

# Demographic models

*Gil McVean*

Department of Statistics, University of Oxford

February 24, 2003

# Contents

<b>1</b>	<b>Demographic modelling in population genetics</b>	<b>2</b>
1.1	Models and parameters . . . . .	2
1.2	Gene genealogies . . . . .	3
1.3	Patterns of polymorphism . . . . .	3
1.4	Estimation . . . . .	4
<b>2</b>	<b>Population growth</b>	<b>5</b>
2.1	A model of population growth . . . . .	5
2.2	Gene genealogies in growing populations . . . . .	5
2.3	Detecting growth . . . . .	7
2.4	Estimating growth parameters . . . . .	8
<b>3</b>	<b>Bottlenecks</b>	<b>9</b>
3.1	Bottleneck parameters . . . . .	9
3.2	Genealogies and patterns of polymorphism . . . . .	11
3.3	Estimating bottleneck parameters . . . . .	12
<b>4</b>	<b>Subdivided populations</b>	<b>13</b>
4.1	The inbreeding effect of population structure . . . . .	14
4.2	Detecting structure from haplotype data . . . . .	17
4.3	Genealogies in subdivided populations . . . . .	18
4.4	Estimation from subdivided populations . . . . .	20
<b>5</b>	<b>Split-time models</b>	<b>21</b>
<b>6</b>	<b>Admixture</b>	<b>24</b>
6.1	Detecting admixture . . . . .	24

# **1 Demographic modelling in population genetics**

Biological populations are nothing like the idealised Wright-Fisher (WF) model in population genetics. Real biological populations may fluctuate in size over time, they may receive immigrants from neighbouring populations, and there is usually a tendency for individuals from the same geographic region to mate with each other. Under some circumstances, for example when there are minor fluctuations in the population size, or some low level of inbreeding, we can still use the WF model to describe patterns of genetic variability, but with a different effective population size. More often, however, we need to incorporate the demographic complication. This lecture is about the effects such complications have on population genetic inference; how such deviations from the null model can be detected and how it may be possible to estimate some of the important quantities relating to the demographic models.

The field of demographic modelling and inference in population genetics is old, vast, and often confusing. In order to create a coherent picture, it is necessary to present a very personal view, in which some areas are grossly under-represented relative to their historical importance, and other areas have, what may seem to some, undue representation. Within this view, there are a few key themes that we will return to throughout the course of this lecture.

## **1.1 Models and parameters**

The first step in demographic inference is to choose the model within which you wish to work. It is clearly impossible to model the full biological complexity of population demography, so we must look for the simplest formulation that captures the relevant features. We need to decide what the key parameters within this model are, and what range of parameter values we might be interested in.

Within this lecture I will consider five caricatures of demographic models

- Population growth

- Population bottlenecks
- Subdivided populations
- Split-time models
- Admixture

Of course, any real population may well have experienced several of these demographic complexities. While it is straight forward to write down such hybrid models, and even relatively easy to simulate data taken from such populations, inference (parameter estimation, hypothesis testing, etc.) within such complex frameworks is a daunting challenge.

## **1.2 Gene genealogies**

Given a model, and a set of parameter values relating to the model, we then need to think about the effect of the demographic complexity on patterns of genetic variability. The approach I will take here is to think about the effect of demography from the genealogical point of view - the coalescent process. The rationale for taking this approach is that if we assume all mutations are neutral then the underlying genealogy contains all possible information about the demographic process of interest. Of course, we do not observe gene genealogies, we only observe mutations that have occurred on the genealogies. However, a genealogical viewpoint provides an intuitive means of predicting the patterns of polymorphism we expect to see under different demographic models.

## **1.3 Patterns of polymorphism**

Each demographic model has its own signature in the patterns of polymorphism we expect to observe. I will try to characterise this signature in two ways; first by presenting an intuitive explanation based around a single data set simulated under the demographic model, and second, by looking at the patterns one expects to see

in various summary statistics of the data. In particular, I will consider the effects each demographic model has on various estimators of the population mutation rate  $\theta = 4N_e m u$ . These are

- Watterson's estimate (Watterson, 1975):  $\hat{\theta}_W = S(\sum_{i=1}^{n-1} 1/i)^{-1}$
- Pairwise differences:  $\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{i,j,i \neq j} k_{ij}$
- The number of external mutations (Fu and Li, 1993):  $\hat{\theta}_e = \eta_e$
- Fu's (1996a) estimator:  $\hat{\theta}_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$

See the previous lecture for details of how to calculate each estimate. The key point is that the different estimators are sensitive to different aspects of the data. For example, pairwise differences are most influenced by intermediate-frequency variants, while the Fu and Li (1993) estimator is influenced by rare mutations, and Fu's (1996a) estimator is heavily influenced by high-frequency derived mutations. The differences between these estimators are used as test statistics to detect departures from the null, WF model in the tests of Tajima (1989), Fu and Li (1993) and Fay and Wu (2000). As a measure of haplotype structure, I will also consider the ratio of the number of haplotypes to the number of segregating sites.

## 1.4 Estimation

By looking at summary statistics, we can get an idea of what form of demographic model may apply to the data. The next step is to estimate the parameters of the model. As when talking about estimation in the WF null model, I will discuss different ways of estimating parameters, notably moment methods (where an analytical, or simulated expectation is equated to the observed value), coalescent-based likelihood methods (where likelihoods are estimated by stochastic simulation) and Bayesian methods.

## 2 Population growth

Population growth is a ubiquitous feature of biological populations. Whenever new populations are founded, these will typically increase in size over time. Notable examples are the human population over the last 10,000 years and many of their commensals, such as rats and disease-causing pathogens. On a shorter time-scale, epidemics such as that of *HIV-1*, are associated with rapid increases in population size. Even the dynamics of a single infection will be associated with rapid population expansion following infection.

### 2.1 A model of population growth

The simplest possible model has three parameters chosen from: the size of the population before expansion, the time of onset of population growth, the rate of population growth (and the form, e.g. exponential, linear, logistic, etc.), and the current population size. The reason that only three of these parameters is needed is that given three, the fourth is uniquely determined. It is also worth pointing out that a simpler model can be constructed with one fewer parameter, by assuming that the population was founded by a single individual at the beginning of the growth phase. It is also convenient to scale time in terms of the current (or ancestral) population size.

### 2.2 Gene genealogies in growing populations

Consider a pair of chromosomes sampled from the current population. Looking back in time, the rate of coalescence for any pair of lineages is proportional to the inverse of the population size at the time;  $1/2N_t$ . So as you go further back in time, the rate of coalescence increases. Suppose that the current effective population size is  $N_0$ , which grew from a single individual at rate  $l$  per generation (i.e. individuals have, on average,  $l$  offspring). Thinking about the process in terms of discrete generations, we can write the probability that the chromosome pair coalesces  $t$

generations ago as the probability that the chromosomes don't coalesce for the first  $t - 1$  generations times the probability that they coalesce in the  $t$ th generation

$$P(\text{coalesce at } t) = \left(1 - \frac{1}{2N_0}\right) \left(1 - \frac{1}{2N_1}\right) \dots \left(1 - \frac{1}{2N_{t-1}}\right) \frac{1}{2N_t}$$

This can be simplified, first by approximating the product of the probabilities of not coalescing as an exponential

$$\left(1 - \frac{1}{2N_0}\right) \left(1 - \frac{1}{2N_1}\right) \dots \left(1 - \frac{1}{2N_{t-1}}\right) \approx e^{-\sum_{i=0}^{t-1} \frac{1}{2N_i}}$$

second, by using the relationship that  $N_t = ((N_0/l)/l) \dots = N_0 l^{-t} \approx N_0 e^{(l-1)t}$ , valid if  $l$  is near to 1. Third by using the approximation

$$\sum_{i=0}^{t-1} \frac{1}{2N_i} \approx \frac{1}{2N_0} \sum_{i=0}^{t-1} e^{(l-1)t} \approx \frac{1}{2N_0} \int_0^t e^{(l-1)t} dt = \frac{e^{(l-1)t} - 1}{2N_0 l}$$

Finally, if we rescale time in terms of  $2N_0$  generations ( $\tau = t/2N_0$ ,  $\lambda = (l - 1) \times 2N_0$ ), the probability density function for the probability of coalescence at time  $\tau$  is

$$e^{-\frac{1}{\lambda}[e^{\lambda\tau} - 1]} \quad (1)$$

compared to  $e^{-\tau}$  in a constant population. In other words, the rate of coalescence increases exponentially. Similar considerations can be used to calculate the coalescent rates for samples of  $n$  sequences. It is worth pointing out that in order to make the above approximations, you have to assume that even as the population gets smaller it remains large. Clearly this is not true if the initial population was a single individual, however in terms of the accuracy of the approximation in finite populations, the error is negligible.

What effect does the acceleration in coalescence rate have on gene genealogies? Relative to the case of no growth with the same current population size, population growth will tend to reduce coalescence times. The other effect is that for a genealogy of  $n$  sequences, the shape of the coalescent tree is very much distorted from the constant population size expectation. In particular, whereas the further back in time you go, the longer the interval between successive coalescent

events becomes in the constant size model, in growing populations, the intervals become shorter. This means that the external branches (those leading to the sampled chromosomes) will tend to be relatively longer than the internal branches. In extreme cases of growth, essentially all coalescent events will occur within a very short space of time leading to star-like genealogies.

### **2.3 Detecting growth**

How will such distorted genealogies be reflected in patterns of polymorphism? Clearly, relative to the case of no growth, expanding populations will tend to have lower rates of polymorphism, because of the shorter genealogies. However, this prediction may not be very useful in practice, because it may not be possible to know how much polymorphism to expect. The only case where a lack of polymorphism can be used to indicate growth, is if the population of interest can be compared to one of similar current size thought to have remained constant in size.

Population growth does, however, also leave specific signatures in patterns of polymorphism. Of particular importance is the effect on the allele frequency spectrum. If you think about throwing mutations onto a genealogy from a growing population, most mutations will land on the long external branches. Such mutations will tend to be at low frequency, leading to relatively higher estimates of  $\theta$  from Watterson's (1975) and Fu and Li's (1993) estimators, and relatively lower estimates from pairwise differences and Fu's (1996a) estimator. Growth will also tend to create a very shattered haplotype structure, in which a large proportion of all mutations create a new haplotype. In terms of the test statistics used to detect departures from the null model, population growth leads to negative values of Tajima  $D$ , Fu and Li  $D$  and Fay and Wu  $H$  statistics, although the power to detect population growth differs between the various statistics (Fu, 1996b)

## 2.4 Estimating growth parameters

Suppose that we have detected the signal of population growth in our data, how can we estimate the relevant parameters? Just as with estimation in the null model, we can choose between moment methods, likelihood and a Bayesian framework. The most important moment method for estimating growth parameters is to use the distribution of pairwise differences between sequences (for DNA sequence data). Rogers and Harpending (1992) were the first to suggest that growing and constant populations may have very different distributions of pairwise differences (although their original conclusion, that growth leads to waves in pairwise diversity, is actually the opposite to the effect of growth).

The central idea is that in a growing population, most coalescent events are clustered over a short period of time, so that the time separating pairs of sequences (hence the expected number of mutations) is fairly constant (in the extreme star-like phylogenies all times are equal). Of course, there will be variation in the number of differences due to the variance of the Poisson mutation process, but the key point is that the distribution of differences will be unimodal with a variance approaching that of the mean (the Poisson expectation in star-like genealogies). In contrast, in constant size populations, the variance in pairwise differences will be considerably greater (equal to  $\theta + \theta^2$ ; see lecture 2), and the distribution is often bimodal because the longest branch in the genealogy is expected to be the last coalescent event, which splits the sampled chromosomes into two parts. Given the distribution of coalescence times for pairs of sequences obtained from (1), we can numerically calculate the expectation and variance of pairwise differences. By equating observed and expected values, we can estimate both the mutation rate  $\theta_0$  and the growth rate  $\lambda$  (Slatkin and Hudson, 1991). Confidence intervals for the point estimates can potentially be obtained by numerical resampling techniques, such as the bootstrap, although to account for variance in the evolutionary process, the bootstrap would have to be carried out across loci. Similar procedures for estimating growth rates can be devised for microsatellites, based on the variance

and kurtosis (fourth central moment) of the distribution of allele size (Goldstein et al., 1996)

While the moment method provides a quick way of estimating the growth rate, the variance of estimates is likely to be high, and the approach throws away a lot of the information in the data. For this reason, full coalescent methods, in either a likelihood, or Bayesian framework are potentially much more powerful. Various methods to estimate growth parameters from DNA sequence data (Kuhner et al., 1998), and microsatellites (Beaumont, 1999) have been developed that use importance sampling, or Monte Carlo Markov Chain methods to simulate genealogies conditional on the data. By carrying out large numbers of such simulations it is possible to estimate likelihoods for specific parameter values.

### **3 Bottlenecks**

The inverse of population growth is population decline, the most severe form of which is called a bottleneck. Bottlenecks occur when there is some major event during which only a small proportion of the previous population persists. The type of event causing a bottleneck may be a catastrophe (for example an ice-age, or volcanic eruption), or the bottleneck may occur in the formation of a new population (for example the out-of-Africa hypothesis for humans posits a bottleneck in the generation of non-African lineages). Either way, bottlenecks can have very considerable consequences for patterns of genetic variation

#### **3.1 Bottleneck parameters**

Bottleneck models are typically formulated as step functions in population size. Although there are likely to have been more gradual changes in reality, it is assumed that the time-scale of the bottleneck is short relative to the history of the population, in which case it is justifiable to ignore the subtleties of exactly how the bottleneck occurred. Such a model has several parameters; the ancestral popula-

tion size,  $N_0$ , the time at which the bottleneck commenced,  $T_b$ , the duration of the bottleneck,  $\tau$ , the relative decrease in population size during the bottleneck,  $f$ , and the population size after the bottleneck  $N_1$ . It is, however, possible to make several simplifying assumptions to this model. The most important simplification is that in terms of the genealogical effects of bottlenecks, the strength of the bottleneck is a function of the decrease in population size and the duration of the bottleneck. Looking back in time, consider a pair of chromosomes at the start of the bottleneck, the probability that they will coalesce during the bottleneck is

$$P(\text{coalesce in bottleneck}) = 1 - e^{-\tau/f}$$

Assuming that no mutations occur during the bottleneck (effectively the same as assuming the time-scale is short relative to the history of the population), the ratio  $\tau/f$  is the important quantity in determining the strength of a bottleneck. A long, mild bottleneck is the same as a brief, severe one. This dependency on the ratio only is also true for the case of  $n$  chromosomes. The probability of  $k$  coalescent events occurring during a bottleneck of severity  $v = \tau/f$  is given by

$$P(k|v, n) = \sum_{i=n-k}^n e^{[-i(i-1)v/2]} \frac{(-1)^{i-n+k} (2i-1)(n-k)_{(i-1)} n_{[i]}}{(n-k)!(i-n+k)!n_{(i)}} \quad (2)$$

Where

$$a_{(j)} = a(a+1) \dots (a+j-1), \quad j \geq 1; \quad a_{(0)} = 1,$$

$$a_{[j]} = a(a-1) \dots (a-j+1), \quad j \geq 1; \quad a_{[0]} = 1.$$

(Griffiths, 1980; Tavaré, 1984). It is also worth noting that when treating bottlenecks as instantaneous, it make more sense to work with the quantity  $(1-f)/f$  rather than  $f$  directly, to emphasize the necessity for  $f < 1$  to make an impact on patterns of variability. Other ways of reducing the parameters one needs to estimate are to assume that the current and ancestral population sizes are identical, or to assume a very recent bottleneck.

### 3.2 Genealogies and patterns of polymorphism

Looking back in time, the effect of a bottleneck is to generate a brief period of rapid coalescence. The assumptions we made about the time-scale of the process in effect mean that bottlenecks can be treated as an instantaneous loss of active lineages in the coalescent. Either side of the event, the standard coalescent process applies. If the bottleneck is very strong, only a very few lineages survive, but each of these ancestral lineages leaves many descendant in the sample. Consequently, mutations that arose on these ancestral lineages will be present on multiple chromosomes in the sample, hence be at intermediate frequency. In terms of the rate of polymorphism, bottlenecks will reduce the total tree length compared to the constant population size, but because most mutations in constant population trees occur when there are fewer active lineages (the expected number of mutations when there are  $k$  ancestral lineages is  $\theta/(k - 1)$ ), and it is the more recent coalescent events that bottlenecks influence, the effects of the rate of polymorphism are not nearly so severe as for population growth.

How do bottlenecks affect haplotype structure? Suppose that the bottleneck were very recent, in which case there has not been enough time for much recombination in the ancestral lineages. The rapid coalescence during the bottleneck will create sets of very closely related haplotypes that are defined by mutations in the ancestral lineages that survive the bottleneck. Even if there was enough recombination in the ancestral lineages to remove all linkage disequilibrium between alleles, a recent bottleneck will generate strong linkage disequilibrium. Perhaps the easiest way to think about this is to assume that the bottleneck was yesterday; if you have a sample of size  $n$  and there were  $a$  ancestral lineages, then the  $n - a$  remaining lineages have to be distributed among the  $a$  ancestral lineages in an urn-model fashion. Specifically this means throwing down the  $n - a$  lineages one at a time, where the probability that an ancestral lineage receives the new lineage is proportional to the number it already has. Such a sampling process naturally leads to large variance in the number of descendants between ancestral lineages, So even

if there were no linkage disequilibrium in the ancestral lineages, there would be in the sample.

In terms of the summary statistics and test-statistics, recent bottlenecks lead to an excess of intermediate-frequency alleles and a dearth of rare mutations, leading to positive Tajima  $D$  and Fu and Li  $D$  statistics. Haplotype structure is also greatly increased, such that each haplotype is defined by several segregating sites (therefore  $K/S$  is reduced). In other words, bottlenecks create almost exactly the opposite signature as population growth. This is perhaps not surprising given that the two processes are effectively the opposites of each other.

### **3.3 Estimating bottleneck parameters**

Because of the conceptual similarities between population growth and population decline, many of the techniques that can be used to estimate growth parameters can also be used to estimate bottleneck parameters. For example, the variance of pairwise differences (and the kurtosis of microsatellite length distributions) increases with bottlenecks, which can be used as the basis of moment estimators (Zhivotovsky et al., 2000). Full-likelihood and Bayesian coalescent-based methods can also be derived (Beaumont, 1999). The major limitation of coalescent approaches is that currently they must assume either no recombination, or free recombination. Partial linkage, such as occurs in most human genes cannot be incorporated due to the computational complexities of jointly estimating recombination rates and demographic parameters.

One possibility for estimating bottleneck parameters that has been suggested recently is to use the pattern of linkage disequilibrium in populations (Reich et al., 2001). As mentioned above, bottlenecks can substantially increase linkage disequilibrium. By comparing the relationship between physical distance between markers and summary statistics of linkage disequilibrium, it may be possible to both detect and estimate parameters, particularly if a reference population can be identified which is thought not to have experienced a bottleneck.

## 4 Subdivided populations

Although population growth and bottlenecks are important deviations from the null model, they both assume that within populations mating is random with respect to geographic location. In many situations this is not true; populations are separated by obstacles such as mountains, rivers and oceans and individuals tend to mate with others from a nearby locality.

Models of population subdivision attempt to account for the complexity of population structure. However, unlike the models reflecting changes in population size, models of population structure typically consider equilibrium situations, such that small populations of constant size exchange migrants with each other over an indefinite period of time.

Within the field of population structure there is a huge variety of models of varying complexity. At the simplest level, and the first model to be analysed from a theoretical population genetics perspective, is the island model of Wright (1931). In this model a finite island population receives immigrants from an infinitely large mainland population. While such a model may be appropriate for many real islands, there are many cases in which we may wish to consider multiple populations exchanging migrants with each other. Within this framework, the  $n$ -island model considers a collection of identical populations that each exchange migrants with all other islands at an identical rate. An important extension of this model, that can have considerable impact on patterns of polymorphism is the metapopulation model. In essence it is like the  $n$ -island model, except that populations occasionally go extinct and are re-founded. Small, fragmented populations, epitomised by the butterfly populations of Finland (Hanski and Ovaskainen, 2000), may be well described by metapopulations.

An important feature of biological reality that is missing from such models of population structure is the notion of isolation by distance. Although islands are isolated in the  $n$ -island model, all islands are equally separated from each other. While this makes the maths easier, it is biologically unrealistic, so there are a number of

models that incorporate the spatial element. The simplest among spatial models is the stepping-stone model (Kimura, 1953) in which an array of populations is connected in a chain by migrants between adjacent populations (often called demes). Variants of this model are the circular stepping-stone model (mainly to make the maths easier) and the two-dimensional lattice model. Ideally, one would like to operate within the framework of a continuous space, in which an arbitrary migration kernel described the distribution of possible migration patterns. Such models are difficult to formulate (Felsenstein, 1975), though recent advances have been made (Barton et al., 2002). It should be pointed out that population genetic inference of a sophisticated nature is only possible within the very simplest models.

#### **4.1 The inbreeding effect of population structure**

Rather than launch directly into explicit models of population structure, it is worth discussing some general features of population structure that apply to all models. The key feature of population structure is that it stratifies individuals into separate reproductive units. Each of these units may behave like a good WF model, but because the evolutionary processes in different populations are only coupled through the exchange of migrants, the stochastic nature of the evolutionary process will inevitably lead to genetic differentiation between populations.

The most important feature of genetic differentiation is that alleles may be at different frequencies in different populations. At the most extreme, each population may be fixed for a different allele. Viewed from a global perspective, differentiation between populations has a direct analogy with inbreeding. Consider a single locus with two alleles, and a number of populations, each with its own allele frequency  $p_i$ . Assuming Hardy-Weinberg equilibrium, the expected frequency of heterozygotes in the combined populations would be  $2\bar{p}(1 - \bar{p})$ , where  $\bar{p}$  is the average allele frequency across populations. But, because of the reproductive stratification, even if each population is in Hardy-Weinberg equilibrium, the observed

frequency of heterozygotes will be

$$f_{Het} = 2\bar{p}(1 - \bar{p}) \left[ 1 - \frac{\sigma_p^2}{\bar{p}(1 - \bar{p})} \right] \quad (3)$$

In other words, the frequency of heterozygotes is reduced by a factor proportional to the variance in allele frequency across populations.

Compare Equation (3) with the proportion of heterozygotes under inbreeding

$$f_{Het} = 2p(1 - p)[1 - F]$$

Where  $F$  is the inbreeding coefficient. Clearly, there is a direct analogy between the inbreeding coefficient and the scaled variance in allele frequency. In fact, there is more than an analogy, there is a specific relationship between inbreeding and population structure. Inbreeding increases identity-by-descent (*ibd*) within individuals (relative to the population), population structure increases *ibd* within populations (relative to the species). *Ibd* measures the degree of relatedness between two alleles because of common ancestry.

We can therefore think of the effects of population structure as generating *ibd*. Furthermore, we consider the processes in terms of a hierarchical model in which there is *ibd* within individuals, subpopulations, populations and species, each measured relative to higher levels in the hierarchy. This way of thinking about population structure was first suggested by Sewall Wright (Wright, 1969, 1978), who introduced measures of *ibd* called  $F$ -statistics. Of these, the most famous is the statistic  $F_{ST}$  which measures the identity within populations relative to the total species range.

Although  $F_{ST}$  in Wright's original formulation is a parameter, the probability of identity-by-descent, it is much more common to see  $F_{ST}$  used as a summary statistic for assessing population differentiation in empirical data. Specifically,  $F_{ST}$  is usually defined as the one minus the proportion of all genetic variation that is found within populations

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T} \quad (4)$$

Where the  $H_s$  represented the average heterozygosity across loci within populations ( $S$ ) and in the total sample ( $T$ ). If  $F_{ST}$  is close to zero, there is little or no genetic differentiation, if  $F_{ST}$  is close to one, differentiation is very high. For a given data set, it is possible to assess whether there is significant differentiation by carrying out a permutation of samples by location.

What does  $F_{ST}$  look like in natural populations? In humans,  $F_{ST}$  is low, roughly 7% amongst the major races irrespective of whether it is calculated using allozymes or SNP data. This means that 93% of all variability in humans is present in every population. Even human tribes thought to have been highly isolated, such as the Yanomama, show comparable levels of differentiation (Nei, 1975). Likewise, human commensals, such as the house mouse and *Drosophila melongaster*, show similar levels of differentiation. There are very few species that show high levels of differentiation.

Before leaving it F-statistics, it is worth pointing out that different researchers have very different uses and interpretations of the same statistics. As I have said, Wright's original formulation treated  $F_{ST}$  as an evolutionary parameter. Generally, however, we wish to think about explicit demographic models (the island-model,  $n$ -island model, etc.). For such models,  $F_{ST}$  in Wright's sense is a function of other model parameters (population sizes, migration rates), so is itself a parameter, and can, in some circumstances be related to other model parameters. For example, in Wright's island model  $F_{ST}$  relates to the migration rate as

$$F_{ST} = \frac{1}{1 + 4N_e m} \quad (5)$$

Where  $m$  is the proportion of the island population replaced by migrants from the mainland each generation. Under such models, it is sometimes possible to derive moment-based estimators of the parameter  $F_{ST}$  which, under certain limits, often take a form similar to that of the summary statistic (4) (Cockerham and Weir, 1987). The analysis of variation within this evolutionary framework is often referred to as Analysis of MOlecular VARIation, or AMOVA (Rousset, 2001; Excoffier, 2001).

However, the relationship between real demographic parameters and  $F_{ST}$  is usually not as simple as (5), and it may not be possible to derive moment estimators within the AMOVA framework for such models. Furthermore, the AMOVA framework makes a number of assumptions about the distribution of allele frequencies (typically, it assumes Normality) which are unlike to hold for rare alleles. For these reasons, I will consider  $F_{ST}$  only as a summary statistic in the form of (4), as a way of partitioning genetic variability into within population and among population effects (Nei, 1973).

## 4.2 Detecting structure from haplotype data

$F$ -statistics provide a way of detecting an summarizing population structure from allele frequency differences between populations. However, by ignoring associations between alleles, we may be throwing away considerable information. Hudson (2000) has suggested a way of detecting population structure from the distribution of haplotypes. The method can potentially be more powerful for detecting population structure than allele frequency base methods, although unlike  $F_{ST}$ , the test statistic,  $S_{nn}$  is not expected to bear any simple relationship to the parameters of models of population structure.

The test statistic is the average proportion of nearest-neighbour haplotypes that are present in the same population. To calculate the statistic, for each chromosome in the data set, identify the set of other chromosomes that have the least nucleotide differences (the nearest neighbours - often there may be other identical haplotypes). Among this set, calculate the proportion which are in the same population as the chromosome of interest. Repeat the process across chromosomes and take an average. For strong differentiation,  $S_{nn}$  is expected to be near one, while under weak differentiation the expected value is 0.5.

### 4.3 Genealogies in subdivided populations

In order to consider the effect of population structure on gene genealogies and patterns of polymorphism I will focus on the case of two identical populations of diploids, each of size  $N_e$  which exchange migrants at a constant rate  $m$  per chromosome per generation. Looking back in time we need to calculate the probability of a coalescent event, and the probability of a migration event. Because coalescent events can only occur within populations, the probability of a coalescent event is the sum of the rates in the two populations

$$P(\text{coalescent}) = \frac{n_1(n_1 - 1)}{2N_e} + \frac{n_2(n_2 - 1)}{2N_e}$$

Where  $n_i$  is the number of chromosomes sampled from population  $i$ . The probability of a migration event is

$$P(\text{migration}) = (n_1 + n_2)m$$

What we have in effect is two coalescent processes that are coupled to each other by migration events. When migration is very rare, the processes are essentially independent, so coalescence proceeds rapidly within each population, but there is a long wait until the MRCA for each population coalesce. When migration is high, the two populations essentially behave as a single unit. As with other situations, the key parameter in determining the effect of migration is the product of the effective population size and the migration rate  $M = 4N_e m$ .

We can get a feel for the effects of population subdivision on patterns of polymorphism by considering the times to coalescence for pairs of chromosomes sampled from within and between populations; we will also scale time in terms of  $2N_e$  generations. Consider a pair within a population, we have to wait an exponentially distributed length of time, with mean,  $1/(1 + M)$ , for an event, which is a coalescence with probability  $1/(1 + M)$ , and a migration with probability  $M/(1 + M)$ . So the expected time to coalescence within a population is

$$E[T_w] = \frac{1}{1 + M}[1 + MT_b]$$

For a pair of chromosomes in different populations, we have to wait an exponentially distributed length of time with mean  $1/M$  for a migration to bring them into the same population. The expected coalescence time for a pair of chromosomes between populations is therefore

$$E[T_b] = \frac{1}{M} + T_w$$

Solving the pair of simultaneous equations gives

$$\begin{aligned} E[T_w] &= 2 \\ E[T_b] &= 2 + 1/M \end{aligned} \tag{6}$$

In other words, the expected time to coalescence for chromosomes within a population, hence the expected number of pairwise differences, is identical to that if there was no population subdivision. In contrast, the expected number of pairwise differences for samples between populations increases for lower migration rates.

The remarkable invariance in the expected number of differences within a population can be generalised to more complex situations (Li, 1976; Slatkin, 1987). The result is, however, slightly misleading, because although the expectation may be unchanged, the distribution of pairwise differences is very different in subdivided populations. Specifically, most chromosomes within a population will tend to be fairly similar, but due to the immigration of chromosomes from the other population, some chromosomes will be very different.

Population subdivision also strong effects on the frequency distribution of segregating sites, such that at low migration rates, many sites segregating in the global sample are due to fixed differences between the populations. Such mutations will be at intermediate frequencies, leading to positive Tajima  $D$  and Fu and Li  $D$  statistics. Strong subdivision also leads to strong haplotype structure, again, because of the prevalence of fixed differences. Consequently, if there is no *a priori* reason to expect population subdivision (that can be tested by  $F_{ST}$  or  $S_{nn}$ ), a positive Tajima  $D$  statistic may be indicative of population structure.

The effects on the frequency distribution of segregating sites for samples within populations is quite different. Rather than lead to an excess of intermediate-frequency mutations, strong population structure leads to an excess of high-frequency derived mutations, because of the rapid coalescence of the majority of chromosomes within the population, followed by a longer period waiting for the coalescence of chromosomes that escaped the population by migration. Consequently, when looking at patterns of variability from within a population, a signal of structure may be a significant Fay and Wu  $H$  statistic (Fay and Wu, 2000).

#### 4.4 Estimation from subdivided populations

As mentioned previously, one approach to parameter estimation in subdivided populations is the AMOVA treatment of  $F_{ST}$ . However, because  $F_{ST}$  as an evolutionary parameter is really a function of more intuitive quantities such as migration rates and population sizes, it may be preferable to use an explicit demographic model for inference. While this area is not as well advanced as inference from constant, growing, or bottlenecked populations, there have been considerable advances in using full coalescent likelihoods for inference (Bahlo and Griffiths, 2000; Beerli and Felsenstein, 1999). The framework is very similar to the standard coalescent (although the algorithmic complexity is much greater), the key difference is that both the time and location of mutations must be estimated. A natural way to represent the structured coalescent is using coloured coalescent graphs in which lineages and mutations are coloured according to the population in which they are found.

Under certain circumstances, it is also possible to use results first derived by Wright on the frequency distribution of alleles under the island-model to estimate parameters of migration. In the island model, it is assumed that there is an infinitely large population with fixed allele frequencies, from which migrants arrive each generation at rate  $m$ . Suppose that the allele frequency on the island is currently  $x$  and on the mainland it is  $x_m$ . The mean change in allele frequency per generation

is determined by the difference in allele frequency between the immigrants and the island

$$M_{\delta x} = m(x_m - x)$$

And the variance in allele frequency change is given by binomial sampling

$$V_{\delta x} = \frac{x(1-x)}{2N_e}$$

Under these conditions, across an infinite collection of replicate island populations, the distribution of allele frequencies on the island at migration-drift equilibrium is

$$\phi(x) = Cx^{Mx_m-1}(1-x)^{Mx_m-1} \quad (7)$$

Where  $C$  is a normalising constant. For a more detailed derivation see Crow and Kimura (1970). When  $M$  is high, allele frequencies on the island are clustered around that on the mainland. When  $M$  is low, the distribution is U-shaped, such that fixation on the island is quite likely.

How can we use this result for inference? Suppose we know the mainland allele frequencies, and have a sample of chromosomes from the island population. By combining (7) with the multinomial sampling distribution for allele counts we can derive the likelihood of observing a set of alleles for a given value of  $M$ . If we have data on multiple unlinked (hence independent) loci, likelihoods can be combined across loci. More generally, if there are multiple alleles at each locus, the distribution of alleles in the island population follows a Dirichlet distribution, and sample counts follow a Dirichlet-multinomial distribution (e.g. Rousset, 2001).

## 5 Split-time models

All the models of population subdivision considered so far share one critical feature; equilibrium. But for the vast majority of biological populations, particularly, human populations, the assumption of equilibrium is untenable. What models of population structure might one use in these circumstances?

The simplest non-equilibrium model of population structure is the split-time model, in which at some point in the past some ancestral population gave rise to two daughter populations which have remained completely isolated ever since. Over time, the populations differentiate from each other at a rate determined by the population size (larger populations differentiate slower). Under this model it is only possible to estimate the time of divergence of the populations scaled by the effective population size of the daughter populations. We will also need to estimate the allele frequencies in the ancestral population, although an alternative would be to assume a Bayesian approach and integrate over a prior for allele frequencies in the ancestral population.

Given a set of parameters (divergence time and allele frequencies), how can we estimate the likelihood? If we were dealing with unlinked loci, one possibility would be to use the diffusion approximation to describe how allele frequencies change over time due to genetic drift (Kimura, 1955), however likelihood estimation within this framework is computationally daunting. An alternative would be to assume that the diffusion approach can be well approximated by a simpler process (e.g. Gaussian), for which likelihood-based inference is fast and efficient (Nicholson et al., 2002). Both these approaches can only be used on unlinked, independent loci. Yet another approach would be to take a genealogical perspective, which can be generalised to the case of linked or unlinked loci.

When the data consist of unlinked SNP markers, it is possible to write down exact likelihoods from a genealogical perspective. Consider data from a single locus where in population  $x$  we have observed  $n_{1x}$  of one allele in a sample of size  $n_x$ , and in population  $y$  we have observed  $n_{1y}$  in a sample of size  $n_y$ . We wish to estimate the divergence time for the population pair in terms of their current effective population sizes. If, as in the case of bottlenecks, we assume that no mutations have occurred since the population split, the only thing we can estimate is the product of the population size and time of divergence, which we shall call a single parameter for each population,  $v_x$  and  $v_y$  (the analogue of branch length in

phylogenetic trees). Looking back in time we can trace the ancestral lineages as they coalesce, until the point of the population split, at which point we shall assume that the ancestral alleles from the two populations were sampled in a binomial fashion from the alleles in the ancestral population. By using equation (2), we can see that all information about the branch length is contained in the number of coalescent events since the split. Also, all information about the ancestral allele frequencies is contained in the distribution of allelic type in the ancestral lineages. Consequently, we can obtain the likelihood for each locus in each population by summing over all possible values of  $a_{1x}$  and  $a_x$ . The distribution of  $a_x$  given  $v_x$  and  $n_x$  is given by (2); the distribution of  $a_{1x}$  given the ancestral allele frequency  $f_1$  is binomial, and the likelihood of the data given  $a_{1x}$  and  $a_x$  is obtained from the Urn model (dropping the population subscripts):

$$P(n_1|a_1, a, n) = \binom{n_1 - 1}{a_1 - 1} \binom{n - n_1 - 1}{a - a_1 - 1} \binom{n - 1}{a - 1}^{-1} \quad (8)$$

Within this framework it is possible to formulate a full-likelihood method to estimating population trees using similar principles to those used in phylogenetic tree reconstruction (Felsenstein, 1981).

Split-time models, and island-models can generate very similar patterns of polymorphism. Certainly, if we were simply to summarise differentiation only in terms of  $F_{ST}$  we would have no way of distinguishing recurrent gene flow from isolated differentiation. Is there any way of distinguishing between the models? Fortunately there are subtle differences in the patterns of polymorphism expected under the different models. For example, the variance in pairwise differences is greater under a model of recurrent migration than a model of divergence (Wakeley, 1996). Furthermore, a recent Monte Carlo Markov Chain method aimed at differentiating between the two models based on coalescent likelihoods claims that the two models can be distinguished by means of a likelihood ratio test (Nielsen and Wakeley, 2001). Where the information comes from for such power is not clear.

## 6 Admixture

In many ways, the opposite process to population splits is admixture, where a new population is formed from two source populations. Such models apply to many human populations, for example the mixed European and African ancestry of Jamaica and South Africa. In terms of the parameters of the model, these are similar to the split-time model; the time of the population foundation, allele frequencies in the founding population and the population sizes of the admixed and source populations. The only additional parameter is the proportion of each source population that contributed to the admixed population.

### 6.1 Detecting admixture

Admixture in human populations is both widespread and important. The main reason for its importance in terms of population genetics is that admixture generates linkage disequilibrium between alleles simply due to differences in allele frequency. For example, if two populations differ in allele frequency at two loci by amounts  $\delta_A$  and  $\delta_B$ , the combination of a fraction  $p$  of one population and  $1 - p$  of the other generates the disequilibrium  $D_0 = \delta_A \delta_B p(1 - p)$ , and this decays by a factor of  $(1 - r)$  each generation, where  $r$  is the recombination rate between the loci. Consequently, admixture can generate considerable linkage disequilibrium which persists over several generations, even for completely unlinked loci.

The reason that linkage disequilibrium being caused by admixture is problematic is that association studies which aim to identify genes involved in susceptibility to common diseases use linkage disequilibrium between markers and the disease phenotype to identify the genes involved. Linkage disequilibrium caused by admixture will generate associations across the entire genome if there is a difference in disease prevalence in the two source populations, making identification of disease-associated loci very difficult.

How can one go about detecting admixture? There are several approaches

that differ in the type of information available and assumptions about the time of admixture. For example, if there is information from putative source populations, even old admixture events can be detected and the relative contribution of source populations estimated (Chikhi et al., 2001). In contrast, if there is no *a priori* choice of specific source populations, then as long as the admixture event was recent, clustering algorithms can be used to group similar classes of individuals within a population, and identify those that are combinations of the two (Pritchard et al., 2000).

## References

- Bahlo M, Griffiths RC (2000). Inference from gene trees in subdivided populations. *Theor. Pop. Biol.* 57:79–95.
- Barton NH, Depaulis F, Etheridge A (2002). Gene frequencies and genealogies in spatially continuous populations. Unpublished.
- Beaumont MA (1999). Detecting population expansion and decline using microsatellites. *Genetics* 153:2013–2029.
- Berli P, Felsenstein J (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763–773.
- Chikhi L, Bruford MW, Beaumont MA (2001). Estimation of admixture proportions: A likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158:1347–1362.
- Cockerham CC, Weir BS (1987). Correlations, descent measures: drift with migration and mutation. *Proc. Natl. Acad. Sci. U.S.A.* 84:8512–8514.
- Crow JF, Kimura M (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

- Excoffier L (2001). Analysis of population subdivision. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, Chapter 10, pp. 271–307. John Wiley & Sons Ltd.
- Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Felsenstein J (1975). A pain in the torus: some difficulties with models of isolation by distance. *Am. Nat.* 109:359–368.
- Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17:368–376.
- Fu YX (1996a). New statistical tests of neutrality for DNA samples from a population. *Genetics* 143:557–570.
- Fu YX (1996b). Statistical tests of neutrality against population growth, hitchhiking and background selection. *Genetics* 147:915–925.
- Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
- Goldstein DB, Zhivotovsky LA, Nayar K, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1996). Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol. Biol. Evol.* 13:1213–1218.
- Griffiths RC (1980). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Pop. Biol.* 17:37–50.
- Hanski I, Ovaskainen O (2000). The metapopulation capacity of a fragmented landscape. *Nature* 404:755–758.
- Hudson RR (2000). A new statistic for detecting genetic differentiation. *Genetics* 155:2011–2014.

- Kimura M (1953). "Stepping-stone" model of population. *Annu. Rept. Natl. Inst. Genet. Jpn.* 3:62–63.
- Kimura M (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. U.S.A.* 41:144–150.
- Kuhner MK, Yamato J, Felsenstein J (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149:429–434.
- Li WH (1976). Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. *Theor. Pop. Biol.* 10:303–308.
- Nei M (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70:3321–3323.
- Nei M (1975). *Molecular Population Genetics and Evolution*. New York: Elsevier.
- Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P (2002). Assessing population differentiation and isolation from single nucleotide polymorphism data. *J. Royal Stat. Soc.* in press.
- Nielsen R, Wakeley J (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158:885–896.
- Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
- Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, et al (2001). Linkage disequilibrium in the human genome. *Nature* 411:199–204.
- Rogers AR, Harpending H (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9:552–569.

- Rousset F (2001). Inferences from spatial population genetics. In D. J. Balding, M. Bishop, and C. Cannings (Eds.), *Handbook of Statistical Genetics*, Chapter 9, pp. 239–269. John Wiley & Sons Ltd.
- Slatkin M (1987). The average number of sites separating DNA sequences from a geographically structured population. *Theor. Pop. Biol.* 32:42–49.
- Slatkin M, Hudson RR (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Tavaré S (1984). Line-of-descent and genealogical processes, and their applications in population genetics processes. *Theor. Pop. Biol.* 26:119–164.
- Wakeley J (1996). Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Pop. Biol.* 49:369–386.
- Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7:256–276.
- Wright S (1931). Evolution in Mendelian populations. *Genetics* 16:97–159.
- Wright S (1969). *Evolution and the Genetics of Populations, Vol. II. The Theory of Gene Frequencies*. Chicago: University of Chicago Press.
- Wright S (1978). *Evolution and the Genetics of Populations, Vol. IV. Variability Within and Among Natural Populations*. Chicago: University of Chicago Press.
- Zhivotovsky LA, Bennett L, Bowcock AM, Feldman MF (2000). Human population expansion and microsatellite variation. *Mol. Biol. Evol.* 17:757–767.