circles $10^{-3}$, and squares $10^{-5}$. $M = 0.1$ in all cases.

### 4.9.4 Maximum Likelihood Estimates

Preliminary studies considered the estimation of $M$ by both $\hat{M}$ and $\hat{M}_s$. Table 4.5 gives the results of such a test for two migration rates. The estimates are very similar for the two estimators. Thus $\hat{M}_s$ will be used as the 'default' maximum likelihood estimation procedure, as this can be easily calculated for all mutation rates.

Figure 4.9 illustrates the effect of the number of populations sampled (assuming a fixed total population size) on estimators for $M$. Here $\hat{M}_R$ is the most effective estimator when the number of populations sampled is small, whereas $\hat{M}_g$ and $\hat{M}_s$ are
more effective when the number of populations sampled is higher. This view is confirmed in Tables 4.6 and 4.7 for low values of M. These tables confirm what was shown for coalescence times in Section 4.9.2, that the optimal sampling strategy for precision in estimates is to sample many populations, with small sample sizes if necessary. But if there are only a small number of samples then $\hat{M}_R$ gives the most precise estimates over both values of M. Estimates using $G_{ST}$ were uniformly poor for all sets of data, and are not included.

<table>
<thead>
<tr>
<th>$F_{ST}$</th>
<th>Demes sampled</th>
<th>Estimated $F_{ST}$ from $\hat{M}$</th>
<th>M.S.E</th>
<th>Estimated $F_{ST}$ from $\hat{M}_S$</th>
<th>M.S.E</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.710</td>
<td>100</td>
<td>0.721</td>
<td>0.012</td>
<td>0.720</td>
<td>0.014</td>
</tr>
<tr>
<td>0.710</td>
<td>50</td>
<td>0.716</td>
<td>0.009</td>
<td>0.718</td>
<td>0.009</td>
</tr>
<tr>
<td>0.710</td>
<td>20</td>
<td>0.710</td>
<td>0.017</td>
<td>0.711</td>
<td>0.017</td>
</tr>
<tr>
<td>0.710</td>
<td>10</td>
<td>0.713</td>
<td>0.025</td>
<td>0.703</td>
<td>0.027</td>
</tr>
<tr>
<td>0.710</td>
<td>5</td>
<td>0.666</td>
<td>0.040</td>
<td>0.645</td>
<td>0.043</td>
</tr>
<tr>
<td>0.710</td>
<td>2</td>
<td>0.545</td>
<td>0.104</td>
<td>0.519</td>
<td>0.113</td>
</tr>
<tr>
<td>0.197</td>
<td>100</td>
<td>0.195</td>
<td>0.013</td>
<td>0.210</td>
<td>0.014</td>
</tr>
<tr>
<td>0.197</td>
<td>50</td>
<td>0.196</td>
<td>0.012</td>
<td>0.213</td>
<td>0.015</td>
</tr>
<tr>
<td>0.197</td>
<td>20</td>
<td>0.201</td>
<td>0.009</td>
<td>0.202</td>
<td>0.009</td>
</tr>
<tr>
<td>0.197</td>
<td>10</td>
<td>0.191</td>
<td>0.013</td>
<td>0.189</td>
<td>0.011</td>
</tr>
<tr>
<td>0.197</td>
<td>5</td>
<td>0.213</td>
<td>0.021</td>
<td>0.206</td>
<td>0.018</td>
</tr>
<tr>
<td>0.197</td>
<td>2</td>
<td>0.168</td>
<td>0.031</td>
<td>0.159</td>
<td>0.026</td>
</tr>
</tbody>
</table>

Table 4.5 Comparison between $F_{ST}$ estimated using $\hat{M}$ and $\hat{M}_S$ based on 100 populations of 100 diploids. Total Sample size = 100 diploids; top half of table M=0.1, bottom half M=1.0. All values are based on 200 replicates.
<table>
<thead>
<tr>
<th>$\mu$</th>
<th>demes</th>
<th>alleles</th>
<th>$\hat{M}_\theta$</th>
<th>$\hat{M}_\theta$ mse</th>
<th>$\hat{M}_s$</th>
<th>$\hat{M}_s$ mse</th>
<th>$\hat{M}_R$</th>
<th>$\hat{M}_R$ mse</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>27.2</td>
<td>0.110</td>
<td>5.10x10^-4</td>
<td>0.112</td>
<td>4.60x10^-4</td>
<td>0.104</td>
<td>4.60x10^-4</td>
<td>0.104</td>
</tr>
<tr>
<td>20</td>
<td>19.6</td>
<td>0.118</td>
<td>8.85x10^-4</td>
<td>0.125</td>
<td>1.02x10^-3</td>
<td>0.114</td>
<td>6.80x10^-4</td>
<td>0.104</td>
</tr>
<tr>
<td>10</td>
<td>13.9</td>
<td>0.122</td>
<td>2.25x10^-3</td>
<td>0.136</td>
<td>2.59x10^-3</td>
<td>0.120</td>
<td>1.28x10^-3</td>
<td>0.127</td>
</tr>
<tr>
<td>5</td>
<td>9.3</td>
<td>0.146</td>
<td>1.83x10^-2</td>
<td>0.159</td>
<td>8.96x10^-3</td>
<td>0.127</td>
<td>2.82x10^-3</td>
<td>0.127</td>
</tr>
<tr>
<td>2</td>
<td>4.8</td>
<td>0.266</td>
<td>0.302</td>
<td>0.391</td>
<td>0.521</td>
<td>0.168</td>
<td>1.08x10^-2</td>
<td>0.168</td>
</tr>
<tr>
<td>100</td>
<td>5.8</td>
<td>0.104</td>
<td>2.18x10^-3</td>
<td>0.105</td>
<td>2.24x10^-3</td>
<td>0.050</td>
<td>3.62x10^-3</td>
<td>0.050</td>
</tr>
<tr>
<td>50</td>
<td>5.3</td>
<td>0.107</td>
<td>1.67x10^-3</td>
<td>0.111</td>
<td>1.86x10^-3</td>
<td>0.051</td>
<td>3.38x10^-3</td>
<td>0.051</td>
</tr>
<tr>
<td>20</td>
<td>4.5</td>
<td>0.110</td>
<td>4.19x10^-3</td>
<td>0.114</td>
<td>3.27x10^-3</td>
<td>0.056</td>
<td>3.28x10^-3</td>
<td>0.056</td>
</tr>
<tr>
<td>10</td>
<td>3.7</td>
<td>0.130</td>
<td>1.52x10^-2</td>
<td>0.140</td>
<td>1.48x10^-2</td>
<td>0.060</td>
<td>3.18x10^-3</td>
<td>0.060</td>
</tr>
<tr>
<td>5</td>
<td>3.1</td>
<td>0.179</td>
<td>6.98x10^-2</td>
<td>0.169</td>
<td>2.67x10^-2</td>
<td>0.074</td>
<td>2.93x10^-3</td>
<td>0.074</td>
</tr>
<tr>
<td>2</td>
<td>2.4</td>
<td>0.394</td>
<td>50.80</td>
<td>5.786</td>
<td>2500</td>
<td>0.281</td>
<td>1.24</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2.4</td>
<td>0.104</td>
<td>3.76x10^-3</td>
<td>0.106</td>
<td>3.94x10^-3</td>
<td>0.046</td>
<td>4.28x10^-3</td>
<td>0.046</td>
</tr>
<tr>
<td>50</td>
<td>2.3</td>
<td>0.105</td>
<td>2.82x10^-3</td>
<td>0.106</td>
<td>2.67x10^-3</td>
<td>0.047</td>
<td>4.01x10^-3</td>
<td>0.047</td>
</tr>
<tr>
<td>20</td>
<td>2.2</td>
<td>0.122</td>
<td>1.86x10^-2</td>
<td>0.119</td>
<td>1.18x10^-2</td>
<td>0.049</td>
<td>4.08x10^-3</td>
<td>0.049</td>
</tr>
<tr>
<td>10</td>
<td>2.2</td>
<td>0.133</td>
<td>2.43x10^-2</td>
<td>0.123</td>
<td>1.20x10^-2</td>
<td>0.048</td>
<td>4.55x10^-3</td>
<td>0.048</td>
</tr>
<tr>
<td>5</td>
<td>2.1</td>
<td>0.255</td>
<td>0.441</td>
<td>0.219</td>
<td>0.349</td>
<td>0.068</td>
<td>4.59x10^-3</td>
<td>0.068</td>
</tr>
<tr>
<td>2</td>
<td>2.1</td>
<td>38.78</td>
<td>2.97x10^-5</td>
<td>11.032</td>
<td>6790</td>
<td>0.152</td>
<td>9.38x10^-2</td>
<td>0.152</td>
</tr>
</tbody>
</table>

Table 4.6 Values generated for a simulation with 100 demes and a constant sample size of 100 diploids. The values given are averages of 200 replicates.
Fig 4.9 estimates of $M$ from $\theta$ (triangles), $\hat{M}_S$ (circles) and $\hat{M}_R$ (squares) against the number of populations sampled (keeping total sample size = 100 diploids). $\mu = 10^{-4}$ and $M = 0.1$ in all cases. All values are based on the mean of 200 replicates.

The comparison between maximum likelihood estimators and $\theta$ over higher values of $M$ is confused by $\theta$ being frequently estimated to be less than 0, making an estimate of $\hat{M}_\theta$ impossible. For the higher migration rates maximum likelihood estimates for $F_{ST}$ and $\theta$ are considered only.
Table 4.7 Values generated for a simulation with 100 demes and a constant sample size of 100 diploids. The values given are averages of 200 replicates. Cells marked with an asterisk denote where θ values are less than 0 so not allowing the estimation of $\hat{\theta}$.

4.9.5 Estimation of $F_{ST}$ using Maximum Likelihood

The previous section showed that estimates of $M$ using $\hat{\theta}$ and $\hat{M}_S$ are comparable when $M$ is low. As a comparison for higher values of $M$, which are likely to be seen in samples from natural populations, maximum likelihood estimators of $F_{ST}$ will be compared with $\theta$. Tables 4.8, and 4.9 compare the various estimators for $F_{ST}$ against the expected values. The results for $\theta$, and $\hat{M}_S$ are again comparable when the number of populations sampled are large. The mean $\theta$ estimate is closer to the
expected value but the mean squared errors are comparable. Figure 4.10 shows a direct comparison of the mean squared errors for the 4 estimators for a small migration rate.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$F_{ST}$</th>
<th>$\Theta$</th>
<th>$G_{ST}$</th>
<th>$G_{ST}$ from $\hat{M}_s$</th>
<th>$F_{ST}$ from $\hat{M}_R$</th>
<th>MSE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$10^{-4}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.193</td>
<td>3.39x10^3</td>
<td>0.594</td>
<td>1.63x10^2</td>
<td>0.186</td>
<td>3.33x10^3</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.197</td>
<td>1.34x10^3</td>
<td>0.393</td>
<td>1.40x10^2</td>
<td>0.188</td>
<td>1.60x10^3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.191</td>
<td>1.39x10^3</td>
<td>0.261</td>
<td>1.43x10^2</td>
<td>0.180</td>
<td>1.46x10^3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.200</td>
<td>2.24x10^3</td>
<td>0.221</td>
<td>1.76x10^2</td>
<td>0.174</td>
<td>2.69x10^3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.202</td>
<td>3.10x10^3</td>
<td>0.186</td>
<td>1.43x10^2</td>
<td>0.157</td>
<td>3.48x10^3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.190</td>
<td>1.22x10^2</td>
<td>0.114</td>
<td>6.80x10^3</td>
<td>0.098</td>
<td>1.10x10^2</td>
</tr>
<tr>
<td></td>
<td>$10^{-5}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.195</td>
<td>8.75x10^3</td>
<td>0.595</td>
<td>1.61x10^3</td>
<td>0.191</td>
<td>7.54x10^3</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.187</td>
<td>6.04x10^3</td>
<td>0.386</td>
<td>3.89x10^2</td>
<td>0.182</td>
<td>5.94x10^3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.199</td>
<td>7.40x10^3</td>
<td>0.269</td>
<td>1.10x10^2</td>
<td>0.192</td>
<td>6.66x10^3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.197</td>
<td>1.04x10^2</td>
<td>0.219</td>
<td>9.12x10^3</td>
<td>0.189</td>
<td>9.54x10^3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.207</td>
<td>1.45x10^2</td>
<td>0.193</td>
<td>1.12x10^2</td>
<td>0.177</td>
<td>1.09x10^2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.170</td>
<td>2.51x10^2</td>
<td>0.106</td>
<td>1.90x10^2</td>
<td>0.114</td>
<td>2.14x10^2</td>
</tr>
<tr>
<td></td>
<td>$10^{-6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.196</td>
<td>1.47x10^2</td>
<td>0.596</td>
<td>1.63x10^4</td>
<td>0.195</td>
<td>1.33x10^2</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.200</td>
<td>1.25x10^2</td>
<td>0.395</td>
<td>4.63x10^2</td>
<td>0.196</td>
<td>1.17x10^2</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.207</td>
<td>9.29x10^3</td>
<td>0.277</td>
<td>1.35x10^2</td>
<td>0.201</td>
<td>9.19x10^3</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.196</td>
<td>1.27x10^2</td>
<td>0.219</td>
<td>1.09x10^2</td>
<td>0.191</td>
<td>1.26x10^2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.208</td>
<td>1.60x10^2</td>
<td>0.194</td>
<td>1.21x10^2</td>
<td>0.213</td>
<td>2.14x10^2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.187</td>
<td>2.90x10^2</td>
<td>0.119</td>
<td>2.08x10^2</td>
<td>0.168</td>
<td>3.06x10^2</td>
</tr>
</tbody>
</table>

Table 4.8 Estimated $F_{ST}$ values from 200 replicates of simulations with $M=1$. All values based on samples of size 200 for a single loci, and the whole population consists of 100 demes each with 100 diploids.
Table 4.9 Estimated $F_{ST}$ values from 200 replicates of simulations with $M=1$. All values based on samples of size 200 for a single loci, and the whole population consists of 100 demes each with 100 diploids.
Figure 4.10 Estimated $F_{ST}$ mean squared errors compared to number of populations sampled. Based on 200 replicates with total sample size 100 diploids for 100 populations with $N = 50$ diploids and $m = 0.001$, and for $\mu = 10^{-4}$ (top figure), $10^{-5}$ (middle), and $10^{-6}$. Hatched bar is for $0$, solid bar $1/(1+4\hat{M}_p)$, open bar $1/(1+4\hat{M}_s)$ and grey bar for $G_{ST}$. 
The behaviour of $\hat{M}_R$ can be seen in Figure 4.11 when there are only two populations sampled. The values estimated using $\hat{M}_R$ are closer to the expected value of 0.71 than those estimated using the observed gene frequencies. The success of this estimator with small migration rates is explained by the fact that the probability of identity once an allele leaves a deme is small so most of the information in the sample is of the distribution of alleles prior to leaving. This probability is essentially what $\hat{M}_R$ estimates.

![Figure 4.11 Comparison of empirical distribution of, $\hat{M}_R$ (top), and $\hat{M}_S$ (bottom) for 2 samples of size 50 diploids out of 100 populations of 100 diploids with $\mu = 10^{-4}$, and $M = 0.1$. Expected $F_{ST} = 0.71$](image-url)

Expected $F_{ST} = 0.71$
4.9.6 More than One Locus

All the above results have been concerned with a single locus. When more than one locus is sampled there is additional information, however, the loci will not necessarily be independent, as mentioned in Section 4.5.3. In this section the effects of additional loci on estimators of $M$ and $F_{ST}$ will be investigated assuming independent loci. In the previous sections it has been shown that values for $F_{ST}$ and $M$ estimated by $\theta$ and $\hat{M}_S$ respectively are equivalent.

Tables 4.9 and 4.10 illustrate the improvement in estimators when the number of loci sampled is increased, when estimating both $M$ (Table 4.9), and $F_{ST}$ (Table 4.10). The estimators are again practically equivalent. In both cases $\hat{M}_S$ has a larger bias than $M$ estimated from $\theta$. However there is little difference in the mean squared errors. The mean squared error decreases approximately as $1/(\text{number of loci sampled})$.

<table>
<thead>
<tr>
<th>loci</th>
<th>$M$</th>
<th>mean $\hat{M}_\theta$</th>
<th>$\hat{M}_\theta$ MSE.</th>
<th>$M$ from $G_{ST}$</th>
<th>MSE</th>
<th>mean $\hat{M}_S$</th>
<th>$\hat{M}_S$ MSE.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1.141</td>
<td>0.226</td>
<td>0.718</td>
<td>0.129</td>
<td>1.231</td>
<td>0.295</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.091</td>
<td>0.119</td>
<td>0.706</td>
<td>0.120</td>
<td>1.175</td>
<td>0.154</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.086</td>
<td>0.110</td>
<td>0.705</td>
<td>0.119</td>
<td>1.165</td>
<td>0.136</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.057</td>
<td>0.067</td>
<td>0.696</td>
<td>0.119</td>
<td>1.140</td>
<td>0.088</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.035</td>
<td>0.050</td>
<td>0.687</td>
<td>0.122</td>
<td>1.119</td>
<td>0.066</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.027</td>
<td>0.042</td>
<td>0.684</td>
<td>0.123</td>
<td>1.110</td>
<td>0.056</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.022</td>
<td>0.033</td>
<td>0.683</td>
<td>0.122</td>
<td>1.105</td>
<td>0.045</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1.018</td>
<td>0.032</td>
<td>0.681</td>
<td>0.123</td>
<td>1.101</td>
<td>0.043</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1.014</td>
<td>0.025</td>
<td>0.680</td>
<td>0.122</td>
<td>1.093</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Table 4.9 Results for estimation procedures for $M$ from simulated data for 200 replicates for 20 samples of 5 individuals with $\mu = 10^{-5}$ and $M = 0.1$. Simulations on 100 populations.
<table>
<thead>
<tr>
<th>loci sampled</th>
<th>M</th>
<th>$E(F_{ST})$</th>
<th>mean $\theta$</th>
<th>$\theta$ MSE</th>
<th>$F_{ST}$ from $\hat{M}_s$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.1</td>
<td>0.197</td>
<td>0.199</td>
<td>4.05x$10^{-3}$</td>
<td>0.187</td>
<td>3.85x$10^{-3}$</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.197</td>
<td>0.197</td>
<td>2.05x$10^{-3}$</td>
<td>0.185</td>
<td>1.96x$10^{-3}$</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.197</td>
<td>0.197</td>
<td>1.91x$10^{-3}$</td>
<td>0.186</td>
<td>1.83x$10^{-3}$</td>
</tr>
<tr>
<td>5</td>
<td>0.1</td>
<td>0.197</td>
<td>0.198</td>
<td>1.38x$10^{-3}$</td>
<td>0.186</td>
<td>1.28x$10^{-3}$</td>
</tr>
<tr>
<td>6</td>
<td>0.1</td>
<td>0.197</td>
<td>0.200</td>
<td>1.07x$10^{-3}$</td>
<td>0.188</td>
<td>9.92x$10^{-4}$</td>
</tr>
<tr>
<td>7</td>
<td>0.1</td>
<td>0.197</td>
<td>0.201</td>
<td>1.00x$10^{-3}$</td>
<td>0.188</td>
<td>9.18x$10^{-4}$</td>
</tr>
<tr>
<td>8</td>
<td>0.1</td>
<td>0.197</td>
<td>0.201</td>
<td>8.28x$10^{-4}$</td>
<td>0.188</td>
<td>7.80x$10^{-4}$</td>
</tr>
<tr>
<td>9</td>
<td>0.1</td>
<td>0.197</td>
<td>0.201</td>
<td>7.43x$10^{-4}$</td>
<td>0.189</td>
<td>6.95x$10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>0.1</td>
<td>0.197</td>
<td>0.201</td>
<td>6.45x$10^{-4}$</td>
<td>0.189</td>
<td>5.86x$10^{-4}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.024</td>
<td>0.023</td>
<td>7.23x$10^{-4}$</td>
<td>0.021</td>
<td>3.66x$10^{-4}$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>5.47x$10^{-4}$</td>
<td>0.018</td>
<td>2.98x$10^{-4}$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>3.49x$10^{-4}$</td>
<td>0.017</td>
<td>2.26x$10^{-4}$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>2.85x$10^{-4}$</td>
<td>0.016</td>
<td>2.05x$10^{-4}$</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>2.76x$10^{-4}$</td>
<td>0.017</td>
<td>1.98x$10^{-4}$</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>1.95x$10^{-4}$</td>
<td>0.017</td>
<td>1.72x$10^{-4}$</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.024</td>
<td>0.023</td>
<td>1.73x$10^{-4}$</td>
<td>0.017</td>
<td>1.62x$10^{-4}$</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>1.48x$10^{-4}$</td>
<td>0.016</td>
<td>1.63x$10^{-4}$</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.024</td>
<td>0.022</td>
<td>1.37x$10^{-4}$</td>
<td>0.016</td>
<td>1.54x$10^{-4}$</td>
</tr>
</tbody>
</table>

Table 4.10 Results of estimation procedures of $E[F_{ST}]$ from simulated data for 200 replicates for 20 samples of 5 individuals with $\mu = 10^{-5}$. Simulations on 100 populations.
4.10 Discussion

In the introduction to this chapter it stated that there are at least two goals in statistical work: to test models, and to estimate parameters. The former was considered in this chapter. Most work performed on population structure in natural populations has also focused on this problem.

The estimation of levels of gene flow between populations has been the goal of many studies of Atlantic salmon genetic variation (e.g. Elo 1993, Jordan 1992). Provided the assumption that populations are in equilibrium holds, then levels of gene flow provide an easily conceptualised guide to the levels of differentiation seen (Slatkin 1989, 1993). The estimations of levels of gene flow has been considered previously using private alleles (Slatkin 1985a), using a maximum likelihood approach (Wehran & Powell 1987), and using estimators of $F_{ST}$ (Cockerham & Weir 1993). Comparative studies have been made of these estimators (Slatkin & Barton 1989).

In Chapters 2 and 3 $F_{ST}$ was defined as a statistic, which had an expectation over the evolutionary process. For estimation purposes the parameter $M$ was estimated, from which an estimator for $E[F_{ST}]$ could be easily constructed. This chapter has further developed a method for the estimation of gene flow using a maximum likelihood approach which has shown to incorporate Slatkin's private alleles method (Slatkin 1985b).

In a previous chapter it was shown that the expected increase in inbreeding due to the population structure could be approximated by an equation of the form $1/(1+a)$ where $a$ depends on the details of the model. In some of these cases the maximum likelihood approach is a valid one for estimating the degree of population differentiation, and may be related to important ecological and evolutionary parameters. This chapter has also shown that the degree of variability for pairwise measures sampled from populations is high, independent of the model for mutation used. This is because of the method by which the variability is produced, namely a very few coalescent events, and mutations. Even when the variability of the mutations is removed there is still considerable variation in estimators using coalescence times.
Work on all estimators here has shown that sampling a small number of individuals from a large number of sites is likely to provide the most accurate estimates of gene flow. A similar argument using a different maximum likelihood approach (Barton & Wilson 1996) gives the optimum sampling strategy to be to sample n sites where \( \frac{1}{2n} = F_{ST} \). When \( F_{ST} \) is expected to be small, the best sampling strategy is then to sample as many sites as possible, in agreement to the results here.

The simulations in this chapter have shown the maximum likelihood estimation procedure presented here provides an estimator for gene flow which is, when the number of populations sampled is reasonable (>5), at worst comparable to \( \theta \) (Weir & Cockerham 1993), which performed best in the study of Slatkin & Barton (1989).

Analytic results due to Weir and Cockerham (1993) suggest that the simulation routines in Slatkin and Barton (1989) may be suspect, as disagreements with simulation results for \( \theta \) were found. The work in this chapter has investigated maximum likelihood estimators, and the results also disagree with those of Slatkin and Barton (1989). This confirms, to some extent, the reservations of Weir and Cockerham (1993). Possible explanations for this are that the estimator presented here incorporates the sampling errors within populations as well as those caused by genetic drift, and the simulation process guarantees that populations will be in equilibrium.

This chapter has ignored the problem of non-independence of loci even when they are physically unlinked. Most previous estimators for \( F_{ST} \) have also ignored this problem. It is well known (Hartl & Clark 1989) that population subdivision with limited migration between populations can cause linkage-disequilibrium, yet the effects of this on estimates of population structure are unquantified.

This maximum likelihood estimator allows the construction of support curves which give a guide to the precision of estimation, for the data studied. The use of likelihood is potentially of more crucial importance for DNA fingerprinting in forensic science where the presence of population subdivision may alter the match probabilities by orders of magnitude (Nichols & Balding 1991). In this case integrating over a support curve may provide a more conservative procedure for calculating match probabilities than the use of a point estimate (Balding & Nichols 1995).
For fisheries management the actual levels of gene flow may be of less importance than the presence of any population subdivision, as the amount of gene flow which causes significant differentiation in gene frequencies is insignificant in comparison to the amounts of fish that commercial fishermen may catch. The next chapter investigates tests for population subdivision, including a test based on $\hat{M}_S$. 
Chapter 5

THE POWER OF STATISTICAL TESTS OF POPULATION DIFFERENTIATION

5.1 Introduction

In the previous chapter, methods for estimating levels of population differentiation in subdivided populations were developed and discussed. These estimates are important for evolutionary studies. However for fisheries management, particularly in the case of marine teleosts and invertebrates where the population sizes involved can be very large, any differentiation due to limited dispersal is likely to be crucial for management. For example if the population size is large, dispersal of 50 individuals per generation between two populations (or stocks in fisheries management terminology) will not lead to large levels of genetic differentiation at neutral loci, and will not necessarily prevent local adaptation. Nevertheless, this amount of dispersal is significant in ecological terms, and overfishing of one stock will not feasibly be compensated for by dispersal from the other (Pitcher & Hart 1982).

When fishing grounds are isolated from spawning grounds it is also important to know if samples are representative of more than a single reproductively isolated population, and the relative contribution of the two stocks, although this will not be considered here. This is of relevance to Atlantic salmon since sampling from natural populations of Atlantic salmon is usually done from a stretch of river so that any finer structuring is lost. Population subdivision is known to cause a deficit of
heterozygotes in a mixed population, the Wahlund effect (Crow & Kimura 1970), so that samples will show Hardy-Weinberg disequilibrium.

In both of the afore-mentioned cases, knowledge of the power of tests for population differentiation is important for the design of sampling strategies.

Asymptotic tests for both population and Hardy-Weinberg disequilibrium are available. However, the use of exact, or permutation tests, are preferable for genetic data (Raymond & Rousset 1995, Guo & Thompson 1992, Hudson et al 1992). The computer power now available allows us to use exact tests based on the distribution of allele, and genotype, frequencies under the null hypothesis of no differentiation, rather than using asymptotic results. When there is too much data to sample every combination Monte-Carlo techniques can be used (e.g. Guo & Thompson 1992).

In this chapter the power of tests for Hardy-Weinberg disequilibrium, and population differentiation will be considered. Asymptotic results on the estimators constructed in Chapter 4, and the variety of tests for independence of contingency tables (Lancaster 1969), will be used to test for population differentiation. The power of tests for Hardy-Weinberg disequilibrium are known to be low, and the power is dependent not only on the allele frequencies but also on the number of segregating alleles (Chakraborty & Zhong 1994). Rousset and Raymond (1995) computed the power of various tests of Hardy-Weinberg disequilibrium for a variety of test statistics, however, these power tests were based on sampling from simulated data with fixed allele frequencies rather than from a population model.

Here, the powers of tests of Hardy-Weinberg disequilibrium will be calculated for data generated from standard multinomial distributions using known values of the inbreeding coefficients and mean gene frequencies and for data from a mixture of populations generated using a coalescent simulation. The advantages of using coalescent simulations were described in Chapter 4. The power of tests for population subdivision is unknown for samples from natural populations, where the number of alleles is unknown. This will also be examined in this chapter.
5.2 Tests for Hardy-Weinberg Disequilibrium

A variety of tests for Hardy-Weinberg disequilibrium have been devised (e.g. Levene 1949, Li & Horvitz 1953, Chapco 1976). These tests fall into two groups: tests based on asymptotic results (e.g. Chi-squared, and likelihood ratio), and exact tests. Tests based on asymptotic results have been criticised when using data with large number of segregating alleles as the results are known to be suspect when there are small cell numbers (e.g. Rice 1988). However, the approaches can be combined by using conventional Chi-squared, G test statistics, or parameter estimators, and then using an algorithm (e.g. Louis & Dempster 1987) to get the distribution of the statistic under the null hypothesis, and hence obtain significance levels (Chakraborty & Zhong 1994). There are thus a variety of possible tests for Hardy-Weinberg disequilibrium, which can be performed exactly.

If a sample of $m$ diploids is taken from an autosomal locus with $k$ observable alleles then the genotypes can be described as a semimatrix $h$.

\[
\begin{array}{c|c|c|c|c|c}
\hline
 & A_1 & A_2 & \ldots & A_k \\
\hline
A_1 & h_{11} & & & \\
A_2 & h_{21} & h_{22} & & \\
\vdots & \vdots & \vdots & \ddots & \\
A_k & h_{k1} & h_{k2} & \ldots & h_{kk} \\
\hline
\end{array}
\]

Define $h_* = h_{ii} + \sum_{j=1}^{k} h_{ij}$ where $h_{ij} = h_{ji}$ for $j > i$, the total number of alleles of type $i$.

The statistic of interest for an exact test of Hardy-Weinberg disequilibrium is (after Levene 1949) the probability of obtaining $h$ under Hardy-Weinberg equilibrium:

\[
Pr(h) = \frac{m!}{(2m)!} \prod_{i=1}^{k} \frac{h_{ii}!}{\prod_{j=i}^{k} h_{jj}!} \left( \frac{1}{2} \sum_{j=i}^{k} h_{jj} \right)^{\frac{1}{2} \sum_{j=i}^{k} h_{jj}}. \tag{5.1}
\]

An exact test for the observed sample $h$ is to calculate
Chapter 5: POWER OF TESTS OF GENETIC DIFFERENTIATION

\[ P = \sum \Pr(g) \text{ where } \Pr(g) \leq \Pr(h) \text{ and } g \text{ has the same allele counts } h, \text{ as } g. \]

\( P \) gives the p-value of the test. This exact test is simply a test for any departure from Hardy-Weinberg proportions. If we have data where we want to test for a specific departure from Hardy-Weinberg proportions then exact tests can be used on alternative statistics to \( \Pr(n) \). When we suspect that any departure from Hardy-Weinberg proportions is due to the Wahlund effect then inbreeding coefficients for all allele combinations are expected to be the same (see Section 2.3.3.1). In this case the likelihood of a set of data is a function of the population gene frequencies \( p \) and \( f \) (the inbreeding coefficient):

\[
L(h, f, p) = \prod_{i,j} \sum_{j=1}^{k} \left[ (p_i^2 + p_i(1 - p_i)f)^{h_{ij}} + \sum_{j=1}^{k} (2p_ip_j(1 - f))^{h_{ij}} \right].
\]

(5.2)

Thus another test may be performed by estimating the inbreeding coefficient, and then testing significance by comparing the estimator to estimators under the null hypothesis. This can be done by either (i) jointly estimating \( p \) and \( f \), or (ii) using the observed allele frequencies \( \hat{p}_i = \left( h_i + \sum_{j=1}^{k} h_{ij} \right) / 2m \), and then estimating \( f \) (as in Chapter 4 for \( \tilde{M} \)). This estimator is designated \( fA \) in Rousset and Raymond 1995, and this notation will be used here. Another test can be performed by using the score of the likelihood (e.g. Stuart & Ord 1991) which is equivalent to \( f_T \) of Robertson and Hill (1984), and is denoted by \( U \) here.

These tests can be performed exactly by obtaining the value of the statistic, or estimator \( T \) for \( h \), where large \( T \) is evidence against the null hypothesis (for a one tailed test) and then calculating

\[ P_T = \sum \Pr(g) \text{ where } T(g) \geq T(h) \text{ and } g \text{ has the same allele counts } g, \text{ as } h. \]

The number of possible semimatrices \( (g) \) increases exponentially with the number of alleles (Hernandez & Weir 1989), and for moderate sample sizes with large numbers of alleles there are too many to make the exact test computationally feasible. This problem can be overcome by approximating the distribution of possible \( g \)'s by taking samples proportional to their probability. In conventional Monte-Carlo exact tests
semimatrices are generated at random using the same marginal allele counts \( (h_*) \) as \( h \). Guo and Thompson (1992) suggest a Markov chain method, sampling from a stationary Markov chain which has the same marginal distribution. This is less computer intensive for data with large number of alleles, but there are problems ensuring that successive samples are independent. Genepop (Raymond & Rousset 1995) uses this algorithm.

The simpler algorithm used in this chapter to generate semimatrices with the same marginal distribution as \( h \) is to randomly permute a linear array of length \( 2r \) constructed from the marginal totals \( h_* \).

The new semimatrix \( m \) is created by taking successive pairs from the permuted array and allocating them to a genotype frequency table. Hence the \( h_* \)'s are conserved.

An algorithm for permuting semimatrices is given in Appendix 7. This method is used here because the standard errors of the p-values can be calculated, as successive values will not be correlated (if the pseudo-random numbers used are uncorrelated).

The results of Rousset and Raymond (1995) suggest that the most powerful tests are those based on (using their notation) either \( f_A \) or \( U \), and these tests will be used for testing simulated data. Tests were performed to provide a check for the algorithms in Appendix 7 against those by Raymond and Rousset (1995) and Guo and Thompson (1992).

5.2.1 Power for 2 Allele Tests

As a first test of the power of these alternate methods samples are generated from a multinomial distribution

\[
M_n(m, p^2(1 + F), 2p(1 - p)(1 - F), (1 - p)^2 (1 + F)).
\]

The power of tests were calculated by:

1) varying the mean gene frequency \( p \),
2) varying the inbreeding coefficient F, and

3) altering the sample size.

Figure 5.1 illustrates the power of two tests, at a 5% significance level for F varying from -0.3 to 0.3. The exact test is less powerful than either 1-tailed test. Powers of tests based on U, and $f_A$ were identical over this range. However, when the mean gene frequency of the most common allele is high, U is more powerful than $f_A$, and has equal power to an exact probability test, shown in Figure 5.2. Figure 5.3 indicates that the power of tests increases approximately linearly with the sample size, over the range of sample sizes here.

Figure 5.1 Power of exact test (diamonds, line with long dashes), and exact U tests for heterozygote excess (triangles, dotted line), and deficit (circles, dashed line) for varying F (at a 5% significance level). Values given as mean ± s.e. Means based on 5000 replicates for p = 0.5, and sample size $m = 100$. Horizontal line indicates the probability of a type 1 error (0.05). Tests based on $f_A$ and U gave almost identical results; tests based on U are given here.
Figure 5.2 Power of exact test for Hardy-Weinberg proportions (diamonds, line with long dashes), and exact tests for heterozygote deficit: U (circles, dashed line), and $f_A$ (down triangles, dotted line) varying the frequency of the most common allele. All tests at a 5% significance level. Values, given as mean ± s.e., based on 5000 replicates for $F = 0.1$ and sample size $n = 100$. Solid line indicates the probability of a type 1 error (0.05).
5.2.2 Power for Simulated Data

The above tests are for data when the values are drawn from known distributions. When we consider data with an unknown population structure then we cannot know in advance the number of alleles or the allele frequencies. The coalescent process, discussed in Chapter 3, is a good method for generated simulated data. In this section data are generated for a set of subpopulations, the alleles within each subpopulation are paired at random to simulate diploids, and the subpopulations merged. Tests for Hardy-Weinberg disequilibrium are performed on this combined data. In this case the inbreeding coefficient $F = E[F_{ST}]$. 

---

Figure 5.3. Power of exact test (diamonds), and U for heterozygote deficit (circles) for varying sample size (at a 5% significance level). Values based on 5000 replicates for $p = 0.5$, and $F = 0.2$ (solid symbols) and 0.1 (open symbols). Values shown as mean ± s.e. Solid line indicates the probability of a type 1 error (0.05).
The influence of sample size, and mean gene frequency were examined in the previous section. Here we will look at the power of tests for Hardy-Weinberg proportions in the following circumstances:

- for varying mutation rate, and
- where the number of populations sampled varies.

These are of most interest, as they represent the questions which must be asked prior to a survey of natural populations. It has been suggested that highly variable loci are likely to uncover more variation, and hence more differentiation than allozyme loci (e.g. O'Connell 1993), but a question of interest is how much more powerful the tests are, and whether using more traditional techniques on larger number of individuals is likely to be more effective?

Table 5.1 gives the power of tests for population structure with a varying mutation rate. In this case tests based on $U$ were somewhat more powerful than exact tests, and tests based on $f_A$. Altering the mutation rate did not have a large effect on the power of tests.

<table>
<thead>
<tr>
<th>mutation rate</th>
<th>Pr</th>
<th>U</th>
<th>$f_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-3}$</td>
<td>0.616</td>
<td>0.660</td>
<td>0.582</td>
</tr>
<tr>
<td>$5 \times 10^{-4}$</td>
<td>0.484</td>
<td>0.484</td>
<td>0.484</td>
</tr>
<tr>
<td>$2.5 \times 10^{-4}$</td>
<td>0.370</td>
<td>0.386</td>
<td>0.382</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>0.274</td>
<td>0.286</td>
<td>0.254</td>
</tr>
<tr>
<td>$5 \times 10^{-5}$</td>
<td>0.250</td>
<td>0.258</td>
<td>0.202</td>
</tr>
<tr>
<td>$2.5 \times 10^{-5}$</td>
<td>0.222</td>
<td>0.222</td>
<td>0.156</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>0.176</td>
<td>0.188</td>
<td>0.120</td>
</tr>
</tbody>
</table>

Table 5.1 Power of tests for Hardy-Weinberg proportions. Values based on 500 replicates with Monte-Carlo tests based on 1000 permutations. Simulations based on 10 populations of 100 diploids, sample size 10 diploids from each of 5 populations with $M = 1$. Tests at a 5% significance level.

The power of tests increases with the number of populations in the mixture when the sample size is constant (Figure 5.4) (a single population was also tested to check the
type-1 errors), so that sampling from a large number of populations is more likely to produce significant deviations from Hardy-Weinberg proportions. Exact tests had higher power when a mixture of few populations was sampled, but tests based on $U$ were more powerful when large numbers of populations were sampled. $f_A$ again proved to be the least powerful estimator.

![Figure 5.4](image)

Figure 5.4. Power of tests for Hardy-Weinberg proportions based on 10 populations of 100 diploids with $M = 1$, and $\mu = 10^3$, and a total sample size of 50 diploids (open symbols), and 100 diploids (solid symbols). Values based on 500 replicates, with 1000 permutations per sample. Powers of exact test (diamonds, line with long dashes), tests using $U$ (circles, line with short dashes), and $f_A$ (triangles, dotted line) are shown at a 5% significance level. Solid line indicates the probability of a type 1 error.

### 5.3 Power of Tests of Genetic Differentiation

Chi-Squared, or G-tests have been used frequently for testing population differentiation. However, as mentioned in Section 5.1, with many alleles at a locus, the numbers in some cells is likely to be small so asymptotic results cannot be used with certainty (Rice 1988). The use of exact, or permutation tests is a solution to
this. In this section different permutation tests for population structure will be considered

5.3.1 Test Statistics

5.3.1.1 Definitions

Consider a sample from \( r \) subpopulations from a single loci with \( k \) alleles \( A_1 \ldots A_k \) and a sample size of \( n_i \) genes from population \( i \). In this case the data is of the form:

<table>
<thead>
<tr>
<th>Deme 1</th>
<th>Deme 2</th>
<th>Deme k</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
<td>( x_{1k} )</td>
</tr>
<tr>
<td>( x_{21} )</td>
<td>( x_{22} )</td>
<td>( x_{2k} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( x_{ri} )</td>
<td>( x_{r2} )</td>
<td>( x_{rk} )</td>
</tr>
</tbody>
</table>

where \( x_{ij} \) denotes the number of alleles of type \( j \) in subpopulation (or deme) \( i \), as in Chapter 4.

5.3.1.2 Chi-Squared Tests

Chi-Squared goodness-of-fit tests for population subdivision are simply tests of independence for contingency tables (Rice 1988). Since we are performing permutation tests corrections for small numbers in cells will not be used, just Pearson’s statistic:

\[
X^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(x_{ij} - E_{ij})^2}{E_{ij}},
\]

where \( E_{ij} = \frac{x_{\ast j} x_{i\ast}}{x_{\ast \ast}} \), \( x_{i\ast} = \sum_{j=1}^{k} x_{ij} \), \( x_{\ast j} = \sum_{i=1}^{r} x_{ij} \), and \( x_{\ast \ast} = \sum_{i=1}^{r} n_i \).

5.3.1.3 G-Test

The G-test for population subdivision is simply the likelihood ratio test (Sokal & Rohlf 1995):
\[-2\log \Lambda = \sum_{i=1}^{r} \sum_{j=1}^{k} x_{ij} \log \left( \frac{x_{ij}}{E_{ij}} \right), \quad (5.4)\]

where \(E\) is as described for Chi-Squared tests.

### 5.3.1.4 Fisher’s Exact Test

For this test conditional inference is used to remove nuisance parameters. The probability of observing a sample \(y\) under the null hypothesis of no differentiation with the same marginal totals \(x_{i*}\) and \(x_{*j}\) as the sample \(x\) is:

\[Pr[y] = \frac{\prod_{j=1}^{k} x_{i*} \prod_{i=1}^{r} x_{*j}!}{x_{*j}! \prod_{i=1}^{r} \prod_{j=1}^{k} y_{ij}!}. \quad (5.5)\]

In this case the statistic of interest is \(Pr[x]\), and its significance is tested by calculating \(P = \sum Pr[y]\) where the set of all possible \(y\)’s is taken with the same marginal totals \(x_{i*}\) and \(x_{*j}\).

Now if we consider a Monte-Carlo test then keeping the marginal totals equal the only changing term is \(d = \left( \prod_{i=1}^{s} \prod_{j=1}^{k} y_{ij}! \right)^{-1}\) so that \(-\sum_{i=1}^{s} \sum_{j=1}^{k} \log(y_{ij}!))\) may be used as a test statistic.

### 5.3.1.5 Maximum Likelihood Estimate for Gene Flow

The maximum likelihood estimator \(\hat{M}_s\) defined in Chapter 4 is a suitable test statistic for a permutation test, analogous to \(f_A\) of Section 5.1.

### 5.3.1.6 Weir and Cockerham’s \(\theta\)

Permutation tests can also be performed, calculating \(\theta_x\) for the data, and comparing this to \(\theta_x\) calculated from permutations of the data, both for a single locus, and combining all loci.

All of the above tests will be used as permutation tests, i.e. the values will be found for the observed data, and then these data will be permuted assuming that the null hypothesis (there is no population subdivision) is true. The maximum likelihood test
is very computer intensive, and will be done only for selected cases. Algorithms for permutation are given in Appendix 7.

5.3.2 Power of Tests for Simulated Data

All tests were performed for preliminary studies of 100 Monte-Carlo permutations per sample.

The power of tests for population subdivision increased with the mutation rate, although this was more apparent for exact $\chi^2$, G tests and tests based on $\hat{M}_s$ than for tests based on $\theta$. The power of tests using a single locus were low for all mutation rates when only two populations were sampled (Table 5.2).

Increasing the number of populations sampled, when the total sample size remained constant, also increased the power, and again $\theta$ had lower power than other tests. $\hat{M}_s$ performed well in both cases, but because of the large amount of computer time used by this test only Pr, $\theta$, $\chi^2$ and G tests were considered further.

<table>
<thead>
<tr>
<th>mutation rate</th>
<th>$\theta$</th>
<th>Pr</th>
<th>$\hat{M}_s$</th>
<th>$\chi^2$</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^{-3}$</td>
<td>0.094</td>
<td>0.11</td>
<td>0.122</td>
<td>0.106</td>
<td>0.11</td>
</tr>
<tr>
<td>$5 \times 10^{-4}$</td>
<td>0.076</td>
<td>0.084</td>
<td>0.096</td>
<td>0.084</td>
<td>0.086</td>
</tr>
<tr>
<td>$10^{-4}$</td>
<td>0.044</td>
<td>0.048</td>
<td>0.07</td>
<td>0.046</td>
<td>0.046</td>
</tr>
<tr>
<td>$5 \times 10^{-5}$</td>
<td>0.054</td>
<td>0.056</td>
<td>0.09</td>
<td>0.05</td>
<td>0.052</td>
</tr>
<tr>
<td>$10^{-5}$</td>
<td>0.042</td>
<td>0.046</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>$5 \times 10^{-6}$</td>
<td>0.04</td>
<td>0.044</td>
<td>0.056</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 5.2. Powers of all tests for population differentiation for a 2 population model with $m=0.05$, and a population size of $N=100$ diploids per population. Samples of 10 diploids taken from each population. All values based on a 5% significance level for 500 replicates, with 100 permutations per sample.
Table 5.3. Powers of all tests for population differentiation for a 5 population model with m=0.05, μ = 10^{-4} and population size of 100 diploids per population. All values based on a 5% significance level for 200 replicates, with 100 permutations per sample.

The power of tests to differentiate populations where M is up to 20 per generation is important. Table 5.4 gives the power of tests for varying M for a single locus sampled. The power at high levels of M (>2) is low, and not much greater than the type-I error. Thus testing for population subdivision under these circumstances using a single locus, or testing loci individually is unlikely to uncover subdivision.

Table 5.4 Power of tests compared with M, the effective migration rate. Based on 500 replicates with 500 permutations per sample for 2 populations of 100 diploids = 100 with sample size 50 diploids per population and μ = 10^{-4}. Significance levels are at a 5% significance level.

The sample size per population affects the power of tests approximately linearly when the power is low (Figure 5.5). Even for large sample sizes per population the
power is still low when \( M \) is 5 per generation. The power of tests based on \( \theta \) is lower than exact tests.

The influence of mutation rate on powers of tests for population differentiation is substantial. Figure 5.6 illustrates this. At extremely high mutation rates, equivalent to \( N\mu = 1 \), the power of tests approaches 1, but at smaller mutation rates the power is considerably less. This value for \( N\mu \) may be approximately what is seen for mini-, or micro-satellites, which thus compare favourably with allozymes.

![Figure 5.5 Power of exact test (circles), and theta permutation test (triangles). All tests based on 500 replicates with 500 permutations per replicate. Powers for 2 populations with 100 diploids per population with \( m = 0.01 \) (solid symbols), and 0.05 (open symbols). Mutation rate = \( 10^{-3} \). All powers at a 5% significance level. Values given as means ± s.e.](image-url)

---

Chapter 5: POWER OF TESTS OF GENETIC DIFFERENTIATION
Figure 5.6 Power of tests for genetic differentiation with varying mutation rates for 500 samples with 500 permutations per sample, based on 5 populations with 10 diploids per population. M = 0.5. Tests based on Pr (diamonds), $\chi^2$ (triangles), and $\theta$ (circles). Values given as means ± s.e.

5.4 Discussion

The detection of distinct populations is important for fisheries management (Pitcher & Hart 1982). Large values of M (equal to Nm), while not necessarily evolutionarily important, are undoubtedly important for management of natural fish populations. The design and implementation of surveys of natural populations require knowledge of the power of statistical analyses to uncover any differentiation. This chapter has described computer programs developed to produce simulated estimates for the power of tests for population differentiation and Hardy-Weinberg proportions, and given examples of the power of tests for varying mutation rate, and sampling strategy.

Previous calculations of the power of tests of Hardy-Weinberg proportions have investigated the performance of tests for known numbers of alleles and allele
frequencies (Raymond & Rousset 1995, Chakroborty & Zhong 1994). In this chapter estimates for the power of tests have been made for unknown allele frequencies. Chakraborty and Zhong (1994) hypothesised that departures from Hardy-Weinberg proportions would be easier to detect with highly mutable STR loci; the work presented here has shown this to be the case, with the power of tests for Hardy-Weinberg proportions increasing with the mutation rate. An extension of the approach here, using a k-allele model to simulate STR loci (Slatkin 1995, Valdes et al. 1993), would provide a more comprehensive account. Of the tests presented here, the test based on the score of the likelihood is more powerful than the exact test for simulated data from population admixture. This is similar to the findings of Rousset & Raymond (1995), although on the parameter values they investigated, $f_A$ was also more powerful than exact tests.

When sampling a mixture of several genetically isolated populations, tests of Hardy-Weinberg proportions are more likely to be significant if the mixture is from many populations.

Little attention has been paid previously to the power of tests for population subdivision. Here the power of Fisher’s exact test, $\chi^2$ tests, G-tests, and also tests based on the parametric estimation of $\theta$ (Weir & Cockerham 1994), and $\hat{M}_s$ (described in the previous chapter) have been investigated.

The estimator described in Chapter 4, $\hat{M}_s$, performed well used as a permutation test for population differentiation, although it is highly computer intensive. Tests based on $\theta$ (Weir & Cockerham 1984) had lower power than exact tests and also tests based on $\hat{M}_s$, in all cases. The power of tests for population differentiation are influenced strongly by mutation rates, with tests performed on highly variable loci being far more powerful.

In the last chapter, it was shown how the precision of estimators of $F_{ST}$ was best when a large number of populations were surveyed, even if the number per population is small. This chapter has confirmed that as a sampling strategy for detecting population differentiation, sampling small numbers of individuals from many populations will detect population subdivision more readily than large sample sizes from few populations.
The results of this chapter combined with the previous imply that surveys of natural populations should attempt to sample from the maximum number of populations, or age groups possible, in order to test for population subdivision.

When testing for an effect which will be constant across loci, such as the Wahlund effect, or differentiation due to limited dispersal between populations, it should be possible to test across several loci by combining estimators across loci, and, permuting to obtain the distribution of the statistic (Raymond & Rousset 1995). However, the influence of a single anomalous loci (due to selection) may unduly influence the statistic. Investigation of such tests was not undertaken here because of computational constraints.
Chapter 6

THE STRUCTURE OF ATLANTIC SALMON POPULATIONS

6.1 Introduction

The range of genetic data on population subdivision in Atlantic salmon (Salmo salar) was introduced in Chapter 1, with an overview of evidence for local adaptation. Chapter 2 showed how the branching pattern of a river system was likely to affect the levels of genetic differentiation seen within river systems. Chapter 3 indicated how the extent of variation in samples from different geographic locations could be explained by a number of different population genetic models, and Chapter 4 presented a method for calculating levels of gene flow. In this chapter the results of previous chapters will be applied to selected data on genetic structure of Atlantic salmon populations within river systems, and compared to direct measures of gene flow from tagging experiments on the Aberdeenshire Dee, Scotland.

Heterozygosity observed in allozymes for Atlantic salmon is known to be low (see Chapter 1 for refs.). Estimation of pairwise $F_{ST}$ values was shown to be particularly poor with low mutation rates, such as those seen for allozymes, in Chapter 4. It has been argued (O’Connell 1993) that this low level of heterozygosity indicates that the levels of differentiation observed must be low; and that loci with more resolving power will show greater levels of differentiation. As shown in Chapter 4 the mean levels of genetic differentiation (as measured by Weir and Cockerham’s $\theta$ (1984) or
N\textsubscript{m}) are almost independent of the mutation rates, however our ability to obtain precise estimates is severely compromised as the mean squared error of estimates are improved by using more variable loci. However the levels of differentiation are due to the underlying genealogical structure of the samples, which is independent of the mutation rates.

Other problems are caused by lack of equilibrium. When gene frequencies are not in equilibrium, due to range expansion, or extinction/recolonisation then we may underestimate the true rates of dispersal between populations. However, \textit{F}\textsubscript{ST} approaches an equilibrium faster than the probabilities of identity by descent within and between populations (Chapter 3, and Latter 1973).

Genetic data on population structure presented here is in two parts: allozyme variation for various numbers of loci and mitochondrial DNA (hereafter mtDNA). These two types of genetic data should give related, but not identical information. Allozyme variation is a result of the interactions of selection, random drift, mutation, assortative mating, and dispersal for both sexes whereas mtDNA, which is maternally inherited, gives information about dispersal and random drift (and possibly selection) in females. Differences between mtDNA variation, and allozyme variation should be informative about any sex-based life history differences.

Sampling points for population structure on rivers are primarily on tributaries away from the main stem (exceptions are Heggberget \textit{et al} 1986, Elo \textit{et al} 1994 which show differentiation along the main stem) because of the practical difficulties associated with sampling from wide and deep water, and because of conflict with fishing interests. Therefore, most information is about genetic differentiation in peripheral populations, and rather less on differentiation in the main stem.

A complicating factor in the population genetic models produced in this thesis is that the models depend on non-overlapping generations, which is untrue for the Atlantic salmon which has a complicated life history. A model for the coalescent with overlapping generations is presented.

In this chapter previously constructed models are used in light of data on genetic variation and ecological data from rivers in Scotland, and Europe. The questions that need to be addressed are:
1. Are the ecological estimates for dispersal, population sizes and life history parameters consistent with the data on allozyme electrophoretic, and mitochondrial DNA variation?

2. What management implications do the levels of genetic differentiation have?

The first question will be covered in this chapter. Question 2 will be discussed in the final chapter.

6.2 $F_{ST}$ with overlapping generations

All models produced in this thesis have concentrated on non-overlapping generations, with individuals breeding only in a single year. Atlantic salmon have overlapping generations. Female salmon generally breed only once, with a small percentage returning to sea after spawning, and returning to freshwater a second time (Mills 1989). Males are more complicated, with precocious maturation in freshwater occurring at a variable frequency throughout the species range (Randall et al. 1986). Further difficulties are caused by the resident parr not generally dispersing away from their natal river, but dispersing widely within their natal river (D. Hay, Alan Youngson pers. comm.). The contribution to the gene pool of precocious parr has been estimated (Myers 1984) and these fish are thought to increase the effective population size (e.g. Jordan 1992). The effective population size for overlapping generations has been calculated previously (e.g. Felsenstein 1971, Hill 1972, Charlesworth 1980) but with variable dispersal rates the effective population size does not give an adequate description of the amount of inbreeding observed.

A simple definition for $F_{ST}$ in terms of the probability of 2 genes sampled at random coalescing prior to leaving the deme was described in Section 3.3.2, and this will be used as a basis for calculating $F_{ST}$ in populations with overlapping generations.

Consider purely maternally inherited genes first. Define $p_i^F$, and $m_i^F$ to be the probabilities that: a 0 year old juvenile salmon is the progeny of a female salmon of age i; and that a female salmon returning at age i is not natal to the subpopulation respectively, where $i=1,...,L$, and L is the maximum generation time. Let $N^F$ be the
population size (such that the probability of coalescence is $1/N^F$ in any spawning year). Then we obtain the following set of equations for the probability of identity within a deme (assuming that the probability of identity for dispersing genes is 0):

$$f_0^F = \frac{1}{N^F} + \frac{N^F - 1}{N^F} \sum_{j=1}^{L} p_j^F (1 - m_j^F) f_j^F$$

and

$$f_k^F = \sum_{j=1}^{L} p_j^F (1 - m_j^F) f_{[k-j]}^F$$

where $f_k^F$ is the probability of identity by descent for genes which are in the spawning population k years apart. If $p_i^F = 1/L \forall i$ and $m_i^F = m \forall i$ then $f_0^F = (1-m(L-1))/(1+Nm(1+L))$. Numerical solutions for Equations 6.1 are given in Figures 6.1, and 6.2 and Table 6.1. With 2 age classes and equal dispersal rates for both age classes, the expected value of $F_{ST}$ seen decreases, with the maximum decrease seen when most return after 2 years (Figure 6.1).

![Figure 6.1. Expected $F_{ST}$ values for population with 2 overlapping generations. Population size is 100 haploids with migration rates $(m_1,m_2)$ of (0.01,0.01) solid line, (0,0.1) line with long dashes, (0.1,0.01) short dashes, and (0.01,0.1) dotted line.](image-url)
The presence of a year class which does not disperse increases the expected values of $F_{ST}$ seen, but as shown in Figure 6.2, if the proportions in the 2 year classes in the same then the value of $F_{ST}$ expected is relatively unaltered. Table 6.1 shows that as the number of different year classes increases, so then the value of $M$ expected increases for the same dispersal rate per individual. However if straying is less intense for individuals which breed later, then the value of $M$ seen decreases.

![Figure 6.2](image)

**Figure 6.2 Expected values of $F_{ST}$ against the dispersal in year 1 for a maximum generation time of 2.** Total dispersal rate is 0.01, and population consists of 100 haploids with proportions returning in years 1, and 2 $(p_1,p_2)$ of $(0.5,0.5)$ solid line, $(0.3,0.7)$ line with long dashes, $(0.7,0.3)$ line with intermediate dashes, $(0.95,0.05)$ short dashes, and $(0.01,0.1)$ dotted line.

For an (extreme) simplification for female Atlantic salmon in the River Dee: assuming a maximum generation time of 6 years and return equally likely in years 4, 5, and 6 then the effective dispersal rate ($M$) is increased by a factor of 4.5, compared to the population sizes and dispersal rates seen in a single year.
Table 6.1 Expected values for the effective migration rate $M$ for overlapping generations.

These figures give a baseline for the values of $M$ expected for maternally inherited mtDNA. When considering variation in diploid nuclear DNA, it is simple to extend the model if there is no differential breeding success between males of differential ages; we simply consider males and females separately. As we go back in time, every time a line-of-descent passes through the spawning population it goes through a female and male line with equal probability (however the amount of time spent by lines-of-descent in male and females may be different), and the probability that a pair of genes are both the descendent of a gene in a male is 1/4, and the same for females. In this case, the equations to be solved are:

$$f_0 = \frac{1}{8N^M} + \frac{1}{8N^F} + \left(1 - \frac{1}{8N^M} - \frac{1}{8N^F}\right) \sum_{j=1}^{L} \left(\frac{p_j^M (1-m_j^M)}{2} f_{j-k} + \frac{p_j^F (1-m_j^F)}{2} f_{j-k}\right)$$

and

$$f_k = \sum_{j=1}^{L} \left(\frac{p_j^M (1-m_j^M)}{2} f_{j-k} + \frac{p_j^F (1-m_j^F)}{2} f_{j-k}\right)$$

(6.2)
where $N^M$, and $N^F$ are numbers of male and female salmon in the spawning population and $p^M$, $p^F$, $m^M$, and $m^F$ are as defined above for males and females respectively.

With precocious male Atlantic salmon parr there are added complications.

- Precocious males may breed in more than one year (Garcia de Leaniz 1990), so that coalescence between two genes does not necessarily imply that they are the offspring of the same individual in a single year as assumed in previous models.

- Some precocious males also migrate to sea. However, the mortality rate of precocious maturers is known to be higher than for male parr which do not mature precociously, through disease, and also injury through conflict with adult male salmon (Garcia de Leaniz 1990).

- There may be more precocious parr in the system than adult males. However, their individual contribution is likely to be small.

- There is evidence that precocious maturers migrate within river systems (Youngson et al 1994, Buck & Youngson 1982 and D. Hay pers. comm.).

The variation in male life history strategies is likely to be due to growth rates, and relative rates of mortality in sea water and freshwater (Simpson 1993, Randall et al. 1986), and are, at least partially, genetically determined (Thorpe et al. 1983) The reproductive success of strategies will also depend on the strategies of other members of the population. An evolutionary stable strategy (Maynard Smith 1982) approach has been used to model the evolution of life history strategies (Myers 1986), an approach beyond the scope of this thesis.

Assume that all precocious parr die before returning again as adults, and that precocious males spawn in year 1, and adults in subsequent years, with success for precocious parr and adult salmon of $\nu$, and $(1-\nu)$.

The model can then be extended to two sexes with effective population sizes $N^M_{parr}$, $N^M_{adult}$, and $N^F$ for the numbers of mature male, mature male adult salmon, and mature female salmon, and the vectors $p^M$, and $m^M$ are changed to take account of the differential success of precocious males and females, so that if the probability that a male matures in year $i$ is $\rho$, then:
Chapter 6: ATLANTIC SALMON POPULATIONS

\[ p^M_k = \frac{\rho_1 \nu}{\rho_1 \nu + (1 - \nu) \sum_{i=2}^{L} \rho_i} \quad k = 2, \ldots, L, \text{ and} \]

\[ p^M_k = \frac{\rho_k \nu}{\rho_1 \nu + (1 - \nu) \sum_{i=2}^{L} \rho_i} \quad k = 2, \ldots, L. \]

Thus the following set of equations are obtained:

\[
\begin{align*}
    f_0 &= \frac{1}{4} \left( \frac{\nu^2}{2 N_{\text{parr}}^M} + \frac{(1 - \nu)^2}{2 N_{\text{adult}}^M} \right) + \frac{1}{8 N^F} \\
    &\quad + \left( 1 - \frac{1}{4} \left( \frac{\nu^2}{2 N_{\text{parr}}^M} + \frac{(1 - \nu)^2}{2 N_{\text{adult}}^M} \right) - \frac{1}{8 N^F} \right) \sum_{j=1}^{L} \left( \frac{p^M_j (1 - m^M_j) f_{[j-k]}}{2} + \frac{p^F_j (1 - m^F_j) f_{[j-k]}}{2} \right), \\
    f_k &= \sum_{j=1}^{L} \left( \frac{p^M_j (1 - m^M_j) f_{[j-k]}}{2} + \frac{p^F_j (1 - m^F_j) f_{[j-k]}}{2} \right). \quad (6.3)
\end{align*}
\]

Hence the effect of precocious maturation is to decrease the probability of coalescence in any year if \( \frac{2 N_{\text{parr}}^M}{N_{\text{adult}}^M + 2 N_{\text{parr}}^M} > \nu \), and if precocious parr survive to maturity this will further decrease the probability of coalescence in any year.

For any reasonable model there will be a functional relationship between \( \nu, \rho_1, N_{\text{parr}}^M, \text{ and } N_{\text{adult}}^M \) but this is also beyond the scope of this thesis. The model presented above does give an appreciation of the qualitative effects of precocious parr on the amount of inbreeding seen within any population.

Table 6.2 gives expected values of \( F_{ST} \) for a variety of parameter values when the maximum generation time, \( L \), is 2. The only large effects on the values of \( F_{ST} \) expected are when a large contribution to spawning is made by precocious parr, and they disperse at a greater rate than the general population, the effects of other variables are minor.
### Table 6.2

<table>
<thead>
<tr>
<th>( \nu )</th>
<th>( p_1 )</th>
<th>( P_1^F )</th>
<th>( m^M )</th>
<th>( m^F )</th>
<th>( N^M_{\text{parc}} )</th>
<th>( N^M_{\text{adult}} )</th>
<th>( N^F )</th>
<th>( E[F_{ST}] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.123</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>(0.0,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.131</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>(0.1,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.082</td>
</tr>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>40</td>
<td>40</td>
<td>50</td>
<td>0.123</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.123</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>(0.0,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.125</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>(0.1,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.095</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>40</td>
<td>40</td>
<td>50</td>
<td>0.011</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.083</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>(0.0,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.128</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>(0.1,0.01)</td>
<td>(0.0,0.01)</td>
<td>750</td>
<td>40</td>
<td>50</td>
<td>0.021</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
<td>(0.01,0.01)</td>
<td>(0.0,0.01)</td>
<td>40</td>
<td>40</td>
<td>50</td>
<td>0.120</td>
</tr>
</tbody>
</table>

Table 6.2: Expected values of F_{ST} over the evolutionary process with precocious maturation. Values from Equations 6.3.

### 6.3 The Aberdeenshire Dee

#### 6.3.1 Description

The Aberdeenshire Dee (shown in Figure 6.3) is one of Scotland’s major salmon rivers. The river’s main stem is c. 126 km in length and the catchment drains an area of c. 2,100 km². There are seventeen major tributaries. The river drains through both granite (acid) and basic (limestone) rocks, with a mean gradient in the main stem of 8.9 m. km⁻¹. Suitable spawning habitat is found throughout the system with spawning taking place from just above the tidal limit to altitudes of 500 m. The river rises in the Highlands, and drains through moorland, where the main activities are deer stalking, grouse shooting, sheep farming, tourism, and angling. In the lower part of the river it flows through cultivated arable land. This provides a variety of habitats for juvenile salmon with a range of productivity (Shackley & Donaghy 1992).
There is differentiation in life history traits between salmon in the upper part of the river, e.g. in the tributaries Geldie, and Baddoch, and in the lower tributaries such as the Sheeoch (J. Webb, A. McLay, E. Verspoor, A.F. Youngson and D. Hay pers. comm.). Salmon typically spawn earliest in the upper parts of the river, with spawning beginning progressively later further downstream (Webb & McLay unpubl. manuscript). Hatching of ova is also structured within the system, with hatching in the lower tributaries occurring earlier, possibly due to differences in water temperature which varies with altitude (Shackley & Donaghy 1992).
These differences are also related to the times of return of fish into the river. Multi-
sea-winter fish and grilse destined to spawn high in the river typically enter the river
between December and May, and May and June respectively. Fish spawning in
lower tributaries are later running (Hawkins & Smith 1986).

Information on return times, and numbers of spawners has been obtained from fish
traps positioned on the Girnock Burn since 1967 (Youngson et al 1994, Buck & Hay
1984, Buck & Youngson 1982), and more recently on the Baddoch Burn.

The Girnock Burn (shown on Figure 6.3) is approximately 9.5 km long, and drains
an area of c. 28 km². The trap on the Burn is 700 metres from the confluence of the
Burn with the main stem. There are c. 7 km of spawning habitat above the fish trap
(A.F. Youngson pers. comm.). The production of smolts from the Girnock is c.
3000 per year, and the number of returning adults is of the order of 50-100 with
around 30% of the spawners coming back as grilse, and the majority of the
remainder as 2 sea-winter salmon (Youngson et al 1994, A. F. Youngson pers.
comm.). The proportion of male parr which mature prior to smolting is thought to be
45-58% (Garcia de Leaniz 1990), and observations have estimated the proportion of
eggs fertilised by precocious parr to be 60% in Newfoundland (Myers 1984).
Although this is a completely different environment, preliminary work on the
Girnoch Burn using DNA fingerprinting suggests that the contribution of precocious
parr may be of this order, at least for some matings (E. Verspoor pers. comm.).
Also, up to 14 precocious parr may be involved in fertilising the eggs of a female
(Verspoor 1995).

6.3.2 Tagging Experiments

A long term tagging program has been run on the River Dee in Aberdeenshire.
However, data from the early studies is suspect because of the large size of external
(Carlin) tags used, as these caused high fish mortalities (Jordan 1992). The most
useful data is that obtained by internal microtagging of smolts on the Girnock and
Baddoch Burns on the River Dee. Youngson et al (1994) estimated that 50% of the
potential spawners in the Girnock Burn were natal to that burn. Subsequent studies
on the Girnock burn have given different estimates, with best estimates for female
salmon being approximately 90%, and for males approximately 50% (D. Hay pers.
The estimates for males, and to a lesser extent females, may be under the true values as the proportion of early spawners natal to the burn is higher than later spawners, and there is evidence that late spawning males may have spawned previously, and are 'spent', making their contribution of genetic material negligible.

Estimates for the Baddoch burn are lower. Around 30% of spawners are natal to the burn, but the structure of the burn is different, as the trap is above a branching of 2 approximately equal sized tributaries (shown in Figure 6.3). However the same reservations apply to these data.

The information about dispersal which can be obtained from this data is useful, but limited. A model was developed to investigate the relationship between mean dispersal distance and tagging returns. Two distributions for dispersal distances between parents and offspring were considered: a normal, and a double exponential distribution \( f(x) = \lambda / 2 \exp(-\lambda |x|) \) for \(-\infty < x < \infty\). The exponential distribution provides larger probabilities of moving both large and small distances. Figure 6.4 illustrates the two distributions, with mean distance between parent and offspring the same.
As a first approximation, assume that the habitat on the ocean side of the trap is of infinite length, and has the same density as above the trap. Then, assuming that the return rate per unit length is \( r \), and there is a length \( m \) of spawning habitat above the trap we have:

\[
R_T = r \int_{0}^{m} (1 - \Phi(-x,d))\,dx \quad \text{and} \quad R_N = r \int_{-\infty}^{0} (1 - \Phi(x,d))\,dx,
\]

where \( R_T \) is the number tagged; \( R_N \) is the number untagged, and \( F(y,\sigma^2) \) is the cumulative density with mean distance moved \( d \).

Hence the expected proportion of tagged fish in the trap can be calculated. There is no simple analytic solution to the above equations for the normal density. With double exponential density the proportion tagged, \( p_T = 2\lambda + \exp(-\lambda m) - 1 \). In the normal case, numerical integration was used.

Figure 6.5 illustrates the proportion of fish from above and below the traps assuming constant habitat above the trap of various lengths, (also assuming that fish who move upwards past the end of the range spawn within the range).

This is obviously an extreme simplification of the real river system. The Girnock Burn trap is 700 metres from the confluence with the main river, and the spawning density within the main river may be higher. Taking this into account, and assuming that there is a length \( b \) habitat before the main river, and that the density in the main river is \( c \) times that in the burn; and that upstream or downstream dispersal in the main river is the same then

\[
R_T = r \int_{0}^{l} (1 - \Phi(x,d))\,dx \quad \text{and} \quad R_N = rc \int_{-b}^{0} (1 - \Phi(x,d))\,dx + r \int_{-\infty}^{-b} (1 - \Phi(x,d))\,dx.
\]
Figure 6.5 The expected proportion of tagged fish returning to a trap for a length of river above the trap of 3000 metres (triangles); 7000 metres (circles); and 10000 metres (diamonds). Values from a normal distribution of dispersal, and double exponential dispersal distances gave very similar results.

Figure 6.6 illustrates the proportions considering the density in the main stem is the same (so multiplied by 2 to account for upstream and downstream) multiplied by 10, and multiplied by 100. In this case the proportion tagged depends on the model for dispersal used. With double exponential dispersal, the mean dispersal distance is smaller than with a normal distribution of dispersal distances to produce the same proportion of fish tagged, when the density in the main stem is large.
Figure 6.6 The expected proportion of tagged fish returning to a trap for a length of river above the trap of 7000 metres with a length of 700 metres below the trap. Relative density in the main stem is 1 (circles), 10 (triangles), and 100 (diamonds). Open symbols for normal distribution, solid symbols for double exponential distribution.

We thus have indirect measures for the amount of gene flow between tributaries and the main river, and an idea about the mean dispersal distances required to produce these levels of gene flow. The models suggest that a simple isolation by distance model is not likely to be a good explanation for the patterns seen, particularly for males, as the mean distance between parent and offspring is extremely large. Only by adding the influence of the main river at high density do the figures for the Girnock Burn become reasonable. However, these figures give information, at most, about the maximum potential gene flow. From radio tagging studies (refs. in Mills 1989), adult salmon are known to travel some distance up rivers before returning to the sea to spawn elsewhere, and the traps tell us only about entries to the Burns, and the traps are not suitable for adult downstream migrants.


6.3.3 Genetic Data

There are two types of data which will be analysed from the River Dee: one on mitochondrial DNA variation, and one for allozymes, (both courtesy of Eric Verspoor, unpublished). The structuring of the River Dee was shown in Figure 6.3; the amount of branching is small so an island model should prove a good approximation (Section 2.7.4).

The maximum likelihood method described in Chapter 4 was used to obtain estimates for the effective migration rate \( M \) for:

1. all allozymes,
2. the \( MEP-2^* \) locus,
3. allozymes not including \( MEP-2^* \), and
4. mitochondrial DNA.

These gave estimates for \( M \) of 20.4, 357.3, 17.4, respectively, for the nuclear loci, and \( M_F \) of 8.6 for mtDNA, where \( M_F = N_F m_F \) for females.

The malic enzyme locus \( MEP-2^* \) was examined separately because of its known selective effects (described in Chapter 1). The values for allozymes and mtDNA are remarkably similar (assuming an effective population size for females of \( N/2 \)), indicating that little difference exists between the sexes. Figure 6.7 illustrates the difference in log likelihoods for different genetic data on the River Dee (where \( 2M_F \) is used for mtDNA). The support curves for levels of gene flow are almost identical for allozymes (except for \( MEP-2^* \)), as is the mtDNA variation.

Comparisons of semimatrices of pairwise measures of \( F_{ST} \), and pairwise measures of Nei’s \( D \), were performed using a Mantel test (Manly 1986) using a Spearman rank correlation (Sokal & Rohlf 1995) as the test statistic, and testing for significance using semimatix permutation under the null hypothesis of no association.

There is no association between the semimatrices of pairwise measures of \( F_{ST} \) (estimated using maximum likelihood) and geographic distance (absolute or in numbers of branch points between sampling locations). There are also no
correlations between pairwise $F_{ST}$ values calculated from the three allozyme data sets described previously, however the levels of heterozygosity are low, so that pairwise statistics are extremely noisy.

Pairwise $F_{ST}$ values calculated from mtDNA variation were also not correlated with distance, similarly for any pairwise $F_{ST}$ values calculated from allozymes.

![Figure 6.7 Difference in log likelihoods from the maximum likelihood estimate for 4 data sets on the River Dee. Solid line represents all allozymes, line with long dashes the same data set removing $MEP-2^*$, line with short dashes $MEP-2^*$ data, and dotted line mtDNA. The line for mtDNA is obtained from $(1-F_{ST})/F_{ST}$ to account for the differences in population size of females.](image)

These analyses are uninformative about the processes occurring in the River Dee. There are differences in gene frequencies between tributaries, but these do not appear to be related to geographic distance either from river length, or the number of branch points. However, with only 6 observations this is not surprising. There is
limited dispersal between populations. As most females are 2-6 years old when they reproduce (Hay 1995), this gives a value of $N_{m}m_{p}$ of around 2 females per year (from the results of section 6.1), but the dispersal seems fairly random from within, and possibly without, the River Dee. There appear to be minimal differences in the amount of differentiation occurring in male, or female salmon, which is somewhat surprising given the differing levels of dispersal for males and females, and precocious mating of males, which should act to increase the value of M seen for allozymes, compared to the mtDNA (see section 6.1). However, the support curves for M, from allozymes and mtDNA (Figure 6.7) suggest that values for M from 10 to 60 for allozymes, and from 5 to 30 for mtDNA are reasonable. Direct estimates of gene flow from tagging studies suggest that, assuming conservatively a breeding population of 40 females per year, that indirect estimates for females are of the order of direct estimates. The direct estimates for allozymes do not seem to agree with the different dispersal rates of males and females, but as noted before, there may be a large number of opportunistic males that contribute little genetic material to the population. These results are compared in the next section to allozyme studies performed on single river systems in Europe.

### 6.4 Other Studies of Genetic Variation

This section will consider other published surveys of genetic variation in Atlantic salmon populations. The lack of consistency in locus, and allele nomenclature (combined with the complexity caused by the tetraploid origin of Atlantic salmon) mean that comparisons between studies are extremely difficult, and no adequate review of the literature is available at present. As a consequence the analyses below are restricted to studies by single groups, on within river variation. McElligott and Cross (1991) examined protein variation in Southern Ireland, including nine tributaries of the River Blackwater for seven enzymes; $sAA4$, $GPI-1$, $IDDH-1$, $IDDH-2$, $IDHP-3$, $MDH-3$, and $MEP-2^*$. The $sAA4$, $IDDH-2$ and $MEP-2^*$ loci were the most polymorphic in the sample from the River Blackwater. Galvin et al. (1994) have investigated variation within the River Shannon in Southwest Ireland. Eleven sites were chosen in the river system, as well as a hatchery population (which is not included in the analyses here).
Verspoor et al. (1991) investigated protein variation within the Kyles of Sutherland river system in North East Scotland. These authors found significant genetic differentiation between tributaries, and also an association within this river system between temperature and frequency of the 125 variant at the MEP-2* locus. A major survey was conducted by Jordan (1992), on rivers within Scotland. Six locations within the Tweed river system were surveyed (Jordan et al. 1992). In these studies the variation both in time (by sampling over successive years) and space were investigated, no temporal difference in allele frequencies were found.

Ståhl (1987) surveyed at the River Kalix in the Baltic at 4 protein loci. This study found significant differences in allele frequencies between tributaries. Elo et al. (1994) investigated protein variation in the north of Norway in two rivers, the Näättämö, and the Teno. These rivers are close to the northernmost range of the Atlantic salmon. Eight polymorphic loci were observed. Within these 2 river systems 93.6% of the total genetic variation was distributed within locations. Other surveys by Ståhl (1987), and Ståhl and Hindar (1988) examined at the Rivers Teno (Tana in this work) and Alta in Norway.

The studies considered are summarised in Table 6.3.

<table>
<thead>
<tr>
<th>Region</th>
<th>River</th>
<th>Sites</th>
<th>Method</th>
<th>Loci</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norway</td>
<td>Teno (a)</td>
<td>23</td>
<td>Allozyme</td>
<td>8</td>
<td>Elo et al. (1994)</td>
</tr>
<tr>
<td>Norway</td>
<td>Näättämö</td>
<td>4</td>
<td>Allozyme</td>
<td>8</td>
<td>Elo et al. (1994)</td>
</tr>
<tr>
<td>Ireland</td>
<td>Shannon</td>
<td>11</td>
<td>Allozyme</td>
<td>8</td>
<td>Galvin et al. (1994)</td>
</tr>
<tr>
<td>Scotland</td>
<td>Tweed</td>
<td>7</td>
<td>Allozyme</td>
<td>6</td>
<td>Jordan et al. (1991)</td>
</tr>
<tr>
<td>Scotland</td>
<td>Kyles of Sutherland</td>
<td>5</td>
<td>Allozyme</td>
<td>7</td>
<td>Verspoor et al. (1992)</td>
</tr>
<tr>
<td>Norway</td>
<td>Teno (b)</td>
<td>10</td>
<td>Allozyme</td>
<td>4</td>
<td>Ståhl &amp; Hindar (1988)</td>
</tr>
<tr>
<td>Norway</td>
<td>Alta</td>
<td>5</td>
<td>Allozyme</td>
<td>4</td>
<td>Ståhl (1987)</td>
</tr>
<tr>
<td>Scotland</td>
<td>Dee</td>
<td>6</td>
<td>Allozyme/mtDNA</td>
<td>10</td>
<td>Verspoor unpubl. data</td>
</tr>
<tr>
<td>Norway/Baltic</td>
<td>Kalix</td>
<td>5</td>
<td>Allozyme</td>
<td>2</td>
<td>Ståhl (1987)</td>
</tr>
</tbody>
</table>

Table 6.3 Allozyme, and mitochondrial DNA studies used in this Chapter.
The values of M estimated using maximum likelihood are shown in Table 6.4. There is a large variation in the values of M estimated so F\textsubscript{ST}, calculated from M, will be used in subsequent analyses.

<table>
<thead>
<tr>
<th>River</th>
<th>Number of loci</th>
<th>Number of sites</th>
<th>Estimated M for all loci</th>
<th>Estimated M for MEP-2*</th>
<th>Estimated M without MEP-2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alta</td>
<td>3</td>
<td>5</td>
<td>14.9</td>
<td>53.9</td>
<td>10.0</td>
</tr>
<tr>
<td>Blackwater</td>
<td>6</td>
<td>9</td>
<td>12.3</td>
<td>23.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Dee</td>
<td>10</td>
<td>6</td>
<td>20.4</td>
<td>357.3</td>
<td>17.4</td>
</tr>
<tr>
<td>Kalix</td>
<td>3</td>
<td>5</td>
<td>14.3</td>
<td>321.0</td>
<td>9.5</td>
</tr>
<tr>
<td>Kyles</td>
<td>7</td>
<td>5</td>
<td>20.9</td>
<td>8.1</td>
<td>28.0</td>
</tr>
<tr>
<td>Naatamo</td>
<td>8</td>
<td>4</td>
<td>10.5</td>
<td>9.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Shannon</td>
<td>8</td>
<td>11</td>
<td>4.9</td>
<td>12.9</td>
<td>4.5</td>
</tr>
<tr>
<td>Teno (b)</td>
<td>4</td>
<td>10</td>
<td>6.3</td>
<td>7.3</td>
<td>6.0</td>
</tr>
<tr>
<td>Teno (a)</td>
<td>6</td>
<td>23</td>
<td>4.6</td>
<td>5.9</td>
<td>4.4</td>
</tr>
<tr>
<td>Tweed</td>
<td>4</td>
<td>7</td>
<td>44.7</td>
<td>33.9</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Table 6.4 Estimated M within river systems for studies used.

Figure 6.8 shows the support curve for F\textsubscript{ST} calculated from the allozyme data not including MEP-2* . There is a significant difference between estimates from the Teno, and the Tweed, and Kyles, although there is little difference between estimates from different studies in the Teno.

The values of F\textsubscript{ST} estimated from allozyme variation not including MEP-2* , and those using MEP-2* alone are correlated, but not significantly (R = 0.503 Pearson, p>0.05). Ordering values of F\textsubscript{ST} estimated using the commonly studied loci MEP-2*, AAT-3, and SORD-1 produced no systematic effects over the studies.
Figure 6.8 Difference in Log Likelihoods from the maximum likelihood estimate against $F_{ST}$ for Kyles of Sutherland (line with short dashes), Tweed (line with long dashes), Teno from Ståhl and Hindar (1988) (dotted line), and Teno from Elo et al. 1994 (solid line).

There is a positive correlation between the number of sites surveyed and the value of $F_{ST}$ estimated using all allozymes ($R = 0.558$, $p<0.05$, Spearman rank), and using all allozymes not including $MEP-2^*$ ($R = 0.509$, $p>0.05$, Spearman rank), (Figure 6.9 illustrates the number of sites plotted against $F_{ST}$). The possible reasons for this are discussed at the end of this chapter.
Figure 6.9 Estimated $F_{ST}$ using maximum likelihood and the number of sites sampled.

Pairwise $F_{ST}$ values were calculated, as in the previous section, for all loci, for $MEP-2^*$, and for all loci except $MEP-2^*$. Mantel tests were performed between the pairwise $F_{ST}$ values calculated from $MEP-2^*$, and for the other loci. Only the Teno study by Elo et al. (1994) was significant at a 5% level ($p<0.001$) when a Bonferroni correction was used.

The largest study on a single river system, conducted to date, was on the Teno in Northern Norway by Elo et al. (1994). A map of which is given in Figure 6.10 with the location of sampling sites.
The value of $M$ estimated for this study is the largest, and sites are drawn uniformly over the whole river system, including the main stem. Repeatedly drawing 10 sampling locations at random from the 23 present did not alter the estimated levels of gene flow drastically. The relationship between geographic and genetic distances was investigated using a Mantel test (Manly 1986) as described above. There is no significant association between geographic distance, measured along the river, and
genetic distances, measured using Nei’s D, or pairwise $F_{ST}$. Grouping sites into upper, lower, and tributary groups gave a marginally insignificant result ($p = 0.055$). However, when the distance between sites was measured simply as the number of branch points between sites (as shown in Figure 6.11) then there was a significant positive association ($p = 0.0015 \pm 1.2 \times 10^{-4}$). Figure 6.12 illustrates the relationship using a linear regression. This regression, although not statistically valid (the points are correlated) provides a reasonable guide.

![Diagram of sampling sites](image)

**Figure 6.11** Distances between sampling sites for tributary distance measure used on Teno system. Bold line indicates main stem.
This relationship suggests that geographic distance may not be important, rather some form of tributary distance could be the factor which determines the gene flow between populations. This is mirrored in deep sea species living close to hydrothermal vents where there is a significant association between genetic distance, and distance along mid-ocean ridges, but no relationship between genetic distance and straight line distance (Tunnicliffe & Fowler 1996).

6.5 Discussion

The biology of the Atlantic salmon, particularly in freshwater, is well documented (e.g. Mills 1989), and a considerable number of surveys of allozyme variation have been conducted. The Girnock Burn, a tributary of the River Dee in Aberdeenshire, has been studied continuously for many years, and genetic data for both allozymes and mtDNA variation have been collected for sites throughout the river. This has presented an opportunity to compare estimated levels of gene flow from genetic
data "indirect estimates" to "direct estimates" of gene flow (terminology from Slatkin 1985b).

In order to compare estimates, the effect of overlapping generations on expected values of F_{ST} was modelled. The effective population size for overlapping generations has been considered previously using classical methods (Felsenstein 1971, Hill 1972, Charlesworth 1980), but here the approximation for F_{ST} derived in Chapter 2 was used to get estimates using the coalescent process. Increasing the length of generations decreased the expected value of F_{ST}. If dispersal was greater for those fish with a short generation time then the values of F_{ST} were decreased further. However this effect was minimal unless there was a large asymmetry in return times and dispersal rates.

Information on return rates of previously tagged fish (Youngson et al. 1994) was used to model the approximate mean dispersal distances. These results indicated that, at least for males, the dispersal distances were very high. Analysis of genetic differentiation within the River Dee produced indirect estimates for gene flow. These estimates were consistent for the maternally inherited mtDNA, which was consistent with equal dispersal for the two sexes. However, direct estimates of gene flow for males were higher, and not consistent with the indirect estimates.

Studies within river systems on allozyme variation were reanalysed using the method developed in Chapter 4. There were large differences between estimates from different river systems, and studies which had larger number of sampling points uncovered more differentiation. This is expected from the properties of estimators (Chapter 4). Nevertheless, the differences were too large to attribute to statistical bias. This may reflect the sampling points; if more samples were they may have been collected where fish densities were lower, and hence random genetic drift may be a stronger force.

Results considering just MEP-2* were equivocal, with no consistent underestimates, or overestimates of gene flow compared to the other loci. This is somewhat surprising as previous studies have shown an association with water temperature (e.g. Verspoor et al. 1991). However information about water temperature was not available so it was not possible to examine this further.
The studies did show genetic differentiation between populations, and in the cases where replicate studies were taken, the results were comparable which suggests that differences are due to limited gene flow between populations. This has management implications which will be discussed in the next chapter.

The large study of Elo et al. (1994) on the River Teno in Northern Norway indicates that dispersal within populations may be a more complicated process to model than any considered here. There is no correlation between geographic distance and genetic distance (measured using Nei's D) or the total number of branching points between sampling locations. However, if we consider a homing process where tributaries are taken as groups, with no distance between points on the same tributary, irrespective of the number of branches, then we get a positive relationship.

As mentioned in Chapter 2, the geographic structure from a salmon's point of view may be radically different to a terrestrial viewpoint. If the hypothesis that homing to tributaries is due to an olfactory sense (evidence reviewed in Stabell 1984), then the tributary distance is likely to be an important factor in homing precision. The implications of this are discussed in the final chapter.
Chapter 7

DISCUSSION

7.1 Theoretical Results

The theoretical work in this thesis has concentrated on results which are applicable to studies of Atlantic salmon in the wild, and more generally to fisheries management. However, the results have broader implications than for just Atlantic salmon.

The anadromous life cycle of the Atlantic salmon provided a rationale for investigating the effects of branching structures in population subdivision as Atlantic salmon are known to home to their natal river, and tributary to some degree within branching river systems. This thesis has shown that the branching structure within which Atlantic salmon breed, and live in freshwater can have important implications for the levels of population differentiation seen. The effects of limited levels of gene flow within branching structures have not previously been studied (but see Sawyer 1978). The patterns of isolation by distance seen in branching structures are qualitatively different from those seen in one and two dimensional habitats. When the branching structure is of sufficient depth the isolation by distance seen is greatly increased compared to that seen in one, and two dimensional habitats, and even with shallow branching patterns the effects of branching are different to standard models. When we have isolation by distance, the branching pattern of populations seems to still have an important effect. This has implications for local adaptation of populations by analogy with models in one and two dimensions (Slatkin 1973, Nagylaki 1975). Many plants and animals live in habitats which are branching to some extent. Rivers branch, but also valleys, and even sea shores have a branching pattern at some scales. Species which live in these habitats are likely to be restricted
in their movement in some way. It is important to investigate, or at least to recognise, that impediments to gene dispersal may exist even in habitats which appear to be uniform. This effect is most visible in hybrid zones; karyotypic races of the Alpine Grasshopper (Barton and Hewitt 1982) were shown to be effectively separated by a small stream (Barton & Gale 1993). The effect of barriers to gene flow has been investigated by Nagylaki et al. (Nagylaki 1988, Nagylaki & Barchilon 1988) in one dimension, who showed that it resulted in a sharp step in the correlation of gene frequencies. This thesis has provided another example of geographic features having large effects on the genetic variation within populations.

The coalescent process has been shown to be a useful technique for modelling the levels of genetic differentiation seen in samples from natural populations. There is a close relationship between the coalescent and the classical theory of identity by descent (Wright 1951, Malecot 1948), and classical results may be used to calculate distributions of coalescence times, for example in the distribution of coalescence times in a one dimensional population. In most circumstances, however, it is preferable to work with the coalescent process directly. The coalescent process should prove of increased value when the use of highly variable STR loci becomes more widespread. With STR loci the length of an allele after a mutation is dependent on its length before the mutation (Valdes et al. 1993). The effects of mutation on the patterns of length polymorphisms can be modelled directly using a coalescent process. This will be of use in studies of population subdivision (O'Connell & Slatkin 1993, Slatkin 1995), and also of use for the correct interpretation of STR loci used in forensic science, where the amount of inbreeding within populations can be of crucial importance (Nichols & Balding 1991, Balding & Nichols 1995).

The coalescent is able to model the vast majority of population processes: limited gene flow (Slatkin 1993, Herbots 1995, Chapter 3, this thesis), selection (Hey 1991), elimination of deleterious mutations (Hudson 1990), population fission, extinction/recolonisation, and range expansion (this thesis). It provides an alternative way of visualising the processes occurring in natural populations, and is suited to simulation experiments. The coalescent process shows that most events which influence patterns of genetic variation occurred in the distant past, and that most individuals, in the absence of extreme inbreeding, are distantly related (Barton & Wilson 1994, 1995, and Chapter 3). This leads us to problems in evaluating the
relationships between populations. While the differences in gene frequencies may enable us to infer differentiation (Chapters 4, and 5) the relationship between populations is determined by very few events, and hence evaluating, for example, which populations are most closely related, is likely to prove difficult (Chapter 6). There is a lack of discriminatory ability in pairwise statistics for determining population structure, with Wright’s $F_{ST}$ (1931) primarily being a measure of the probability of coalescence within the deme prior to migrating out of it, and Nei’s $D$ (Nei 1987) being a measure of the processes which occur when genes are in different subpopulations.

This lack of discriminatory ability is due to the structure of genetic data. However, better survey design may help us to increase our ability to obtain precise estimates of gene flow. Chapter 4 investigates the use of a multinomial-Dirichlet distribution for the estimation of population genetic parameters, and investigates the influence of sampling strategy on the precision of estimators for population structure. The multinomial-Dirichlet distribution provides a good approximation for the patterns of genetic variation seen in samples from an island type distribution, and the support curve for parameter estimates is informative about the strength of evidence for a particular value. This type of estimation procedure utilises more of the information in a sample than rare alleles (Slatkin 1995a), or Weir and Cockerhams θ (1984). It was also shown that if mutation is a powerful enough force to ensure that lines-of-descent which leave the deme are not identical by descent to those which coalesce within a deme, then an estimate of gene flow can, in principle, be obtained with data from a single deme. This is implicit in Balding and Nichols (1995).

Another method for estimating levels of gene flow using the spatial autocorrelation of allele frequencies has been investigated by Epperson (1993). This method was developed for two dimensional populations, and uses both the variance of allele frequencies, and covariance from different locations to attempt to estimate both the neighbourhood size, and the dispersal rate of individual gene. However this is only practical if sampling locations are distant from each other, (much greater than the standard deviation of dispersal distances). This approach is inappropriate for Atlantic salmon, because the potential dispersal distances are so large both within, and between river systems (Youngson et al. 1994). Also, river systems do not conform to standard one- or two-dimensional populations.
Computer programs for the calculation of the expected powers of tests for population subdivision, and Hardy-Weinberg proportions were developed. Tests for Hardy-Weinberg proportions were shown to have low power for all mutation rates, although the power increased for high mutation rates, as predicted by Chakraborty and Zhong (1994). The power of tests for population differentiation for a variety of parameters are investigated in Chapter 5. These tests are of limited importance for evolutionary studies, but for stock differentiation in fisheries they are crucial. Low levels of gene flow between populations, in terms of numbers of individuals per generation, may not allow substantial genetic differences to develop between populations (Wright 1943), but in fisheries management where the populations sizes may be extremely large, evolutionarily high numbers of individuals per generation may be insufficient to restock severely depleted local populations. Genetic panmixia is not equivalent to actual population panmixia. This thesis has shown that high mutation rate loci have more power in detecting a lack of panmixia, for both tests of population differentiation, and Hardy-Weinberg proportions.

7.2 Management Implications

Atlantic salmon (*Salmo salar* L.) are an economically important species, both in culture, and in the wild. They have a complicated life history, which involves two major migrations: from the river in which they were spawned to ocean feeding grounds, and a return to freshwater to spawn (Mills 1989). There are three major commercial enterprises in Scotland based on this fish: the fish farming industry, the off and inshore net fishery, and the recreational rod fishery. Large, unpolluted rivers, such as the Spey, Tay, and Aberdeenshire Dee, have the most important sports fisheries in Scotland, and these rivers have large numbers of tributaries which support salmon stocks. It is thought (Hawkins & Smith 1986, Hay 1995, A.F. Youngson, J. Webb and E. Verspoor pers. comm.) that tributaries high in the river are those which provide the most economically important fish, spring running salmon. Over the last 10 years the proportion of spring running salmon in the Aberdeenshire Dee has dropped (Hay 1995), and this is mirrored in the Rivers Spey, and Tay. The reasons for this are unclear, and likely to be complex with both marine and freshwater factors (A.F. Youngson pers. comm.). In order for these fisheries to be managed effectively detailed information about the stocks contained within these
rivers is required. This is dependent on knowing the scale of local populations (Ståhl 1987, Altukov & Salmenkova 1987), and in order to achieve this detailed knowledge of the patterns of dispersal, and local adaptation within rivers is required. Additionally, there is controversy about the effects of escaped farmed salmon (Hansen & Bakke 1989, Bentson 1994) the resolution of which requires information on local adaptation of salmon stocks. This thesis has examined the evidence for population differentiation in Atlantic salmon; models have been constructed to explain this differentiation; statistical techniques for determining the levels of differentiation have been considered; and these techniques used to look at selected studies of genetic differentiation within the Atlantic salmon.

Verspoor (1995) suggested that management of salmon stocks within rivers should occur on a tributary basis. The theoretical work of Chapter 3 has also concluded that tributaries are likely to contain discrete stocks, as genetic differentiation is increased by branch points. Previous surveys of genetic variation have produced genetic data which has confirmed the presence of substantial differentiation between stocks, with effective numbers of migrants per generation being of the order of 2-20 for most river systems. There is no obvious relation between the location of the river and the amounts of differentiation. However, the differentiation was significantly correlated with the number of sites sampled, a slightly confusing relationship. The presence of 2 samples from the Teno, a river with a large amount of differentiation, may have skewed the relation, however, or conceivably, when a large number of samples are taken then locations with low fish numbers may be sampled, so increasing the relative strength of population subdivision.

The results of Chapter 2, on gene flow in branching systems, combined with the reanalysis of the Teno data (Elo et al. 1994) suggest that tributaries are the important factors in stock structure, rather than isolation by distance between habitats. These tributaries are the most heavily fished as the salmon destined to spawn there enter the river earlier (Hawkins & Smith 1986) and are thus in the river for the whole of the fishing season. This leads to potential management difficulties. The most prized sporting fish (and hence the most valuable commercially), are also most at risk from over-fishing. This has been recognised in the River Dee where, in the last year, voluntary catch/release schemes have been operated, and the start of the fishing season has been delayed by one month.
One particular question of interest highlighted by this study is the apparent contradiction between the levels of gene flow measured using mtDNA, and allozymes in the River Dee. In Chapter 6 the importance of life history on the expected amount of inbreeding observed within natural problems was investigated. The concordance between estimates based on maternally inherited DNA, and nuclear DNA, while at first glance gratifying, is slightly puzzling. There are large numbers of precocious male parr, and their dispersal in the spawning season has been observed (D. Hay 1995), therefore the estimates of gene flow using nuclear DNA should be inflated. Accurate measurements for the genetic contribution of male parr, and their dispersal rates would be essential in order to attempt to unravel this problem.

The lack of discriminatory ability of genetic data presents problems for the design of surveys of natural populations. Using more variable loci will provide more precise estimates for gene flow. Perhaps a better use for these loci would be to examine the effects of local adaptation with small, semi-closed systems such as the Girnock and Baddoch Burns on the River Dee.

Local adaptation of genotypes is an important factor in:

- artificial restocking of rivers,
- using hatcheries to supplement local runs of salmon,
- the effects of escapees of farmed origin, and possibly
- stocking of rivers with fish with 'desirable traits'.

The influence of local adaptation is a difficult question to address in the absence of any great quantity of hard evidence. As stated in the introduction to this thesis there is a plethora of unsatisfactory anecdotes, and a few more convincing indications for local adaptation. However it is extremely difficult to disentangle environmental effects from local adaptation of genotypes. Garcia de Leaniz et al. (1989), have shown that stocking rivers using non-native salmon is likely to be unsuccessful.

To get any additional insights more information on life-histories, local adaptation and dispersal is required. The problem with local adaptation is that the lifetime reproductive success of a fish is extremely difficult to measure. The use of RFLPs (restriction fragment length polymorphisms) allows us to isolate families of Atlantic
salmon, and to ascribe paternities, and maternities (Pemberton et al. 1992, Taggart and Ferguson 1994), and may also be useful as genetic tags (Verspoor et al. in prep.). In this case long term studies in the wild would be possible however they would, by necessity, be on an extremely long time scale, and a complete study on lifetime reproductive success in, for example, the River Dee in Aberdeenshire would take 7 years to be completed, and be prohibitively expensive.

The implications of this study for escapes of cultured salmon are unclear. It could be argued that since observed levels of dispersal between tributaries are high, yet actual differentiation is higher than would be expected (at least for males) this implies that male dispersers between tributaries in the wild are unsuccessful. However, this ignores any potential dangers from disruption of locally adapted gene complexes. The evidence for these local gene complexes is extremely slim anyway. As is always the case, more studies are required on the levels of local adaptation, before we may be sure about potential effects.
References


References


References


References


References


Myers, R.A. 1986. Game theory and the evolution of Atlantic salmon (*Salmo salar*) age at maturity. In *Salmonid age at maturity* (Ed. Meerburg, D.J.) Canadian Special publication of Fisheries and Aquatic Sciences **89**.


Randall, R.G., Thorpe, J.E., Gibson, R.J. & Reddin, D.G. 1986. Biological factors affecting age at maturity in Atlantic salmon (Salmo salar). In Salmonid age at maturity (Ed. Meerburg, D.J.). Canadian Special publication of Fisheries and Aquatic Sciences 89.


Appendix 1: Algorithms for Simulations on Tree Structures

```c
/* Structure binary_node is used in the simulation function */

#define DIMENSION 3
struct binary_node {
    int gene_number;
    int previous_gene_number;
    struct binary_node *children[DIMENSION];
};
float mig,mu;
int N, no_genes, neigh, height, initial_gene_number;

void recur_maketree(struct binary_node *any_node, struct binary_node *ancestor, int depth, long *count) {
    int i;
    any_node->previous_gene_number = initial_gene_number;
    *count += 1;
    if (depth != 1){
        for (i=0;i<DIMENSION;i++){
            if ((any_node->children[i] = (struct binary_node *)malloc(sizeof(struct binary_node))))
                ==NULL) {printf("error in allocating memory");exit(0);} recur_maketree(any_node->children[i],any_node,depth-1,count);}
    else{
        for (i=0;i<DIMENSION;i++)
            any_node->children[i]=NULL;}
    any_node->children[DIMENSION-1] = ancestor;
}

void generation(struct binary_node *any_node, int root_neighbours) {
    int i, tot=0;
    if (any_node->children[0] == NULL){
        any_node->gene_number = mat_drift(DIMENSION*root_neighbours,
                                         any_node->previous_gene_number);
    }
    else {
        for (i=0;i<DIMENSION;i++) tot += any_node->children[i]->previous_gene_number;
        any_node->gene_number = mat_drift(root_neighbours+tot,
                                           any_node->previous_gene_number);
        for (i=0;i<DIMENSION-1;i++)
            generation(any_node->children[i],any_node->previous_gene_number);
    }
}
```
Appendix 2: Algorithms for Quick Random Drift Calculations

/* Global variables */
extern int N, no_genes, no_neigh;
extern float mig, mu;
long seed;
float ***p;
void set_up_matrix(void)
{
    int i,j,k;
    double prob,temp;

    float log_bico(int n,int k); /* function which returns the logarithm of the binomial coefficient */

    p = (float ***)malloc((no_genes+1)*sizeof(float **));
    if (!p){printf("allocation error 1");exit(0);} }

for (i=0;i<=no_genes;i++)
    p[i] = (float **)malloc((no_neigh+1)*sizeof(float *));
    if (!p[i]){printf("allocation error 2");exit(0);} }

for (i=0;i<=no_genes;i++)
    for (j=0;j<=no_neigh;j++)
        temp =0.0;
        prob = (1.0-mig)*(1.0-mu)*((float)i/(float)no_genes) + mig*(1.0
mu)*((float)/((float)(no_neigh)) + (1-mig)*mu*(1-(float)i/((float)no_genes) +
mig*mu*(1.0-((float)/((float)(no_neigh));
        for (k=0;k<=no_genes;k++)
            temp += exp(log_bico(no_genes,k) + (double)k*log(prob) + (double)(no_genes-
            k)*log(1.0-prob));
        p[i][j][k] = (float)temp;
        p[i][j][no_genes] = 2.0;
    }

printf("p allocated");
}

int mat_drift(int neighbours, int number)
{
    float prob;
    int where, count=0;
    int rand(long *seed);

    where = number;//an educated guess to reduce the number of operations
    prob = rand(seed);
    for (;;){
        if (prob <= p[number][neighbours][where]){
            if (where==0) return 0;
            if (prob > p[number][neighbours][where-1]) return where;
            else{
                where--;
            }
        }
    }
}
Appendix 2

```c
count++;}
else{
  where++; count++;
  if (where >= no_genes) return no_genes;
  if (count > no_genes) printf("error\n");
}
```
Appendix 3: Algorithms for Generation of Trees

/* Structures needed for the simulation procedures */

struct model_type {
    float mig;
    int n_pops;
    float pop;
    float *v_pop; // or other information
    int *sample_size;
    int sampled_pops;
    int total_sample;
    int mig_type;
    float **mig_matrix; // or other information
    float mutation_rate;
    int mutation_type;
};

struct node {
    int n_muts;
    double time;
    int allele;
    int location;
    struct node *desc1;
    struct node *desc2;
    struct node *ancestor;
    int n_desc;
};

/* Algorithms require function rand(*long) which returns a floating point pseudo-random number uniformly distributed between 0 and 1*/

void make_tree_island(struct node *tree, int sample_size, int *where, float pop, float mig, int n_pops, long *seed) // type 1
    /* This routine is of type 1 in discrete generations */
{
    int in, how_many_left, j, k, from, to;
    float t = 0.0, prob;
    struct node **list; /* a list of pointers to pointers to nodes for book-keeping */
    float rand(long *i cli ,,,);

    /* Initialise things*/
    how_many_left = sample_size;
    if ((list = (struct node **malloc(sample_size*sizeof(struct node *)) == NULL){
        printf("error in allocating list"); exit(0); } /* allocate memory for list */
    for (in = 0; in < sample_size; in++){
        tree[in].time = 0.0;
        tree[in].n_desc = 1;
        tree[in].desc1 = tree[in].desc2 = NULL;
        tree[in].location = where[in] + 1;
        tree[in].n_muts = 0;
        list[in] = tree + in;)
    /* Main program loop, continues until one line-of-descent remains */
    for (; ; ){
        if (how_many_left == 1) break; /* if one line-of-descent left breaks the loop */
        t += 1.0; /* adds one time unit */
        /* Is there a migration event? */
        for (j=1;j<=how_many_left;j++) {
            if (rand((long *)&j) <= mig)
                /* Do migration */
                // Various operations on `tree'
            ++how_many_left;
            // Update tree
        }
    } /* End main loop */
} /* End make_tree_island */
prob = rand(seed);
if (prob < mig) {
    from = list[j-1]->location;
    to = rand(seed)*(n_pops-1)+1;
    if (to>=from) to++;
    list[j-1]->location = to;}
}
/* Are there any coalescent events */
for (j=0;j<=how_many_left-1;j++){
    for (k=j+1;k<=how_many_left-1;k++){
        if (list[j]->location==list[k]->location) /* are the lines in the same location */
            if (rand(seed) > (1.0-1.0/(2.*population))) { /* is there a coalescent event ? */
                tree[(2*sample_size)-how_many_left].time = t; /* allocate time of ancestor */
                tree[(2*sample_size)-how_many_left].n_desc = list[j]->n_desc+
                    list[k]->n_desc; /* how many descendents does the node have */
                tree[(2*sample_size)-how_many_left].location=list[j]->location;
                list[j]->ancestor = tree+(2*sample_size)-how_many_left; /* sets ancestor */
                list[k]->ancestor = tree+(2*sample_size)-how_many_left; /* for nodes */
                list[j] = list[how_many_left-1];/* how_many_left -= 1; */
            }
    }
}
free(list);

/**************************************************************/
void make_tree_island2(struct node *tree, int sample_size, int *where,
    float pop, float mig, int n_pops, long *seed) //type 2
{
    /* This routine is for type 2 for continuous generations */
    int in, how_many_left, c1, c2, from, to, *left_where;
    float t=0.0, prob, rate, rate_mig, *rate_coal;
    struct node **list;
    float rand(long *idum);

    /* initialise things */
    how_many_left = sample_size;
    if ((list = (struct node **)malloc(sample_size*sizeof(struct node *))
        == NULL) printf("error in allocating list");exit(0);)
    if ((left_where = (int *)malloc(n_pops+1*sizeof(int)))
        == NULL) printf("error in allocating left where");exit(0);)
    if ((rate_coal = (float *)malloc(n_pops+1*sizeof(float)))
        == NULL) printf("error in allocating left where");exit(0);)
    for (in=1,in<=n_pops,in++) left_where[in] = 0;
    for (in = 0; in < sample_size; in++) {
        tree[in].time = 0.0;
        tree[in].n_desc = 1;
        tree[in].desc1 = tree[in].desc2 = NULL;
        tree[in].location = where[in+1];
        tree[in].n_muts = 0;
        list[in] = tree + in;
        left_where[where[in+1]]+=1;
    }
    /* main loop of routine */
    for (;;) { /* Finished */
        if (how_many_left == 1) break;
        rate_mig = (float)how_many_left*mig; /* rate of migrations */
rate_coal[0]=0.0;
for (in=1;in<=n_pops;in++) {
    if (left_where[in] > 1)
        rate_coal[in] = (float)left_where[in]*(left_where[in]-1)/(4.*pop);
    else rate_coal[in]=0.0;
    rate_coal[0]+= rate_coal[in]; } /* rate of coalescences */
rate = rate_coal[0]+rate_mig; /* rate until first event */
t += -log(1-rand(seed))/rate; /* exponential distribution at rate */
if (rand(seed)<rate_mig/rate) { /* migration occurs first */
    j = rand(seed)*(how_many_left)+1; /* picks out the migrant */
    from = list[j-1]->location;
    to = rand(seed)*(n_pops-1)+1;
    if (to>from) to++;  /* picks out the migrant */
    list[j-1]->location = to; /* Swaps the locations */
    left_where[from]--; left_where[to]++;}
else /* coalescence */
    rate_coal[1] = rate_coal[0];
for (in=2;in<=n_pops;in++)
    rate_coal[in] = rate_coal[in-1] + rate_coal[in]/rate_coal[0]; /* relative probabilities */
prob = rand(seed);
to=0;
for (; ; ){
    to += 1;
    if (prob <= rate_coal[to]) break;
    if (left_where[to]==2){c1=1; c2=2;}
    else /* which of them */
        c1 = rand(seed)*(left_where[to])+1;
        c2 = rand(seed)*(left_where[to]-1)+1;
        if (c2>=c1) c2++;
/* now need to find the required lines of descent */
    for (in=0;in<=how_many_left;in++)
        if (list[in]->location==to){
            if (--cl==0) j = in; /* the location of the first */
            if (--c2==0) from=in; /* second */
            if (c1<0 && c2<=0) break;}
    tree[(2*sample_size)-how_many_left].time = t;
    tree[(2*sample_size)-how_many_left].n_desc = list[j]->n_desc+list[from]->n_desc;
    tree[(2*sample_size)-how_many_left].location = list[j]->location;
    list[from]->ancestor = tree+(2*sample_size)-how_many_left;
    list[from]->ancestor = tree+(2*sample_size)-how_many_left;
    tree[(2*sample_size)-how_many_left].desc1 = list[j];
    tree[(2*sample_size)-how_many_left].desc2 = list[from];
    list[from]=tree+(2*sample_size)-how_many_left;
    list[j]=list[how_many_left-1];
    how_many_left--; left_where[to]--;
}
free(list);
free(left_where);
free(rate_coal);
Appendix 4: Range Expansion Simulation

```c
#include <math.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include "qstat.h"
#include "solveex.h"
#include "nrutil.h"

/* a program to get the mean times of coalescence for pairs of genes sampled from x and y from a linear array of length k for mig, R,P. Requires uniform and exponential random number generators */

float nomig(int posi, int pos2, int k, float N, float S, float R);
float disp(int posi, int pos2, int k, float N, float S, float R, float mig);
void migratesame(int *pos1, int *pos2, int nleft);
void migratediff(int *pos1, int *pos2, int neigh, int nleft);
float simplepairf(float mean0, float mean1);
float pair_f(float **theor, int i, int j);
float pair_f_times(float **theor, int i, int j);

long seed = -345;

void main(int argc, char *argv[])
{
    long n_samples;
    struct qstats **times;
    int neigh, is_theory, k, i, j;
    FILE *output;
    float **theor_times, **theor_f, **range_times, **range_f, u, mig, N, S, R, t, mean, se;

    if (argc!=11){
        printf("Error in input
Correct usage:
 n'9;
printf("range filename mig k N S R n samples seed theory u
exit(0);
    if ((output = fopen(argv[1],"w")) ==NULL){ printf("error opening output file"); exit(0);}

    mig=atof(argv[2]);k=atoi(argv[3]);N=atof(argv[4]);S = atof(argv[5]);
    R = atof(argv[6]);n_samples=(long)atof(argv[7]);i2 = abs(atoi(argv[8]))%100;
    is_theory=atoi(argv[9]); u = atof(argv[10]);

    theor_f = matrix(1,k,1,k);theor_times = matrix(1,k,1,k);
    range_f=matrix(1,k,1,k);range_times=matrix(1,k,1,k);

    tbar(theor_times,k,mig,N);
    get_f(theor_f,k,mig,N,u);
    range_e(range_times,range_f,k,N,R,S,u);

    if (is_theory==1){/*do we only want the theoretical results?*/
        for (i=1;i<=k;i++)
            for (j=1;j<=k;j++){
                printf(output,"%d %d %d %f %f %f %f %f %f %f %f %f %n",i,j,abs(i-j),theor_f[i][j],theor_times[i][j],range_f[i][j],range_times[i][j],pair_f_times(theor_times,i,j),range_times[i][j],pair_f_times(range_times,i,j));
        exit(0);}

    if ((times = (struct qstats **)malloc((k+1)*sizeof(struct qstats*)))
```

---

This code is a program designed to calculate the mean times of coalescence for pairs of genes sampled from two linear arrays of genes. It requires uniform and exponential random number generators and is used to simulate range expansion scenarios in genetics.
Appendix 4

```c
for (i=1;i<=k;i++)
    if ((times[i] = (struct qstats *)malloc((k+1)*sizeof(struct qstats)))
        ==NULL) {printf("error in allocating memory");exit(0);}
    for (j=1;j<=k;j++) reset_stats(&times[i][j]);

for (i=1;i<=k;i++)
    for (j=1;j<=k;j++)
        for (sample=1;sample<=n_samples;sample++)
            if (mig<0.0)
                t = nomig(i,j,k,N,S,R);
                add_value(&times[i][j],t);
            else
                t = disp(i,j,k,N,S,R,mig);
                add_value(&times[i][j],t);
printf("finished the simulations - now get results\n");
for (i=1;i<=k;i++)
    for (j=1;j<=k;j++)
        times[i][j].x = times[j][i].x;
        times[i][j].n = times[j][i].n;
        times[i][j].ss = times[j][i].ss;

for (i=1;i<=k;i++)
    for (j=1;j<=k;j++)
        fprintf(output, "%d %d %d 
",i,j,abs(-i));
        fprintf(output, "%f%f%f%f 
", calcmean(times[i][j]),
                times[i][j].ss/times[i][j].n,sqrt(calcvar(times[i][j])/times[i][j].n),
                simplepairf(calcmean(times[i][j])/2.0+calcmean(times[i][j])/2.0,calcvar(times[i][j]));
        fprintf(output, "%f%f%f%f 
", theor_times[i][j],pair_times(theor_times,i,j);
        fprintf(output, "%d%d%d%d 
", range_times[i][j],pair_times(range_times,i,j));

free(times);
free_matrix(theor_times,1,k,1,k);
free_matrix(range_times,1,k,1,k);
}

*******************************************************************************
float nomig(int pos1, int pos2, int k, float N, float S, float R)
/* the time until the next event when there is no migration */
{
    float time;
    int temp;
    time =0.0;
    if (pos2<pos1){temp=pos2;pos2=pos1;pos1=temp;}
    if (pos1==pos2) time = S + (k+1-pos1)*R;
    time += expdev(1./(2.*N));
    return time;
}
*******************************************************************************
float recol(float S, float R, float k, float n_left)
{
    return S + R*(float)(k+1-n_left);
}
*******************************************************************************
```
float disp(int pos1, int pos2, int k, float N, float S, float R, float mig)
{
    int n_left = k, neigh;
    float time = 0.0, t, next_recol = S + R;
    for (; ;)
    {
        if (n_left == 1) return time + expdev(1.0 / (2.0 * N));
        if (pos1 == pos2)
        {
            neigh = (pos1 > 1) + (pos1 < n_left);
            t = expdev(1.0 / (2.0 * N) + mig * (float)neigh);
            if (time + t < next_recol)
            {
                if (rand(&seed) < 1.0 / (1.0 + 2.0 * mig * (float)neigh))
                    return time + t;
                else
                    time += t;
            }
            else
            {
                time = next_recol;
                n_left -= 1;
                next_recol = recol(S, R, k, n_left);
                if (pos1 == n_left) pos1 = 1; pos2 = 1;}
        }
        else
        {
            time = next_recol;
            n_left -= 1;
            next_recol = recol(S, R, k, n_left);
            if (pos1 == n_left) pos1 = 1;
            if (pos2 == n_left) pos2 = 1;}
    }
}

void migratesame(int *pos1, int *pos2, int neigh, int n_left)
{
    float p = rand(&seed);
    if (neigh == 2)
    {
        if (p < 0.25) *pos1 = 1;
        else if (p < 0.5) *pos1 = 1;
        else if (p < 0.75) *pos2 = 1;
        else *pos2 += 1;
    }
    else if (*pos1 == 1)
    {
        if (p < 0.5) *pos1 = 1;
        else *pos2 += 1;
    }
    else if (*pos1 == n_left)
    {
        if (p < 0.5) *pos1 = 1;
        else *pos2 = 1;
    }
    else
    {
        printf("error");
    }
}

void migratediff(int *pos1, int *pos2, int neigh, int n_left)
{
    float p;
    p = rand(&seed);
    if (neigh == 4)
    {
        if (p < 0.25) *pos1 = 1;
        else if (p < 0.5) *pos1 = 1;
        else if (p < 0.75) *pos2 = 1;
        else *pos2 += 1;
    }
else if (neigh == 3)
    if (*pos1 == 1)
        if (p < 1.3) *pos1 += 1;
        else if (p < 2.3) *pos2 += 1;
        else *pos2 += 1;
    else if (*pos1 == n_left)
        if (p < 1.3) *pos1 -= 1;
        else if (p < 2.3) *pos2 -= 1;
        else *pos2 += 1;
    else if (*pos2 == 1)
        if (p < 1.3) *pos1 += 1;
        else if (p < 2.3) *pos1 -= 1;
        else if (*pos2 != 1)
            else if (p < 0.5) *pos1 += 2;
            else if (p < 0.5) *pos1 -= 1;
            else *pos1 += 1;
        else if (p < 0.5) *pos1 -= 1;
        else *pos1 += 1;
    else {
        printf("error'");
    }
}

float simplepairfsf(float mean0, float mean1)
{
    return (mean1 - mean0) / (mean0 + mean1);
}

float pair_j(float **theor, int i, int j)
{
    float f0 = (theor[i][i] + theor[j][j]) / 2.;
    return (f0 / 2. - theor[i][i] / (1. - f0 / 2. - theor[i][i] / 2.));
}

float pair_j_times(float **theor, int i, int j)
{
    float t0 = (theor[i][i] + theor[j][j]) / 2.;
    return (theor[i][i] / 2. - t0 / 2. / (t0 / 2. + theor[i][i] / 2.));
}
Appendix 5: Maximum Likelihood Estimation Using SYSTAT

Commercially available packages such as SYSTAT have the facility for non-linear maximisation. As an example I show the estimation of gene flow using SYSTAT for data from a single locus from a simulated dataset. The data consists of samples from 12 demes (out of 40 total) with population size 500 diploids in each and migration and mutation rates of 0.001 and 5E-6 respectively. SYSTAT can be used to get maximum likelihood estimates of gene flow by using the NONLIN option. As an example consider the computer generated dataset below where there are 4 alleles at a locus.

<table>
<thead>
<tr>
<th>ALL</th>
<th>AL2</th>
<th>AL3</th>
<th>AL4</th>
<th>N</th>
<th>DUMMY</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>12</td>
<td>19</td>
<td>0</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>19</td>
<td>0</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>79</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>29</td>
<td>0</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>63</td>
<td>5</td>
<td>14</td>
<td>0</td>
<td>82</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>1</td>
<td>5</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>27</td>
<td>22</td>
<td>0</td>
<td>60</td>
<td>0</td>
</tr>
</tbody>
</table>

AL1..AL4 are the number of each allele at each location. N is the sample size at the location. The column DUMMY is needed by SYSTAT for the calculations.

Now to get maximum likelihood estimates for M and the \( p_i \)'s the loss function is set to be

\[
-(\text{LGM}(N+1)-\text{LGM}(AL1+1)-\text{LGM}(AL2+1)-\text{LGM}(AL3+1)-\text{LGM}(AL4+1)-
\text{LGM}(m+n)+\text{LGM}(AL1+m*p1)+\text{LGM}(AL2+m*p2)+\text{LGM}(AL3+m*p3)+\text{LGM}(AL4+m*(1-p1-p2-p3))+\text{LGM}(m)-\text{LGM}(m*p1)-\text{LGM}(m*p2)-\text{LGM}(m*p3)-\text{LGM}(m*(1-p1-p2-p3))
\]

and the model a dummy one namely

DUMMY=DUMMY

The output is

DEPENDENT VARIABLE IS DUMMY

FINAL VALUE OF LOSS FUNCTION IS 83.461

PARAMETER ESTIMATE

M 1.682
These are the maximum likelihood estimates for $4M$ and the $p_i$'s.

Now the likelihood for a null hypothesis of no subdivision i.e. assuming that the data for all populations follow the same multinomial distribution is calculated using the loss function

$$-(\text{LGM}(N+1)-\text{LGM}(\text{AL}1+1)-\text{LGM}(\text{AL}2+1)-\text{LGM}(\text{AL}3+1)-\text{LGM}(\text{AL}4+1) + \text{AL}1 \times \log(p1) + \text{AL}2 \times \log(p2) + \text{AL}3 \times \log(p3) + \text{AL}4 \times \log(1-p1-p2-p3))$$

which returns

FINAL VALUE OF LOSS FUNCTION IS 278.394

PARAMETER  | ESTIMATE
---|---
P1  | 0.554
P2  | 0.276
P3  | 0.162

Now use a likelihood ratio test, namely $2(278.4 - 83.5)$ vs Chi-squared on 1 d.f.

A 1 dimensional maximisation can be performed by using the observed gene frequencies instead of variables $P1$, $P2$ and $P3$.

FINAL VALUE OF LOSS FUNCTION IS 83.780

PARAMETER  | ESTIMATE
---|---
$M$  | 1.669

The value of $4M$ calculated is close to that by 4-dimensional maximisation. The support curve of a one-dimensional maximisation is shown below in figure A5. The expected value of $M$ is within the approximate 95% confidence limits calculated.
Figure A5 Support Curve for the data analysed in appendix 5 (Solid Line). The dashed line is the maximum value for the log likelihood at $4M = 1.669$. 
Appendix 6 Mutation Algorithms

```c
int mutate(struct node *tree, struct model_type model, long *seed)
/* simple procedure to produce mutations on a tree */
{
    int alls, muts;
    struct node *root;
    void inf alleles(struct node *node, float mutation_rate, int *total_muts, int *total_alleles, long *seed);
    void kalleles(struct node *tree, float mutation_rate, int *total_muts, int *total_alleles, long *seed);

    root = tree + (2 * model.total_sample - 2); /* root points to the earliest node */
    root->allele = 1; /* root is given allele 1 */
    alls = 1; /* number of alleles = 1 */

    if (model.mutation_type == 1)
        inf alleles(root, model.mutation_rate, &muts, &alls, seed);
    else
        kalleles(root, model.mutation_rate, &muts, &alls, seed);
    return alls; /* returns the number of alleles */
}

/* function for an infinite allele mutation model uses rand(*long) and poidev(float xm, *long) which returns a Poisson random variable with mean xm */

void inf alleles(struct node *node, float mutation_rate, int *total_muts, int *total_alleles, long *seed)
{
    int nmuts;
    float t;
    float rand(long *jc/jidum);
    float poidev(float xm, long *idum);

    /* first descendent 1 */
    t = (node->time) - (node->desc1->time); /* time for mutation to occur */
    nmuts = poidev(mutation_rate * t, seed); /* number of mutations */
    *total_muts += nmuts;
    node->desc1->n_muts = nmuts;
    if (nmuts > 0) {
        *total_alleles += 1;
        (node->desc1->allele) = *total_alleles; /* change the allele number */
    } else
        (node->desc1->allele) = (node->allele); /* has the same allele as the ancestor */

    /* then descendent 2 */
    t = (node->time) - (node->desc2->time);
    nmuts = poidev(mutation_rate * t, seed);
    node->desc2->n_muts = nmuts;
    *total_muts += nmuts;
    if (nmuts > 0) {
        *total_alleles += 1;
        (node->desc2->allele) = *total_alleles;
    } else
        (node->desc2->allele) = (node->allele);

    if (node->desc1->desc1 != NULL)
        inf alleles(node->desc1, mutation_rate, total_muts, total_alleles, seed);
    if (node->desc2->desc1 != NULL)
        inf alleles(node->desc2, mutation_rate, total_muts, total_alleles, seed);
}

void kalleles(struct node *tree, float mutation_rate, int *total_muts, int *total_alleles, long *seed)
{
    /* this is a more complex algorithm for producing mutations */
    /* it uses the Poisson distribution for the number of mutations */
    /* and the random number generator for the allele distribution */
    /* it also takes into account the relationship between the ancestors */
}
```

Appendix 6

257

/* function for a k-allele mutation model uses rand(*long) and poidev(float xm, *long) which returns a Poisson random variable with mean xm */

void kalleles(struct node *tree, float mutation_rate, int k, long *seed)
{
    int nmut, allele;
    float t;
    float rand(long *idum);
    float poidev(float xm, long *idum);
    
    t = (tree->time) - (tree->desc1->time);
    nmut = poidev(mutation_rate*t,seed);
    tree->desc1->n_muts = nmut;
    if (nmut > 0) {
        allele = rand(seed)*(k-1)+1;
        if (allele >= tree->desc1->allele) allele ++;
        (tree->desc1->allele) = allele; }
    else (tree->desc1->allele) = (tree->allele);
    t = (tree->time) - (tree->desc2->time);
    nmut = poidev(mutation_rate*t,seed);
    tree->desc2->n_muts = nmut;
    if (nmut > 0) {
        allele = rand(seed)*(k-1)+1;
        if (allele >= tree->desc2->allele) allele ++;
        (tree->desc2->allele) = allele; }
    else (tree->desc2->allele) = (tree->allele);
    if (tree->desc1->desc1 != NULL)
        kalleles(tree->desc1,mutation_rate,k,seed);
    if (tree->desc2->desc1 != NULL)
        kalleles(tree->desc2,mutation_rate,k,seed);
}

/******************************************************************************/
Appendix 7: Permutation Algorithms

```c
void permute_semi_matrix(int **new_mat, int n, int sample_size, long *seed)
/* permutes a semi-matrix A[i,j] representing genotype frequencies
assuming that marginal frequencies are constant */
{
    int i,j,k,count=1,*itemp;
    void permute_integers(int *data, int n, long *seed); /* function for permuting an array */
    itemp = ivector(2*sample_size); /* allocates and initialises a temporary integer vector */
    for (i=1;i<=n;i++)
        for (j=1;j<i;j++)
            for (k=1;k<=new_mat[i][j];k++)
                itemp[count++] = i;
    permute_integers(itemp,2*sample_size,seed); /* permutes the integers */
    /* now make the new matrix */
    for (i=1;i<=n;i++)
        for (j=1;j<=i;j++)
            new_mat[i][j]=0; /* initialise the new matrix to 0 */
    for (count=1;count<=2*sample_size;)
        if (itemp[count-1]==itemp[count+1])
            new_mat[itemp[count]][itemp[count+1]]+=1;
        else
            new_mat[itemp[count]][itemp[count+1]]+=1;
        count+=2;
    free_ivector(itemp); /* frees memory */
    return;
}

/* function to randomly permute a matrix, keeping row and column sums constant */
void permute_matrix(int **new_mat, int rows, int columns, int total, int *n, long *seed)
{
    int j,k,l,*itemp, count;
    void permute_integers(int *data, int n, long *seed); /* permutes an array of integers of length n.
 requires random number generator rand() which takes a pointer to a long integer and returns
 a float between 0 and 1 */
    count = 1;
    itemp = ivector(total); /* allocates temporary vector */
    for (j=1;j<=rows;j++)
        for (k=1;k<=columns;k++)
            for (l=1;l<=new_mat[j][k];l++)
                itemp[count++] = k;
    permute_integers(itemp,total,seed); /* permutes the integers */
    count = 1;
    for (j=1;j<=rows;j++)
        for (k=1;k<=columns;k++)
            new_mat[j][k] = 0;
        for (k=1;k<=n[j];k++)
            new_mat[j][itemp[count++]] = 1;
    free_ivector(itemp); /* frees temporary vector */
    return;
}

/* */
void permute_integers(int *data, int n, long *seed)
/* permutes an array of integers of length n. Requires random number generator rand() which takes a
pointer to a long integer and returns a float between 0 and 1 */
{
    int temp,i,which;
    float rand(long *seed);
```
for (i=1;i<=n;i++) {
    which = 1+(rand(seed)*n); /*integer between 1 and n*/
    temp = *(data+i);
    *(data+i) = *(data+which);
    *(data+which)=temp;
}
/*****************************/
Appendix 8: Genealogies and geography

Corrected Proofs


Genealogies and geography

N. H. Barton and I. Wilson

3.1 Introduction

Any set of homologous genes can be traced back to a single common ancestor, and so their evolutionary relationship can be described by a genealogy, such as that shown in Fig. 3.1a. With asexual reproduction, every gene has the same ancestry. For example, since human mitochondrial DNA is inherited maternally, there is no opportunity for recombination, and so all mitochondria must trace back to one maternal ancestor. By contrast, sexual reproduction allows recombination, so that different genes, or even different segments of the same gene, have different genealogies (Fig. 3.1b). The ancestral lineages of each gene might coincide, so that even in a sexual population, one individual might by chance contribute the entire future ancestry (top of Fig. 3.1b). In the more distant past, all segments of the genome presumably trace back to one ancestral gene, via successive duplications. Here, however, we will be concerned with the immediate genealogy, and what this tells us about current evolutionary processes. The discordant genealogies generated by recombination allow natural selection to select efficiently on individual loci, and also give us very much more information about the whole population. However, it makes it difficult (if not impossible) to estimate the whole set of genealogies that give the relationships between genomes, and greatly complicates the inferences that can be made from that set. In this chapter we analyse the structure of genealogies where genes diffuse through a two-dimensional habitat, and discuss how genealogical information might be used to estimate the rate and pattern of gene flow.

Until recently, population genetics was based on allele frequencies, or occasionally on genotype frequencies. In particular, the relative rates of gene flow and genetic drift have traditionally been estimated using the standardized variance of allele frequencies across subpopulations (Fst—Wright et al. (1942)). DNA sequencing now presents us with data which are more naturally represented by genealogies. Every sequence may be unique, in which case allele frequencies tell us nothing; all the information is contained in the genealogical relationship between sequences. If recombination is rare enough relative to mutation, the genealogy can be seen more or less directly (as for example in bacteria where recombination occurs via occasional transformation (Maynard Smith 1990) or in viruses, where mutations are frequent (Sharp et al. 1996)).
Fig. 3.1 (a) A typical genealogy for one locus in one panmictic population. Initially there are many pairs of lineages, and so coalescence events are very frequent; as lineages coalesce, further coalescence becomes rarer. For this example, it takes $2.77N$ generations for 20 lineages to coalesce into two, and then a further $3.14N$ generations for these two to meet in the common ancestor. Note that time is plotted on a square-root scale; on a linear scale, early coalescences are indistinguishable. (b) Genealogy for two loci in one panmictic population, illustrating the discordance between genealogies caused by recombination at a rate $r$. In each generation, there is a chance $r$ that genes carried together will have been inherited from different parents (full circles). Lineages coalesce at random, and so may come back into association (full squares). All genes at the first locus become identical by descent at time $5.91N$, whilst those at the second become identical at $4.10N$. Further back in time, lineages at the two loci will sporadically come together on one chromosome for expected times of $1/r$ generations, separated by intervals of $2N$ generations (top of graph).
Genealogies and geography

Even if we cannot be sure of the relationships, the theory may still best be described in terms of the underlying genealogies. For example, if we can calculate the likelihood of parameters such as the population size or the rate of gene flow for a particular genealogy, the overall likelihood can be found as a weighted sum over the set of plausible genealogies (Felsenstein 1992). Even if our statistical inferences are based on allele frequencies, it may still give us more insight to derive the theoretical predictions via a genealogical approach.

Here we will assume that we have the full genealogy of each segment of the genome, including the times at which different lines of descent coalesce. Of course, estimating the genealogy itself is hard. However, by considering the ideal data we can at least set an upper limit on what can be inferred about the processes of evolution. How much more valuable is it to use information from the full set of genealogies, rather than from allele frequencies, or pairwise relationships between genes? What is the best sampling scheme—is it better to have small genealogies across many loci, or genealogies for a few loci, but each containing many individuals? We concentrate on the specific question of how to analyse genealogies with geography, and in particular, how to estimate rates of gene flow and genetic drift. We first show the close relationship between the classical theory of identity by descent and the genealogical structure, and then develop a simple diffusion approximation which describes how lines of descent coalesce. This approximation applies to a variety of local population structures, and so could be used to make robust inferences about effective population densities and rates of gene flow. However, we show that simple estimators based on the rate of dispersion of lineages over time can be misleading, and that in a two-dimensional population, the genealogical structure depends primarily on long-term history. Finally, we discuss possible ways of making inferences about this history.1

This chapter is to a large extent a synthesis of existing analyses: the classic work of Wright (1943b) and Malecot (1948) on identity by descent; Slatkin’s (1991) work on the structure of genealogies in stepping-stone models, and its application to measuring gene flow (Slatkin and Maddison 1990; Hudson et al. 1992); Felsenstein’s (1992) likelihood methods; and Neigel, Ball, and Avise’s (Ball et al. 1990; Neigel et al. 1991; Neigel and Avise 1993) use of genealogies to infer population histories. The methods here aim primarily at estimating rates of gene flow and genetic drift, rather than distinguishing qualitatively different population histories. Thus, they complement Crandall and Templeton’s methods, which make qualitative distinctions between alternative hypotheses (Crandall and Templeton 1993; Templeton 1992).

3.2 The coalescent process

First, consider a single panmictic population containing \(2N\) genes. If a small fraction of these genes are sampled, then their relationship can be approximated

1 A condensed version of this chapter is to be found in Barton and Wilson (1995).
very simply: there is a probability $1/2N$ that any two lines of descent will coalesce in a common ancestor in each generation (Kingman 1982). Thus, if there are $k$ genes, there are $k(k-1)/2$ pairs which might coalesce, and the time back to the first coalescence is exponentially distributed with expectation $4N/k(k-1)$ generations. There are then $(k-1)$ lines of descent remaining, and so the expected time back to the previous coalescence is $4N/(k-1)(k-2)$ generations. The expected age of the whole genealogy (that is, the expected time back to the common ancestor) is thus $4N[1/k(k-1)+1/(k-1)(k-2)\ldots+1/2]=4N(1-1/k)$ generations, which tends to $4N$ generations for a large sample (Felsenstein 1992).

This calculation shows that the structure of any genealogy is very variable. Looking back, the average time taken for a large number of lineages to coalesce to two ancestors is the same as the time these two remaining lineages take to merge (Fig. 3.2a). Since the latter time is exponentially distributed, and since any two randomly chosen genes have a chance of $1/3$ of being related via the last common ancestor, the average divergence time between randomly chosen pairs of genes has a high variance, regardless of how large a sample is taken. Felsenstein (1992) uses this argument to show that pairwise statistics are much less efficient than estimators which include the genealogical structure. Essentially the same consideration applies with geographical structure, and so we summarize the argument here. The population size could be estimated from the average pairwise divergence time $\bar{t}$, using the relationship $\hat{N}=E(\bar{t})/2$. (In practice, $\bar{t}$ could itself be estimated from the average pairwise sequence divergence.) However, the variance of $\hat{N}$ tends to $2N^2/9$ for large samples, rather than to zero (Felsenstein 1992, eqn 18). By contrast, the maximum likelihood estimator (m.l.e.) is

$$\hat{N}=\sum_{j=2}^{k} j(j-1) t_j/4(k-1),$$

where $t_j$ is the time for which there are $j$ lines of descent. The variance of the m.l.e. is $N^2/(k-1)$, which does tend to zero as sample size increases (Felsenstein 1992, eqn 7). This basic argument suggests that genealogies will also give much more information about spatial structure than will pairwise measures such as Wright’s $F_{is}$, or the statistics proposed by Slatkin and co-workers (Slatkin 1989; Slatkin and Hudson 1991; Hudson et al. 1992). However, it also suggests that information from any one genealogy may be misleading, so that reliable estimates may require data from many loci.

This theory of the coalescent process is equivalent to the classical theory of identity by descent. Two genes are said to be identical by descent, relative to an ancestral population $t$ generations in the past, if they derive from the same gene in that population (Malecot 1948). Clearly, the probability that two genes coalesce at time $t$, $f_t$, is just the difference between the probability of identity by descent relative to ancestral populations at times $t$ and $t-1$ ($f_t=F_t-F_{t-1}$, where $F_t$ denotes the probability of identity by descent via a population $t$
Fig. 3.2 (a) A genealogy connecting 10 genes, plotting time on one axis, and one spatial axis on the other. These trees are drawn assuming random reproduction, with no local density regulation. Time is drawn on a square-root scale. (b) The same tree in two dimensions, plotting locations in space, and drawing lines to indicate the relationships. Black circles indicate genes, and the open circle the common ancestor. The species' range is assumed to be infinite, corresponding to zero density and zero neighbourhood size. With finite range, and hence non-zero density and neighbourhood size, these plots would be wrapped around, confusing the relation between geography and genealogy for large times.
It is important to realize that the probability of identity by descent is purely a description of the genealogy, and does not depend on the allelic state of the genes. The latter is described by the probability of identity in state, which is the chance that two randomly chosen genes share the same allele. This depends on how alleles are identified, and on the mutational process, but is not defined relative to any base population. The probabilities of identity by descent and identity in state are closely related, and indeed are often confused with each other. If there are infinitely many alleles, and a rate $\mu$ per generation of mutation to a novel allele, then the probability of identity in state is

$$F = \sum_{i=1}^{\infty} (1-\mu)^{2i}(\bar{F}_i - \bar{F}_{i-1}) = \sum_{i=1}^{\infty} (1-\mu)^{2i}f_i.$$  

(Throughout, we assume discrete generations.) Equation (3.1) applies to any population structure, and shows the close relation between identity in state ($F(\mu)$), identity by descent ($\bar{F}_i$), and the distribution of coalescence times ($f_i$). Moreover, eqn (3.1) shows that if $F$ is considered as a function of $z = (1-\mu)^2$, then $F(z)$ is the generating function for the distribution of coalescence times. Then, $\bar{F}/\bar{z} = f_0$, and $\bar{F}/\bar{z} = k^t \tilde{E}(t^k)$. The wealth of existing results on identity in state in spatially structured populations therefore leads directly to the distribution of coalescence times.

### 3.3 Gene flow and the coalescent process

The pattern of neutral genetic variation can be found by superimposing random changes in allelic state on the genealogy. This is possible because, by definition, neutral mutations do not affect reproduction, and so are independent of the genealogy. It is tempting to treat the movement of genes in an analogous way: just as genes change their allelic state through random mutation, so their location changes as the organisms which carry them disperse. This idea was introduced by Cavalli-Sforza and Edwards (1964) as the 'Brownian/Yule process'; Edwards (1970) discusses the joint inference of ancestral locations and times. (The process was originally conceived as a model of the whole population, run forward in time, in contrast to the present application.) Here we use this model to derive a simple expression for the joint distribution of geographic location and genealogical structure. However, we will see that this expression cannot describe natural populations because it implicitly ignores local regulation of population density (Felsenstein 1975). Nevertheless, it leads us to a diffusion approximation which does apply to a variety of actual population structures.

Consider a large population of $2N$ genes, which are distributed over a two-dimensional habitat with area $A$. A sample of genes is taken at random from the population, without regard to their location. If we make the apparently
Genealogies and geography

innocent assumption that reproduction is independent of the position of the genes in the sample (or their ancestors), then their relationship should have the same distribution as in a single panmictic population, the intervals between coalescence times $t_j$ being exponentially distributed with expectation $4N/j(j-1)$. If we now assume that genes move in a Gaussian random walk, with variance $\sigma^2$ per generation along each of the two axes, we can write the joint distribution of coalescence times and locations. For example, the probability that two randomly chosen genes are related through a common ancestor $t$ generations ago, and are at positions $x, x'$ is:

$$\psi(x, x', t) = \frac{1}{2t Nb} \frac{dxdx'}{A^2} (1-1/2N) \exp(-|x-x'|^2/4\sigma^2)$$

where $Nb = 4\pi \rho \sigma^2$ is Wright's (1943b) neighbourhood size, and $\rho = N/A$ is the population density. Neighbourhood size plays a crucial role in determining the relative rates of gene flow and genetic drift in a two-dimensional population; roughly speaking, it is the number of individuals within one generation's dispersal range. The expression for a set of $k$ genes would involve a similar product of the exponential distribution of coalescence times, with the Gaussian distribution of locations, given those times.

In fact, genes are sampled from particular locations. We therefore require the distribution of coalescence times conditional on location: $\psi(t|x, x') = \psi(x, x', t) / \psi(x, x')$, where $\psi(x, x')$ is the chance that genes will be found at $x, x'$. Summing over time:

$$\psi(x, x) \frac{dxdx'}{Nb} \frac{A^2}{dxdx'}$$

for $|x-x'| = 0, N \gg 1$ (3.3a)

$$\psi(x, x') \frac{dxdx'}{Nb} = \frac{K_0(|x-x'|/\sqrt{2N\sigma^2})}{\sqrt{A^2}}$$

for $|x-x'| \gg \sigma$ (3.3b)

$$\approx \frac{\log(\sqrt{2N/|x-x'|})}{\sqrt{A^2}}$$

for $\sigma \gg \sqrt{2N/|x-x'|} \gg \sigma$ (3.3c)

where $K_0$ is the modified Bessel function.

This expression raises two difficulties. First, it implies extreme clumping: eqn (3.3a) shows that the density near to a randomly chosen individual is increased by a factor $\approx \log(\sqrt{2N})$ above the average. This tendency for randomly dispersing and reproducing populations to become clumped was first emphasized by Felsenstein (1975) as a criticism of classical models of identity by descent. It can be seen as a consequence of the heterogeneous structure of
random genealogies discussed above. On average, any large genealogy takes $2N$ generations to coalesce into two clades, and then a further $2N$ generations for the remaining two lineages to meet. Hence, the population tends to evolve into distantly related and widely scattered clusters.

The second difficulty is that we have not properly accounted for the finite range of the species. At equilibrium, a random set of genes will have spread over an area of $\approx 8N\sigma^2 = (2N\pi A)$. Thus, unless neighbourhood size is very small, most sets of genes will be related by lineages that have crossed the species’ range several times. This is true however large the range, because for given neighbourhood size, divergence time increases in proportion to area. Hence, to be consistent, we must allow for either the finite spatial range, or the finite age, of the species.

It is traditional to allow for finite range by assuming that the habitat is wrapped over the surface of a torus of size $D \times D = A$, so that a movement $x$ is equivalent to a movement of $(x + JD)$, where $i$ represents a pair of integers $(i_1, i_2)$. Now, eqn (3.2) must be replaced by a sum over all equivalent locations:

$$\psi (x, x', t) \frac{dx dx'}{2t N b A^2} = \left(1 - \frac{1}{2N}\right)^t \sum_{j \neq 0} \exp \left(-\frac{1}{4} \frac{x - x' + jD}{\sigma^2 t}\right).$$

Assuming a toroidal habitat should be seen as a convenient approximation to more realistic models, with complicated boundaries, and where individuals near the edge might be reflected back into the range. A square habitat with reflecting boundaries would be described in essentially the same way, by folding over the square rather than by wrapping around the torus.

The spatial distribution is found by summing over time, and applying the same approximations that led to eqns (3.3):

$$\psi (x, x') \frac{dx dx'}{N b A^2} = \left[\log(2\sqrt{N}) + \sum_{j \neq 0} K_0(\sqrt{\frac{|x|^2 + 2\pi j}{2N b}})\right]$$

for $|x' - x'| = 0 \Rightarrow 2N \gg 1$ (3.5a)

$$\psi (x, x') \frac{dx dx'}{N b A^2} \approx \left[\log(\sigma \sigma / |x - x'|) + \sum_{j \neq 0} K_0(\sqrt{\frac{|x'|^2 + 2\pi j}{2N b}})\right]$$

for $\sigma \gg 2N \gg |x - x'| \gg \sigma$. (3.5b)

For large neighbourhood size, the sum over Bessel functions can be approximated by an integral, and tends to 1. Figure 3.3 shows examples of the clumping

---

2 Recall that $\sigma^2$ is the variance of distance moved along each of the two axes.
Appendix 8: GENEALOGIES AND GEOGRAPHY

Fig. 3.3 (a) The correlation in density as a function of distance, with random reproduction \( A^2 \psi(x, x') \) from eqn (3.5). This gives the density at a point, conditional on the presence of a gene a distance \( |x - x'| \) away. Neighbourhood size is \( Nb = 10 \). The upper curve is for a species' range of \( A = 10^6 \sigma^2 \), whilst the lower curve is for \( A = 10^8 \sigma^2 \).

(b) Conditional distribution of coalescence times for genes separated by \( |x - x'| = 0, 5 \sigma, 10 \sigma \) (from the ratio of eqns (3.4) and (3.5)).

produced by random reproduction (Fig. 3.3a), and the distribution of coalescence times, conditional on location (Fig. 3.3b). Since the degree of clumping is small for large neighbourhood size, the conditional distribution does not differ much from the raw distribution.

Sedentary organisms may not have had time to diffuse over the species' range over the time since they occupied it. For example, the flightless alpine grasshopper *Podisma pedestris* has a dispersal range of \( \sigma \approx 20 \text{ m/generation}^{1/2} \) (Barton and Hewitt 1982) and an annual life cycle. It has occupied its present range since the glaciers retreated \( (T < 10^4 \text{ generations}) \). In that time, genes would diffuse only \( \approx 2 \text{ km} \). In such cases \( (\sigma^2 T \ll A) \) clustering is limited by finite age rather than finite range. In general, we should consider a sporadic series of
extinctions and recolonizations over the whole history of the organism. Instead, we make the simpler assumption that the range was occupied by \( N \) diploid individuals \( T \) generations ago, distributed uniformly without regard to their relationship. The joint distribution of locations and genealogies is then given by equ (3.2) (or its multigene extension) for \( t \ll T \), and by a uniform spatial distribution for more ancient relationships \( (t > T) \). For two genes

\[
\psi(\bar{x}, \bar{x}', t) d\bar{x} d\bar{x}' = \frac{1}{2t N b} \frac{d\bar{x} d\bar{x}'}{A^2} (1 - 1/2N) e^{-(|\bar{x} - \bar{x}'|/4\sigma^2 t)} t < T \tag{3.6a}
\]

\[
\psi(\bar{x}, \bar{x}', t) d\bar{x} d\bar{x}' = g(t) \frac{d\bar{x} d\bar{x}'}{A^2} t \geq T
\]

\[
(1 - 1/2N) \sum_{i=1}^{\infty} g(i) = (1 - 1/2N)^N \tag{3.6b}
\]

where \( g(t) \) is the arbitrary distribution of relationships among pairs of colonists. Summing over time to obtain the distribution of spatial locations

\[
\psi(\bar{x}, \bar{x}', t) d\bar{x} d\bar{x}' = \frac{d\bar{x} d\bar{x}'}{A^2} \left[ (1 - 1/2N)^T + \frac{1}{2Nb} \sum_{i=1}^{T} (1 - 1/2N) e^{-(|\bar{x} - \bar{x}'|/4\sigma^2 t)} \right] \tag{3.7a}
\]

\[
\approx \frac{d\bar{x} d\bar{x}'}{A^2} \left[ 1 + \log T + \frac{T}{2Nb} \right] \quad \bar{x} = \bar{x}', \; T \ll 2N \tag{3.7b}
\]

\[
\approx \frac{d\bar{x} d\bar{x}'}{A^2} \left[ 1 + \log(4\sigma^2 T / |\bar{x} - \bar{x}'|^2 / 4\sigma^2) + \gamma \right] \quad \bar{x} = \bar{x}', \; T \ll 2N \tag{3.7c}
\]

where \( \gamma = 0.5772 \) is Euler's constant. Hence, the conditional distribution of coalescence times is

\[
\psi(t|\bar{x}, \bar{x}') \approx \frac{1}{t(2Nb + \log T + \gamma)} T \ll 2N \tag{3.8a}
\]

\[
\psi(t|\bar{x}, \bar{x}') \approx \frac{\exp(-|\bar{x} - \bar{x}'|^2 / 4\sigma^2 t)}{t(2Nb + \log(4\sigma^2 T / |\bar{x} - \bar{x}'|^2 / 4\sigma^2) + \gamma)} \sigma \ll |\bar{x} - \bar{x}'| \ll \sigma \sqrt{T} \tag{3.8b}
\]
These models of random reproduction give a consistent conclusion for times which are short compared with the age of the population or the time taken for genes to diffuse across the species' range \((t < T, A/\sigma^2)\). The distribution of coalescence times is then given by 
\[
\frac{\lambda}{t} \exp\left(-\frac{(x-x')^2}{4t}\right)/(2tN)
\]
where \(N_0(jx-x')\) is the effective neighbourhood size, which is increased somewhat above \(4\pi p\sigma^2\) by the clustering that occurs in the absence of local density regulation. While we have given explicit results only for pairs of genes, it is easy to generate the joint distribution of genealogies and locations for arbitrary numbers of genes, or to draw random realizations of this distribution (Fig. 3.2).

### 3.4 A diffusion approximation to the coalescent process in two dimensions

The model developed in the previous section is simple, but unrealistic: in nature, fitness must decrease with local density. Densities may vary from place to place, but in response to local carrying capacity, rather than to the unchecked accumulation of demographic fluctuations. The model of a locally unregulated population might describe the random evolution of morphology, and would give an explanation for the clustering of asexual phenotypes into species (Higgs and Derrida 1991). However, it is implausible as a description of population dynamics in two spatial dimensions.

There have been extensive treatments of the extreme case where individuals are grouped into demes whose density is absolutely regulated. Analyses of such stepping-stone models have dealt primarily with identity in state (Wright 1943b; Malecot 1948; Kimura and Weiss 1964; Maruyama 1972; Felsenstein 1976; Nagylaki 1974, 1988). However, since the identity in state is the generating function for the distribution of coalescence times eqn (3.1), it leads immediately to the distribution of coalescence times (see below). Slatkin (1991) and Nei and Takahata (1993) have derived the mean coalescence time. However, the results would be cumbersome to extend to higher moments. Here we develop a simple diffusion approximation which applies to all but local scales, and which extends to whole genealogies.

The approximation is based on Wright's (1943b) argument that ancestors can be considered as being drawn from a neighbourhood whose size increases with time into the past. Thus, the probability that two nearby genes are identical by descent in the previous generation is \(1/2N\); in two dimensions, the pool of ancestors is spread over an area whose area increases linearly with time, and so the probability of identity by descent \(t\) generations back is \(1/2tN\).

Let \(f(t,x'\mid x')\) be the probability that genes at \(x\) and \(x'\) are identical by descent via an ancestor \(t\) generations back, who lived at \(x\). (With two genes, we will not need to keep track of the positions of ancestors; however, this is necessary to extend the argument to more genes; see eq (3.14).) Let \(g(t,\hat{y})\) be the chance that a gene at \(x\) derived from an ancestor at \(y\) in the previous generation. We assume that \(g\) depends only on the distance between parent and offspring, has zero mean, and has variance \(\sigma^2\) along each of the two axes. Let
Appendix 8: GENEALOGIES AND GEOGRAPHY

N. H. Barton and I. Wilson

\[ f(1, y|x, x') = \frac{1}{2N_b} h(\tau, x, x') = \frac{g(\tau, x)g(\tau, x')}{2\rho} \]

where \( h(\tau, x, x') = \frac{1}{2N_b} \int g(\tau, x)g(\tau, x')d\tau \).

This argument requires several approximations. It is assumed that the lines descend independently, so that the joint probability of movement of two genes is \( g_1(\tau, x)g_2(\tau, x') \). It is assumed that one can choose an area \( d\tau \) large enough that \( 1/2\rho\delta\tau < 1 \), but small enough that \( f \) is approximately constant within it. These assumptions hold for a demic structure with strict density regulation, in which case equations (3.9) give the coalescence times exactly. They are approximations to models of truly continuous populations; in the cases we consider, the approximation is remarkably good (see Fig. 3.4).

Fig. 3.4 Comparison of simulated distributions of coalescence times from a two-dimensional stepping-stone with 10 diploid individuals per deme and a migration rate of 0.05 in each direction (dotted curves), to a Gaussian approximation for the distribution of times (eqns (3.10); full curves). Simulations were based on a grid of 100 by 100 demes over 2000 generations. The simulated cumulative frequency is based on 8000 replicates for each distance. The distances are zero demes \((d=|x-x'|=0)\), one deme in both directions \((d=|x-x'|=4.47\rho)\), two demes in both directions \((d=|x-x'|=8.95\rho)\), and five demes in both directions \((d=|x-x'|=22.2\rho)\). Neighbourhood size \(=6.28\).
Equations (3.9) give a recursion across one generation which relates the probability of coalescence at time \( t \) to that at time \( (t - 1) \). They can be rewritten as a recursion across many generations, which leads naturally to the diffusion approximation. Let \( g_t(y, \bar{x}) \) be the chance that a gene at \( \bar{x} \) descended from an ancestor at \( y \) \( t \) generations back. Extend the definition of \( h \) to

\[
\frac{1}{2N_b} h_t(y', \bar{x}, \bar{x}') = \frac{g_t(y, \bar{x}) g_t(y, \bar{x}')}{2t}.
\]

This is the chance that two genes both descend from an ancestor at \( y \), \( t \) generations back, ignoring the possibility of more recent coancestry. Applying eqn (3.9b) recursively:

\[
f(t, z|\bar{x}, \bar{x}') = \frac{1}{2N_b} \left[ h_t(y, \bar{x}, \bar{x}') - \sum_{i=1}^{t-1} \int f(t-i, z|y, \bar{x}) h_i(y, \bar{x}, \bar{x}') \, dy \right].
\]

This is the chance that the two genes descend from a common ancestor at \( z \), subtracting the probabilities that they were identical by descent in any of the intervening generations.

By the central limit theorem, \( g_t(y, \bar{x}) \) tends to a Gaussian with variance \( \sigma^2 t \) for large \( t \); \( h_t \) tends to a Gaussian with variance \( \sigma^2 t/2 \), being the distribution of locations of the common ancestor. If we average over the location of the ancestors \( (f(t|\bar{x}, \bar{x}')) \), and use the fact that \( f(t|\bar{x}, \bar{x}') \) is independent of \( \bar{x} \), eqn (3.10b) simplifies to

\[
f(t|\bar{x}) = \frac{1}{2N_b} \left[ 1 - \sum_{i=1}^{t-1} \frac{f(t-i|\bar{x}, \bar{x})}{i} \right]
\]  
(3.11a)

\[
f(1|\bar{x}) = \frac{1}{2N_b}
\]  
(3.11b)

\[
f(2|\bar{x}, \bar{x}') = \frac{1}{2N_b} \left[ \frac{1}{2} - \frac{1}{2N_b} \right]
\]  
(3.11c)

\[
f(3|\bar{x}, \bar{x}') = \frac{1}{2N_b} \left[ \frac{1}{3} - \frac{1}{2N_b} - \frac{1}{2N_b} \left[ \frac{1}{2} - \frac{1}{2N_b} \right] \right]
\]  
(3.11d)
Appendix 8: GENEALOGIES AND GEOGRAPHY

36 N. H. Barton and I. Wilson

The distributions of coalescence times predicted by this Gaussian approximation are shown in Fig. 3.4, together with simulation results. Agreement is close over all but short times and nearby genes. In most natural populations, \( F_{st} \) is low (< 0.10, say; Slatkin, 1987), implying that neighbourhood size is large. In this limit, the distribution of coalescence times in this model of absolute density regulation converges to that developed in the previous section for the case of no local regulation.

These recursions for the distribution of coalescence times lead to parallel recursions for the identity in state. Applying eqn (3.1) to eqn (3.9b):

\[
F(x, x') = (1 - \mu)^2 \int f(y, y') f(y', x') \left[ F(y, y') + \frac{1}{2\rho} (1 - f(0)) \delta(y - y') \right] dy dy' \tag{3.12}
\]

where \( \delta(y) \) is the Dirac delta function, and the probability that nearby genes are identical has been rewritten as \( F(x, x) = F(0) \) to emphasize that it is independent of location. Equation (3.12) is identical to Malecot's (1948) model if (as Malecot assumed) the dispersal distribution is Gaussian. It applies exactly to stepping-stone models if the integrals are replaced by sums, and is an approximation to continuous models which neglect the interactions between nearby genes caused by local density dependence. The Appendix shows that for arbitrary dispersal distributions eqn (3.12) is approximated by

\[
F(0) = \left( 1 + \frac{Nb}{\log(\sigma/\sqrt{1 - (1 - \mu^2)})} \right)^{-1}
\]

\[
F(x, x') = \frac{1 - F(0)}{Nb} K_0 \left( \frac{|x - x'|}{\sigma \sqrt{1 - (1 - \mu^2)}} \right)
\]

for \(|x - x'| > \sigma, K, \text{ and } \mu \ll 1\). \( \tag{3.13b} \)

Equations (3.13) give the probability of identity in state, assuming infinitely many alleles; the same expressions give \( F_{st} \) measured from the variance in allele frequencies at loci with a finite number of alleles. \( K \) is a characteristic scale which depends on the local structure of the population. For a stepping-stone model with nearest-neighbour migration on a square grid, \( K = (\text{deme spacing})/\sqrt{32} \). For Malecot's model of Gaussian dispersal, \( K = \sigma \). If mutation rates are low (\( \mu \ll 1 \)), \( 1 - (1 - \mu^2) \approx 2\mu \), and eqns (3.13) depend on the scale \( l = \sigma/\sqrt{2\mu} \). There will be significant fluctuations over local scales \((\approx \sigma, K)\), but there will be correlations between allele frequencies over the much longer scale \( l \). Equation (3.13b) is a close approximation to continuously distributed populations, and breaks down only over local scales (Fig. 3.5). Agreement is similarly close for stepping-stone models, even for neighbouring demes. The distribution of coalescence times can be calculated by differentiating eqn (3.13b) with respect
to \( z = (1 - \mu)^2 \). The breakdown of eqn (3.13b) for small \( |x - x'|/\sigma \) corresponds to the breakdown of the Gaussian approximation to eqn (3.9b) for small coalescence times (Fig. 3.4).

![Fig. 3.5 Identity by descent in a continuously distributed population. The full curve shows the Bessel function approximation of eqns (3.13), whilst the bars show simulation results. The mutation rate is \( \mu = 0.01 \). Haploid individuals disperse in a Gaussian distribution with variance \( \sigma^2 \). They reproduce in discrete generations, with a number of offspring drawn from a Poisson distribution with mean \( 4(1 - \rho/n) \), where \( \rho = 0.16 \). This ensures density-dependent regulation towards \( \rho/k = 0.12 \). The actual density averaged 0.106, and the effective neighbourhood size (based on identity by descent in the previous generation) was 0.66. \( \rho \) is the local density, measured by weighting neighbours according to a Gaussian curve with variance \( \sigma^2 \). Simulations were on a torus, \((1000)^2\), for two replicates of 150 generations.

The approximation that led to eqns (3.10) and (3.11) extends to many genes. It becomes more complicated, however, because the recursions involve the locations of all the ancestors. Usually, we are concerned only with the present locations of the genes, and so must average over the locations of the ancestors. This involves an integral over a Gaussian, which can be carried through using a symbolic algebra package such as Mathematica (Wolfram 1991). By the same arguments that led to eqns (3.10):

\[
\begin{aligned}
f_{ij}(t, z_{ij}) &= \frac{1}{2Nb} \left[ h_{ij}(t, z_{ij}) - \sum_{k=1}^{i-1} \sum_{l=1}^{j-1} \left( f_{ij}(t - \tau_{ij}, z_{ij}) \frac{k_{ij}(t, z_{ij})}{\tau} \right) \right] \\
\end{aligned}
\tag{3.14}
\]

Here, \( x \) represents a set of \( n \) locations \( \{x_1, x_2, \ldots, x_n\} \). \( z_{ij} \) represents a set of \( n \) locations, with \( z_i = z_j = z_{ij} \). \( f_{ij}(t, z_{ij}) \) is the chance that out of a set of \( n \) genes at locations \( x \), the first coalescence is in generation \( t \), between lineages \( i \) and \( j \), at
locations $\zeta_{ij}^v$: $h_{ij}(t,\zeta_{ij}^v|x)/2Nb$ is defined in the same way as in eqn (3.9a), as the chance that two lineages, starting at $x_1$ and $x_h$, coalesce $t$ generations back, at $\zeta_1=\zeta_h$, and that the other lineages at $\zeta_k$ descend from $\zeta_k$. For large $t$, it tends to a Gaussian:

$$h_{ij}(t,\zeta_{ij}^v|x) = \exp\left[-\frac{(\zeta_y-x_y)^2}{\sigma^2t}\right] \prod_{k \neq i,j} \exp\left[-\frac{(\zeta_k-x_k)^2}{2\sigma^2t}\right]$$ \hspace{1cm}(3.15)

where $x_y=(x_1+x_h)/2$. Though cumbersome, this recursion could be carried through symbolically.

### 3.5 Estimating the rate of gene flow

The recursions developed in the preceding sections allow calculation of the distribution of coalescence times amongst genes sampled from a two-dimensional population. However, it would not be easy to use this distribution to make statistical estimates. Here we outline possible methods for estimating the rate of gene flow ($\sigma$), assuming that genes diffuse through a stable and homogeneous population. The traditional method for inferring population structure from genetic data was introduced by Wright (Wright et al. 1942; Wright 1943a). Geographical variation in allele frequency is generated by genetic drift, and reduced by gene flow. In the island model, the balance between these processes is determined by $Nm$, the number of migrants exchanged between demes; the standardized variance of allele frequencies across demes is $Fst=1/(1+4Nm)$ for large deme size and low migration rates. In two dimensions (though not in one), the relationship is similar, with $Fst$ decreasing with neighbourhood size ($Nb=4\pi r^2$; eqns (3.13)).

Almost all analyses of population structure infer $Nm$ or $Nb$ from $Fst$, or from some equivalent measure of variation in allele frequency, such as Slatkin’s private allele method (Slatkin 1985; Barton and Slatkin 1986). However, the spatial pattern of allele frequencies also contains information. If (as is usually the case) $Fst$ is small, allele frequencies will fluctuate rather little, and the whole distribution can be approximated by a multivariate Gaussian, which is defined by its mean and covariance. This covariance is given by eqns (3.13), and depends on two scales: the scale which describes local population structure, and the scale $l=\sigma/\sqrt{2\mu}$, which describes the balance between mutation and gene flow. Thus, if the mutation rate is known, the rate of gene flow ($\sigma$) can be estimated. This approach was first used by Sokal and Wartenberg (1983), and has been explored more recently by Epperson (1989, 1993). It is important to realize that it gives estimates of both the number of migrants or neighbourhood size ($Nm$ or $4\pi r^2$), and the proportion of migrants or rate of diffusion ($m$ or $\sigma$): confusingly, both are referred to as the ‘rate of gene flow’.

To illustrate this approach, and to provide a benchmark for comparison with
genealogical methods, consider a set of \( k \) samples, each of \( 2n \) genes, spread over a two-dimensional area. The frequency of alleles at \( j \) independent loci is measured, giving a set of \( jk \) allele frequencies. (Since we will only consider the limit of large sample size, the variance of the estimates scales inversely with the number of loci.) Assume that samples are large enough, and \( F_{st} \) small enough, that the joint distribution of allele frequencies is approximately Gaussian. Samples are taken from points separated by much more than one dispersal range, so that the approximation of eqns (3.13) is accurate, regardless of the local population structure. First, consider estimation of \( F_{st} \), disregarding spatial information for the moment. The maximum likelihood estimate is

\[
\hat{F}_{st} = \frac{(\text{var}(p)/pq - 1/2n)}{\text{var}(p)/pq - 1/2n}.
\]

Since the coefficient of variation (c.v.) of an estimate of a variance tends to \( \sqrt{2/jk} \) for large samples, the c.v. of \( \hat{F}_{st} \) tends to \( \sqrt{2/jk(1 + 1/2nF_{st})} \) in large samples. For a given total number of genes \( (2njk) \), the best sampling scheme is to set \( 1/2n = F_{st} \); then, the c.v. is \( \approx \sqrt{8/jk} \). The heavy line in Figs. 3.6a and 3.6b shows the coefficient of variation of \( \hat{F}_{st} \), plotted against the number of sample sites; for these parameters, the greatest accuracy is when there are 100 samples of \( 1/F_{st} \approx 10 \) genes. (There has been considerable discussion as to the best estimators of \( F_{st} \) and \( Nm \) (Weir and Cockerham 1984; Slatkin and Barton 1990; Cockerham and Weir, 1993); here, we consider only large samples, in which case the maximum likelihood estimate has minimum variance, and bias which tends to zero.)

An estimate based solely on the variance among samples neglects the correlations among samples, and cannot distinguish the three parameters \( Nb, K, \) and \( l = 4\sigma \sqrt{2\mu} \). Figure 3.6 shows the c.v. of estimates based on fitting the full covariance matrix, calculated using eqns (3.13), and plotted against neighbour- hood size \( (Nh) \). Figure 3.6a shows results when samples are taken over an area large compared with \( l \); in this case, all three parameters can be distinguished, although large samples are needed to give accurate estimates of the two scales \( K \) and \( l \). Estimates of neighbourhood size are more accurate; their coefficient of variation is only slightly lower than \( \sqrt{8/jk} \), because correlations between samples are weak. These calculations are encouraging, in that they show that data on allele frequencies can (with large enough samples, and with a time-scale calibrated by knowledge of \( \mu \)) give separate estimates of neighbourhood size and dispersal rate. However, the method assumes that the population has been at equilibrium for at least \( 1/\mu \) generations, which is unlikely for most organisms. One should therefore sample over scales smaller than \( l \) and \( \sigma \sqrt{T} \). Then, the distribution of allele frequencies becomes independent of mutation rate and long-term history. However, only two parameters can now be estimated, neighbourhood size and the local scale (Fig. 3.6b). Because the relation between \( K \) and \( \sigma \) depends in a complicated way on local population structure, there is a fundamental limitation to the use of allele frequencies to estimate gene flow.

Neigel, Ball, and Avise propose a straightforward method for estimating the long-term rate of gene flow from genealogies (Neigel et al. 1991; Neigel and Avise 1993). They suggest plotting the squared distance between genes against the time since they diverged, for all pairs in the sample, and using the slope of
Fig. 3.6 The accuracy of estimates made from allele frequencies. 1000 diploid individuals were taken from a two-dimensional square grid of sampling locations. Ten loci were scored, each with two neutral alleles that mutate at $\mu = 5 \times 10^{-5}$; hence, the scale is $L = c/\sqrt{2\mu} = 100c$. Neighbourhood size is $Nb = 4\pi a^2 = 80$, and we assume Malecot's (1948) model, corresponding to $K = a$. Hence $F_{st} = 0.054$. The matrix of covariances amongst allele frequencies sampled from these locations was calculated using eqn (3.13b). This was used to derive the expected information matrix, and hence the asymptotic variance of the maximum likelihood estimates. The heavy line shows the coefficient of variation (i.e. the standard deviation/mean) of $F_{st} = \log(L/K)/(Nb + \log(L/K))$, estimated without taking into account any spatial autocorrelations. This is plotted as a function of the number of sampling locations. (a) The coefficients of variation of the maximum likelihood estimates of the three parameters $Nb$ (thin line), $L$ (dotted line), and $K$ (dashed line), assuming that samples are taken from a square of size $(160c \times 160c)$. (b) Here, samples were taken from a square, $(20c \times 20c)$, much smaller than the scale $L$. Now, only $Nb$ (thin line) and $K$ (dashed line) can be estimated. As in (a) the heavy line shows the coefficient of variation of estimates of $F_{st}$, made without using any spatial information.
this relationship to estimate $4\sigma^2$ (Fig. 3.7). (In the notation used in that work, the squared distance in two dimensions increases as $2\sigma^2$ per unit time; two genes that shared a common ancestor $t$ generations back are separated by $2t$ generations). Neigel et al. (1991) simulate a set of genes forward in time, assuming random dispersal over a square of size $(10,000\sigma)^2$. Overall population size is regulated at 1000 individuals, so that the model corresponds to that described by eqn (3.2). Statistics are based on samples taken from this set, and support a linear relation between the variance of dispersal distance and lineage age, at least for closely related genes. Neigel and Avise (1993) run a wider range of simulations, some of which include local population regulation.

![Fig. 3.7 Simulation results from a stepping-stone model compared with Neigel and Avise's (1993) results (straight line). The simulated results are based on all coalescence times (circles), times less than 100,000 generations (triangles), and times less than 10,000 generations (crosses). The distances were taken from a linear array of demes in a two-dimensional stepping-stone model of 100 by 40 demes with 20 diploid individuals per deme and a migration rate of 0.05 in both directions. Distances are in standard deviations, $\sigma$.](image)

This approach is supported by the analytic results given here, which are also based on the idea that pairs of genes diffuse apart at a rate proportional to $\sigma^2$. However, there are several problems, which suggest some possible improvements. First, because the simulations are run forward in time, Neigel et al. must necessarily follow a small number of individuals, of which a relatively large proportion are sampled. Moreover, the neighbourhood size is much smaller than is typical of natural populations; for example, Fig. 5 of Neigel and Avise (1993) shows results from a population of $N=10,000$ genes, dispersing over an area of $A=10^6\sigma^2$. This corresponds to a neighbourhood
size of \( Nb = 4\pi \rho \sigma^2 = 4\pi N_0 \sigma^2 / A = 0.13 \). Hence most lineages coalesce before they have had time to spread over the whole species range, and nearby genes are likely to be close relatives. Thus, it is not clear how far their results would extend to small samples taken from a large population with moderate to large neighbourhood size. Simulations of the coalescent process for \( Nb = 6.28 \) show that although \( |x - x'|^2 \) does increase approximately linearly with time, at least initially, the slope of the relation is much smaller than \( 4\sigma^2 \), and even nearby genes are separated by a long coalescence time (Fig. 3.7).

Secondly, Neigel et al. suggest regressing \( |x - x'|^2 \) against time. Now, the distribution of \( |x - x'|^2 \) is determined by the sampling geometry, and so \( |x - x'|^2 \) should be treated as the independent variable; the coefficient of regression of \( |x - x'|^2 \) against time will depend on the distribution of distances sampled, whereas the converse regression of time against \( |x - x'|^2 \) would not. Neigel et al. (1991, p. 425) avoid this difficulty by sampling uniformly over the species' range; however, this would not be easy in practice. Thirdly, because the distribution of coalescence times decreases with \( 1/t \) in two-dimensional populations, the expected coalescence time at equilibrium is infinite, for any \( |x - x'| \). (This can be shown by treating \( F \) in eqns (3.13) as the generating function for \( f(t) \); its differential with respect to \( (1 - \mu)^2 \) at \( \mu = 0 \) is infinite.) In any actual sample the mean must be finite; however, if the distribution has infinite moments, the results will be very variable. This paradoxical behaviour of the distribution of coalescence times only applies to an infinitely large and indefinitely old population. If the population in fact colonized its present range \( T \) generations back, it would be reasonable to use only pairs that share a common ancestor before that time. However, the mean coalescence time would depend strongly on \( T \) (cf. Fig. 3.7), and would be very variable.

One might try to improve accuracy using the m.l.e., derived from the actual distribution of coalescence times, rather than the mean. If neighbourhood size is large, the distribution can be approximated by \( f(t) = \exp(-|x - x'|^2/4\sigma^2 t)/(2Nbt) \). Then, the maximum likelihood estimate of \( \sigma^2 \) (counting only pairs up to time 7) is given by the solution to

\[
\frac{|x - x'|^2}{t} = \frac{\sum_{i=1}^{T} \exp(-|x - x'|^2/4\sigma^2 t)/t}{\sum_{i=1}^{T} \exp(-|x - x'|^2/4\sigma^2 t)/t} \tag{3.16}
\]

This equation gives the m.l.e. for \( \sigma^2 \), given a single pair of genes, separated by \( |x - x'| \) and related at time \( t \). An estimate for many pairs of genes could be found by summing both sides of eqn (3.16). However, this will not be an m.l.e. because the pairs are not independent. We have run simulations of this estimator, and find that it performs poorly. The reason seems to be that the approximation \( f(t) = \exp(-|x - x'|^2/4\sigma^2 t)/(2Nbt) \) breaks down for times beyond \( t \approx Nb \). Using the full distribution/eqns (3.9) would not be feasible.
In a single panmictic population, pairwise estimates are inefficient because they include a disproportionately large contribution from a single event—the coalescence of two clades into the one common ancestor (Felsenstein 1992). This problem is less severe in samples from two dimensions, because the present rate of gene flow can be estimated only from closely related pairs of genes; the locations of more distantly related genes depend on the ancient history of the species. Unless neighbourhood size is unusually low, only a small proportion (≈ 1/Nb) of pairs of genes will be closely related. Hence, most information will come from independent pairs of genes rather than related clusters of genes. This raises a difficulty, however, in that most of the genealogy does not tell us about the rate of diffusion, but rather about the more distant history of the species. Genealogical data are therefore a fundamentally inefficient source of information about gene flow. The ideal solution to the problem would be to make an m.l.e. based on the structure of the whole tree, rather than on pairs of genes. Since only the recent part of the tree, which evolved after the species occupied its present range, can be used to infer $\sigma^2$, this method would only need to be applied to small clusters of genes. Nevertheless, it presents a daunting computational problem. In the next section we discuss ways of analysing the bulk of the tree, which tells us about the long-term history of the species.

3.3 Extinction and recolonization

Over long time-scales, populations cannot be adequately described by the uniform diffusion of genes from place to place. We know that (at least in temperate regions) most species have suffered drastic range changes in the last few thousand generations as a result of climatic change. If genes diffused at the current dispersal rate, they would not have had time to spread to fill the present range of the species. If lineages coalesced at the slow rate implied by current population sizes, then genealogies would be much deeper than is seen, and neutral heterozygosity would be much higher. The above analysis shows that unless neighbourhood size is very small, most of the information in a tree comes from distant relationships: only a small fraction $\approx 1/Nb$ of gene pairs are close relatives. The prime need in genealogical analysis is thus to find statistical methods for using this information to infer the distant history of the population, and the processes responsible for that history. In particular, we must explain how genes spread faster, and become more closely related, than is possible by diffusion through a large and stable population.

If the whole range were rapidly colonized from a randomly mating source at time $T$, there would be no association between genealogy and geography before that time. At the other extreme, spatial relations might be preserved despite expansions and contractions of the range. This is plausible if populations are adapted to a climatic gradient, and shift with that gradient (Coope 1979; Atkinson et al. 1987). Between these extreme possibilities, there might be expansion from a number of refugia, so that before time $T$ genealogies would
only reflect the source of the ancestral population and not more detailed spatial relations. We can imagine fitting data to a variety of such particular historical scenarios, and indeed, this is the usual approach to ‘phylogeography’ (Avise 1991). However, unless these scenarios can be constrained or corroborated by independent evidence, there is a danger of being able to explain too much. It is therefore attractive to seek ways of representing drastic changes as a statistical process—for example, by supposing that there is a low rate of expansions, in which the population in some area $A_1$ is replaced by individuals drawn from a smaller area $A_2$. The way allele frequencies are affected by random extinctions and recolonizations has received considerable attention, though mainly for the simplest case of the island model (for example Slatkin 1977; Wade and McCauley 1988; Whitlock and McCauley 1990; McCauley 1991). However, this theory has not led to ways of distinguishing random drift from random extinction (Slatkin 1987); the question is whether genealogical data may be more informative.

Changes in the species' distribution will cause older lineages to spread over larger areas than expected with diffusion alone, and cause lineages to coalesce faster than expected from the current population density. One could make an ad hoc estimate of some effective diffusion rate, $a^2$, and effective neighbourhood size, $N_{b_e}$, as a function of time, by adapting the methods discussed above. Naively, an increase in $a^2$ would lead to an increase in the scale over which allele frequency fluctuations are correlated, or equivalently, an increase in the rate of dispersion of lineages with age. Genetic distances do indeed often increase over scales much larger than can be explained by simple isolation by distance, suggesting sporadic range expansions. For example, in the alpine grasshopper *Podisma pedestris*, allele frequencies are correlated over all scales from 50 m to 3 km—a much flatter relationship than is consistent with isolation by distance with a $a$ rate of $20 \text{ m year}^{-1/2}$ and $T \approx 8000$ years (Fig. 3.8).

However, diffusion rates do not actually increase back into the past: what is needed is a model of the extinction/recolonization process itself. The crucial feature of this process is that it involves concerted movements, such that all the genes within an area tend to move together. This correlation across genes in turn generates correlations between the relationships of genes at different loci. How does this affect genetic variation? First, consider allele frequencies. Isolation by distance alone (i.e. diffusive gene flow and sampling drift), causes fluctuations which are independent across loci; by contrast, range expansions tend to produce correlations between the patterns at different loci. This idea has been applied with particular success to the reconstruction of the history of human populations in Europe. Consistent patterns across loci have shown that linguistic boundaries coincide with genetic boundaries (Sokal et al. 1990), and that genes from the originators of agriculture spread into the native European population with a degree of intermingling ('demic diffusion'; Ammermann and Cavalli-Sforza 1981; Sokal et al. 1991).

What degree of concordance is to be expected between genealogies at different loci? Even if dispersal is solely by the independent diffusion of genes, and even if
Fig. 3.8 Correlations in allele frequency versus geographical distance in *Podisma petraea*, estimated from 1815 individuals from 273 sites, scored for 5 enzyme loci with 18 alleles (Est-1, Est-2, Amy-1, Amy-2, ME—data from Halliday et al. (1983, 1984)). \( F_{st} \) was calculated from the covariance of square-root transformed allele frequencies between pairs of sites in different distance classes, corrected for sampling error. Significant correlations are found, over all scales up to \( \approx 3 \) km.

Linkage is loose, genealogies at different loci are expected to be to some degree parallel, because they will mirror the geographical location of the samples. In contrast, allele frequencies should be independent across loci if linkage disequilibria are negligible. In order to interpret genealogical data from multiple loci, we therefore need to understand the outcomes expected under the null hypothesis of isolation by distance. Slatkin (1989) and Slatkin and Maddison (1990) use the proportion of concordant genealogies to set a bound on the number of migrants between two demes; here we illustrate the same idea for a population in two dimensions. It is simplest to find the chance that a genealogy will be concordant with the geographic location of the samples, since this determines the concordance of genealogies derived from independent genes with each other. Figure 3.9a shows the three types of relationship between two pairs of genes, drawn from two locations. The genealogy may perfectly match the geography (Type 1); one pair may match (Type 2); or no pair may match (Type 3). If the sample locations are very close, or if the neighbourhood size is very high, there will be no relation between geography and genealogy, and the three types of tree will be in the proportions 1:2:6. Figure 3.9b shows how the degree of concordance increases as the genes move further apart, whilst Fig 3.9c shows how concordance falls with increasing gene flow. When gene flow is low, and the samples are far apart, the concordance between geography and genealogy reaches a plateau, which depends on neighbourhood size. This is because in two dimensions only a proportion \( \approx 1/Nb \) of lineage pairs coalesce early enough to preserve spatial information. Thus, under isolation by distance, only a fraction of the more closely related genealogies will be concordant with geography. Spatial patterns, and concordance across loci, among more distantly related genes therefore indicate large-scale changes in population structure.
(a) N. H. Barton and I. Wilson

Type 1

A

A

B

Probability 1/9

Type 2

A

A

B

B

Probability 2/9

Type 3

A

B

A

B

Probability 2/3

(b)

0 0.05 0.1 0.15

migration rate

0 1

1

tiugratlorl rate
Fig. 3.9  (a) The three possible concordances for two pairs of genes; with no population subdivision, the null probabilities of the trees are (1/9, 2/9, 2/3). (b) The frequencies of the three different types of concordance possible for two sets of two genes. Simulations are from a stepping-stone model with 20 diploid individuals per deme. Pairs of genes are sampled from demes a distance two demes apart in both directions. Migration rates given are in both directions. (c) Simulation results for pairs of genes sampled from a stepping-stone model with 10 diploid individuals per site, and a migration rate of 0.06 per generation in both x and y directions. The distance is given in standard deviations of distance moved, σ. For distance 0 the results are exact.

The power of analyses of allele frequencies comes from having data from many samples and many loci. The same may be true for genealogies. As noted above, strong concordance with geography over large scales, and concordance between loci, indicate the degree of large-scale population movement as opposed to independent diffusion. However, discordance may arise for a variety of reasons, reducing the power of this approach. Contraction of the range into refugia will only leave a genetic trace if the populations are small enough, for long enough, for there to be appreciable coalescence. Otherwise, the only effect will be a randomization of ancestral locations. There are inevitably errors in estimating the tree: it is disturbing that even when using mitochondrial DNA to estimate the relationships among ten major groups of vertebrates, at least 8000 contiguous bases of sequence are needed to give a 95 per cent chance of inferring the correct tree (Cummings et al. 1992). Discordance may also arise through selection on particular loci. This may be a particular problem in using mitochondrial DNA for within-species analyses, since the genealogy can be distorted by selection on any of the genes it carries (Thorpe et al. 1996), or on other elements that are inherited maternally, such as Wolbachia (Turelli et al. 1992). Drosophila mitochondrial sequences show significant deviations from neutral expectations (Ballard and Kreitman 1994; Rand et al. 1994), and the frequent introgression of mitochondrial genomes across boundaries demarcated by nuclear alleles may also be a sign of selection on the mitochondrial genome (Harrison 1989). It remains to be seen whether the more detailed information that is contained in genealogies will compensate
for the much smaller sample sizes, and whether the concordance between the
genealogies for a few loci could allow similar inferences to those based on
allele frequencies at many more loci.

### 3.7 Hybrid zones and barriers to gene flow

Species are often divided into a mosaic of distinct geographic races, which are
separated by narrow zones of hybridization. Such parapatric distributions may
be first identified through particular character—for example, different warning
colours in *Heliconius* butterflies (Mallet 1993), different chromosome arrange-
ments, as in the grasshopper *Podisma pedestris* (Barton and Hewitt 1981), or
different plumage, such as define many avian subspecies (Remington 1968; Hall
and Moreau 1970). However, these races are usually found to differ in many
other respects; the average Nei's genetic distance between 34 examples of hybrid
zones is $D = 0.26$ (Barton and Hewitt 1989). In most cases, the present
distribution is due to secondary contact between distinct populations, and
the extensive genetic divergence reflects the long time over which the source
populations have been diverging (Barton and Hewitt 1985; Hewitt 1993a,b).

Thus, parapatric distributions are the most striking examples of concordant
geographical patterns across many loci. Until recently, they have been studied
primarily through analysis of Mendelian markers and quantitative traits. Here,
we discuss whether genealogical information could, in principle, tell us more
about the origins of hybrid zones, and about their effect on gene flow.

The boundary between hybridizing populations is often much narrower than
would be expected if genes had simply been diffusing freely since the time of
contact. For example, the two chromosome races in *Podisma pedestris* are
thought to have met after the last glaciation, $T \approx 8000$ years ago. Random
movement at the present annual rate of $\sigma = 20$ m/generation$^{-1/2}$ would give a
cline $\sqrt{2/\pi\sigma^2 T} \approx 4.5$ km wide; yet, in most places the chromosomal cline is only
$\approx 800$ m wide (Barton and Hewitt 1981). Natural selection must act to balance
diffusion and maintain the sharp boundary between the hybridizing popula-
tions. This selection acts to produce a barrier to gene exchange that impedes the
flow of genes that are not themselves under selection: in order to pass from one
population to the other, they must recombine onto the new genetic background
before being eliminated by selection against the genes with which they are
associated. Such barriers to gene flow have the same effect as physical obstacles
to dispersal; both can be described by a parameter $B$, which has the dimensions
of a distance. A barrier produces a sharp step in allele frequency, proportional
to the gradient in allele frequency ($\Delta p = B(dp/dx)$; Fig. 3.10a), and reduces the
correlation between allele frequencies on either side (Fig. 3.10b). Both effects
can be used to give an estimate of barrier strength. For example, the two
chromosomal races of *Podisma pedestris* meet abruptly at Lac Autier, and are
separated by a stream a few metres wide. The pattern of chromosome
frequencies, and of alleles at six enzyme loci, combine to give an estimate of
Genealogies and geography

$B \approx 1.5$ km (Fig. 3.10). This is partly due to reduced dispersal across the stream and partly to a genetic barrier caused by selection against hybrids (Jackson 1992; Barton and Gale 1993).

![Graph showing the frequency of the chromosomal fusion across a transect at Lac Autier in the Alpes Maritimes.](image)

![Graph showing the covariance between fluctuations in allele frequency at six polymorphic enzyme loci.](image)

Fig. 3.10 (a) The frequency of the chromosomal fusion which distinguishes two races of the grasshopper *Podisma pedestris*, across a transect at Lac Autier in the Alpes Maritimes, plotted on a logistic scale. (Data courtesy of D. Currie; see Barton and Gale (1993)). In most places the cline is smooth, and would give a single straight line on this scale. Here, however, the frequency changes sharply across a stream. This could simply be due to a physical barrier to dispersal, disrupting a cline maintained by selection against chromosomal heterozygotes (straight lines). However, this requires a physical barrier an order of magnitude stronger than is measured directly from grasshopper movements ($B = 1390$ m for flow to the left, 2050 m to the right, versus 150 m measured directly). A model in which the physical barrier is augmented by selection against introgressing alleles at 100 loci predicts a further change just near the stream (curved lines; $S/R = 0.26$ to the left, 0.61 to the right, giving net barriers of 1670 m on the left, 3870 m on the right). (b) The covariance between fluctuations in allele frequency at six polymorphic enzyme loci, plotted against distance. The full circles are for sites on the same side of the stream, and the open circles are for sites on opposite sides. The net barrier estimated from the difference between the two lines is 1.5 km (support limits 500 m – 600), consistent with the estimate from the cline in the chromosomal fusion (Fig. 3.10a).
The effect of a local barrier on genealogical relationships can be derived by treating the probability of identity by descent as a generating function (eqn 13.1). Nagylaki (1988) and Nagylaki and Barchilon (1988) have analysed the effect of a barrier on identity by descent in one dimension. Extending these results to two dimensions shows that fluctuations in allele frequency are more localized, and the effect of a local barrier is consequently greater (Fig. 3.11a). Successive differentiation of the expression for identity by descent gives the distribution of coalescence times. Figure 3.11b shows the difference in the distributions for genes immediately adjacent to the barrier, either on the same side or on different sides, for the limit of large neighbourhood size. Although there can be a substantial effect on coalescence times, this is only apparent for the first few generations. Since in practice genealogies can only be resolved over long time-scales, they are unlikely to be much help in detecting even strong barriers. By contrast, fluctuations in allele frequencies can reveal the immediate effects of quite weak barriers (Fig. 3.10b).

Over the long time-spans which are reflected in most genealogies, the barriers to gene flow generated by selection in hybrid zones are unlikely to have much evolutionary consequence. For example, the pattern of allozyme genotypes gives an estimate of a barrier to gene flow from the toad Bombina bombina into B. variegata of $\approx 50$ km, relative to a dispersal rate of $\approx 1$ km/generation$^{-1/2}$ (Szymura and Barton 1991). This would restrict the diffusion of neutral alleles for $(B/a)^2 \approx 2500$ years since the toads met in their current distribution. However, the broad distribution of gene frequencies across Europe is dominated by successive expansion and contraction of populations over the $3 \times 10^6$ years for which these two taxa have been diverging. Thus, genealogical information would tell us about this long-term history rather than about current patterns of gene flow.

### 3.8 Conclusions

The main purpose of this chapter has been to emphasize the close relation between genealogical descriptions of spatially structured populations and the classical theory of identity by descent. A naive model of an unregulated population leads to an explicit formula for the joint distribution of locations and relationships, but also leads to unreasonable clumping. For populations subject to strict density regulation, we develop a diffusion approximation for the relation between genes. Both approaches give approximately the same distribution of coalescence times, $f_t = \exp\left(-|x-x'|^2/(4\sigma^2 t)\right)/(2\pi \sigma^2 t)^{1/2}$, though only for short times ($t \approx Nb$; see eqns (3.11)). This mathematical complexity makes it hard to develop sound statistical estimators which make full use of genealogical information. For example, Neigel, Ball, and Avise's suggestion that the rate of dispersal of lineages with time gives the rate of gene flow fails for populations with large neighbourhood size.
Fig. 3.11  (a) The variance of allele frequency fluctuations ($F$) immediately to the left of a barrier of strength $B$ (upper curve), and the covariance of allele frequencies between genes immediately on either side (lower curve). Results are for the limit of large neighbourhood size, and so are given as the product $(Nb F)$. The mutation rate is $\mu = 5 \times 10^{-5}$. (b) The distribution of coalescence times up to $t = 8$ generations, calculated by differentiating $F$ with respect to $z = (1 - \mu)^2$ at $\mu = 0$. Full symbols are for points on the same side of the barrier, and open symbols are points separated by the barrier. Circles: no barrier; squares: $B = 5\sigma$; triangles: $B = 10\sigma$. 
These difficulties arise in part from the mathematical and computational complexities. However, there are fundamental problems in making inferences about evolutionary processes from genealogies and geography—even if the ideal data were available. First, in two dimensions only a small fraction of gene pairs are likely to be closely related, and so most of the information in the tree is about sporadic events in the distant past. This contrasts with the simpler case, where the population is divided across a few islands, all of which can be sampled (cf. Slatkin and Maddison 1990). Inferences may then be tested against geological history (for example Thorpe et al. 1996). Secondly, sporadic events are hard to fit to any quantitative model, and so we are left with the difficult task of judging the relative merits of a multitude of possible histories, rather than estimating any well-defined parameters. Thirdly, genealogies derived from one or a few loci can only inform us of the history of the whole population if they are all affected by population structure in the same way. If species really consist of competing geographic races, which hardly recombine, then genealogies may well be largely concordant. However, whether this is so is at present obscure. Here, we have sketched some possible solutions to the simplest case of isolation by distance. There is an urgent need for a better theoretical and empirical understanding of the distribution of genealogies across multiple loci, and of the effects of large-scale population restructuring.

Acknowledgements

This work was supported by BBSRC grant GR/H/09928, and by a MAFF studentship. We would like to thank A. W. F. Edwards and S. Otto for their helpful comments.

Appendix: Convergence to the Wright/Malecot model of identity in state

We define the Fourier transform of $f(x)$ as $\hat{f}(\xi) = (1/2\pi) \int f(x) \exp(ix \cdot \xi) dx$, where $x$ is a two-dimensional vector. Then, if the dispersal is the same everywhere ($g(Y,x) = g(Y-x)$) the Fourier transform of eqn. (3.12) is:

$$\hat{F}(\xi, \xi') = \left[1 - \mu^2 (2\pi g_1(\xi))^2 \right] \left[ F(\xi, \xi') + \delta(\xi + \xi') (1 - F(0)) \right].$$

Hence

$$\hat{F}(\xi, \xi') = \frac{(1 - \mu^2) \delta(\xi + \xi') (1 - F(0))}{2\rho \left[1/(2\pi g_1(\xi))^2 - (1 - \mu)^2\right]}.$$  (3.A2)

The Dirac delta function $\delta(\xi + \xi')$ arises because $F(x, x')$ depends only on the
displacement \((\mathbf{x} - \mathbf{x'})\). Since \(g_1\) is a probability density, and is assumed to have mean zero, \(2\pi g_1(\mathbf{x})\) tends to \((1 - \mathbf{e}^2/2)\) as \(\sigma|x|\) tends to zero. For small \(\mu\)

\[
F(\mathbf{x}, \mathbf{x'}) = \frac{2\pi \delta(\mathbf{x} + \mathbf{x}') (1 - F(0))}{Nb(|\mathbf{x}|^2 + 2\mu/\sigma^2)}.
\]

(3.4a)

High spatial frequencies (\(\sigma|\mathbf{x}| \approx 1\)) make a negligible contribution to the inverse Fourier transform for \(|\mathbf{x} - \mathbf{x'}| \gg \sigma\). Hence

\[
F(\mathbf{x}, \mathbf{x'}) \approx \frac{1 - F(0)}{Nb} K_0 \left( \frac{\sqrt{2\mu} |\mathbf{x} - \mathbf{x'}|}{\sigma} \right).
\]

(3.4b)

To find the probability of identity in state between nearby genes \((F(0))\), we must integrate the full expression (eqn. (3.4a)):

\[
F(0) = \frac{1}{1 + \left[Nb/\log(\sigma/K_0/2\mu)\right]}
\]

where

\[
\log(K/\sigma) = \frac{\sigma^2}{2\pi} \int \left[ \frac{1}{\exp(|\mathbf{x}|^2/\sigma^2) - 1} - \frac{1}{[1/2\pi g_1(\mathbf{x})]^2 - 1} \right] d\mathbf{x}
\]

(3.5)

\(K\) depends only on the local dispersal distribution; if it is Gaussian \(K = \sigma\).

References


Appendix 8: GENEALOGIES AND GEOGRAPHY


Jackson, K. S. (1992). The population dynamics of a hybrid zone in the alpine
Appendix 8: GENEALOGIES AND GEOGRAPHY

Genealogies and geography


Appendix 8: GENEALOGIES AND GEOGRAPHY 294

N. H. Barton and I. Wilson


Appendix 9: Genealogies and Geography (condensed version)


Genealogies and geography

N. H. BARTON AND I. WILSON

Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, U.K.

SUMMARY

Any sample of genes traces back to a single common ancestor. Each gene also has other properties: its sequence, its geographic location and the phenotype and fitness of the organism that carries it. With sexual reproduction, different genes have different genealogies, which gives us much more information, but also greatly complicates population genetic analysis. We review the close relation between the distribution of genealogies and the classic theory of identity by descent in spatially structured populations, and develop a simple diffusion approximation to the distribution of coalescence times in a homogeneous two-dimensional habitat. This shows that when neighbourhood size is large (as in most populations) only a small fraction of pairs of genes are closely related, and only this fraction gives information about current rates of gene flow. The increase of spatial dispersion with lineage age is thus a poor estimator of gene flow. The bulk of the genealogy depends on the long-term history of the population; we discuss ways of inferring this history from the concordance between genealogies across loci.

INTRODUCTION

Until recently population genetics was based on allele frequencies or occasionally on genotype frequencies. In particular, the relative rates of gene flow and genetic drift have traditionally been estimated by using the standardized variance of allele frequencies across subpopulations (\( F_{ST} \), Wright et al. 1942). DNA sequencing now presents us with data that are more naturally represented by genealogies. Every sequence may be unique, in which case allele frequencies tell us nothing; all the information is contained in the genealogical relationship between sequences. If recombination is rare enough relative to mutation, the genealogy can be seen more or less directly, as for example in bacteria where recombination occurs via occasional transformation (Maynard Smith 1990) or in viruses where mutations are frequent (Sharp, this volume). Even if we cannot be sure of the relationships, the theory may still best be described in terms of the underlying genealogies. For example, if we can calculate the likelihood of parameters such as the population size or the rate of gene flow for a particular genealogy, the overall likelihood can be found as a weighted sum over the set of plausible trees (Felsenstein 1992). Even if our statistical inferences are based on allele frequencies, it may still give more insight to derive the theoretical predictions via a genealogical approach.

Here we shall assume that we have the full genealogy of each segment of the genome, including the times at which different lines of descent coalesce. How can we make inferences about the processes of evolution from such ideal data? How much more valuable is it to use information from the full set of genealogies, rather than from allele frequencies, or from pairwise relationships between genes? What is the best sampling scheme: is it better to have small genealogies across many loci, or genealogies for a few loci, but each containing many individuals? We concentrate on the specific question of how to analyse genealogies with geography and, in particular, how to estimate rates of gene flow and genetic drift. We first show the close relation between the classical theory of identity by descent and the genealogical structure, and then develop a simple diffusion approximation that describes how lines of descent coalesce. This approximation applies to a variety of local population structures, and so could be used to make robust inferences about effective population densities and rates of gene flow. However, we show that simple estimators based on the rate of dispersion of lineages over time can be misleading and that in a two-dimensional population the genealogical structure depends primarily on long-term history. Finally, we discuss possible ways of making inferences about this history. (A more extensive version of this paper is to be found in Barton & Wilson 1995), which includes more detailed derivations and a discussion of the effects of barriers to gene flow.)

This paper is to a large extent a synthesis of existing analyses: the classic work of Wright (1943) and Malecot (1948) on identity by descent; Slatkin’s (1991) work on the structure of genealogies in stepping-stone models and its application to measuring gene flow (Slatkin & Maddison 1989; Slatkin & Maddison 1990; Hudson et al. 1992); Felsenstein’s (1992) likelihood methods; and the use by Neigel, Ball and Avise (Ball et al. 1990; Neigel et al. 1991; Neigel & Avise 1993) of genealogies to infer population histories. The methods here aim primarily at estimating rates of gene flow and genetic drift rather than distinguishing qualitatively different population histories. Thus, they complement...
the methods of Crandall and Templeton, which make qualitative distinctions between alternative hypotheses (Templeton 1992; Crandall & Templeton 1993).

THE COALESCENT PROCESS

First, consider a single panmictic population containing $2N$ genes. If a small fraction of these genes are sampled, then their relationship can be approximated very simply: there is a probability $1/2N$ that any two lines of descent will coalesce in a common ancestor in each generation (Kingman 1982). Thus, if there are $k$ genes, there are $k(k-1)/2$ pairs that might coalesce, and the time back to the first coalescence is exponentially distributed with expectation $4N/k(k-1)$ generations. There are then $k-1$ lines of descent remaining, and so the expected time back to the previous coalescence is $4N/(k-1)(k-2)$ generations.

The expected age of the whole genealogy (that is, the expected time back to the common ancestor) is thus $4N(1/k(k-1) + 1/(k-1)(k-2) + 1/2) = 4N(1-1/k)$, which tends to $4N$ generations for a large sample (Felsenstein 1992).

This calculation shows that the structure of any genealogy is very variable. Looking back, the average time taken for a large number of lineages to coalesce to two ancestors is the same as the time these two remaining lineages take to merge (figure 1a). Since the latter time is exponentially distributed, and since any two randomly chosen genes have a chance of 1/3 of being related via the last common ancestor, the average divergence time between randomly chosen pairs of genes has a high variance, regardless of how large a sample is taken. Felsenstein (1992) uses this argument to show that pairwise statistics are much less efficient than estimators that include the genealogical structure. Essentially the same consideration applies with geographic structure, and so we summarize the argument here. The population size could be estimated from the average pairwise divergence time $\bar{t}$, by using the relation $\bar{N} = E(\bar{t})/2$. (In practice, $\bar{t}$ could itself be estimated from the average pairwise sequence divergence.) However, the variance of $\bar{N}$ tends to $2N^2/9$ for large samples, rather than to zero (Felsenstein 1992, equation 18). In contrast, the maximum likelihood estimator (MLE) is $\hat{N} = \sum_{j=2}^n (j-1) t_j/4(k-1)$, where $t_j$ is the time for which there are $j$ lines of descent. The variance of the MLE is $N^2/(k-1)$, which does tend to zero as sample size increases (Felsenstein 1992, equation 7). This basic argument suggests that genealogies will also give much more information about spatial structure than will pairwise measures such as Wright’s $F_\text{st}$ or the statistics proposed by Slatkin et al. (Slatkin & Hudson 1991; Hudson et al. 1992). However, it also suggests that information from any one genealogy may be misleading, so that reliable estimates may require data from many loci. Clearly, the probability that two genes coalesce at time $t$, $P(t)$, is just the probability that two genes coalesce at time $t$ and $t-1$ ($P(t) = F_t - F_{t-1}$, where $F_t$ denotes the probability of identity by descent via a population $t$ generations back). It is important to realize that the probability of identity by descent is purely a description of the genealogy and does not depend on the allelic state of the genes. The latter is described by the probability of identity in state, which is the chance that two randomly chosen genes share the same allele. This depends on how alleles are identified and on the mutational process but is not defined relative to any base population. The probabilities of identity by descent and identity in state are closely related and indeed are often confused with each other. If there are infinitely many alleles and a rate $\mu$ per generation of mutation to a novel allele, then the probability of identity in state is

$$F = \sum_{t=1}^{\infty} (1-\mu)^t (F_t - F_{t-1}) = \sum_{t=1}^{\infty} (1-\mu)^t f_t.$$  

(Throughout, we assume discrete generations.)

Equation (1) applies to any population structure and shows the close relation between identity in state ($F$) and identity by descent ($F_t$) and the distribution of coalescence times ($f_t$). Moreover, equation (1) shows that, if $F$ is considered as a function of $z = (1-\mu)^t$, then $F(z)$ is the generating function for the distribution of coalescence times. Then, $dF/dz\bigg|_{z=0} = tf_t$ and $d^2F/dz^2\bigg|_{z=1} = k(E\hat{f})$. The wealth of existing results on identity in state in spatially structured populations therefore leads directly to the distribution of coalescence times.

GENE FLOW AND THE COALESCENT PROCESS

The pattern of neutral genetic variation can be found by superimposing random changes in allelic state on the genealogy. This is possible because, by definition, neutral mutations do not affect reproduction and so are independent of the genealogy. It is tempting to treat the movement of genes in an analogous way: just as genes change their allelic state through random mutation, so their location changes as the organisms that carry them disperse. This idea was introduced by Cavalli Sforza & Edwards (1964), as the ‘Brownian/Yule process’; Edwards (1970) discusses the joint inference of ancestral locations and times. (The process was originally conceived as a model of the whole population, run forwards in time, in contrast to the present application.) Here we use this model to derive a simple expression for the joint distribution of geographic location and genealogical structure. However, we shall see that this expression cannot describe natural populations because it implicitly ignores local regulation of population density (Felsenstein 1975). Nevertheless, it leads us to a diffusion approximation that does apply to a variety of actual population structures.
Consider a large population of $2N$ genes, which are distributed over a two-dimensional habitat with area $A$. A sample of genes is taken at random from the population, without regard to their location. If we make the apparently innocent assumption that reproduction is independent of the position of the genes in the sample (or their ancestors), then their relationship should have the same distribution as in a single panmictic population, the intervals between coalescence times $l_j$ being exponentially distributed with expectation $4N/j(j-1)$. If we now assume that genes move in a gaussian random walk, with variance $\sigma^2$ per generation along each of the two axes, we can write the joint distribution of coalescence times and locations. For example, the probability that two randomly chosen genes are related through a common ancestor $t$ generations ago and are at positions $x$, $x'$ is

$$\psi(x,x',t) \, dx \, dx' = \frac{1}{2t \, Nb \, A^2} \left( 1 - \frac{1}{2N} \right) \exp \left( -\frac{(x-x')^2}{4\sigma^2t} \right),$$

where $Nb = 4\pi\rho\sigma^2$ is Wright's (1943) neighbourhood size and $\rho = N/A$ is the population density. Neighbourhood size plays a crucial role in determining the relative rates of gene flow and genetic drift in a two-dimensional population; roughly speaking, it is the number of individuals within one generation's dispersal range. The expression for a set of $k$ genes would involve a similar product of the exponential distribution of coalescence times, with the gaussian distribution of locations, given those times.

In fact, genes are sampled from particular locations. We therefore require the distribution of coalescence times conditional on location, $\psi(l|x,x') = \psi(x,x',t) / \psi(x,x')$, where $\psi(x,x')$ is the chance that genes will be found at $x$, $x'$. By summing over time,

$$\psi(x,x') \, dx \, dx' = \frac{\log \sqrt{2N} \, dx \, dx'}{Nb} \, \frac{1}{A^2}$$

for $|x-x'| = 0$, $2N \gg 1$, \hspace{1cm} (3a)

$$\psi(x,x') \, dx \, dx' = -\frac{K_0}{Nb} \left( \frac{|x-x'|}{\sqrt{2N} \sigma^2} \right) \, dx \, dx'$$

for $|x-x'| < \sigma$, \hspace{1cm} (3b)

$$\approx \log \left[ \frac{\sigma(2N)}{|x-x'|} \right] \, dx \, dx'$$

for $\sigma \ll \sqrt{\frac{2N}{|x-x'|}}$, \hspace{1cm} (3c)

where $K_0$ is the modified Bessel function.

This expression raises two difficulties. First, it implies extreme clumping: equation (3a) shows that the density near to a randomly chosen individual is increased by a factor of about $\log \sqrt{2N}$ above the average. This tendency for randomly dispersing and reproducing populations to become clumped was first emphasized by Felsenstein (1975) as a criticism of classical models of identity by descent. It can be seen as a consequence of the heterogeneous structure of random genealogies discussed above. On average, any large genealogy takes $2N$ generations to coalesce into two clades and then a further $2N$ generations for the remaining two lineages to meet. Hence, the population tends to evolve into distantly related and widely scattered clusters.

The second difficulty is that we have not properly accounted for the finite range of the species. At equilibrium, a random set of genes will have spread over an area of ca. $8N\sigma^2 = (2Nb/\pi)A$. (Recall that $\sigma^2$ is the variance of distance moved along each of the two axes.) Thus, unless neighbourhood size is very small, most sets of genes will be related by lineages that have crossed the species' range several times. This is true however large the range, because, for given neighbourhood size, divergence time increases in proportion.
The spatial distribution is found by summing over time, and applying the same approximations that led to equations (3),

\[
\psi(x, z) \, dx \, dz' = \frac{dx \, dz'}{N b \, A^2} \left( \log \sqrt{2N + \sum_{r=0}^{\infty} K_0 \left( \frac{\sqrt{2 \pi} \, 2r}{Nb} \right) } \right)
\]

for \(|x - z'| = 0, 2N \gg 1, \quad (5a)

\[
\psi(x, z) \, dx \, dz' \approx \frac{dx \, dz'}{N b \, A^2} \left( \log \sqrt{2N + \frac{\sigma \, 2N}{|x - z'|}} \right) + \sum_{r=0}^{\infty} K_0 \left( \frac{\sqrt{2 \pi} \, 2r}{Nb} \right) \quad \text{for } \sigma \sqrt{2N} \gg |x - z'| \gg \sigma. \quad (5b)
\]

For large neighbourhood size, the sum over Bessel functions can be approximated by an integral and tends to 1. Figure 2 shows examples of (a) the clumping produced by random reproduction, and b) the distribution of coalescence times, conditional on location. Since the degree of clumping is small for large neighbourhood size, the conditional distribution does not differ much from the raw distribution.

Sedentary organisms may not have had time to diffuse over the species' range over the time since they occupied it. For example, the flightless alpine grasshopper *Podisma pedestris* has a dispersal range of \(\sigma \approx 20\) m generation\(^{-1}\) (Barton & Hewitt 1982) and an annual life cycle. It has occupied its present range since the glaciers retreated \((T < 10^6\) generations\). In that time, genes would diffuse only ca. 2 km. In such cases \((\sigma^2 \, T \ll A)\) clustering is limited by finite age rather than finite range. In general, we should consider a sporadic series of extinctions and recolonizations over the whole history of the organism; for simplicity, we suppose that the range was occupied by \(X\) diploid individuals \(T\) generations ago, distributed uniformly without regard to their relationship. The joint distribution of locations and genealogies is then given by equation (2) (or its multigene extension) for \(t < T\), and by a uniform spatial distribution for more ancient relationships \((t > T)\). For two genes,

\[
\psi(x, z, t) \, dx \, dz' \approx g(t) \frac{dx \, dz'}{A^2} \left( 1 - \frac{1}{2N} \right) \exp \left( -\frac{|x - z'|^2}{4\sigma^2_t} \right) \quad 0 \leq t < T. \quad (6a)
\]

\[
\psi(x, z, t) \, dx \, dz' = g(t) \frac{dx \, dz'}{A^2} \sum_{t_i = T}^t g(t) = \left( 1 - \frac{1}{2N} \right)^T, \quad (6b)
\]

where \(g(t)\) is the arbitrary distribution of relationships among pairs of colonists. By summing over time to obtain the distribution of spatial locations,

\[
\psi(x, z') \, dx \, dz' = \frac{dx \, dz'}{A^2} \left( 1 - \frac{1}{2N} \right)^T + \frac{1}{2Nb} \sum_{t=i}^{T} \left( 1 - \frac{1}{2N} \right)^t \exp \left( -\frac{|x - z'|^2}{4\sigma^2_t} \right) \quad \text{for } \sigma \sqrt{2N} \gg |x - z'| \gg \sigma. \quad (7a)
\]

\[
\approx \frac{dx \, dz'}{A^2} \left( 1 + \frac{\log (T + \gamma)}{2Nb} \right) x = z', \quad T \ll 2N \quad (7b)
\]
A DIFFUSION APPROXIMATION TO THE COALESCENT PROCESS IN TWO DIMENSIONS

The model developed in the previous section is simple but unrealistic: in nature, fitness must decrease with local density. Densities may vary from place to place, but in response to local carrying capacity rather than to the unchecked accumulation of demographic fluctuations. The model of a locally unregulated population might describe the random evolution of morphology and would give an explanation for the clustering of asexual phenotypes into species (Higgs & Derrida 1991). However, it is implausible as a description of population dynamics in two spatial dimensions.

There have been extensive treatments of the extreme case where individuals are grouped into demes whose density is absolutely regulated. Analyses of such stepping-stone models have dealt primarily with identity in state (Wright 1943; Malecot 1948; Kimura & Weiss 1964; Maruyama 1972; Nagylaki 1974; Felsenstein 1976; Nagylaki 1986). However, since the identity in state is the generating function for the distribution of coalescence times (equation (1)), it leads immediately to the distribution of coalescence times (see below). Slatkin (1991) and Nei & Takahata (1993) have derived the mean coalescence time. However, the results would be cumbersome to extend to higher moments. Here, we develop a simple diffusion approximation that applies to all but local scales and that extends to whole genealogies.

The approximation is based on Wright’s (1943) argument that ancestors can be considered as being drawn from a neighbourhood whose size increases with time into the past. Thus, the probability that two nearby genes are identical by descent in the previous generation is (by definition) $1/2Nb$; in two dimensions, the pool of ancestors is spread over an area that increases linearly with time, and so the probability of identity by descent $t$ generations back is $1/2tNb$.

Let $f(t|z,x,x')$ be the probability that genes at $z$ and $x'$ are identical by descent via an ancestor $t$ generations back who lived at $z$. (With two genes, we shall not need to keep track of the position of ancestors. However, this would be necessary to extend the argument to more genes. See Barton & Wilson (1995).) Let $g_1(y,x)$ be the chance that a gene at $x$ derived from an ancestor at $y$ in the previous generation. We assume that $g_1$ depends only on the distance between parent and offspring, has zero mean and has variance $2\sigma^2$ along each of the two axes. Let $f(1|y,x,x') = h(y,x,x')/2Nb$ be the chance that two genes at $y$, $x'$ were identical by descent through an ancestor at $y$ in the previous generation. There is a chance $g_1(y,x)h_1(y,x,x')\delta y_2$ that two lines of descent come from some small area $\delta y$ in the previous generation; if they do, there is a chance $(1/2p\delta y)$ that they will be identical and a chance $(1-1/2p\delta y)$ that they are not but instead are identical via some more distant ancestor. Hence

\[
\begin{align*}
\frac{df(1|y,x,x')}{dy} &= \frac{1}{2Nb}h_1(y,x,x') = \frac{g_1(y,x)g_1(y,x')}{2p} & (9a) \\
\frac{df(t|z,x,x')}{dy} &= \int(f(t-1|z',y,y')g_1(y,x)h_1(y,x,x'))dy' \\
&= -\frac{1}{2Nb}\int(f(t-1|y,y')h_1(y,x,x'))dy. & (9b)
\end{align*}
\]

This argument requires several approximations. It is assumed that the lines descend independently, so that the joint probability of movement of two genes is $g_1(y,x)g_1(y,x')$. It is assumed that an area $\delta y$ can be chosen large enough that $1/2p\delta y < 1$ but small enough that $f$ is approximately constant within it. These assumptions hold for a demic structure with strict density regulation, in which case equations (9) give the coalescence times exactly. They are approximations to models of truly continuous populations: in the cases we consider, the approximation is remarkably good (see figure 3).

Equations (9) give a recursion across one generation that relates the probability of coalescence at time $t$ to that at time $t-1$. It can be rewritten as a recursion across many generations, which leads naturally to the diffusion approximation. Let $g_1(y,x)$ be the chance that a gene at $x$ descended from an ancestor at $y$, $t$ generations back. Extend the definition of $h_1$ to

\[
(1/2Nb)h_1(y,x,x') = g_1(y,x)g_1(y,x')/2p. & (10a)
\]

This is the chance that two genes both descend from an ancestor at $y$, $t$ generations back, the possibility of more recent coancestry being ignored. By applying equations (9) recursively,

\[
\begin{align*}
\frac{df(t|z,x,x')}{dy} &= \frac{1}{2Nb}\left(h_1(y,x,x')/t\right) \\
&= -\sum_{i=1}^{t-1}\int(f(t-i|z,y,y')h_1(y,x,x'))dy. & (10b)
\end{align*}
\]

This is the chance that the two genes descend from a common ancestor at $z$, after subtraction of the
probabilities that they were identical by descent in any of the intervening generations.

By the central limit theorem, \( g_i(y, x) \) tends to a gaussian with variance \( \sigma^2 t \) for large \( t \); \( h_i \) tends to a gaussian with variance \( \sigma^2 t^2/2 \), being the distribution of locations of the common ancestor. If we average over the location of the ancestors \( (f(t|x, x')) \) and use the fact that \( f(t|x, x') \) is independent of \( x \), equations (10) simplify to

\[
\begin{align*}
  f(t|x, x') &= \frac{1}{2N_b} \left( \sum_{i=1}^{L-x'} f(t-i|x, x') \right), \\
  f(t|x, x) &= \frac{1}{2N_b} (1 - (1 - \mu)^2), \\
  f(2|x, x) &= \frac{1}{2N_b} \left( \frac{1}{2} - \frac{1}{N_b} \right) \left( \frac{1}{2} + \frac{1}{N_b} \right), \\
  f(3|x, x) &= \frac{1}{2N_b} \left( \frac{1}{3} - \frac{1}{2N_b} + \frac{1}{N_b} \right). 
\end{align*}
\]

The distributions of coalescence times predicted by this gaussian approximation are shown in figure 3, together with simulation results. Agreement is close over all but short times and nearby genes. In general \( E_0 \) is low (< 0.10, say (Slatkin 1987)), implying that neighbourhood size is large. In this limit, the distribution of coalescence times in this model of absolute density regulation converges to that developed in the previous section for no local regulation.

These recursions for the distribution of coalescence times lead to parallel recursions for the identity in state. By applying equation (1) to equations (9),

\[
F(x, x') = (1 - \mu)^2 \int g_i(y, x) g_i(y', x') [F(y, y') + \frac{1}{2\rho} (1 - F(0)) \delta(y - y')] dy dy',
\]

where \( \delta(y) \) is the Dirac delta function, and the probability that nearby genes are identical has been rewritten as \( F(x, x') = F(0) \) to emphasize that it is independent of location. Equation (12) is identical to Malecot's (1948) model if (as Malecot assumed) the dispersal distribution is gaussian. It applies exactly to stepping-stone models if the integrals are replaced by sums and is an approximation to continuous models that neglects the interactions between nearby genes caused by local density dependence. In the appendix of Barton & Wilson (1995), it is shown that for arbitrary dispersal distributions equation (12) is approximated by

\[
F(x, x') = \frac{1}{N_b} K_w \left( \frac{|x - x'|}{\sigma} \right) \sqrt{1 - (1 - \mu)^2}
\]

for \( |x - x'| \gg \sigma, \) \( K \) and \( \mu \ll 1 \). (13b)

Equation (13) gives the probability of identity in state, based on the assumption of infinitely many alleles; the same expression gives \( D_2 \) measured from the variance in allele frequencies at loci with a finite number of alleles. \( K \) is a characteristic scale which depends on the local structure of the population. For a stepping-stone model with nearest-neighbour migrational boundaries, \( K = \text{(deme spacing)} / \sqrt{32} \). For Malecot's model of gaussian dispersal, \( K = \sigma \). If mutation rates are low \( (\mu \ll 1) \), \( 1 - (1 - \mu)^2 \approx 2\mu \), and equations (13) depend on the scale \( \ell = \sigma / \sqrt{2\mu} \). There will be significant fluctuations over local scales \( (\approx \sigma, K) \), but there will be correlations between allele frequencies over the much longer scale \( \ell \). Equations (13) are a close approximation to continuously distributed populations and break down only over local scales (figure 5 of Barton & Wilson 1995). Agreement is similarly close for stepping-stone models, even for neighbouring demes. The distribution of coalescence times can be calculated by differentiating equations (13) with respect to \( z = 1 - \mu^2 \). The breakdown of equation (13b) for small \( |x - x'| / \sigma \) corresponds to the breakdown of the gaussian approximation to equation (9) for small coalescence times (figure 3).

**ESTIMATING THE RATE OF GENE FLOW**

The recursions developed in the preceding sections allow calculation of the distribution of coalescence times among genes sampled from a two-dimensional population. However, it would not be easy to use this distribution to make statistical estimates. Here, we outline possible methods for estimating the rate of gene flow (\( \sigma \)), assuming that genes diffuse through a stable and homogeneous population. The traditional method for inferring population structure from genetic data was introduced by Wright (1942, 1943). Geographic variation in allele frequency is generated by genetic drift and reduced by gene flow. In the island model, the balance between these processes is determined by \( Nm \), the number of migrants exchanged between demes; the standardized variance of allele frequencies across demes is \( F_{st} = 1/(1 + 4Nm) \) for large deme size and low migration rates. In two dimensions (though not in one), the relationship is similar, with \( F_{st} \) decreasing with neighbourhood size \( Nb = 4\pi \rho \sigma^2 \); equations (13). Almost all analyses of population structure infer \( Nm \) or \( Nb \) from \( F_{st} \), or from some equivalent measure of variation in allele frequency, such as Slatkin's private allele method (Slatkin 1985: Barton & Slatkin 1986). However, the spatial pattern of allele frequencies also contains information. If as is usually the case) \( F_{st} \) is small, allele frequencies will fluctuate rather little, and the whole distribution can be approximated by a multivariate gaussian, which is defined by its mean and covariance. This covariance is given by equations (13) and depends on two scales: the scale that describes the local population structure and the scale \( \ell = \sigma / \sqrt{2\mu} \), which describes the balance between mutation and gene flow. Thus, if the mutation rate is known, the rate of gene flow (\( \sigma \)) can be estimated. This approach was first used by Sokal & Wartenberg (1983) and has been explored more recently by Epperson (1989, 1993). It is important to realize that it gives estimates of both the number of migrants or neighbourhood size \( (Nm \) or \( 4\pi \rho \sigma^2) \) and the proportion of migrants or rate of diffusion (\( m \) or \( \sigma \): confusingly, both are referred to as the 'rate of gene flow'. This approach is discussed in more depth in Barton & Wilson (1995).
in which it is shown that $\sigma$ could be estimated, but only if the population had been in equilibrium for about $1/\mu$ generations and the mutation rates were known. Even then large samples would be needed.

Neigel, Ball and Avise propose a straightforward method for estimating the long-term rate of gene flow from genealogies (Neigel et al. 1991; Neigel & Avise 1993). They suggest plotting the squared distance between genes against the time since they diverged, for all pairs in the sample, and use of the slope of this relation to estimate $4\sigma^2$ (figure 4). (In the notation used in this paper, the squared distance in two dimensions increases as $2\sigma^2$ per unit time; two genes that shared a common ancestor $t$ generations back are separated by $2t$ generations.) Neigel et al. (1991) simulate a set of genes forward in time, assuming random dispersal over a square of size $(10000 \sigma)^2$. Overall population size is regulated at 1000 individuals, so that the model corresponds to that described by equation (2) above. Statistics are based on samples taken from this set and support a linear relation between the variance of dispersal distance and lineage age, at least for closely related genes. Neigel & Avise (1993) run a wider range of simulations, some of which include local population regulation.

This approach is supported by the analytic results given here, which are also based on the idea that pairs of genes diffuse apart at a rate proportion to $\sigma^2$. However, there are several problems, which suggest some possible improvements. First, because the simulations are run forward in time, they must necessarily follow a small number of individuals, of which a relatively large proportion is sampled. Moreover, the neighbourhood size is much smaller than is typical of natural populations; for example, figure 5 of Neigel & Avise (1993) shows results from a population of $N = 100000$ genes dispersion over an area of $A = 10^6 \sigma^2$. This corresponds to a neighbourhood size of $Nb = 4\pi \sigma^2 = 100000 \sigma^2 / A = 0.13$. Hence most lineages coalesce before they have had time to spread over the whole species range, and nearby genes are likely to be close relatives. Thus it is not clear how far their results would extend to small samples taken from a large population with moderate to large neighbourhood size. Simulations of the coalescent process for $Nb = 6.28$ show that although $|x-x'|^2$ does increase approximately linearly with time, at least initially, the slope of the relation is much smaller than $4\sigma^2$ and even nearby genes are separated by a long coalescence time (figure 4).

Second, Neigel et al. suggest regression of $|x-x'|^2$ against time. Now, the distribution of $|x-x'|^2$ is determined by the sampling geometry, and so $|x-x'|^2$ should be treated as the independent variable; the coefficient of regression of $|x-x'|^2$ against time will depend on the distribution of distances sampled, whereas the converse regression of time against $|x-x'|^2$ would not. Neigel et al. (1991, p. 425) avoid this difficulty by sampling uniformly over the species’ range; however, this would not be easy in practice. Third, because the distribution of coalescence times decreases with $1/t$ in two-dimensional populations, the expected coalescence time at equilibrium is infinite, for any $|x-x'|$. (This can be shown by treating $F$ in equations (13) as the generating function for $f(t); its differential with respect to $(1-\mu)^2$ at $\mu = 0$ is infinite.) In any actual sample, the mean must be finite; however, if the distribution has infinite moments, the results will be very variable. This paradoxical behaviour of the distribution of coalescence times only applies to an infinitely large and indefinitely old population. If the population in fact colonized its present range $T$ generations back, it would be
reasonable to use only pairs that share a common ancestor before that time. However, the mean coalescence time would depend strongly on $T$ (cf. figure 4) and would be very variable.

In a single panmictic population, pairwise estimates are inefficient because they include a disproportionately large contribution from a single event, the coalescence of two clades into the one common ancestor (Felsenstein 1992). This problem is less severe in samples from two dimensions, because the present rate of gene flow can be estimated only from closely related pairs of genes; the locations of more distantly related genes depend on the ancient history of the species. Unless neighbourhood size is unusually low, only a small proportion (ca. $1/Nb$) of pairs of genes will be closely related. Hence, most information will come from independent pairs of genes, rather than related clusters of genes. (Slatkin & Maddison 1990 make the same point, by showing that branch lengths are so long in two dimensions as to be uninformative.) This raises a difficulty, however, in that most of the genealogy does not tell us about the rate of diffusion, but rather about the more distant history of the species. Genealogical data are therefore an inefficient source of information about gene flow. The ideal solution to the problem would be to make a maximum likelihood estimate based on the structure of the whole tree, rather than on pairs of genes. Since only the recent part of the tree, which evolved after the species occupied its present range, can be used to infer $\sigma^2$, this method would only need to be applied to small clusters of genes. Nevertheless, it presents a daunting computational problem. In the next section, we discuss ways of analysing the bulk of the tree, which tells us about the long-term history of the species.

EXTINCTION AND RECOLONIZATION

Over long timescales, populations cannot be adequately described by the uniform diffusion of genes from place to place. We know that (at least in temperate regions) most species have suffered drastic range changes in the last few thousand generations as a result of climatic change. If genes diffused at the current dispersal rate, they would not have had time to spread to fill the present range of the species. If lineages coalesced at the slow rate implied by current population sizes, then genealogies would be much deeper than is seen, and neutral heterozygosity would be much higher. The above analysis shows that, unless neighbourhood size is very small, most of the information in a tree comes from distant relationships: only a small fraction (ca. $1/Nb$) of gene pairs are close relatives. The prime need in genealogical analysis is thus to find statistical methods for using this information to infer the distant history of the population and the processes responsible for that history. In particular, we must explain how genes spread faster, and become more closely related, than is possible by diffusion through a large and stable population.

If the whole range were rapidly colonized from a randomly mating source at time $T$, there would be no association between genealogy and geography before that time. At the other extreme, spatial relations might be preserved despite expansions and contractions of the range. This is plausible if populations are adapted to a climatic gradient and shift with that gradient (Coope 1979; Atkinson et al. 1987). Between these extreme possibilities, there might be expansion from a number of refugia, so that before time $T$ genealogies would only reflect the source of the ancestral population and not more detailed spatial relations. We can imagine fitting data to a variety of such particular historical scenarios and indeed this is the usual approach to "phylo-geography" (Avise 1991). However, unless these scenarios can be constrained or corroborated by independent evidence, there is a danger of being able to explain too much. It is therefore attractive to seek ways of representing drastic changes as a statistical process, for example by supposing that there is a low rate of expansions, in which the population in some area $A_1$ is replaced by individuals drawn from a smaller area $A_2$. The way allele frequencies are affected by random extinctions and recolonizations has received considerable attention, though mainly for the simplest case of the island model (Slatkin 1977; Wade & McCauley 1988; Whitlock & McCauley 1990; McCauley 1991). However, this theory had not led to ways of distinguishing random drift from random extinction (Slatkin 1987; Slatkin & Maddison 1989, figure 9); the question is whether genealogical data may be more informative.

Changes in the species' distribution will cause older lineages to spread over larger areas than expected with diffusion alone and cause lineages to coalesce faster than expected from the current population density. One could make an ad hoc estimate of some effective diffusion rate, $\sigma^2$, and effective neighbourhood size $N_b$, as a function of time, by adapting the methods discussed above. Naively, an increase in $\sigma^2$ would lead to an increase in the scale over which allele frequency fluctuations are correlated, or equivalently, an increase in the rate of dispersion of lineages with age. Genetic distances do indeed often increase over scales much larger than can be explained by simple isolation by distance, suggesting sporadic range expansions. For example, in the alpine grasshopper *Podisma pedestris*, allele frequencies are correlated over all scales from 50 m to 3 km, a much flatter relation than is consistent with isolation by distance with a rate of expansion of 20 m year$^{-1}$ and $T \approx 8000$ years (figure 8 of Barton & Wilson 1995).

However, diffusion rates do not actually increase back into the past: what is needed is a model of the extinction/recolonization process itself. The crucial feature of this process is that it involves concerted movements, such that all the genes within an area tend to move together. This correlation across genes in turn generates correlations between the relations of genes at different loci. How does this affect genetic variation? First, consider allele frequencies. Isolation by distance alone (i.e. diffusive gene flow and sampling drift) causes fluctuations that are independent across loci; in contrast, range expansions tend to produce correlations between the patterns at different loci. This idea has been applied with particular success to the recon-
What degree of concordance is to be expected between genealogies at different loci? Even if dispersal is solely by the independent diffusion of genes, and even if linkage is loose, genealogies at different loci are expected to be to some degree parallel, because they will mirror the geographic location of the samples. In contrast, allele frequencies should be independent across loci if linkage disequilibria are negligible. To interpret genealogical data from multiple loci, we therefore need to understand the outcomes expected under the null hypothesis of isolation by distance. Slatkin & Maddison (1989) use the proportion of concordant genealogies to set a bound on the number of migrants between two demes, and give simulation results that allow estimates of neighbourhood size from two samples embedded in one- and two-dimensional stepping stone models (Slatkin & Maddison 1990).

Here, we give a simple illustration of this idea for a two-dimensional population. It is simplest to find the chance that a genealogy will be concordant with the geographic location of the samples, since this determines the concordance of genealogies derived from independent genes with each other. Figure 5a shows the three types of relation between two pairs of genes, drawn from two locations. The genealogy may perfectly match the geography (type 1), one pair may
match (type 2), or no pair may match (type 3). If the sample locations are very close, or if the neighbourhood size is very high, there will be no relation between geography and genealogy, and the three types of tree will be in the proportions 1:2:6. Figure 5c shows how the degree of concordance increases as the genes move further apart, while figure 5b shows how concordance falls with increasing flow. When gene flow is low and the samples are far apart the concordance between geography and genealogy reaches a plateau, which depends on neighbourhood size. This is because in two dimensions only a proportion of approximately $1/N_b$ of lineage pairs coalesce early enough to preserve spatial information. Thus, under isolation by distance, only a fraction of the more closely related genealogies will be concordant with geography. Spatial patterns and concordance across loci among more distantly related genes therefore indicate large-scale changes in population structure.

The power of analyses of allele frequencies comes from having data from many samples and many loci. The same may be true for genealogies. As noted above, strong concordance with geography over large scales and concordance between loci indicate the degree of large-scale population movement, as opposed to independent diffusion. However, discordance may arise for a variety of reasons, reducing the power of this approach. Contraction of the range into refuges will only leave a genetic trace if the populations are small enough, for long enough, for there to be appreciable coalescence. Otherwise, the only effect will be a randomization of ancestral locations. There are inevitably errors in estimating the tree: it is disturbing that, even when using mitochondrial DNA to estimate the relationships among ten major groups of vertebrates, at least 8000 contiguous bases of sequence are needed to give a 95% chance of inferring the correct tree (Cummings et al. 1995). Discordance may also arise through selection on particular loci. This may be a particular problem in using mitochondrial DNA for within-species analyses, since the genealogy can be distorted by selection on any of the genes it carries (Thorpe et al., this volume) or on other elements that are inherited maternally, as in Walbachia (Turelli et al. 1992). Drosophila mitochondrial sequences show significant deviations from neutral expectations (Ballard & Kreitman 1994; Rand et al. 1994), and the frequent introgression of mitochondrial genomes across boundaries demarcated by nuclear alleles may also be a sign of selection on the mitochondrial genome (Harrison 1989). It remains to be seen whether the more detailed information that is contained in genealogies will compensate for the much smaller sample sizes, and whether the concordance between the genealogies for a few loci could allow similar inferences to those based on allele frequencies at many more loci.

CONCLUSIONS

The main purpose of this paper is to emphasize the close relation between genealogical descriptions of spatially structured populations and the classical theory of identity by descent. A naive model of an unregulated population leads to an explicit formula for the joint distribution of locations and relationships, but also leads to unreasonable clumping. For populations subject to strict density regulation, we develop a diffusion approximation for the relation between genes. Both approaches give approximately the same distribution of coalescence times, $f_1 = \exp(-|x-y|^2/4\sigma^2 t)/2Nh_t$, though only for short times ($t \approx Nb$; see equations 11). This mathematical complexity makes it hard to develop sound statistical estimators that make full use of genealogical information. For example, the suggestion by Neigel, Ball & Avise (1991) that the rate of dispersion of lineages with time gives the rate of gene flow fails for populations with large neighbourhood size.

These difficulties arise in part from the mathematical and computational complexities. However, there would be fundamental problems in making inferences about evolutionary processes from genealogies and geography, even if the ideal data were available. First, in two dimensions only a small fraction of gene pairs are likely to be closely related, and so most of the information in the tree is about sporadic events in the distant past. This contrasts with the simpler case, where the population is divided across a few islands, all of which can be sampled (cf. Slatkin & Maddison 1989). Inferences may then be tested against geological history (see, for example, Thorpe et al., this volume). Second, sporadic events are hard to fit to any quantitative model, and so we are left with the difficult task of judging the relative merits of a multitude of possible histories, rather than estimating any well defined parameters. Third, genealogies derived from one or a few loci can only inform us of the history of the whole population if they are all affected by population structure in the same way. If species really consist of competing geographic races, which hardly recombine, then genealogies may well be largely concordant. However, whether this is so is at present obscure. Here, we have sketched some possible solutions to the simplest case of isolation by distance. There is an urgent need for a better theoretical and empirical understanding of the distribution of genealogies across multiple loci, and of the effects of large-scale population restructuring.

This work was supported by BBSRC grant GR/H/09928 and by a Scottish Office studentship. We thank A. W. F. Edwards and S. Otto for their helpful comments.

REFERENCES


Ballard, W. O. & Kreitman, M. 1994 Unraveling selection...


Wright, S. 1943 An analysis of local variability in flower color in \textit{Linanthus parryae}. \textit{Genetics, Princeton} 28, 139–156.
