DESISLAVA PETKOVA

INFERRING EFFECTIVE MIGRATION FROM GEOGRAPHICALLY INDEXED GENETIC DATA

THE UNIVERSITY OF CHICAGO

Contents

List of Figures

Genetic data often exhibit patterns that are broadly consistent with "isolation by distance" — a phenomenon where genetic similarity tends to decay with geographic distance. In a heterogeneous habitat, decay may occur more quickly in some regions than others: for example, barriers to gene flow such as mountains or deserts could accelerate the genetic differentiation between neighboring groups. In this thesis we present a method to quantify and visualize variation in effective migration across the habitat, and, under further assumptions, to infer the presence or absence of barriers to migration, from geographically indexed large-scale genetic data.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Matthew Stephens, for his guidance and encouragement throughout the development of this project. His tremendous support made this formidable journey less complicated, if not easier.

I am also grateful to my colleagues at the Departments of Statistics and Human Genetics for their inspiring companionship, for their collective critical eye, but above all, for their advice, assistance and guidance on many difficult problems.

I am indebted to so many but I especially wish to acknowledge the efforts, love and encouragement of my parents. They watched me from a distance while I worked towards my degree. The completion of this thesis would mean a lot to them, so I dedicate this project to my parents, Ema and Ivo.

And finally, I would like to thank my sister and my friends for allowing me to realize my own potential and without whose love, affection and encouragement this thesis and many other pursuits would not have been successful.

1
Population Structure in Genetic Variation

The term "population structure" is used to describe nonrandom patterns of genetic similarity (or alternatively, dissimilarity) between individuals from the same species. One task is to detect such patterns; this is often done in association studies because systematic ancestry differences between cases and controls that are not genetic risk factors for the disease can bias the results of the study [Price et al., 2010]. A more challenging task is to explain population structure as the outcome of events in the evolutionary history of the species such as splits or admixture (events in time) and/or migration (events in space) [Lawson and Falush, 2012].

Admixture is partial ancestry from two or more distinct subpopulations as the result of interbreeding.

Two widely used approaches for inferring genetic ancestry are principal components analysis and model-based clustering. In both cases, interpretation of results and inference of demography are founded on the assumption that sample structure is evidence for population structure, to the exclusion of other possible sources such as family structure, cryptic relatedness or sample processing artifacts.

Principal component analysis (PCA) was first used in population genetics to summarize human genetic variation across continents [Menozzi et al., 1978, Cavalli-Sforza et al., 1994]. Their synthetic maps of allele frequency variation show gradients that could support hypotheses for specific migration events such as the spread of Neolithic farming. This interpretation of PCA maps is not universally accepted because PCA can produce similar wave patterns in simulated spatial data, where gradients result from local dispersal and not directed migration [Novembre and Stephens, 2008]. However, even though PCA might not explain what processes generated the structured variation in genetic data, the method has been successfully applied to detect population stratification and infer genetic ancestry. For example, the top principal components of the sample covariance matrix across a large number of (randomly selected) SNPs align well with geographic distribution in some datasets [Novembre et al., 2008, Wang et al., 2012].

Alternatively, population structure can be analyzed with a model-based clustering approach. For example, STRUCTURE [Pritchard et al., 2000] assigns individuals into K genetically homogeneous subpopulations [i.e., random mating and hence under Hardy-Weinberg equilibrium], with individual-specific ancestry proportions. As a clustering algorithm, STRUCTURE assumes the number of clusters is known. Even more importantly, it uses a discrete model of population structure that is most applicable where high level of divergence have resulted into well differentiated clusters.

Both PCA and STRUCTURE can produce results that are difficult to interpret. For example, STRUCTURE can fail if the population consists of distinct groups characterized by small differences in allele frequencies, or a single population where the distribution of allele frequencies varies continuously across space. In both cases, it is hard for a clustering algorithm to distinguish between clusters, or find the correct number of clus-

ters. The study design can also influence the extent of observed "clusteredness" as many datasets consist of multiple observations from few locations [Serre and Pääbo, 2004]. If the sample configuration does not represent the geographic distribution of the species, much naturally occurring genetic variation remains unobserved. And indeed the genetic differentiation of a widely distributed species such as humans is likely to exhibit evidence for both clusters, which correspond to discontinuous jumps in allele frequencies across large barriers such as oceans or the Himalayas, and clines, which reflect smooth gradations in allele frequencies across unbroken geographic regions [Rosenberg et al., 2005].

In the case of low differentiation between clusters, STRUCTURE results can be improved with a stronger, more informative prior on cluster membership. For example, [Hubisz et al., 2009] introduces a prior that places more weight on cluster assignments that are correlated with sampling locations (because origin is often informative about ancestry). In another modification of STRUCTURE that incorporates geographic information, [Guillot et al., 2005] explicitly models the distribution of clusters across the habitat to encourage spatially continuous clusters (because subspecies often occupy locally connected areas).

In the case of smoothly varying population structure, it is not appropriate to assign individuals to a fixed number of distinct clusters, even if the clustering method allows fractional membership. PCA is effective in presenting continuous variation and PC projections are related to the underlying genealogical process [McVean, 2009]. However, the algorithm is not based on a population genetics model, so it does not estimate relevant demographic parameters, and its results are strongly affected by uneven sampling.

Genetic data often exhibit patterns that are broadly consistent with "isolation by distance" [Weiss and Kimura, 1965, Rousset, 1997] where genetic similarity tends to decay with geographic distance. That is, a population in which the exchange of migrants is constant in both space and time still has structure as individuals that are close together are, on average, more genetically similar than individuals that are far apart [if reproduction and dispersal tend to occur locally/over small distances in every generation].

In a heterogeneous habitat, genetic similarity may decrease faster in some regions than others because a barrier to migration could accelerate the genetic differentiation between neighboring groups — thus creating patterns of population structure that are not consistent with uniform migration. Here we develop a method aimed at investigating this kind of scenario. Specifically, we introduce a parametric model for genetic structure that attempts to explain the spatial structure observed in geographically indexed large-scale genetic data in terms of effective rates of migration. We say "effective" because the model's applicability to genetic data is motivated under of series of assumptions [most importantly, equilibrium in time] that mean estimated rates cannot be interpreted as actual rates of migration unless the assumptions are reasonably satisfied. However, even when estimated population parameters are not directly interpretable in terms of demographic history, our method provides an intuitive and informative way to quantify and visualize spatial patterns of population structure.

2

Population Structure due to Migration

In this background chapter we explain how population structure is reflected in observed genetic data via the genealogy of the sample and review briefly a mathematical model for spatially structured populations.

In natural populations, mating is not random due to a complex mixture of evolutionary and ecological factors. Non-random mating creates structure in genetic variation as closely related individuals tend to be more similar genetically than distantly related individuals. Thus shared ancestry leads to genetic similarity [Section 2.1].

An important factor for non-random mating is geographic distance as two individuals located close in space are more likely to reproduce than two individuals far apart. Thus geographic proximity leads to genetic similarity — a phenomenon called *isolation by distance* [Section 2.2].

A population genetics model that exhibits isolation by distance is Kimura's stepping-stone model [Section 2.3]. It represents a spatially distributed population as a graph where vertices are groups of randomly mating individuals (called *demes*) and edges are direct routes of migration. Thus demes that are closer together in the graph tend to be more similar.

In fact, the stepping-stone model can capture the effect of both geographic distance and heterogeneous habitat on genetic similarity as edges can have different migration rates to reflect heterogeneity in gene flow. This weighted population graph describes precisely what it means for two demes to be "close together" [Section 2.4].

2.1 Pairwise expected coalescence times explain population structure

The more closely related two individuals are, the more genetically similar they are. Therefore, the genetic similarities observed in a sample contain information about the evolutionary processes undergone by the entire population. In this section we explain the connection between genealogical histories and genetic similarities; the review is largely based on [McVean, 2009].

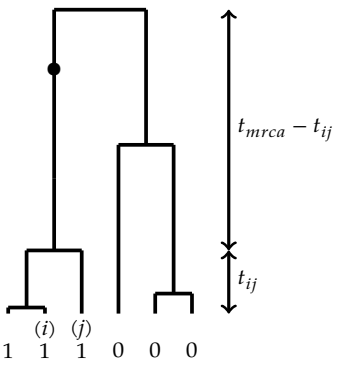


Figure 2.1: This genealogy specifies one possible history for a sample of size 6. Since exactly one mutation occurs, denoted by •, we observe a pattern of both 0s and 1s. Regardless of which branch carries the mutation, the events 'only 0s' and 'only 1s' are excluded.

Let $z_1, \ldots, z_n$ be the genotypes of n individuals at a single segregating locus. For simplicity, assume the genetic markers are biallelic (e.g., SNPs): each individual carries either the ancestral allele, labeled '0', or the derived allele, labeled '1'.

Although life occurs forward in time and in discrete generations, it is often more convenient to model the ancestry of a sample backwards in time using a continuous-time process called the *coalescent* [Kingman, 1982b,a] that traces the lineages backwards in time until their convergence into a single common ancestor. Thus the coalescent constructs the history of the sample, at a single locus, in the form of a genealogical tree [Figure 2.1]. The most important demographic functions of the genealogy are

- the time to the most recent common ancestor t_{mrca} [the height of the tree];
- the total size of the tree t_{tot} [the sum of all branches];
- the pairwise time to coalescence t_{ij} for every pair (i, j) [the length of the path from i, or equivalently from j, to the most recent common ancestor of i and j].

With a slight abuse of notation, let K_t denote the number of mutations that occur on a path with length t. If mutations are generated by a Poisson process with intensity [mutation rate] θ, the probability that the path accumulates a mutation depends only on its relative length, not on its position within the genealogy. In particular, $\mathrm{P}\{K_t = 0\} = \mathrm{E}\{e^{-\theta t}\}$. Similarly, let $K_{t_{tot}}$ denote the number of mutations that occur throughout the genealogy. Thus $\{K_{t_{tot}} > 0\}$ is the event that the site segregates in the sample. For a fixed mutation rate θ, the probability that at least one mutation occurs on a path with length t and none in the rest of the tree is

$$\mathrm{P}\{K_t > 0, K_{t_{tot}-t} = 0\} = \mathrm{E}\{(1 - e^{-\theta t})e^{-\theta(t_{tot}-t)}\}. \tag{2.1}$$

Similarly, the probability that the site segregates is

$$\mathrm{P}\{K_{tot} > 0\} = \mathrm{E}\{1 - e^{-\theta t_{tot}}\}, \tag{2.2}$$

where the expectation is with respect to all possible genealogies of the sample.

[Nielsen, 2000] argues that if we assume the mutation rate is low and condition on the site segregating in the sample, then the mutation rate θ is of little interest and so it can be treated as a nuisance parameter. Following [Nielsen, 2000] we can eliminate θ from the analysis by taking the limit $\theta \to 0$. Under the infinitely-many-sites model [Kimura, 1969], the event of at least one mutation is equivalent to the event of exactly one mutation. Therefore, $\mathrm{P}\{t = 0\}$ and $\mathrm{P}\{t = 1\}$ are complementary events. Together with the low mutation limit $\theta \to 0$, this implies

Alternatively, without explicitly making the infinitely-many-sites assumption, we can ignore the probability of event $\{K_t > 1\}$ if the mutation rate θ is very low.

$$\mathrm{P}\{K_t = 1 | K_{tot} = 1\} = \frac{\mathrm{P}\{K_t > 0, K_{tot} > 0\}}{\mathrm{P}\{K_{tot} > 0\}} = \frac{\mathrm{P}\{K_t > 0, K_{t_{tot}-t} = 0\}}{\mathrm{P}\{K_{tot} > 0\}} \tag{2.3a}$$

$$= \lim_{\theta \to 0} \frac{\theta^{-1}\mathrm{E}\{e^{-\theta(t_{tot}-t)} - e^{-\theta t_{tot}}\}}{\theta^{-1}\mathrm{E}\{1 - e^{-\theta t_{tot}}\}} \tag{2.3b}$$

Interchange limit and expectation [valid if $\mathrm{E}\{t_{tot}\} < \infty$] and use the Taylor approximation $e^{-x} \sim 1 - x$.

$$= \frac{\mathrm{E}\{t_{tot} - (t_{tot} - t)\}}{\mathrm{E}\{t_{tot}\}} = \frac{\mathrm{E}\{t\}}{\mathrm{E}\{t_{tot}\}} \equiv \frac{T}{T_{tot}}, \tag{2.3c}$$

where for convenience we denote the expectation of coalescence time t by T.

Therefore, for biallelic markers and under the conditions specified above, there is a relationship between expected coalescence times and the probability that a particular branch in the genealogy carries the derived allele. We will use it to derive the first two moments of the genotype vector $Z = (Z_1, \ldots, Z_n)$.

Proposition 2.1 *Suppose that a sample of size n is collected from a population that evolves according to the neutral infinitely-many-sites model, where mutations are generated by a Poisson process with low mutation rate. At segregating sites where exactly one mutation occurs in the sample, the allele carried by individual i, denoted by Z_i, is a binary random variable such that*

$$\mathrm{E}^*\{Z_i\} = \frac{T_{mrca}}{T_{tot}}. \tag{2.4}$$

Furthermore, for two distinct individuals i and j,

$$\mathrm{E}^*\{Z_i Z_j\} = \frac{T_{mrca} - T_{ij}}{T_{tot}}. \tag{2.5}$$

Here the symbol $*$ indicates that the expectation is with respect to all possible sample genealogies with exactly one mutation.

Proof. In a genealogical tree with exactly one mutation, the ith lineage carries the derived allele if the mutation occurs anywhere on the path from ith external branch to the most recent common ancestor of the entire sample. This path has length t_{mrca} for all i; its average length is T_{mrca}. Therefore, the conditional probability of observing the derived allele is the same for every individual and

$$\mathrm{E}^*\{Z_i\} = \mathrm{E}\Big\{K_{t_{mrca}} = 1 \,|\, K_{tot} = 1\Big\} = \frac{T_{mrca}}{T_{tot}}. \tag{2.6}$$

That is, the genotypes at a biallelic marker are Bernoulli random variables with frequency T_{mrca}/T_{tot}. Furthermore, since the genotypes are binary, the event $\{Z_i = 1, Z_j = 1\} \Leftrightarrow \{Z_i Z_j = 1\}$ implies that the mutation occurs on the branch from the pair's most recent common ancestor to the most recent common ancestor of the sample. This ancestral branch has length $t_{mrca} - t_{ij}$. Therefore, the conditional expectation that two individuals i and j carry a common mutation at a biallelic marker is given by

$$\mathrm{E}\Big\{Z_i Z_j \,|\, K_{tot} = 1\Big\} = \mathrm{P}\Big\{K_{t_{mrca}-t_{ij}} = 1 \,|\, K_{tot} = 1\Big\} = \frac{\mathrm{E}\{t_{mrca}\} - \mathrm{E}\{t_{ij}\}}{\mathrm{E}\{t_{tot}\}}. \tag{2.7}$$

□

Thus the individual genotypes have the same marginal distribution: the Z_is are identically distributed but not independent. Finally, in equations (2.4) and (2.5) the expected coalescence times T_{mrca}, T_{tot}, T_{ij} are marginal expectations with respect to all possible histories [genealogies] of the sample, not only histories that can induce the observed pattern of 0s and 1s.

The principle behind equation (2.5) states that the more history two individuals share, the more genetically similar they are. Here we should interpret "shared history" precisely as "common ancestral branch" in the genealogy rather than broadly as a "demographic past" in the sense of evolutionary history. Different models can produce the same expected genealogy. For example, a long branch separating two samples could correspond to a split into distinct subpopulations some time in the past or constant migration between two locations at a low rate. Conversely, without further assumptions, patterns of similarities and differentiation observed in genetic data reveal information about the underlying genealogies, and hence, indirectly, about the demographic model that generated them. In this thesis we average observed genetic similarities across markers; thus we ignore information (e.g., the variance) that could in principle improve the ability to distinguish demographic models.

2.1.1 Bias due to SNP ascertainment

Ascertainment bias refers to systematic deviations in the SNP discovery process where a small number of individuals are used to find sites polymorphic in the entire population [Clark et al., 2005]. In particular, rare SNPs are harder to ascertain and more likely to be underrepresented. Furthermore, the genetic variation in a geographic region could be misrepresented in a panel with unbalanced sample configuration. [McVean, 2009] observes that two samples are effectively involved in ascertainment — first a panel to discover SNPs for genotyping on a microchip and then a sample to genotype. We condition on sites that segregate in both samples and this can distort (the average shape of) the observed genealogies and thus produce misleading results. In this thesis we ignore SNP ascertainment as a potential source of sample structure.

2.2 *Isolation by distance in a spatially distributed population*

Geographic separation can act as a genetic barrier because in a natural population migration tends to be local rather than long-range. If long-distance migration events are rare, a mutation that arises in one area might take a long time to spread throughout the habitat (if at all). Consequently, individuals that are closer together tend to be more similar genetically than those that are far apart. This phenomenon is known as isolation by distance. However, the relationship between geography and genetic similarity also depends on dispersal. If the habitat is homogeneous and migration is characterized by the same dispersal density everywhere, genetic similarity decreases as a function of relative distance.

The effect of subdivision on population structure is often quantified in terms of a statistic called F_{ST} that measures the genetic variation among subpopulations relative to the total genetic variation. Several definitions of F_{ST} have been proposed [Wright, 1943, Cockerham, 1969, Nei, 1973]. We use Nei's definition where F_{ST} is a function of the probabilities of identity within and between subpopulations. [Two lineages are identical, at a given locus, if they carry the same allele.] The F-statistic is defined as

[Wright, 1943] introduces F_{ST} as the statistic $\mathrm{var}\{p\}/[\bar{p}(1-\bar{p})]$ where $\mathrm{var}\{p\}$ is the variance in allele frequency among subpopulations and $\bar{p}$ is the overall mean allele frequency in the population. Intuitively, F_{ST} is high when individuals are similar within subpopulations and different between subpopulations.

$$F_{ST} = \frac{\phi_0 - \phi}{1 - \phi}, \tag{2.8}$$

where ϕ is the probability of identity for two individuals chosen at random without reference to geography, and ϕ_0 is the probability of identity for two individuals chosen at random from the same subpopulation.

As a measure of genetic differentiation, the F-statistic is related to coalescence times because identity means neither lineage accumulates a mutation in the time to most recent common ancestor. If the mutation process is Poisson with low mutation rate θ,

$$\phi(\theta) = \mathrm{E}\left\{e^{-\theta t}\right\} \approx 1 - \theta \mathrm{E}\{t\}. \tag{2.9}$$

In this case, by substituting $\phi_0 = 1 - T_0$ and $\phi = 1 - T$ in equation (2.8), [Slatkin, 1991] derives the approximation

$$F_{ST} \approx \frac{T - T_0}{T}, \tag{2.10}$$

where T_0, T are the expected coalescence times for a pair of distinct lineages sampled at random from the same subpopulation and from the entire population, respectively. The coalescent-based approximation to the F-statistic is very general: [Slatkin, 1991] derives it in the low mutation limit but otherwise makes no assumptions about the demographic model. Thus, the approximation holds under a subdivided population at equilibrium, a growing population, or a population that has undergone a split some time in the past.

By analogy, [Rousset, 1997] considers the F-statistic for two demes separated by distance x,

$$F_{ST}(x) = \frac{\phi_0 - \phi_x}{1 - \phi_x} \approx \frac{T_x - T_0}{T_x}, \tag{2.11}$$

as well as the linearized F-statistic given by

$$\frac{F_{ST}(x)}{1 - F_{ST}(x)} \approx \frac{T_x - T_0}{T_0}. \tag{2.12}$$

[Rousset, 1997] analyzes the relationship between genetic differentiation, F_{ST}, and geographic distance, x, in a spatially-homogeneous stepping-stone model where demes

are equally sized and regularly spaced (on a ring in one dimension and a torus in two dimensions), and migration is determined by a symmetric dispersal kernel. The important demographic parameters are the effective population density D per length/area unit and the mean squared dispersal distance σ^2, which determines the speed at which two lineages with a common ancestor move away from each other in a generation. By symmetry, the probability of identity for two randomly sampled individuals is also a function of the relative distance x. [Rousset, 1997, 2004] derives the following large-distance approximations to the linearized F_{ST},

$$\frac{F_{ST}(x)}{1-F_{ST}(x)} \approx \frac{x}{4D\sigma^2} + C_1; \qquad \text{(in one dimension)} \qquad (2.13a)$$

$$\frac{F_{ST}(x)}{1-F_{ST}(x)} \approx \frac{\ln(x/\sigma)}{4\pi D\sigma^2} + C_2; \qquad \text{(in two dimensions)} \qquad (2.13b)$$

where the constants C_1 and C_2 depend on the population density and the dispersal distribution but not on the population sizes or the mutation rate.

Therefore, if migration is uniform, the linearized F_{ST} increases with geographic distance. This relationship is appropriate only for homogeneous habitats as it ignores the effect of barriers (or corridors) to migration: two demes separated by a barrier would appear to be more genetically dissimilar than relative distance would suggest. In other words, we need a measure of *effective* distance to describe the patterns of movement across the habitat.

2.3 *The stepping-stone model of population subdivision*

Section 2.1 explains that coalescence times represent population structure because genetically similar individuals are likely to have a recent common ancestor and thus shorter coalescence time. The relationship between genetic correlation and coalescence times in equation (2.5) is very general. For example, [McVean, 2009] uses as an example a model of population split in which groups derived from a common ancestor do not exchange migrants and thus develop independently after the split. In this thesis we aim to analyze the spatial structure of genetic variation, and therefore, we need to model [and apply equation (2.5) to] a spatially distributed population.

Kimura's stepping-stone model [Kimura and Weiss, 1964] represents a population across the span of its habitat as a connected grid of panmictic (randomly mating) demes (colonies) which exchange migrants in a fixed pattern. For simplicity, in this chapter we consider a haploid population. To extend the framework, a diploid individual can be represented as the sum of two independent haplotypes, one from each parent.

A haploid organism has a single copy of its genome; a diploid organism has two copies, one inherited from the father and the other from the mother.

The stepping-stone model makes the following assumptions:

- There are d demes and deme α consists of N_α randomly mating individuals. The total population number is $N_T = \sum_\alpha N_\alpha$ and the average deme size is $N_0 = N_T/d$. The demes remain constant in size and $N_\alpha \sim O(N_0)$ for all α.

- The mutation rate per site per generation is u and the scaled mutation rate for two distinct lineages in N_0 generations is $\theta = 2N_0u$.

The ancestral process develops backwards in time, from the present towards the past. A coalescence event means that two individuals have the same parent and a migration event means that an individual from α has a parent from β.

- The coalescence rate for a pair of distinct lineages drawn at random from deme α is $q_\alpha = N_0/N_\alpha \sim O(1)$. Two lineages coalesce when they merge into a common ancestor and in a single generation this event has probability $1/N_\alpha$.

- The migration rate for a lineage to move from deme α to deme $\beta \neq \alpha$ is $m_{\alpha\beta} \sim$

$O(1)$. The migration matrix $M = (m_{\alpha\beta})$, where $M_{\alpha\alpha} = -\sum_{\beta:\beta\neq\alpha} m_{\alpha\beta}$, describes the transition process of a single lineage backwards in time.

All rate parameters are constant in times and on the scale of N_0 generations. The assumptions $q_\alpha \sim O(1)$ for every deme α and $m_{\alpha\beta} \sim O(1)$ for every pair $(\alpha \neq \beta)$ imply that migration is *weak*. That is, the probability of multiple migration and/or coalescence events occurring in the same generation [before scaling by N_0] is $O(N_0^{-2})$ and can be ignored.

The stepping-stone model describes how a spatially distributed population evolves under equilibrium in time, i.e., under the condition that both migration and coalescence rates are the same in every generation. Therefore, the model can characterize systematic differences between the groups due to gene flow but not due to splits or admixture events. In other words, the stepping-stone model can represent population structure in space but not in time. [As we show through simulations in Chapter 5, temporal structure can be explained as spatial structure, in terms of effective rates of migration.]

If demes of constant size exchange migrants at fixed rates as required under equilibrium, the number of individuals to emigrate is equal the number of individuals to immigrate, i.e., migration is *conservative* [Nagylaki, 1980]. Mathematically,

$$\sum_{\beta:\beta\neq\alpha} m_{\alpha\beta}/q_\alpha = \sum_{\beta:\beta\neq\alpha} m_{\beta\alpha}/q_\beta \qquad \Leftrightarrow \qquad M'q^{-1} = 0 \tag{2.14}$$

where $q^{-1} = (q_\alpha^{-1}) = (N_\alpha/N_0)$ is the vector of coalescence rates.

In a general stepping-stone model, migration is not necessarily symmetric. However, in this thesis we assume that $m_{\alpha\beta} = m_{\beta\alpha}$ for all edges (α, β). The condition that migration is both symmetric and conservative implies that all demes have the same size: on one hand, $Mq^{-1} = M'q^{-1} = 0$, and on the other, $M1 = 0$ as M is a Laplacian matrix; hence $q \propto 1$. Thus the average deme size $N_0 = N_T/d$ is a convenient choice for the coalescent timescale.

The stepping-stone model characterizes dispersal not in terms of an explicit dispersal density but indirectly through the combined effect of the graph topology and the migration rates. It may not seem natural to represent the geographic distribution of organisms with a graph. However, discrete models for migration are common in population genetics. In fact, a continuous model of isolation by distance (with normal dispersal and continuous spatial distribution) can lead to inconsistencies [Felsenstein, 1975].

2.3.1 Expected coalescence times in a subdivided population

In Section 2.1 we described how the probability that two individuals both carry the derived allele is related to the expected coalescence time to their most recent common ancestor. We will use this connection between genetic similarity and shared ancestry to analyze the spatial structure in genetic variation, and in particular, to estimate migration rates. The inference procedure requires that we express pairwise coalescence times as functions of migration rates.

The coalescent process can be extended to represent the ancestry of a sample from the stepping-stone model [Notohara, 1990, 1993]. This version, called the *structured coalescent*, describes the movement of lineages between demes as well as their coalescence into common ancestors. We can use the properties of the structured coalescent [as we do in Appendix 7.1] to derive the following system of linear equations for the pairwise expected coalescence times $T = (T_{\alpha\beta})$ as a function of the coalescence rates $q = (q_\alpha)$

and the migration rates $M = (m_{\alpha\beta})$:

$$\text{diag}\{q\}\,\text{diag}\{T\} - MT - TM' = 11'. \tag{2.15}$$

Furthermore, if the migration rates are symmetric, as we assume throughout, then there is no variation in coalescence rates across demes, i.e., $q = 1, M = M'$ and

$$\text{diag}\{T\} - MT - TM = 11'. \tag{2.16}$$

In equation (2.15) $T_{\alpha\beta}$ is the expected coalescence time between two randomly chosen lineages, one from α and the other from β. In equation (2.5) T_{ij} is the expected coalescence times between two sampled individuals $i \in \alpha_i$ and $j \in \alpha_j$. Crucially, the pairwise coalescence times do not depend on the sample configuration, $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)$, because individuals are exchangeable within each deme. Therefore, the expected coalescence time for an observed pair $(i \in \alpha_i, j \in \alpha_j)$ is the same as the expected coalescence time for any pair $(i' \in \alpha_i, j' \in \alpha_j)$ from the subdivided population:

Individuals are exchangeable *within* demes but not *across* demes because the sample location is informative about the alleles an individual carries.

$$T_{ij} = T_{\alpha_i\alpha_j}. \tag{2.17}$$

Notation: We use Greek letters $[\alpha, \beta]$ to denote subpopulations and Latin letters $[i, j]$ to denote sampled individuals. And we will distinguish between the population matrix $T = (T_{\alpha\beta} : \text{demes } \alpha, \beta)$ and the sample matrix $\underline{T} = (T_{ij} : \text{individuals } i, j)$ where $\underline{T} = T(\underline{\alpha}) - \text{diag}\{T(\underline{\alpha})\}$. The diagonal is subtracted because coalescence time with self is always 0.

In any population graph, $T_{\alpha\beta} > T_{\alpha\alpha}$ because coalescence is possible only for lineages in the same deme. However, if α and β are separated by a barrier, fewer migrants move between α and β, and so the pairwise coalescence times $T_{\alpha\beta}$ would be larger than the time expected under isolation by distance, i.e., uniform migration. Thus, the matrix of pairwise coalescence times $T = (T_{\alpha\beta})$ would contain evidence for habitat heterogeneity.

Since longer coalescent time mean less genetic similarity, coalescence times are a natural measure of genetic dissimilarity and hence population structure. For the stepping-stone model we can compute the matrix of expected coalescence times, T, given the migration rates M and the coalescence rates q using equation (2.15). Alternatively, there exists a computationally efficient method to approximate T, which we discuss next.

2.4 Isolation by resistance is a metric for gene flow

Isolation by resistance (IBR) [McRae, 2006, McRae et al., 2008] draws an analogy between a subdivided population in which neighboring demes exchange migrants and an electrical network in which current flows through conductors. [Or in other words, between Kimura's stepping-stone model and an undirected random walk.] To understand the analogy better, concepts in electrical networks can be given population genetic interpretation [Table 2.1]. Using this correspondence between population genetics and circuit theory, McRae develops IBR to test whether putative barriers to genetic flow affect genetic differentiation.

Isolation by resistance predicts effective distances from a raster grid of landscape resistance (or friction): each cell in the grid specifies how difficult it is for an animal to migrate locally and these values are assigned based on expert knowledge about the species and the habitat. If the effective distances agree with the observed genetic dissimilarities, then the hypothesized grid explains the data well. Such a raster map could be hard to produce, especially at fine scales, and if the agreement is low, there is no

Electrical term	Ecological interpretation
conductance $c_{xy} : \forall (x, y) \in E$	direct migration $m_{\alpha\beta}$: the number of migrants exchanged between two neighboring demes α and β in a single generation. (On the coalescent timescale $m_\alpha = N_0 \dot{m}_{\alpha\beta}$ where $\dot{m}_{\alpha\beta}$ is the probability that a lineage in α has a parent from β.)
resistance $c_{xy} = 1/c_{xy}$	cost $1/m_{\alpha\beta}$: measure of local landscape friction in the direction from α to β. If migration is symmetric, $m_{\alpha\beta} = m_{\beta\alpha}$. (Since $N_\alpha = N_0$, the $m_{\alpha\beta}$s are comparable across the habitat.)
effective conductance $C_{xy} : \forall (x, y) \in V \times V$	effective migration $M_{\alpha\beta}$: the number of migrants that would produce the same level of genetic differentiation between α and β if these two demes made up a two-deme system.
effective resistance $R_{xy} = 1/C_{xy}$	distance metric R_{xy}: quantifies the genetic differentiation between a pair of demes (α, β) by taking into account the existence of multiple pathways between them.

Table 2.1: Circuit theory concepts and their ecological interpretation, adapted from [McRae et al., 2008]. McRae specifies the edge conductances as $c_{\alpha\beta} = m_{\alpha\beta}/q_\alpha$. However, it is natural to define conductances only in terms of the migration process because lineages cannot coalesce until they meet.

method to facilitate improving the map of resistances. However, IBR does provide an useful and efficient approximation to expected coalescence times.

To begin with, consider a stepping-stone model that has only two demes, α and β, with equal size and a single edge with migration rate $m_{\alpha\beta}$. In this population, also known as a two-island model,

$$m_{\alpha\beta} = \frac{(T_{\alpha\alpha} + T_{\beta\beta})/8}{T_{\alpha\beta} - (T_{\alpha\alpha} + T_{\beta\beta})/2}. \tag{2.18a}$$

[This follows from the system of linear equations (2.15).] The two-island model is a very special case and the equation (2.18a) does not hold more generally. In fact, unless the population graph is fully connected, many pairs of demes might not exchange migrants directly and then $m_{\alpha\beta} = 0$. However, [McRae, 2006] extends the relevance of the relationship (2.18a) to the general stepping-stone model by introducing the concept of effective migration $M_{\alpha\beta}$ between α and β. It is given by

$$M_{\alpha\beta} \equiv \frac{(T_{\alpha\alpha} + T_{\beta\beta})/8}{T_{\alpha\beta} - (T_{\alpha\alpha} + T_{\beta\beta})/2}. \tag{2.19}$$

That is, the effective migration $M_{\alpha\beta}$ is the number of migrants that would produce the actual genetic differentiation between α and β in a hypothetical two-island system. Since two lineages take the same time to reach their common ancestor, $T_{\alpha\beta} = T_{\beta\alpha}$ and the definition (2.19) implies that effective migration is always symmetric even though the underlying true migration patterns might not be symmetric.

It is natural to relate the concept of effective migration in a subdivided population, $M_{\alpha\beta}$, and the concept of effective conductance in an electrical circuit, $C_{\alpha\beta}$. In circuit theory, $C_{\alpha\beta}$ is the conductance in a two-node, single-conductor network required to produce the same amount of current between α and β as in the original network.

Proposition 2.2 *Consider a population graph (V, E) with symmetric migration rates $\{m_{\alpha\beta} : \forall (\alpha, \beta) \in E\}$. This corresponds to a circuit network (V, E) with conductances $\{c_{\alpha\beta} = m_{\alpha\beta}\}$. For every pair $(\alpha, \beta) \in V \times V$, the effective conductance $C_{\alpha\beta}$ in the circuit is a measure of the effective migration $M_{\alpha\beta}$ in the population:*

$$M_{\alpha\beta} \approx C_{\alpha\beta}. \tag{2.20}$$

And thus the resistance distance $R_{\alpha\beta} = 1/C_{\alpha\beta}$ is a measure of genetic differentiation.

Proof. The relationship between effective migration and effective conductance is exact only if migration is isotropic, i.e., demes are equivalent with respect to the size and pattern of movement. Here we assume only that migration is symmetric and conservative.

The migration process can be represented as a continuous-time discrete-space random walk on an undirected graph [Levin et al., 2008]. Then $M = (m_{\alpha\beta})$ is the transition kernel of the embedded jump chain, which determines the sequence of locations occupied by the lineage, and $m_\alpha = \sum_{\beta:\beta\neq\alpha} m_{\alpha\beta}$ are the rates of the holding distributions, which determine the waiting times before jumps. Let $m = (1/d)\sum_\alpha m_\alpha$ be the average holding rate.

Since migration is symmetric and conservative, the demes have the same size N_0, which is also a convenient choice for the coalescent timescale. Let T_0 be the average within-deme expected coalescence time. Then by Strobeck's theorem [see equation (7.8) in Appendix 7.1],

$$T_0 \equiv \sum_\alpha T_{\alpha\alpha}/d = d. \tag{2.21}$$

Thus T_0 does not depend on the migration process.

Furthermore, let $\tau_{\alpha\beta}$ be the expected time for two lineages, one from α and the other from β, to occupy the same deme. Then

$$(T_{\alpha\alpha} + T_{\beta\beta})/2 \approx T_0, \tag{2.22a}$$

$$T_{\alpha\beta} - (T_{\alpha\alpha} + T_{\beta\beta})/2 \approx \tau_{\alpha\beta}. \tag{2.22b}$$

These two approximations are exact if migration is isotropic: since the demes are equivalent with respect to the migration process, the within-deme coalescence times $T_{\alpha\alpha}$ must be equal by symmetry. Hence, $T_{\alpha\alpha} = T_0$, $T_{\alpha\beta} = \tau_{\alpha\beta} + T_0$ and once the lineages meet for the first time, we can restart the random walk with two lineages in the same deme chosen at random.

Under the coalescent process, two lineages — one from α and another from β — move simultaneously until they coalesce into a common ancestor. Suppose that they meet for the first time in deme γ. Together the paths $\alpha \to \gamma$ and $\beta \to \gamma$ have half the length of a commute between α and β that passes through γ. Therefore, the expected time to first meet, $\tau_{\alpha\beta}$, can be related to the expected commute length, $K_{\alpha\beta}$, in the corresponding circuit network:

$$\tau_{\alpha\beta} \approx K_{\alpha\beta}/(4m), \tag{2.23}$$

where $K_{\alpha\beta}$ is the expected number of jumps in a random walk that starts at α, visits β and returns to α, and $1/(2m)$ is the average waiting time before either lineage jumps. The relationship is approximate because the waiting time varies across vertices.

Finally, by [Chandra et al., 1996] for a undirected graph [whether isotropic or not],

$$K_{\alpha\beta} = c_G R_{\alpha\beta} = c_G / C_{\alpha\beta}, \tag{2.24}$$

where $R_{\alpha\beta}$ is the effective resistance between nodes α and β, $C_{\alpha\beta}$ is the effective conductance, and c_G is the total conductance of the network given by

$$c_G = \sum_\alpha \sum_{\beta:\beta\neq\alpha} m_{\alpha\beta} = \sum_\alpha m_\alpha = dm. \tag{2.25}$$

Therefore,

$$M_{\alpha\beta} = \frac{(T_{\alpha\alpha} + T_{\beta\beta})/8}{T_{\alpha\beta} - (T_{\alpha\alpha} + T_{\beta\beta})/2} \approx \frac{T_0/4}{\tau_{\alpha\beta}} \approx \frac{(d/4)}{R_{\alpha\beta}(dm)/(4m)} = C_{\alpha\beta} \tag{2.26}$$

□

Essentially, McRae's approximation splits the between-deme coalescence time, $T_{\alpha\beta}$, into the time to first meet, $\tau_{\alpha\beta}$, and the average within-deme coalescence time, T_0. However, since the population graph is not necessarily symmetric, not every deme γ is equally likely to be the deme where two lineages, starting from α and β, meet for the first time. And furthermore, the within-deme coalescence times are not necessarily equal. Therefore, the effective resistance metric reflects the migration process accurately but ignores the fact that the lineages do not necessarily coalesce on their first opportunity. On the other hand, the coalescence time metric correctly captures the effect of both processes because Kingman's coalescent models migration and coalescence by explicitly tracking both lineages until their common ancestor. Since higher rates imply faster mixing, we can conclude that the higher migration rates are, the better McRae's approximation is. See Figures 2.2 and 2.3.

2.4.1 *Effective resistance approximates expected coalescence time*

McRae's method approximates the ancestral process of two lineages evolving simultaneously in terms of one lineage evolving at twice the rate. However, one random walk cannot represent a coalescence event where two lineages merge into their most recent common ancestor. Thus, while effective resistance, $R_{\alpha\beta}$, provides a measure for the genetic differentiation between demes, it does not capture the genetic differentiation between individuals from the same deme [$R_{\alpha\alpha} = 0$ for every deme α]. However, it follows directly from McRea's approximation that

$$T_{\alpha\beta} \approx \tau_{\alpha\beta} + T_0 \approx T_0(R_{\alpha\beta}/4 + 1), \tag{2.27}$$

or equivalently in matrix notation,

$$T \approx T_0(R/4 + 11'). \tag{2.28}$$

The main advantage of approximating coalescence times in terms of effective resistances is computational efficiency. To compute T, we solve a linear system of equations $Ab = x$ with $d(d+1)/2$ unknowns that corresponds to eq. (2.16). In this problem A is sparse (because the population graph G is sparse) and positive definite, and so we can use an iterative preconditioned gradient method. There are several methods to compute R; we use a method that inverts the $d \times d$ matrix $M + 11'$ [Babić et al., 2002]. Since A is of higher order than M, it is more efficient to compute R. Furthermore, R gives a very good approximation to T when migration rates are high and it is more appropriate than other distance metrics such Euclidean distance and least-cost path. Therefore, effective resistance offers a compromise between accuracy of representation and efficiency of computation.

In this chapter we introduced two important components of our method for analyzing spatial population structure: the stepping-stone model and the effective resistance metric. In the next chapters we describe how we can estimate and visualize effective rates of migration from geographically referenced genetic data.

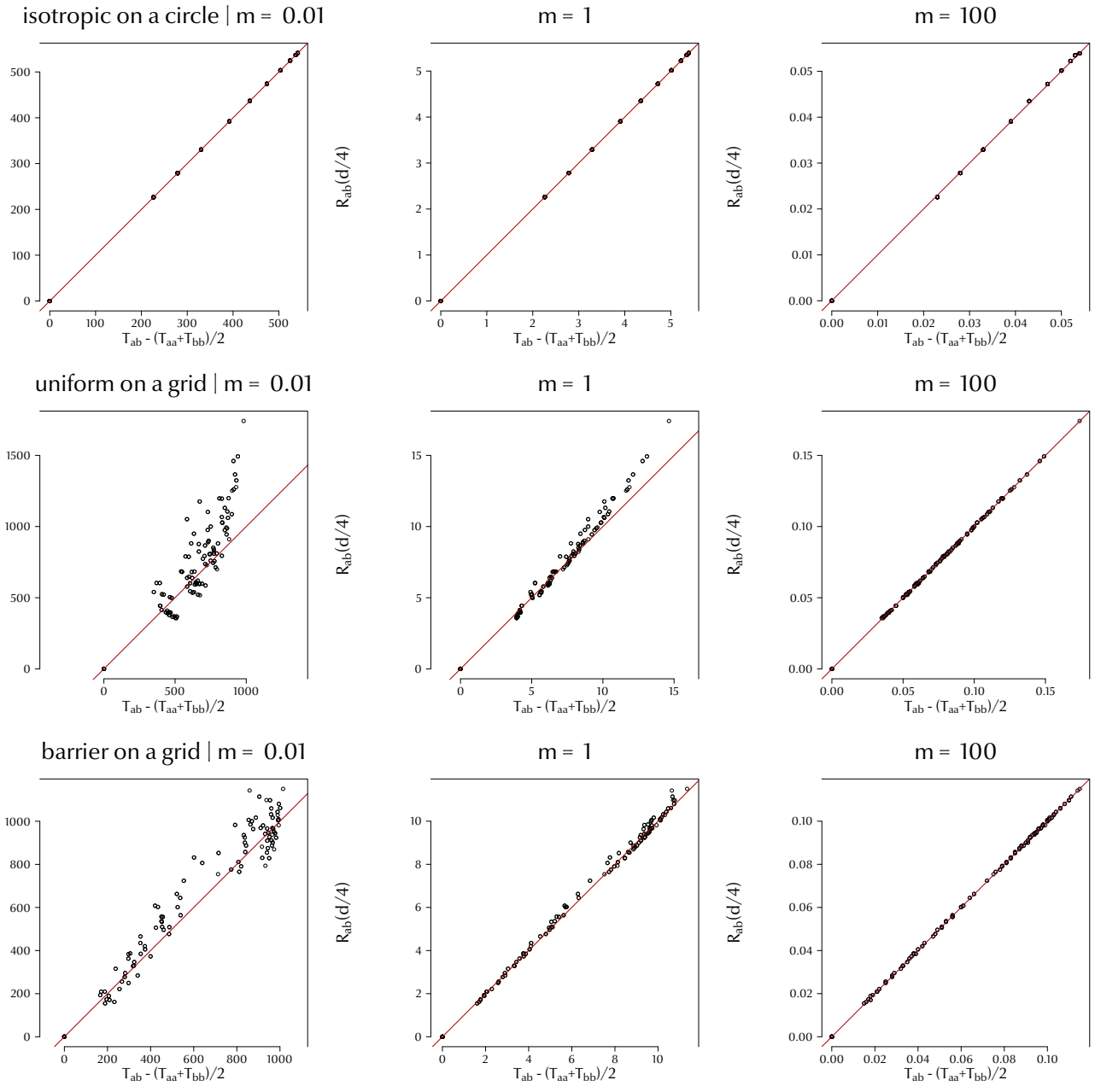


Figure 2.2: On the x-axis, $T_{\alpha\beta} - (T_{\alpha\alpha} + T_{\beta\beta})/2$ is the expected time to reach the same deme; on the y-axis, $R_{\alpha\beta}(d/4)$ is the (appropriately scaled) effective resistance. As the migration rate increases, $R_{\alpha\beta}$ becomes a better approximation of the expected time to first meet, $\tau_{\alpha\beta}$, even if migration is not isotropic. [Results for a 5×4 regular triangular grid with uniform migration rate $m = 0.01, 1$ or 10.]

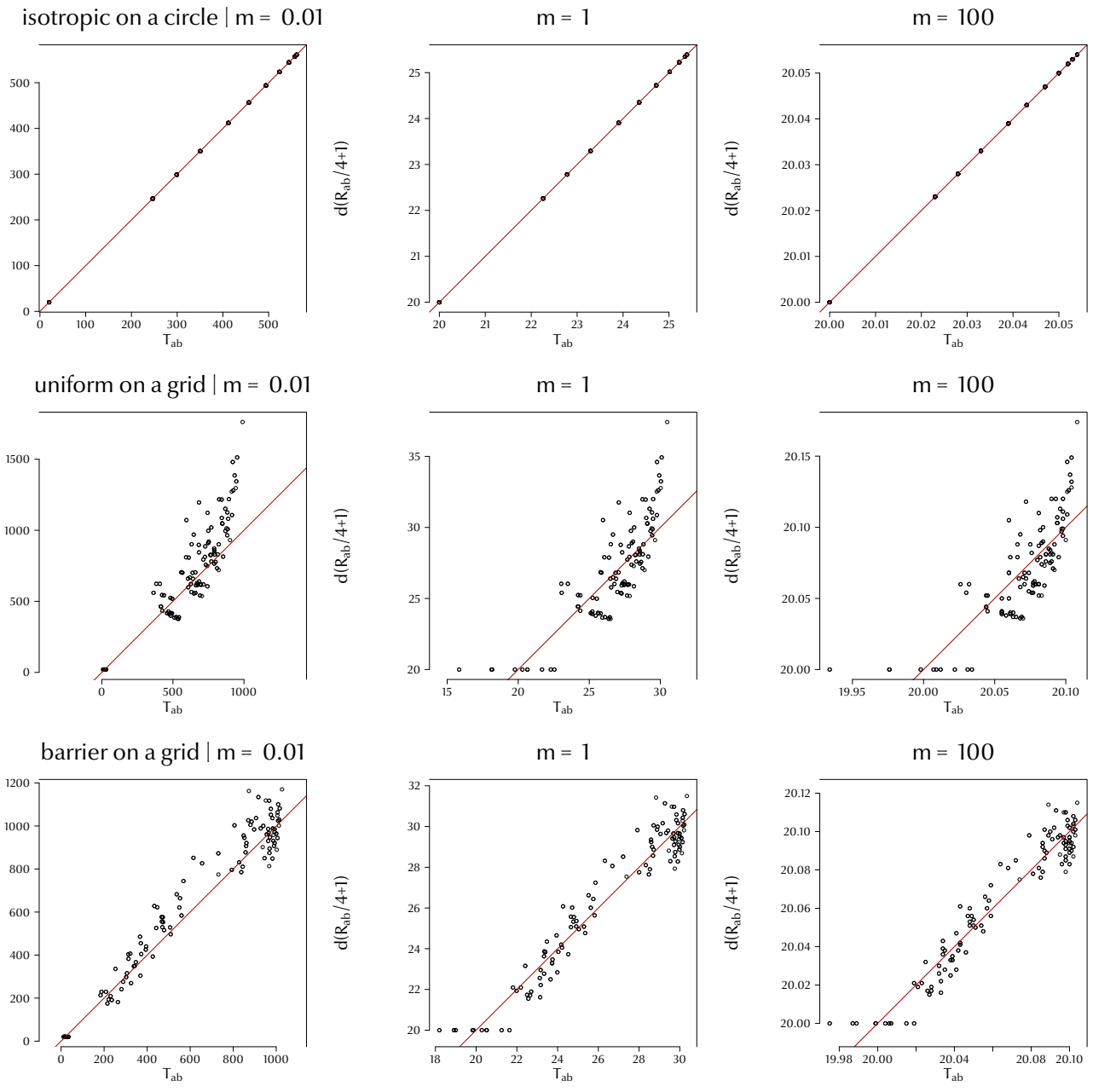


Figure 2.3: On the x-axis, $T_{\alpha\beta}$ is the expected time to coalescence; on the y-axis, $d(R_{\alpha\beta}/4 + 1)$ is the IBR approximation. The approximation to the within-deme coalescence times, $T_{\alpha\alpha}$, is always $T_0 = d$; there are the points closest to the origin at $T_0 = 20$ in a 5×4 grid. Although the pattern does not change as the migration rate increases, the relative error $\Delta T_{\alpha\beta}/T_{\alpha\beta}$ decreases.

3

Genetic Dissimilarities and Distance Matrices

Habitat heterogeneity can shape genetic variation by reducing or increasing gene flow. The stepping-stone model is a natural representation of a spatially distributed population and the effects of gene flow on its genetic structure. In this thesis a population is a graph $G = (V, E, M)$ comprised of vertices V [randomly mating demes of equal size], edges E [symmetric routes of migration between neighboring demes] and a weight function $M : E \to \mathbb{R}_+$ that specifies the rates at which migrants are exchanged.

Throughout, we will assume that the population graph G is embedded in a two-dimensional habitat, with the vertex set V and the edge set E both fixed. In practice, this graph is not known and does not necessarily exist. For example, it might not be possible to split the population into distinct groups that satisfy the random mating assumption. Instead, we cover the habitat with a regular triangular grid in which vertices do not represent actual colonies. This simplification indicates that we should interpret the migration parameters carefully — as *effective* rather than *actual* rates of migration.

Thus the topology of the graph is determined by the shape of the habitat [and the somewhat arbitrary choice that the graph is triangular and regularly spaced] and not the sample configuration or the sample "clusteredness". And so we construct the graph differently from methods that aim to subdivide the population into clusters that are similar within and dissimilar between. However, if we make the grid (V, E) sufficiently fine, we can reasonably assume that each vertex represents a randomly mating group without further structure. In this case, individuals would be similar within demes but not necessarily dissimilar between demes.

In a habitat with uniform migration, the genetic differentiation between individuals from the same species is positively correlated with the distance between their origin; in a heterogeneous habitat, landscape features such as barriers or corridors create spatial structure in genetic variation. For example, individuals separated by a barrier are less closely related, and therefore less genetically similar, than if the barrier were absent. The stepping-stone model can represent such effects because some edges in the population graph can have high migration rates and others — low. In this thesis we develop a Bayesian procedure to estimate the effective migration rates in a fixed grid (V, E) of equally sized demes, from geographically indexed genetic data. The function M measures the relative rate at which two connected demes exchange migrants; we call M a *migration surface*.

To analyze population structure, we will assume that all genotyped sites develop under the same evolutionary process which determines the expected structure in genetic correlations (or equivalently, genetic distances). In contrast, many methods for association testing assume that individuals are independent while sites are correlated. (In population genetics, the systematic association between loci is called *linkage disequilib-*

rium.) The problem at hand determines which assumption is appropriate to make. To find associations between disease status and genetic makeup, it is reasonable to assume that the disease develops under the same mechanism in all sampled cases but not all sites contribute to the disease and not with equal effect. To analyze population structure, it is reasonable to assume that the same evolution process underlies all genotyped sites but not all sampled individuals are genetically similar to equal degree.

The PCA decomposition of the observed covariance matrix XX' can be used to correct for population stratification [Price et al., 2006] incorporate the leading eigenvectors in a regression analysis that tests for association between sites and disease.

In this chapter, let $Z = (z_i : i = 1, \dots, n)$ be a vector of n genotypes at a single polymorphic site. We will consider multiple sites in the next chapter. Also let $\underline{\alpha} = (\alpha_1, \dots, \alpha_n)$ denote the sample configuration, in which α_i is the sampling location of the ith haplotype.

3.1 Mean and covariance of genotype vectors: SNPs

First we consider the simplest case — a haploid population from which we have a sample of n individuals genotyped at a single nucleotide polymorphism (SNP). Following [McVean, 2009], we make the following assumptions:

A1. SNPs are identically distributed: Since all sites evolve under the same demographic model, the observed genotype z_i at any SNP is a realization of the same random variable Z_i.

A2. SNPs segregate in the sample: Since exactly one mutation occurs in every sampled genealogy, we observe both the ancestral allele '0' and the derived allele '1' at every site.

A3. The scaled mutation rate θ is low: Since A2 and A3 together imply θ is a nuisance parameter [Nielsen, 2000], we can take the limit $\theta = 2N_0u \to 0$ and thus ignore small differences in mutation rate across SNPs.

Under these assumptions, the probability that individuals i and j share the derived mutation at a randomly chosen segregating site is given by

$$\mathrm{E}^*\{Z_iZ_j\} = \frac{T_{mrca} - T_{ij}}{T_{tot}}, \tag{3.1}$$

where T_{mrca} and T_{tot} are the height and the size of the expected genealogy of the sample, and T_{ij} is the expected time for i and j to coalesce in a sample of size 2 [McVean, 2009]. The symbol $*$ indicates the condition that both 0s and 1s are observed, i.e., the expectation on the left in equation (3.1) is with respect to all possible genealogies (observed or not) with exactly one mutation. The expectations on the right in equation (3.1) are unconditional. The relevance is that for Kimura's stepping-stone model there is an explicit formula for pairwise coalescence times, T_{ij}, and a good approximation in terms of effective resistances, R_{ij}.

Furthermore, since the Z_is are binary random variables and the time to coalescence with self is always 0,

$$\mathrm{E}^*\{Z_i\} = \mathrm{E}^*\{Z_i^2\} = \frac{T_{mrca}}{T_{tot}}. \tag{3.2}$$

Therefore, the expected genealogy fully specifies the first two moments of the allele count vector $Z = (Z_i)$ at a particular segregating SNP. In matrix notation,

$$\mathrm{E}^*\{Z\} = \frac{T_{mrca}}{T_{tot}}1 \equiv \mu 1, \tag{3.3a}$$

$$\mathrm{var}^*\{Z\} = \frac{T_{mrca}}{T_{tot}}\left(1 - \frac{T_{mrca}}{T_{tot}}\right) - \frac{1}{T_{tot}}\underline{T} \equiv \sigma^2(11' - \lambda\underline{T}). \tag{3.3b}$$

For sample with configuration $\underline{\alpha}$ from a population with model G, the parameters are given by

$$\mu = \frac{T_{mrca}}{T_{tot}}, \qquad \sigma^2 = \frac{T_{mrca}}{T_{tot}}\left(1 - \frac{T_{mrca}}{T_{tot}}\right), \qquad \lambda\sigma^2 = \frac{1}{T_{tot}}, \tag{3.4}$$

where $\underline{T} = (T_{ij})$ is the matrix of expected pairwise coalescence times between sampled individuals. That is, T_{ij} is the expected time to coalescence between $i \in \alpha_i$ and $j \in \alpha_j$ in a sample of size 2, regardless of the composition of the entire sample $\underline{\alpha}$. Since T_{ij} does not depend on the sample configuration or even the sample size n, it is completely determined by the population model G. However,

- The expected height and size of the sample genealogy, T_{mrca} and T_{tot}, depend on both the population model G and the sample configuration $\underline{\alpha}$. In particular, they are strongly influenced by uneven sampling. Therefore, T_{mrca}/T_{tot} and $1/T_{tot}$ are nuisance parameters because it would be very hard to decouple the effects of population structure from the effects of uneven sampling. The confounding of population and sample-specific information also makes it difficult to interpret PCA projections in terms of a (historic) demographic process [Novembre et al., 2008, McVean, 2009].

- The matrix $\underline{T} = (T_{i,j} : \text{individuals } i, j)$ describes the expected genetic differentiation in the sample and has a block structure which depends on how many individuals, if any, we observe from each deme. On the other hand, $T = (T_{\alpha\beta} : \text{demes } \alpha, \beta)$ specifies how genetic variation increases with geographic distance for all pairs of demes, whether they are sampled from or not. Thus T is a dissimilarity matrix that characterizes the entire population. Although $\underline{T}$ is a function of the sample configuration, it depends on $\underline{\alpha}$ in a straightforward way:

  $$\underline{T} = JTJ' - \text{diag}\{JTJ'\}, \tag{3.5}$$

  where $J \equiv J(\underline{\alpha}) = (J_{i\alpha}) \in \mathbb{Z}^{n\times d}$ is an indicator matrix such that $J_{i\alpha} = 1$ if $i \in \alpha$ and 0 otherwise. And we remove the diagonal because the coalescence time with self is always 0.

The demographic model G, which describes the population, determines the coalescent process and hence the expected pairwise coalescence times $T_{\alpha\beta}$ for all deme pairs (α, β). On the other hand, both the model G and the configuration $\underline{\alpha}$ determine the genealogical statistics T_{mrca} and T_{tot} which are generally not of interest as the goal is to estimate population-level features of G — such as the migration rates between pairs of connected demes — while accounting for the sample specific features of $\underline{\alpha}$. In this thesis $G = (V, E, M)$ is always a population graph (V, E, M) with equally sized demes V, undirected edges E and effective migration rates $M : E \to \mathbb{R}^+$.

We have shown that the expected mean and variance of a genotype vector are computable functions of the effective migration rates M. Next we derive similar expressions for the mean and the variance as functions of expected coalescence times in the case of diploid SNPs and microsatellites.

3.1.1 The case of diploid data

Since a diploid individual is the offspring of a pair of diploid parents, we can represent the genotype of a diploid as the sum of two haploids, each drawn randomly from the same location, i.e., $X_i = Z_i^{(1)} + Z_i^{(2)} \in \{0, 1, 2\}$ where the superscript indicates one of two haplotypes. However, since we do not distinguish between the haplotype inherited

from the mother and the haplotype inherited from the father, this assumption is reasonable only for autosomal SNPs (and not for sex-linked ones) in outbred individuals.

A sample $X_1, \dots, X_n$ of n diploid individuals is polymorphic if

$$\begin{aligned}&\left\{X_1, \dots, X_n : \text{at least one } X_i \geq 1\right\}\\ &\Leftrightarrow \left\{Z_1^{(1)}, Z_1^{(2)}, \dots, Z_n^{(1)}, Z_n^{(2)} : Z_i^{(1)} = 1 \text{ or } Z_i^{(2)} = 1\right\}.\end{aligned} \tag{3.6}$$

That is, a segregating SNP in a diploid sample of size n is equivalent to exactly one mutation in a haploid sample of size $2n$. [This excludes the possibility that all individuals carry the same allele, either ancestral or derived.]

Furthermore, at a segregating site in a diploid sample, the copies $Z_i^{(1)}$ and $Z_i^{(2)}$, which constitute X_i, are not independent — the event that one carries the mutation but not the other is informative for the time to their most common ancestor. Therefore,

$$\mathrm{E}^*\{X_i\} = \mathrm{E}^*\left\{Z_i^{(1)}\right\} + \mathrm{E}^*\left\{Z_i^{(2)}\right\} = 2\mathrm{E}^*\{Z_i\} = 2\mu \tag{3.7a}$$

$$\mathrm{var}^*\{X_i\} = 2\mathrm{var}^*\{Z_i\} + 2\mathrm{cov}^*\left\{Z_i^{(1)}, Z_i^{(2)}\right\} = 4\sigma^2 - 2\lambda\sigma^2 T_{ii} \tag{3.7b}$$

$$\mathrm{cov}^*\{X_i, X_j\} = 4\mathrm{cov}^*\{Z_i, Z_j\} = 4\sigma^2 - 4\lambda\sigma^2 T_{ij} \tag{3.7c}$$

where the symbol $*$ indicates the condition that there is exactly one mutation in a sample of $2n$ haplotypes [and T_{ii} is the expected coalescence time for two distinct lineages with the same origin as individual i]. In matrix notation,

$$\mathrm{E}^*\{X\} = 2\mu 1, \qquad \mathrm{var}^*\{X\} = 4\sigma^2(11' - \lambda \underline{T}_2), \tag{3.8}$$

where

$$\underline{T}_2 = JTJ' - \frac{1}{2}\,\mathrm{diag}\,\{JTJ'\}. \tag{3.9}$$

The subscript 2 indicates that the matrix of pairwise coalescence times corresponds to a diploid population. Here the mean does not depend on the location. (This is the case for haploid data as well.) However, the variance $\mathrm{var}^*\{X_i\}$ can vary with location unless the demographic model implies $T_{\alpha\alpha} = T_0$ for all demes α, i.e., isotropic migration.

3.2 Mean and covariance of genotype vectors: microsatellites

Microsatellites (also called short tandem repeats) are repeating sequences of a particular short DNA segment. Mutation can increase or decrease the number of repeats k, and each k corresponds to an allele.

To model microsatellites, we assume that a locus s evolves from its ancestral allele A_s according to a symmetric stepwise mechanism where mutations occur with rate θ_s and each mutation increases or decreases the number of repeats by exactly one, with equal probability. Here we consider the evolution at a particular site, and for simplicity of notation, we omit the subscript s in the rest of this section.

The ancestral allele A and the mutation rate θ are unknown site-specific parameters while the genealogy $\mathcal{T}$ has a distribution determined by Kingman's coalescent. As we did for SNPs, we assume that the microsatellites are neutral and hence their genealogies are identically distributed. On the other hand, microsatellites are usually highly variable markers (i.e., with high mutation rates), so we cannot take the low-mutation limit.

Conditional on the mutation rate θ and the genealogical tree $\mathcal{T}$ of the sample, mutations occur independently and the number of mutations on a branch with length t is

a Poisson random variable with mean θt. This follows from the assumption that mutations are generated by a Poisson process with intensity [mutation rate] θ. For example, the total number of mutations is

$$K_{tot} \,|\, \theta, \mathcal{T} \sim \mathrm{Po}(\theta t_{tot}), \tag{3.10}$$

while the number of mutations carried by individual i is

$$K_i \,|\, \theta, \mathcal{T} \sim \mathrm{Po}(\theta t_{mrca}). \tag{3.11}$$

All lineages share the same Poisson mean parameter because every branch from a lineage to the most common ancestor of the entire sample has length t_{mrca}.

Let $\mathcal{K}$ denote the set of all mutations that occur in the genealogy, with $|\mathcal{K}| = K_{tot}$. Also, let $\mathcal{K}_i \subset \mathcal{K}$ denote the set of mutations carried by individual i, with $|\mathcal{K}_i| = K_i$. Since each mutation is equally likely to decrease or increase the allele length by 1, the ith allele is

$$Z_i = A + \sum_{k \in \mathcal{K}_i} S_k, \tag{3.12}$$

where $S_k = \pm 1$ with probability $1/2$ and thus $\mathrm{E}\{S_k\} = 0$ and $\mathrm{var}\{S_k\} = \mathrm{E}\{S_k^2\} = 1$.

First we derive the mean and variance of allele Z_i given the mutation rate, the ancestral allele and the genealogy. The binary variables, S_k, are independent of the sample history, so $\mathrm{E}\{S_k \,|\, \theta, A, \mathcal{T}\} = \mathrm{E}\{S_k\}$ and $\mathrm{var}\{S_k \,|\, \theta, A, \mathcal{T}\} = \mathrm{var}\{S_k\}$. And furthermore, conditional on the number of mutations, the S_ks are mutually independent. Therefore,

$$\mathrm{E}\{Z_i \,|\, \theta, A, \mathcal{T}\} = A + \mathrm{E}\Big\{\mathrm{E}\Big\{ \sum_{k \in \mathcal{K}_i} S_k \,|\, K_i \Big\}\Big\} = A + \mathrm{E}\Big\{ \sum_{k=1}^{K_i} \mathrm{E}\{S_k\} \Big\} = A, \tag{3.13a}$$

Since the mutations are independent, $\mathrm{E}\{S_k S_{k'}\} = \mathrm{E}\{S_k\}\mathrm{E}\{S_{k'}\} = 0$ for $k \neq k'$.

$$\mathrm{var}\{Z_i \,|\, \theta, A, \mathcal{T}\} = \mathrm{E}\Big\{ \sum_{k=1}^{K_i} \mathrm{E}\{S_k^2\} \Big\} + \mathrm{E}\Big\{ \sum_{k \neq k'} \mathrm{E}\{S_k S_{k'}\} \Big\} = \mathrm{E}\{K_i\} = \theta t_{mrca}, \tag{3.13b}$$

because K_i is a Poisson random variable with mean θt_{mrca} by equation (3.11).

Let $\mathcal{K}_{i \oplus j}$ be the set of mutations that occur in one lineage but not the other, with $|\mathcal{K}_{i \oplus j}| = K_{i \oplus j}$. Such mutations occur on the branch from i to $mrca(i,j)$ or on the branch from j to $mrca(i,j)$. Therefore, $K_{i \oplus j}$ has mean $2\theta t_{ij}$. Similarly, let $\mathcal{K}_{i \setminus j}$ be the set of mutations carried by i but not j.

Again, the cross terms are 0 by mutual independence.

$$\mathrm{E}\Big\{(Z_i - Z_j)^2 \,|\, \theta, A, \mathcal{T}\Big\} = \mathrm{E}\Big\{\Big(\sum_{k \in \mathcal{K}_{i \setminus j}} S_k - \sum_{k \in \mathcal{K}_{j \setminus i}} S_k \Big)^2\Big\} = \mathrm{E}\Big\{ \sum_{k=1}^{K_{i \oplus j}} \mathrm{E}\{S_k^2\} \Big\} = \mathrm{E}\{K_{ij}\} = 2\theta t_{ij}, \tag{3.14a}$$

$$\mathrm{cov}\{Z_i, Z_j \,|\, \theta, A, \mathcal{T}\} = \mathrm{var}\Big\{Z_i \,|\, \theta, A, \mathcal{T}\Big\} - \frac{1}{2}\mathrm{E}\Big\{(Z_i - Z_j)^2 \,|\, \theta, A, \mathcal{T}\Big\} = \theta t_{mrca} - \theta t_{ij}. \tag{3.14b}$$

Now we have expressions for the mean, variance and covariance of the genotypes at a particular microsatellite, given the site-specific mutation rate θ, ancestral allele A and genealogy $\mathcal{T}$. We treat θ and A as nuisance parameters to be estimated and we marginalize the genealogy out. The goal is to express the model in terms of the expected coalescence times rather than the coalescence times at a particular site. We took the same approach for SNP data but in the former case, $A = 0$ for every segregating site and θ is eliminated in the small mutation limit $\theta \to 0$. Finally,

$\mathrm{E}\{X\} = \mathrm{E}\big\{\mathrm{E}\{X \,|\, Y\}\big\}$

$$\mathrm{E}\{Z_i \,|\, \theta, A\} = \mathrm{E}\{A \,|\, \theta, A\} = A, \tag{3.15a}$$

$\mathrm{var}\{X\} = \mathrm{E}\big\{\mathrm{var}\{X \,|\, Y\}\big\} + \mathrm{var}\big\{\mathrm{E}\{X \,|\, Y\}\big\}$

$$\mathrm{var}\{Z_i \,|\, \theta, A\} = \mathrm{E}\{\theta t_{mrca} \,|\, \theta, A\} + \mathrm{var}\{A \,|\, \theta, A\} = \theta T_{mrca}, \tag{3.15b}$$

$\mathrm{cov}\{X, Z\} = \mathrm{E}\big\{\mathrm{cov}\{X, Z \,|\, Y\}\big\} + \mathrm{cov}\big\{\mathrm{E}\{X \,|\, Y\}, \mathrm{E}\{Z \,|\, Y\}\big\}$

$$\mathrm{cov}\{Z_i, Z_j \,|\, \theta, A\} = \mathrm{E}\{\theta t_{mrca} - \theta t_{ij} \,|\, \theta, A\} + \mathrm{var}\{A \,|\, \theta, A\} = \theta(T_{mrca} - T_{ij}). \tag{3.15c}$$

In the case of microsatellites, we do not condition on observing variability in the sample, i.e., on the event $\{K_{tot} > 0\}$ as microsatellites have higher mutation rates and we can estimate the parameter rather than take its limit to 0. For SNPs such that we observe exactly one mutation at every site, the "variability" condition is explicitly modeled because it modifies the genealogy distribution. Intuitively, it "stretches" the tree and thus changes (proportionally) all branches $t \in \mathcal{T}$.

Therefore, the genotype vector of n sampled individuals at a particular microsatellite has mean and variance

$$\mathrm{E}^*\{Z\} = \mu 1, \qquad \mathrm{var}^*\{Z\} = \sigma^2(11' - \lambda \underline{T}) \tag{3.16}$$

where the symbol $*$ indicates conditioning on the ancestral allele A and the mutation rate θ, and the parameters are given by

$$\mu = A, \qquad \sigma^2 = \theta T_{mrca}, \qquad \lambda = \frac{1}{T_{mrca}}. \tag{3.17}$$

As for SNP data, the mean and the variance of genotypes at a particular locus do not depend on the origin of an individual. However, for microsatellite data, the mean and the variance vary across sites because the ancestral allele A and the mutation rate θ are both site-specific parameters. On the other hand, the scale λ is shared across sites and therefore every site has the same correlation matrix $\Sigma \equiv 11' - \lambda \underline{T}$.

With this parametrization, the demographic parameters are estimable up to a proportionality constant. If we multiply the migration and coalescence rates by 2, we speed up the structured coalescent process by a factor of 2, and hence, we decrease the expected coalescence times by 2. However, the covariance matrix Σ remains unchanged because the dissimilarity matrix $\underline{T}$ is appropriately scaled.

3.3 *Effective migration can explain spatial structure in genetic variation*

In the previous section, we discussed how to specify the mapping from the stepping-stone model $G = (V, E, M)$ to the genetic covariance matrix $\mathrm{cor}\{Z\} = \Sigma$, for both SNP and microsatellite data. Briefly, we followed three steps. First, $G = (V, E, M)$ determines $T = (T_{\alpha\beta})$ through the system of linear equations (2.15). Then, in turn, the expected coalescence times between demes, T, determine the expected coalescence times between sampled individuals, $\underline{T}$, through equation (3.5). Finally, the distance matrix determines the correlation matrix $\Sigma = 11' - \lambda \underline{T}$ by equation (3.3b) where λ is an appropriately chosen scalar parameter that guarantees Σ is positive definite.

Our goal is to estimate the effective migration rates M across the habitat; these are sample-independent (population-level) features of the population graph G. The mean μ and the variance σ^2 of derived alleles as well as the scale factor λ of expected coalescence times can be treated as nuisance parameters because they are sample-dependent and shared by all individuals in the sample. For example, for haploid SNPs the overall mean is $\mu = T_{mrca}/T_{tot}$ [with $\sigma^2 = \mu(1-\mu)$] and the scale factor is $\lambda = 1/T_{tot}$, so (μ, σ^2, λ) contain some information about G. Although the scalars T_{tot} and T_{mrca} are, formally, functions of the effective migration rates M they are very difficult to compute.

On the other hand, the matrix $\underline{T} = (T_{ij})$ of pairwise coalescence times is a computable function of M. This matrix is also a pairwise dissimilarity (distance) matrix [and formally, a semivariogram]: the more genetically dissimilar two individuals are, the longer the time to their most recent common ancestor because the probability that the branch T_{ij} accumulates a mutation is proportional to its relative length in the average genealogy tree. The property that $\underline{T}$ is a distance matrix is important because it can

explain genetic dissimilarities (correlations) as a linear function of distances between locations. Expected coalescence time is a particular choice of distance metric motivated by coalescent theory [McVean, 2009]. We can consider other metrics such as effective resistance [McRae, 2006].

$W \in \mathbb{S}^d$ is a symmetric matrix of weights.

$\Delta \in \mathbb{D}^d$ is the population distance matrix.

$\underline{\Delta} \in \mathbb{D}^n$ is the sample distance matrix.

$\Sigma \in \mathbb{V}^n$ is the sample covariance matrix.

$$W \equiv \left\{ \begin{array}{lcll} M & = & \{ \text{ migration rates } & m(e) \ \} \\ C & = & \{ \text{ conductances } & c(e) \ \} \end{array} : \forall e \in E \right\}$$

$$\xrightarrow{(1)} \quad \Delta \equiv \left\{ \begin{array}{lcll} T & = & \{ \text{ coalescence times } & T_{\alpha\beta} \ \} \\ R & = & \{ \text{ effective resistances } & R_{\alpha\beta} \ \} \end{array} : \forall (\alpha, \beta) \in V \times V \right\}$$

$$\xrightarrow{(2)} \quad \underline{\Delta} \equiv \left\{ \begin{array}{lcll} \underline{T} & = & \{ \text{ coalescence times } & T_{ij} \ \} \\ \underline{R} & = & \{ \text{ effective resistances } & R_{ij} \ \} \end{array} : \forall (i, j) \in \underline{\alpha} \right\}$$

$$\xrightarrow{(3)} \quad \Sigma \equiv 11' - \lambda \underline{\Delta}$$

The first step, denoted by $\xrightarrow{(1)}$, is to compute all $d(d+1)/2$ pairwise distances between d demes. This operation is expensive even for medium-size grids. However, the covariance matrix Σ is a function of the sample distance matrix $\underline{\Delta}$, not the population distance matrix Δ. That is, in principle, we could avoid computing the full $d \times d$ dissimilarity matrix, especially for sparsely sampled habitats. [This is the advantage of $\underline{R}$ over $\underline{T}$.]

In a certain sense, T is an "appropriate" dissimilarity measure for population structure as genetically similar individuals are likely to have a recent common ancestor and thus shorter coalescence time. For the stepping-stone model we can obtain the matrix of pairwise coalescence times T exactly or approximate it with the matrix of effective resistances, R. However, the stepping-stone model itself does not represent the true history of the population — the grid is placed arbitrarily and there are underlying assumptions, including equilibrium in time, low mutation rate and no selection. Therefore, in a manner similar to McRae's definition of the effective migration rate, $m_{\alpha\beta}$, for a pair of demes, we should interpret the migration rate function $M = \left\{ m_{\alpha\beta} : (\alpha, \beta) \in V \times V \right\}$ as *effective migration surface* because it would produce the observed patterns of genetic differentiation if the population were evolving under the stepping-stone model.

3.4 *Related methods for analyzing population structure*

We have shown that genetic correlations can be modeled in terms of a distance matrix. This representation is motivated by the relationship between genetic similarities and expected coalescence times. However, we can consider other distance metrics (on the population graph) as long as they capture relevant features of a spatially heterogeneous habitat, and effective resistance is particularly useful because it approximates the coalescent-based metric and is efficient to compute.

Here we discuss briefly two related methods for analyzing spatially distributed populations.

3.4.1 *MIGRATE*

[Beerli and Felsenstein, 2001] develop an approach to estimate migration rates among demes, and more generally, to compare and rank structured population models. Their

method MIGRATE is also based on the structured coalescent but it makes different assumptions about the spatial distribution and the migration pattern.

In MIGRATE the demes are sampling locations and all demes potentially exchange migrants, so the population graph is constructed without explicit geographic information. [Some edges can be excluded to test and compare various migration patterns.] Every deme in the resulting graph has a size parameter and every edge has two migration parameters. [MIGRATE allows asymmetric gene flow.] Thus for a graph with d demes, the most complex model to test has $d(d-1)$ migration rates and d deme sizes.

In contrast, our method uses a regular triangular grid constructed independently of the sampling configuration [or an *a priori* grouping of individuals into subpopulations]. Migration is symmetric and constrained to occur only between neighboring demes but not all demes need to be sampled. And edges are grouped via a Voronoi tessellation of the habitat to encourage parameter sharing and locally constant migration. This representation is flexible and the number of (unique) migration rates varies with the number of tiles.

A Voronoi tessellation of a Euclidean space is a partition into T convex polygons (tiles) generated by T distinct points (centers). The region associated with the tth center u is the set of points closer to u than any other center. Boundary points are equidistant to two centers. [Okabe et al., 2000].

3.4.2 GENELAND

[Guillot et al., 2005] also uses Voronoi tiling to model the spatial structure in genetic variation but their method GENELAND is cluster-based and thus best suited to analyze discrete structure. Since individuals sampled from geographically close locations are more likely to come from the same subpopulation, GENELAND attempts to find clusters that are both genetically and geographically coherent. Compared with a spatial representation in terms of a population graph, such clusters can correspond to single demes in the graph (e.g., if migration is low and even demes close in space are clearly differentiated); or they can correspond to groups of demes where allele frequency distributions are indistinguishable (e.g., if gene flow is high so that a mutation that arises in one deme can quickly "spread" to nearby locations).

4

Estimating Effective Rates of Migration

In this chapter we introduce a likelihood function and prior distributions to perform Bayesian inference for the effective migration surface M based on the similarities observed in georeferenced genetic data. The posterior estimate of M can represent graphically population-level features such as barriers to migration, or more generally, the combined effect of evolutionary processes on genetic differentiation.

Our method assumes that we have data for n individuals sampled from a spatially distributed population at locations $(x_1, y_1), \dots, (x_n, y_n)$ and genotyped at p loci, either SNPs or microsatellites. The geographic information is used to assign individuals to the closest deme in the population graph (V, E); this defines the sample configuration $\underline{\alpha} = (\alpha_1, \dots, \alpha_n)$. Given $G = (V, E, M)$ with symmetric migration rates $M = (m_{\alpha\beta})$ we can compute the pairwise distance matrix for entire population $\Delta = (\Delta_{\alpha\beta})$; given Δ and the deme indicators $\underline{\alpha}$ we can obtain the expected pairwise distances for the observed sample $\underline{\Delta} = (\Delta_{ij})$. Notation: Here we discuss the likelihood of the sample, so we will write simply Δ throughout as there is no need to distinguish between the population and the sample distance matrices.

In the previous chapter we derived expressions for the mean and variance of the allele count vector $Z = (Z_i)$ at a segregating site [eq. (3.3) for single nucleotide polymorphisms; eq. (3.16) for microsatellites]. Recall that

$$\mathrm{E}\{Z\} = \mu 1, \qquad \mathrm{var}\{Z\} = \sigma^2\big(11' - \lambda\Delta\big), \tag{4.1}$$

where μ is the allele frequency and σ^2 is the variance in allele frequency [in the sample, not the population]. It is convenient to normalize Δ so that $1'\Delta^{-1}1 = 1$; then the correlation matrix $\Sigma = 11' - \lambda\Delta$ is positive definite for $\lambda \in (0, 1)$ [Appendix 7.2].

Recall further that neutral sites (not under selection) develop under the same coalescent process, and therefore, the genotype vectors $Z = (Z^1, \dots, Z^p) \in \mathbb{Z}^{n\times p}$ at p segregating sites have the same correlation matrix Σ. The scalar parameters μ, σ^2 can vary across sites. For microsatellites μ is the ancestral allele and σ^2 depends on the mutation rate θ, and both are site specific. For SNPs μ is the expected allele frequency if the derived allele is coded as 1; but the labels might not be consistent as usually the minor allele is coded as 1.

Our aim here is to incorporate these expressions for the mean and variance into a likelihood function in order to infer effective migration rates from observed data. Note that every individual has mean μ regardless of location; intuitively, the shared parameter μ contains little information about patterns of genetic differentiation between individuals, as we discuss in Section 4.4. So, to simplify, assume that we observe the pairwise differences, $Z_i - Z_j$, rather than the allele counts Z_i. Equivalently, assume that we observe LZ where $L \in \mathbb{R}^{(n-1)\times n}$ is a basis for contrasts, e.g., $L =$

$(e_2 - e_1, e_3 - e_1, \ldots, e_n - e_1)'$ where e_i is the standard basis vector with 1 in the ith coordinate and 0 otherwise. Note that

$$\mathrm{E}\{LZ\} = 0, \qquad \mathrm{var}\{LZ\} = -\sigma^* L\Delta L', \tag{4.2}$$

where we define $\sigma^* = \lambda\sigma^2$ because the variance and the scale are longer identifiable. The matrix $-L\Delta L'$ is positive definite, and thus a valid covariance matrix, because the distance matrix Δ is nonnegative definite on contrasts and Lv is a contrast for every $v \in \mathbb{R}^{n-1}$.

Therefore, it might be natural to assume a Normal likelihood for the pairwise differences,

$$LZ \,|\, \sigma^*, \Delta \sim \mathrm{N}_{n-1}\big(0, -\sigma^* L\Delta L'\big). \tag{4.3}$$

Suppose further that the genotyped markers are independent; then it is straightforward to extend the Normal likelihood (4.3) for one locus to multiple loci. In particular, for SNP data where usually there are many more SNPs than individuals and mutation rates are low, let $S = ZZ'/p$ be the observed similarity matrix averaged across p SNPs. Then LSL' is a scatter matrix of pairwise differences and

$$LSL' \,|\, \sigma^*, \Delta \sim \mathrm{W}_{n-1}\Big(p, -\frac{\sigma^*}{p}(L\Delta L')\Big), \tag{4.4}$$

where the degrees of freedom are the number of independent SNPs and and the scale parameter σ^* is shared. Therefore, by considering the pairwise differences, we avoid estimating a nuisance parameter μ with dimensionality that grows with the number of markers p. In practice we also gain efficiency with faster MCMC convergence.

4.1 Effective degrees of freedom for SNP data

So far we have considered the case where the p genotyped markers are independent (unlinked). The assumption of independence between loci is very strong and likely to be violated. In particular, SNPs in close proximity are often associated (in linkage disequilibrium) because individuals inherit long segments of unbroken DNA from their parents. For this reason, SNPs data is often "thinned" by removing SNPs in high LD. We propose an alternative method to correct for model mis-specification due to both dependence between SNPs and non-normality of genotypes.

In the Wishart likelihood (4.3) the scatter matrix of contrasts, LSL', has known degrees of freedom p. However, instead of fixing the degrees of freedom to the number of genotyped SNPs, we can estimate this parameter. The likelihood for the scatter matrix becomes

$$LSL' \,|\, k, \sigma^*, \Delta \sim \mathrm{W}_{n-1}\Big(k, -\frac{\sigma^*}{k}(L\Delta L')\Big), \tag{4.5}$$

with degrees of freedom $k \in (n, p)$. Both Wishart likelihoods (4.3) and (4.5) imply $\mathrm{E}\{LSL'\} = -\sigma^* L\Delta L'$. Therefore, estimating the degrees of freedom does not affect the expected pairwise differences as a function of effective migration. However, the Wishart variance is proportional to $(\sigma^*)^2/k$, so it we infer $k \in (n, p)$ rather than set $k = p$, the model variance increases as we would expect if the data contain less information than the sample size suggests, or more generally, if the model is mis-specified. Under normality, $k = p$ implies that all sites are independent; otherwise, the variance increases by a factor of p/k.

4.2 Prior on migration surface represented as a Voronoi tessellation

We have proposed a model for population structure in terms of expected pairwise distances on a population graph $G = (V, E, M)$ where (V, E) is a rectangular grid and M assigns effective migration rates to edges in the graph. The goal is to estimate the effective migration surface M so that the demographic model G explains the observed genetic dissimilarities. The grid is fixed; the likelihood is defined in the previous section. Here we consider prior specification for M.

The regular grid (V, E) is not determined by the sampling locations and it yields a high-dimensional, flexible representation so that fine features in the effective migration surface can emerge if supported by the data. To take advantage of this flexibility, we organize the edges in terms of a Voronoi tessellation of the habitat. Statistically, the Voronoi decomposition offers the advantages of parameter sharing and a locally smooth migration surface. Previous applications of Voronoi tiling in population genetics include [Guillot et al., 2005] and [Wasser et al., 2004].

A Voronoi tessellation of the migration surface M is fully specified by the number of tiles T, their locations $\underline{u}$ and migration rates $\underline{m}$. Thus $\underline{m} = \{m_t : t = 1, \dots, T\}$ is the set of effective migration rates for the T tiles in the partition. Furthermore, let edge $(\alpha, \beta) \in E$ have migration rate

$$m_{\alpha\beta} = \tfrac{1}{2} m_{t_\alpha} + \tfrac{1}{2} m_{t_\beta}, \tag{4.6}$$

where t_α denotes the tile deme α falls into. That is, the rate of an edge is the average rate of the two tiles it connects.

Migration rates are naturally positive and therefore we parametrize them on the log scale as differences from the overall mean rate $\ell\bar{m}$,

$$\log_{10}(m_t) = \ell\bar{m} + e_t. \tag{4.7}$$

If the effect of distance on differentiation is space-homogeneous and the tile-specific effects e_t are (close to) 0, the migration pattern thus produced would correspond to isolation by distance.

Therefore, our model has the following parameters:

1. parameters of interest Θ_1 that determine the effective migration rates M and thus the effective pairwise distances Δ. These are
   - $(T, \ell\bar{m}, \sigma_m^2)$: number of tiles, mean and variance of tile migration rates on the log (base 10) scale.
   - $\{(e_t, u_t) : t = 1, \dots, T\}$: relative effect and center location for each Voronoi tile t. The dimensionality of this group of parameters changes with the number of Voronoi tiles T.
   - k: effective degrees of freedom for SNP data where we observe more sites than individuals, i.e., $p > n$.
2. nuisance parameters Θ_0 that do not depend on the demographic model. For SNP data this is the scale parameter σ^*; for microsatellite data each site has its own scale parameter σ_s^* because mutation rates vary across sites and under the stepwise mutation model the scale σ_s^* is the mutation rate θ_s.

Using the Voronoi tessellation $\mathcal{V}(T, \underline{u}, \underline{e})$ to represent M, we can have fewer than $|E|$ rate parameters to estimate but we do not know how many tiles we need and where their centers are. This depends on the patterns of genetic differentiation across the habitat.

To complete the Bayesian specification we place priors on the model parameters:

$$\text{(number of Voronoi tiles)} \qquad T \mid \nu \sim \text{Po}(\nu), \tag{4.8a}$$

$$\text{(tile locations)} \qquad \underline{u} \mid T \overset{\text{iid}}{\sim} \text{U}(\mathcal{H}), \tag{4.8b}$$

$$\text{(tile effects)} \qquad \underline{e} \mid \sigma_e^2, T \overset{\text{iid}}{\sim} \text{N}(0, \sigma_e^2). \tag{4.8c}$$

The hyperparameter ν controls how much spatial heterogeneity the effective migration surface exhibits. The rate hyperparameters are

$$\text{(overall migration rate)} \qquad \ell\bar{m} \sim \text{U}(lob, upb), \tag{4.9a}$$

$$\text{(tile variance)} \qquad \sigma_m^2 \sim \text{Inv-G}(a/2, b/2). \tag{4.9b}$$

[For all results we report here $a = 6, b = 3$.] The lower and upper bounds on the mean log rate are chosen so that the mean migration rate varies in the range $[1/300, 300]$ on the original scale. The bounds are somewhat arbitrary but based on simulations of genetic data with `ms` [Hudson, 2002]. Restricting the support is necessary because the model is not numerically stable at the two extremes:

- When migration rates are very small (relatively to coalescence rates), it takes very long time on average for two lineages from different demes to coalesce. In the limit, the population is a collection of unrelated subpopulations that evolve independently.
- When migration rates are very large (relatively to coalescence rates), the time it takes to move from one deme to another is negligible compared to the coalescence times. In the limit, the population behaves like a panmictic population without any structure.

The prior on the effective degrees of freedom is uniform on the log scale:

$$\text{(degrees of freedom)} \qquad \pi(k) \propto \frac{1}{k}. \tag{4.10}$$

The prior is proper because k is bounded: $k > n$ because k is the degrees of freedom parameter in a Wishart distribution, and $k < p$ because k should not exceed the number of observed sites (features). [The normalizing constant is therefore $\log(p) - \log(n)$.]

$$\text{(scale parameter)} \qquad \sigma^* \sim \text{Inv-G}(c/2, d/2). \tag{4.11}$$

[For all results we report here $c = 1, d = 1$.]

We use reversible-jump MCMC to estimate T as the dimension of both $\underline{u}$ and $\underline{e}$ changes as the number of tiles T increases or decreases. Full details about the MCMC implementation are given in Appendix 7.6.

4.3 Likelihood for distance matrices

The Wishart likelihood (4.5) is given in terms of the contrast basis L but it does not depend on the choice of L. Instead, it can be written in terms of the observed similarities $S = ZZ'/p$, the model distances Δ and its orthogonal projection Q given by

$$Q = I - \frac{11'\Delta^{-1}}{1'\Delta^{-1}1}. \tag{4.12}$$

In Appendix 7.4 we show that the Wishart log likelihood that corresponds to the model (4.5) can be written as

$$\ell(k,\sigma^*,\Delta) = \begin{cases} + & [k/2]\operatorname{logdet}\left\{-(L\Delta L')^{-1}/\sigma^*\right\} \\ - & [k/2]\operatorname{tr}\left\{-(L\Delta L')^{-1}LSL'/\sigma^*\right\} \\ + & [(k-n)/2]\operatorname{logdet}\{LSL'\} \\ + & [k(n-1)/2]\log(k/2) \\ - & \log\Gamma_{n-1}(k/2) \end{cases} = \begin{cases} + & [k/2]\log\operatorname{Det}\left\{-\Delta^{-1}Q/\sigma^*\right\} \\ - & [k/2]\operatorname{tr}\left\{-(\Delta^{-1}Q)S/\sigma^*\right\} \\ + & [(k-n)/2]\operatorname{logdet}\{S\} \\ + & [k(n-1)/2]\log(k/2) \\ - & \log\Gamma_{n-1}(k/2) \\ + & [(k-n)/2]\log\left(\frac{1'S^{-1}1}{1'1}\right) \\ - & [n/2]\operatorname{logdet}\{LL'\} \end{cases} \tag{4.13}$$

The only term that involves the residual basis L is $(n/2)\operatorname{logdet}\{LL'\}$. Regardless of the choice for L, this terms does not depend on the parameters (k,σ^*,Δ). Full details about the likelihood computation are given in Appendix 7.5.

4.3.1 *Related model*

This is the marginal likelihood for distance matrices introduced in [McCullagh, 2009]. Let D the $n \times n$ pairwise dissimilarity matrix given by

$$D = 1\operatorname{diag}(S)' + \operatorname{diag}(S)1' - 2S. \tag{4.14}$$

The orthogonal projection $Q = I - 11'\Sigma^{-1}/(1'\Sigma^{-1}1)$ satisfies

$$Q'\Sigma^{-1} = Q'\Sigma^{-1}Q = -\lambda Q'\Delta^{-1}Q = -\lambda Q\Delta^{-1} \tag{4.15}$$

since $\ker\{Q\} = \{1\}$ and thus $Q1 = 0$. Similarly, $QDQ' = -2QSQ'$. Therefore, for fixed $k = p$ and after we ignore all terms that do not involve Δ or σ^*,

$$\ell(\sigma^*,\Delta;S) \propto \ell(\sigma^2,\Sigma;D) \propto \frac{p}{2}\log\operatorname{Det}\left\{-\Delta^{-1}Q/\sigma^*\right\} - \frac{p}{2}\operatorname{tr}\left\{-\Delta^{-1}QS/\sigma^*\right\} \tag{4.16a}$$

$$= \frac{p}{2}\log\operatorname{Det}\left\{\Sigma^{-1}Q/\sigma^2\right\} + \frac{p}{4}\operatorname{tr}\left\{\Sigma^{-1}QD/\sigma^2\right\} \tag{4.16b}$$

where $\sigma^* = \lambda\sigma^2$.

Recently, [Hanks and Hooten, 2013] build this likelihood into a parametric model for isolation by resistance [McRae, 2006]. Briefly, the genetic data is generated by a Gaussian Markov random field on an undirected graph (circuit). The covariance structure is given by an intrinsic conditional autoregressive model, i.e., the conditional distribution of each node given the rest of the graph is normal with mean and variance that depend on its first-degree neighbors only. [Hanks and Hooten, 2013] specify the model so that the expected square differences between nodes are exactly effective resistance distances on the population graph. In our notation, let $\Delta = R$ be the matrix of effective

resistances. [Note that this is slightly different from the IBR-based approximation to expected coalescence times $\Delta = \underline{T}$.] [Bapat, 2004] shows that

$$R^{-1} = -\frac{1}{2}L + \tau\tau' \tag{4.17}$$

where L is the Laplacian of the graph $G = (V, E, M)$ and $\tau\tau'$ is a rank-one update. Then

$$R^{-1}Q = R^{-1} - \frac{R^{-1}11'R^{-1}}{1'R^{-1}1} = \Big(-\frac{1}{2}L + \tau\tau'\Big) - \frac{(\tau(1'\tau))(\tau(1'\tau))'}{(1'\tau)^2} = -\frac{1}{2}L \tag{4.18}$$

$$\begin{aligned}(R + 11')^{-1} &= R^{-1}Q + \frac{R^{-1}11'R^{-1}}{1'R^{-1}1} - \frac{R^{-1}11'R^{-1}}{1 + 1'R^{-1}1} \\ &= R^{-1}Q + \frac{R^{-1}11'R^{-1}}{(1'R^{-1}1)(1 + 1'R^{-1}1)} \\ &= -\frac{1}{2}L + \frac{\tau\tau'}{1 + (1'\tau)^2}\end{aligned} \tag{4.19}$$

$$Q[R + 11'] = I - \frac{1\tau'(1'\tau)}{1 + (1'\tau)^2}\frac{1 + (1'\tau)^2}{(1'\tau)^2} = I - \frac{1\tau'}{1'\tau} \tag{4.20}$$

$$(R + 11')^{-1}Q = -\frac{1}{2}L + \frac{\tau\tau'}{1 + (1'\tau)^2} - \frac{\tau\tau'}{1 + (1'\tau)^2} = -\frac{1}{2}L \tag{4.21}$$

That is, $B^{-1}Q[B]/4 = R^{-1}Q[R]$ where $B = R/4 + 11'$. [Hanks and Hooten, 2013] represent conductances between connected nodes as a function of landscape features, e.g., elevation. Instead we represent conductances [i.e., migration rates] through a colored Voronoi tessellation and estimate edge weights without reference to available ecological variables.

Modeling the dissimilarity matrix D instead of the raw allele counts Z is convenient because

- Suppose that O is an orthogonal transformation (rotation or reflection). Then

$$S^O = (ZO)(ZO)' = ZOO'Z' = ZZ' = S$$

- Suppose T is a translation by $\mu = (\mu_1, \dots, \mu_p)'$. Then

$$D^T_{ij} = \langle (z_i - \mu) - (z_j - \mu)\rangle^2 = \langle z_i - z_j\rangle^2 = D_{ij}$$

Although the transformation from the entire data Z to the summary D is not a one-to-one transformation, we do not lose information about relative distances, i.e., population structure. Instead we lose information about some nuisance parameters. For example, $S \to D$ makes the labeling of the alleles irrelevant.

4.4 What do we lose from ignoring the means?

We can use the marginal likelihood (4.3) because sampled individuals are equally distant from the most recent common ancestor of the sample [the root of the genealogy tree], and therefore, share a common mean. Thus $K = 1$ is a basis for the mean space. [Recall that L is a basis for the residual space of pairwise differences.] Therefore,

$$\begin{pmatrix} 1' \\ L \end{pmatrix} Z \sim \mathrm{N}_n\left(\begin{pmatrix} \mu \\ 0 \end{pmatrix},\ \sigma^2\begin{pmatrix} 1'\Sigma 1 & 1'\Sigma L' \\ L\Sigma 1 & L\Sigma L' \end{pmatrix}\right) \tag{4.22}$$

Let $T = 1'Z = \sum_{i=1}^{n} Z_i$ and $Y = LZ$.

$$Y = LZ \mid \mu, \Sigma \sim \mathrm{N}_{n-1}(0, \sigma^2 L\Sigma L') \tag{4.23a}$$

$Q = \Sigma L'(L\Sigma L')^{-1}L$

$= I - 1(1'\Sigma^{-1}1)^{-1}1'\Sigma^{-1}$

$1'\Sigma^{-1}1 = 1'\Delta^{-1}1/(1-\lambda)$

$$\begin{aligned} T \mid Z, \Sigma &\sim \mathrm{N}\Big(\mu + 1'\Sigma L'(L\Sigma L')^{-1}Y, \sigma^2[1'\Sigma 1 - 1'\Sigma L'(L\Sigma L')^{-1}L\Sigma 1]\Big) \\ &= \mathrm{N}\Big(\mu + 1'QZ, \sigma^2 1'1(1'\Sigma^{-1}1)^{-1}1'1\Big) \\ &= \mathrm{N}\Big(\mu + 1'QZ, (1-\lambda)n^2\sigma^2\Big) \end{aligned} \tag{4.23b}$$

The conditional distribution of T depends on Δ only through the bias term $1'QZ$. Therefore we choose to ignore it and use only the marginal distribution of Y to infer Δ.

4.5 Standardizing genotype data

Before performing PCA analysis for population structure it is common to standardize SNPs and to set the missing alleles to the observed average at the corresponding marker [McVean, 2009, Price et al., 2006]. The motivation is that without normalization SNPs with higher variance contribute more to the scatter matrix ZZ'. Therefore, the procedure tends to up-weight the influence of rare variants. Here we discuss why neither centering the genotypes to have mean 0 nor standardizing the variance is appropriate when analyzing population structure.

In matrix notation, let $C = I - 11'/n$ be the centering matrix for n observations. Then multiplying by C removes the mean:

$$X = CZ \overset{\text{iid}}{\sim} \mathrm{N}_n(\mu C1, \sigma^2 C\Sigma C) = \mathrm{N}_n(0, -\sigma^* C\Delta C), \tag{4.24}$$

This operation is convenient because XX' has central Wishart distribution. It also makes the labelling of alleles as ancestral or derived ['0' or '1'] irrelevant, up to a change in sign. Suppose that we "flip" the $0/1$ labels at a particular site, i.e., $z^* = 1 - z$. Then $x^* = C(1-z) = -Cz = -x$ because $C1 = (I - 11'/n)1 = 0$.

However, centering with C assumes the individuals are independent and identically distributed, i.e., no population structure: If $\Sigma = I$, then the projection Q onto the space of contrasts is $Q = I - 11'\Sigma^{-1}/(1'\Sigma^{-1}1) = C$. Since our model assumes the individuals are coupled with correlation given by $11' - \lambda\Delta$, it is not appropriate to naively center the genotypes to have mean 0 or to substitute the average allele frequency for missing values. For SNP datasets, it is better to impute missing SNP values before analyzing population structure. There are various imputation algorithms but they all would take into account similarities across observed alleles to impute missing ones. For microsatellite datasets, which are usually much smaller and harder to impute, we use the likelihood for the observed pairwise distances only. [So that the sample configuration $\underline{\alpha}$ is really site-specific.]

Furthermore, it might not be appropriate to standardize SPNs to have the same variance precisely because this up-weights the contribution of rate alleles [McVean, 2009]. A mutation in effect splits sampled individuals into two groups that are slightly different genetically — those that carry the mutation versus those that do not. Intuitively, newer and especially private mutations, which are carried by a single individual, are informative for structure that is too fine to represent with a model at the level of demes.

5

Simulations of Structured Genetic Data

In this chapter we describe several simulated scenarios that we use to evaluate the performance of our method for estimating effective migration as well as to illustrate some of its properties. We use the program `ms` [Hudson, 2002] to simulate genetic data under the structured coalescent. Given the model parameters (deme sizes and migration rates) and the sample configuration, `ms` first generates a random genealogy, which describes the history of the sample from the present to its most recent common ancestor, and then places a Poisson number of mutations uniformly (and independently of each other) on the tree.

We use `ms` to simulate independent and identically distributed genealogies under the stepping-stone model $G = (V, E, M)$ with conservative migration $M = (m_{\alpha\beta})$. Therefore, the iid assumption across sites holds but the normality assumption is violated. In all examples, we generate $p = 3000$ single nucleotide polymorphisms for $n = 300$ haploid individuals on a 12×8 regular triangular grid. [The corresponding `ms` commands, with detailed explanations, are given in Appendix 7.8.]

To generate histories with exactly one mutation, we choose a small mutation rate θ and discard genealogies that carry zero or multiple mutations.

5.1 *Spatial structure due to constant migration*

First we generate data under different patterns of migration — either uniform or with a barrier — to confirm that the method performs accurately when the underlying demographic model is correct. That is, the population does evolve on a known grid (V, E) of equally sized demes and unknown migration rates. In these simulations, therefore, effective migration rates are true migration rates [up to a constant of proportionality that depends on the coalescent timescale N_0. We set up the simulations so that this constant is 1.]. We report migration rates, as they are parametrized, on the log (base 10) scale, and the blue/brown color scheme is based on [Brewer et al., 2003].

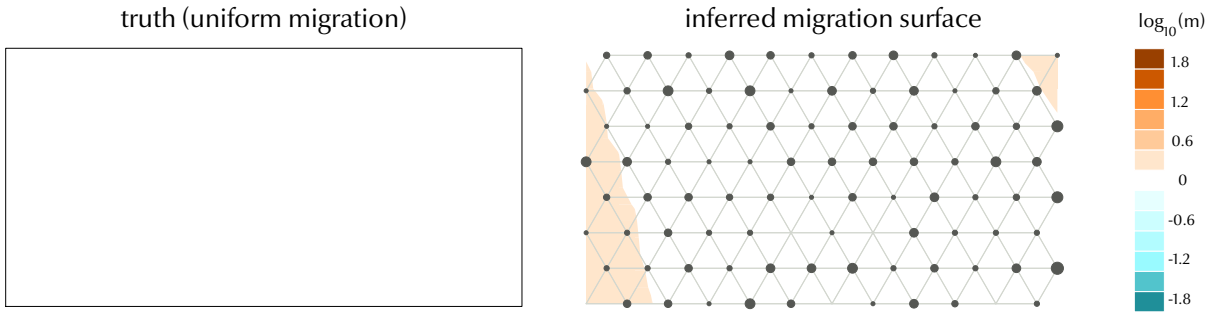


Figure 5.1: Uniform migration rates and equal deme sizes: $q_\alpha = 1$ for all $\alpha \in V$ and $m_{\alpha\beta} = 1$ for all $(\alpha, \beta) \in E$. The size of the gray circles indicates the number of individuals sampled from the corresponding deme.

In Figures 5.1 and 5.2 we directly compare the truth (left) with the posterior mean (right). Not every deme in the population graph is necessarily observed but sampling is

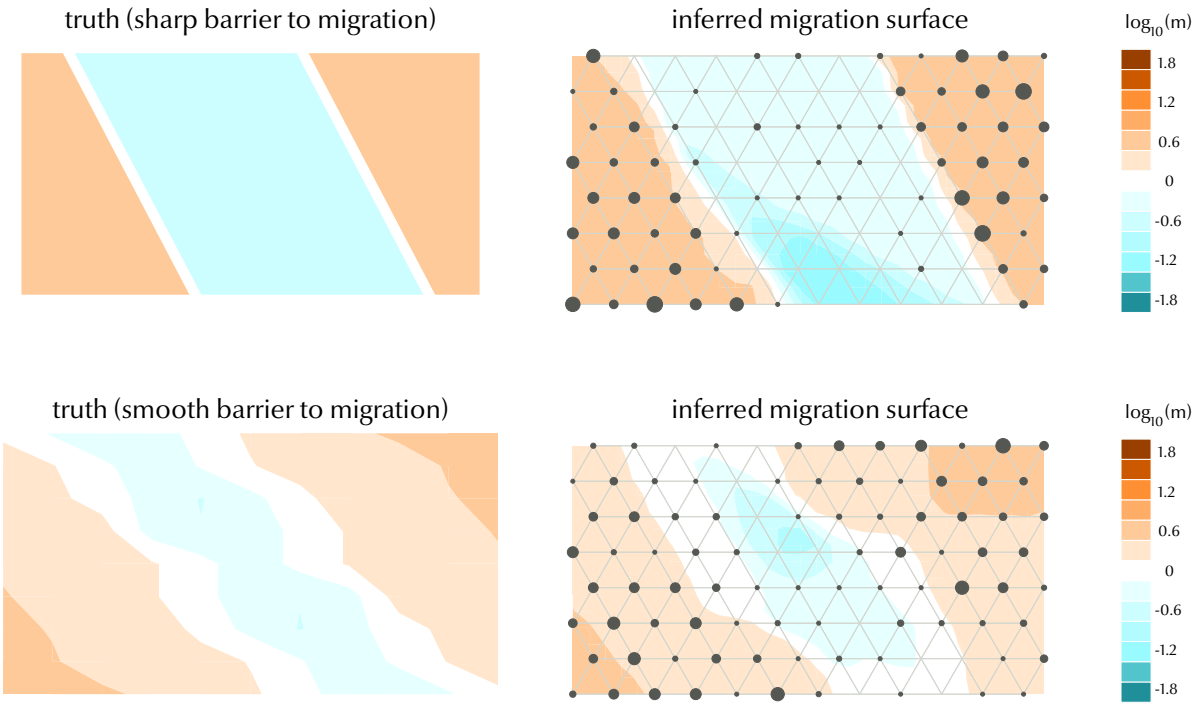


Figure 5.2: Barrier to migration and equal deme sizes: migration rates vary between high, $m_{\alpha\beta} = 3$, and low, $m_{\alpha\beta} = 1/3$, in either a sharp or smooth pattern.

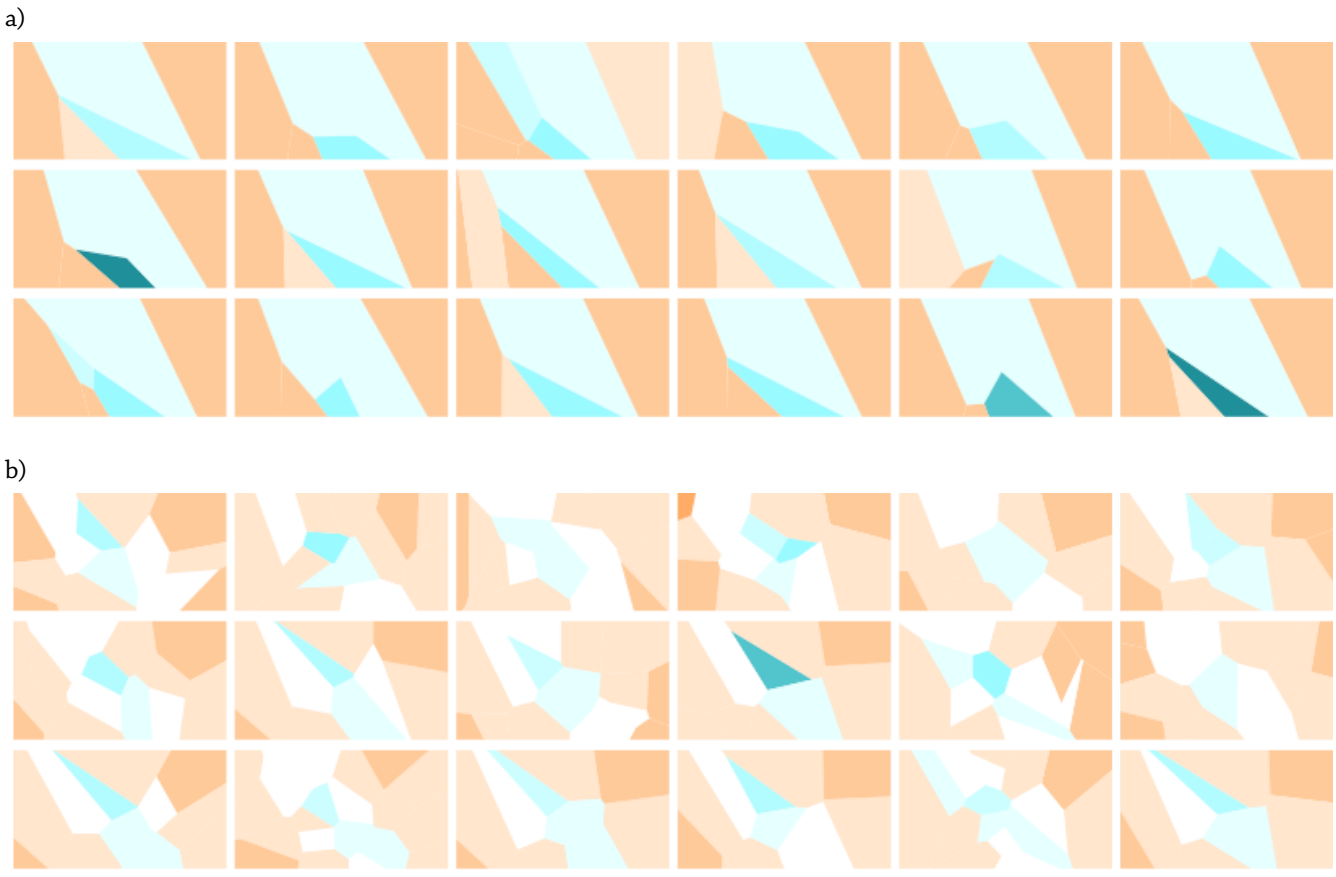
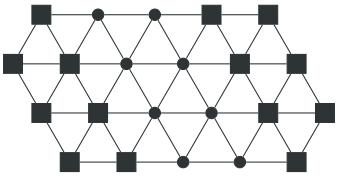


Figure 5.3: Draws from the posterior distribution of effective migration. a) sharp barrier to migration; b) smooth barrier to migration.

balanced because areas with higher migration are sampled with higher probability. In all three cases our method correctly captures the qualitative pattern of migration. And in Figure 5.3 we show samples from the posterior distribution on the colored Voronoi tessellation, to illustrate the uncertainty in the estimated effective migration surface.

5.2 *Spatial structure due to variation in diversity*

The next set of simulations demonstrate that effective migration reflects the combined effect of demographic processes on genetic differentiation. In particular, we use two examples to show how differences in effective population size can influence effective migration rates. In the first example, lower migration rates cancel the effect of bigger demes sizes, to produce uniform effective migration. In the second example, only deme sizes vary to produce the effect of a barrier to migration.



To describe the simulations, consider the example graph with two groups of demes, A (circles, smaller in size) and B (squares, bigger in size), with deme sizes N_A and N_B, respectively. Let m_{AA} be the migration rate of all $A-A$ edges and m_{BB} be the migration rate of all $B-B$ edges. We assign migration rates to the "across" edges $A-B$ and $B-A$ so that migration is conservative and deme sizes are constant over time, as required by the stepping-stone model. Formally,

$$\sum_{\gamma:\gamma\sim\alpha,\gamma\in A} m_{AA}N_A + \sum_{\gamma:\gamma\sim\alpha,\gamma\in B} m_{AB}N_A = \sum_{\gamma:\gamma\sim\beta,\gamma\in A} m_{BA}N_B + \sum_{\gamma:\gamma\sim\beta,\gamma\in B} m_{BB}N_B \quad (5.1)$$

by definition (2.14) where the coalescence rate is $q_A = 1/N_A$. A sufficient condition for conservative migration is that

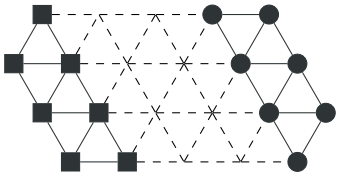
$$m_{AB}N_A = m_{BA}N_B. \quad (5.2)$$

This condition preserves the symmetry as much as possible because the number of migrants from $\alpha \in A$ to $\beta \in B$ is equal to the number of migrants from β to α, i.e., migration is balanced across every edge. Therefore, given the deme sizes N_A and N_B, we let $m_{AB} = 1/N_A$, $m_{BA} = 1/N_B$. [Or more generally, we can let $m_{AB} = m_C/N_A$, $m_{BA} = m_C/N_B$ for a given between-group rate m_C.]

In the first example, bigger demes exchange the same number of migrants as smaller demes. To achieve this, we set $N_B = 5N_A$, $m_{BB} = m_{AA}/5$ and thus $m_{AA}N_A = m_{BB}N_B$. All demographic parameters are scaled by the coalescent timescale N_0, so the effective migration rate of both $A-A$ and $B-B$ edges is approximately $m_{AA}N_A/N_0 = 1/N_0$. That is, differences in population size are canceled by differences in migration rate. Consequently, we expect the migration surface to be uniform, and indeed, this is what we observe in Figure 5.5 a).

In the second example, bigger deme exchange more migrants. To achieve this, we set $N_B = 5N_A$, $m_{BB} = m_{AA}$ and thus $m_{BB}N_B > m_{AA}N_A$. Since migration rates are relative to deme sizes, at the same migration rate bigger demes exchange a higher number of migrants which results in higher *effective* migration. Therefore, coalescence times between B demes [on the same side of the barrier but not across it] are shorter than coalescence times between A demes. Genealogies with such topology are consistent with higher migration *at equal coalescence rates* because lineages that transition more often between demes have fewer chances to coalesce. Consequently, we expect a barrier to effective migration, and indeed, this is what we observe in Figure 5.5 b).

5.3 *Spatial structure due to a split event*



The final sequence of examples produces a barrier effect from a past demographic event that removes edges in the graph and thus splits the habitat into two regions that no longer exchange migrants.

To describe the simulations, consider the example graph with two groups of demes, A (circles) and B (squares), with the same deme size. [The demes in the middle, C, are part of the population but we collect no samples from that area which remains "unobserved".] There are also two types of edges: the solid ones have constant migration rate m; the dashed ones have migration rate 0 for x units of time (measured in N_0 generations) and migration rate m from then on. Since Kingman's coalescent develops backwards in time, this setup simulates a *recent* barrier to migration from the present to point x in the past. Beyond time x the population graph is connected and has uniform migration rate m.

In Figure 5.6 we increase the time of the split event from $x = 0.3$ to $x = 9$ units of time. If the split is too recent on the relative scale of the other parameters, its effect is hard to detect and the effective migration surface is uniform. Otherwise, the split is detected as a barrier to effective migration. [The truth is a temporary barrier, the method infers a constant barrier.] In these simulations, an equilibrium phase of high migration followed by a recent interval of no migration produces genealogies that are dominated by a long branch between the common ancestor of A lineages and that of B lineages. Such topology is consistent with constant migration at low rates across the area separating A and B.

5.4 *The effect of SNP ascertainment*

In this example we simulate the effect of ascertainment bias due to a very small discovery panel (Figure 5.4). In this case there is a true barrier to migration but the discovery panel comes from a very limited area on one side of the barrier. This skews the observed genealogies as we observe only sites that are polymorphic in both the ascertainment sample (in red) and the analysis sample (in black). This example shows that ascertainment — which is not an evolutionary process — can have an effect of the inferred effective migration, especially if the discovery panel is not representative of the population.

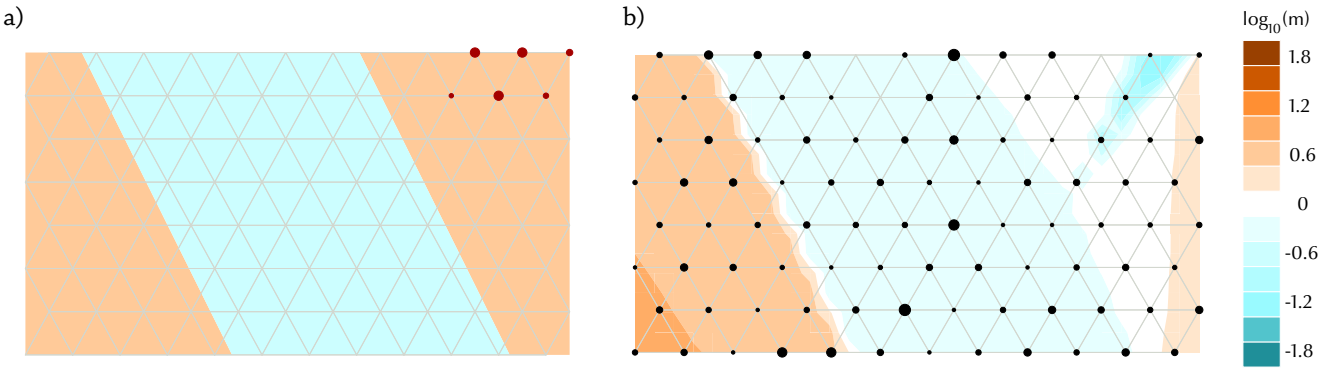


Figure 5.4: Barrier to migration with ascertainment bias. a) True migration pattern and the discovery panel in red; b) Estimated effective migration and the sample in black.

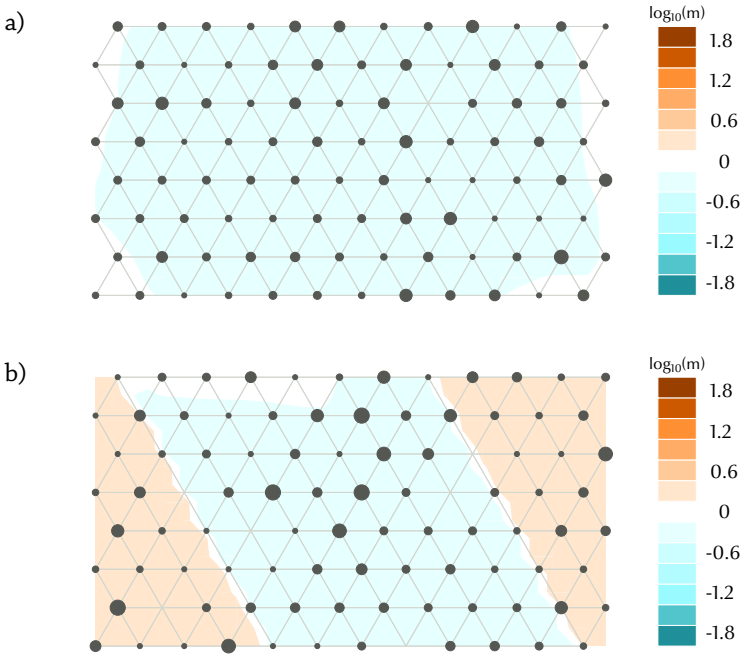


Figure 5.5: Population structure due to differences in deme size. In a) bigger demes exchange migrants at a lower rate and hence there is no variation in effective migration. In b) smaller demes exchange fewer migrants and hence there is an effective barrier to migration.

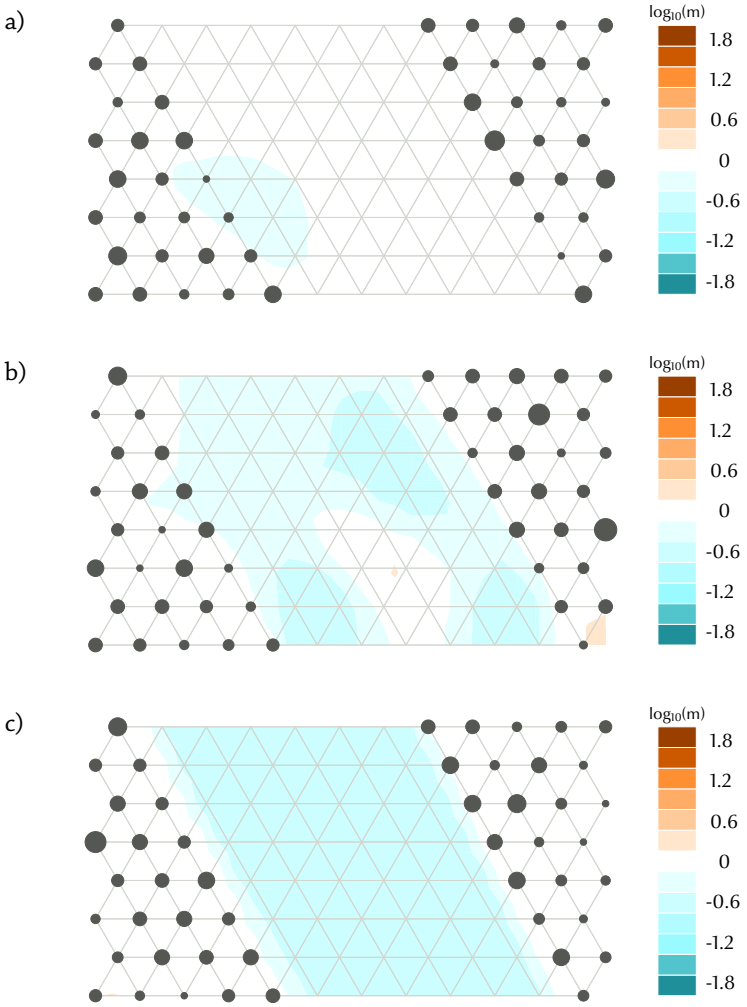


Figure 5.6: A past demographic event results in a barrier to effective migration. Here an ancestral population splits into subpopulations A (east) and B (west) at point x in the past. The further back in time this event occurs, the more differentiated A and B are. a) $x = 0.3$; b) $x = 3$; c) $x = 9$ units of time which is measured in N_0 generations.

5.5 *The effect of uneven sampling on PCA projection and effective migration*

It is well known that PCA projections are heavily influenced by irregular sampling [McVean, 2009]. To examine the impact of sample composition on effective migration, we simulate genetic data under the same barrier pattern as in Figure 5.2 but with various sampling schemes. We compare our method of estimating effective migration and PCA analysis of the observed covariance matrix in Figure 5.7. Even if sampling is biased towards one area of the habitat, the presence and location of the barrier are correctly detected as long as there are observations on both sides. On the other hand, the overall pattern of the PCA projections changes considerably. [Our method can be sensitive to the placement and coarseness of the population grid.]

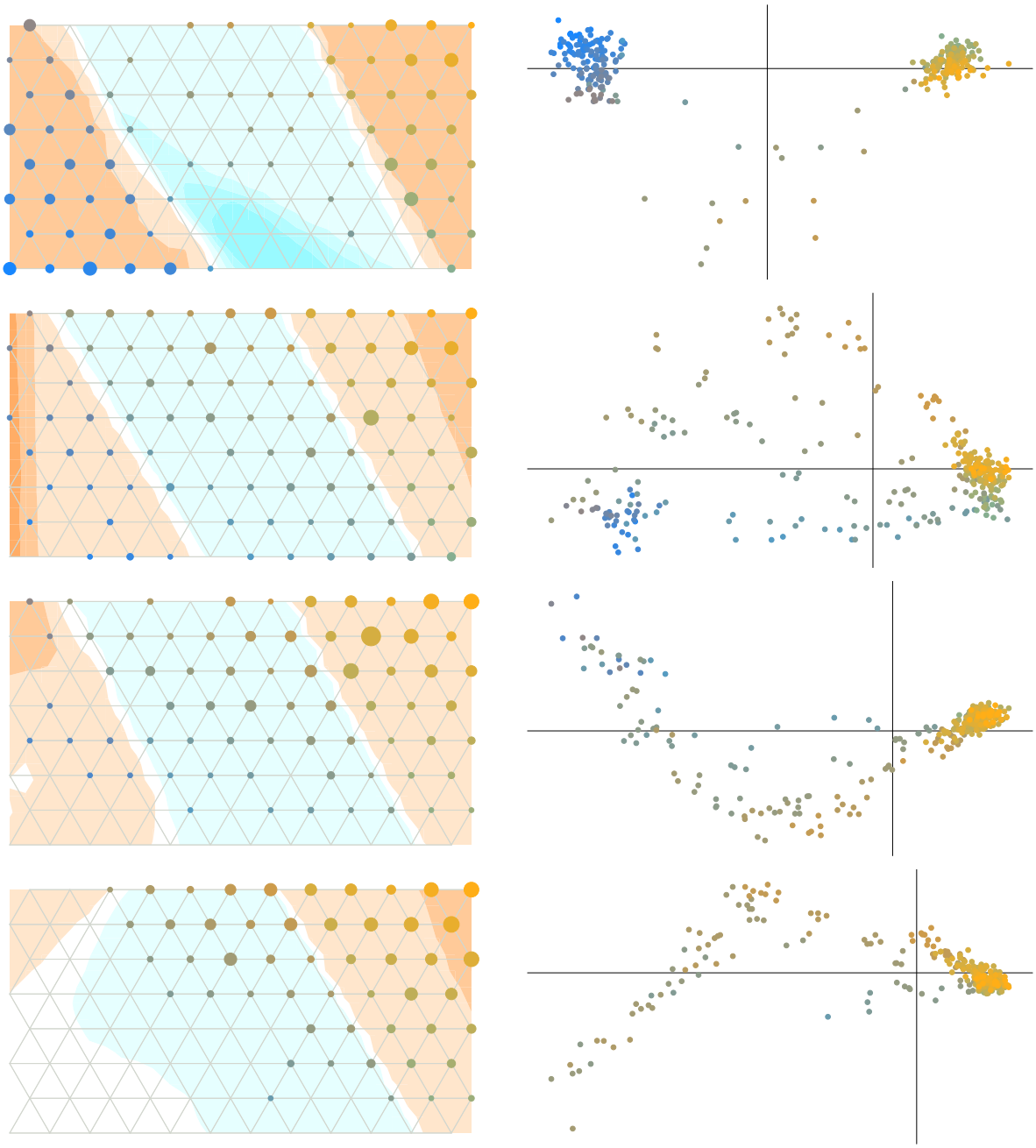


Figure 5.7: Barrier to migration with uneven sampling; colors indicate sampling location. The final example illustrates that naturally the method cannot detect variation from uniform migration in areas where no genetic data is observed.

5.6 Summary

The simulations in this chapter illustrate that effective migration can represent the combined effect of various demographic processes and events on genetic similarity and that our method is robust to uneven sampling but not ascertainment bias. However, effective migration does not help us to distinguish among possible histories as in this framework population structure is always explained with a stepping-stone model on a fixed grid of equally sized demes.

The examples also underline why it is difficult to interpret effective migration in terms of actual evolutionary history. As [McVean, 2009] emphasizes, very different demographic processes can produce very similar average genealogies. Our method, just like PCA, uses the information contained in pairwise comparisons averaged across sites and discards the sequential information contained in the ordering of sites along chromosomes, which can be helpful in selecting between possible histories.

6
Empirical Results

In this chapter we apply out method on four empirical datasets [three consist of SNPs and one of microsatellites] and we further demonstrate that effective migration rates can explain the spatial structure in genetic variation.

6.1 Red-backed fairywrens in Australia

First we present results for a sample of red-backed fairywrens (*Malurus melanocephalus*), a small passerine bird endemic to Australia [Figure 6.1]. The RBFW dataset was collected to study its population structure and demographic history across the Carpentarian barrier. Sampling and genotyping procedures as well as cluster-based analysis of population structure are described in [Lee and Edwards, 2008].

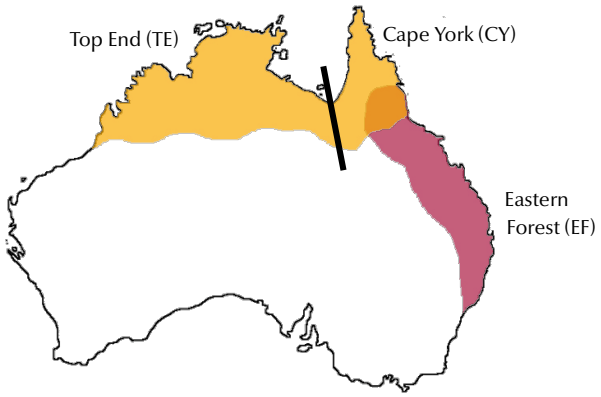


Figure 6.1: Habitat of the red-backed fairywren (*Malurus melanocephalus*), with the Carpentarian barrier (the black bar), in northern and eastern Australia. The map shows the ranges of two subspecies: *M. m. cruentatus* in yellow, *M. m. melanocephalus* in pink, and a broad hybrid zone in orange. The map also shows three major biogeographic regions: Top End (TE), Cape York (CY) and Eastern Forest (EF). The map is modified from [Lee and Edwards, 2008].

The Carpentarian barrier in northern Australia is a semi-arid region, roughly 150 km wide and extremely poor in vegetation [Lee and Edwards, 2008]. It has been argued that this region has had a primary effect on species distribution in northern and eastern Australia by acting as a barrier to migration, with secondary barriers along the coast. [Lee and Edwards, 2008] choose to study the demographic structure of the red-backed fairywren because its taxonomy, which is based mainly on plumage color, is not consistent with the Carpentarian hypothesis. The species has been traditionally categorized into two subspecies but their ranges do not lie on either side of the Carpentarian barrier, as we would expect if it has been the major barrier contributing to their divergence.

The dataset was made available to us by S. Edwards.After initial data processing, the RBFW dataset consists of $n = 27$ diploid individuals genotyped at $p = 1190$ bi-allelic, polymorphic SNPs. [As a reference to the original data, we remove 3 out of 30 individuals because most of their genotypes are missing and we also exclude monomorphic and tri-allelic SNPs.]

Throughout we will refer to three subpopulations of red-backed fairywrens as identified according to location in [Lee and Edwards, 2008]: Top End (TE) in northern Australia to the west of the Carpentarian barrier, Cape York (CY) in northeastern Australia to the east of the Carpentarian barrier and including the hybrid zone, and Eastern Forest (EF) in eastern Australia to the south of the hybrid zone [Figure 6.1].

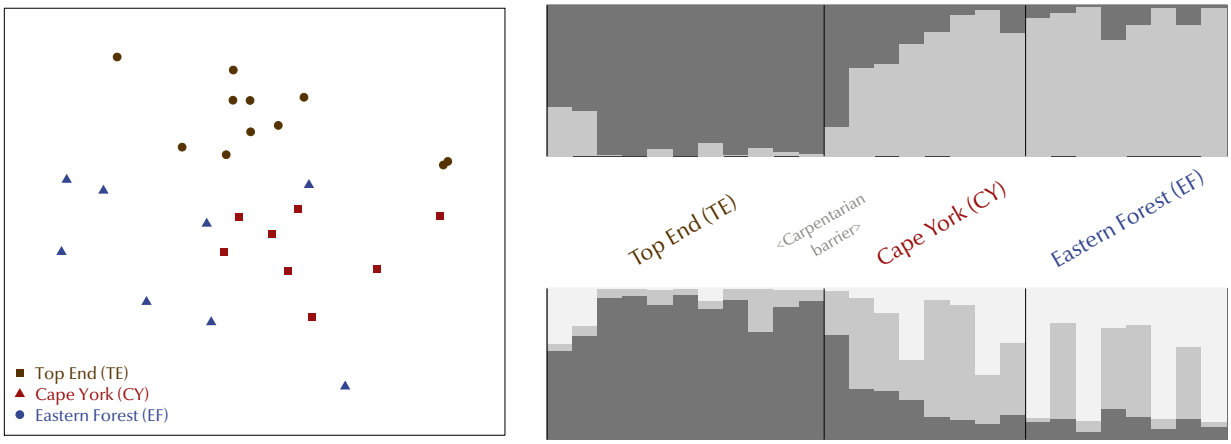


Figure 6.2: PCA and STRUCTURE analysis of the red-backed fairywren (RBFW) data.

First we perform principal components analysis (PCA) and cluster-based analysis (STRUCTURE). In Figure 6.2 (left) we plot the leading two principal components of the genetic covariance matrix, which explain 55% and 10% of the variance, repectively. PCA detects population structure but the results are difficult to interpret in terms of the evolutionary history of the species: there is some differentiation between the three subpopulations [in particular, Top End (TE) is better separated from Cape York (CY) than Eastern Forest (EF)] but there are no clearly delineated clusters. Although the three biogeographic groups are about equally represented, the sample is very small and much of the observed variation is between individuals within groups.

In Figure 6.2 (right) we report STRUCTURE results with two and three clusters, and using the sampling locations as prior information [Pritchard et al., 2000, Hubisz et al., 2009]. As observed in [Lee and Edwards, 2008], if we use STRUCTURE to assign the samples into two distinct clusters, Cape York (CY) and Eastern Forest (EF) are grouped together, which possibly indicates that the Carpentarian barrier has played a role in shaping the genetic differentiation of the red-backed fairy wren. When we use STRUCTURE to assign the samples into three distinct clusters, CY and EF individuals are estimated to be admixtures (with different proportions) of two "ancestral" populations. This suggests migration across the hybrid zone.

In our analysis the data has a slightly higher likelihood with three clusters rather than two as in [Lee and Edwards, 2008].

While both STRUCTURE plots might be interpreted to provide support for the Carpentarian hypothesis, STRUCTURE does not model the geographic distribution of samples across the habitat and thus cannot account for isolation by distance. In a homogeneous habitat, where the population is spatially distributed with uniform migration, genetic differentiation tends to increase with geographic distance. The RBFW data exhibits the isolation by distance property, at least at short to medium distances. The relationship between geographic and genetic distances appears to plateau as the Euclidean distance increases [Figure 6.3].

Cluster-based methods, such as STRUCTURE [Pritchard et al., 2000] and GENELAND [Guillot et al., 2005], attempt to find sharp boundaries between clusters, to maximize similarity within versus between clusters, in terms of allele frequency distributions. [These methods can assign individuals to multiple clusters according to individual-specific fractional membership, but again the differences between clusters must be sharp in or-

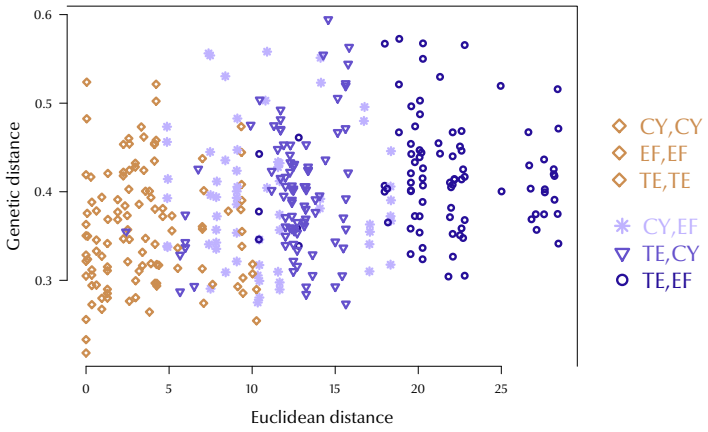
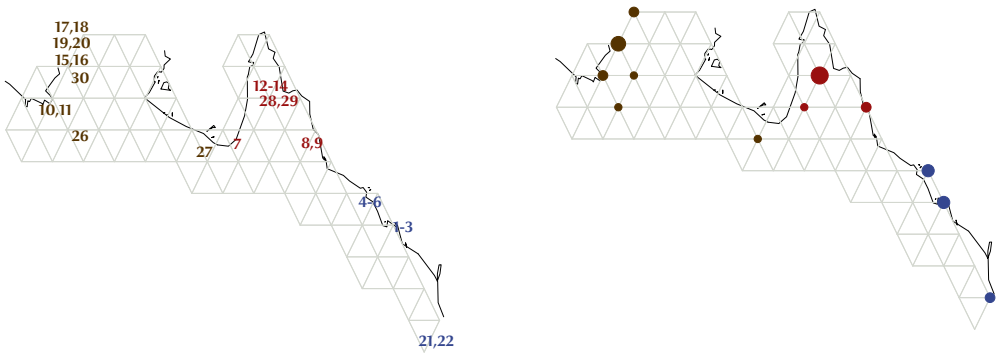


Figure 6.3: Observed genetic distance vs. Euclidean distance for the $\binom{n}{2} = 351$ pairs in the RBFW dataset. Each point is colored according to group membership to emphasize that on average Cape York (CY) is closer to Eastern Forest (EF) than to Top End (TE).

der to estimate these proportions with certainty.] Therefore, cluster-based methods are better suited to analyzing discrete population structure. However, genetic variation can exhibit continuous structure as genetic similarities tend to decay with distance and the decay can be gradual rather than sharp as in Figure 6.3. In this case STRUCTURE effectively separates those individuals that are farthest apart in space as Top End (TE) and Eastern Forest (EF) are assigned to different clusters.

The spatial structure of genetic variation in the RBFW data is continuous and therefore it can be partially explained with isolation by distance. However, since the Carpentarian barrier may reduce gene flow between the TE and CY groups, we estimate the patterns of migration rather than assume genetic differentiation increases as a function of the Euclidean distance between sampling locations [or equivalently, migration is uniform].

Figure 6.4: Irregular triangular grid (V, E) spanning the habitat of the red-backed fairywren. Samples are assigned to the closest deme. The method allows that sampling be both sparse and uneven. If the geographic information is coarse, it is appropriate to choose a coarse grid.



To apply our method for estimating effective migration rates, we first construct an irregular triangular grid (V, E) to cover the known range of the red-backed fairywren [Figure 6.4]. After running the MCMC chain from multiple random starting points to monitor convergence and averaging the runs, we report the posterior mean of the effective migration rates $M = (m_{\alpha\beta})$ in Figure 6.5, on the log base 10 scale, with low migration in blue and high migration in brown. For this small dataset, it is computationally feasible to use the coalescent-based distance matrix (i.e., the expected coalescence times $T_{\alpha\beta}$) as well as its approximation in terms of effective resistances $R_{\alpha\beta}$. The two metrics

produce very similar posterior estimates of the effective migration surface, shown in Figure 6.5 a) and b), respectively.

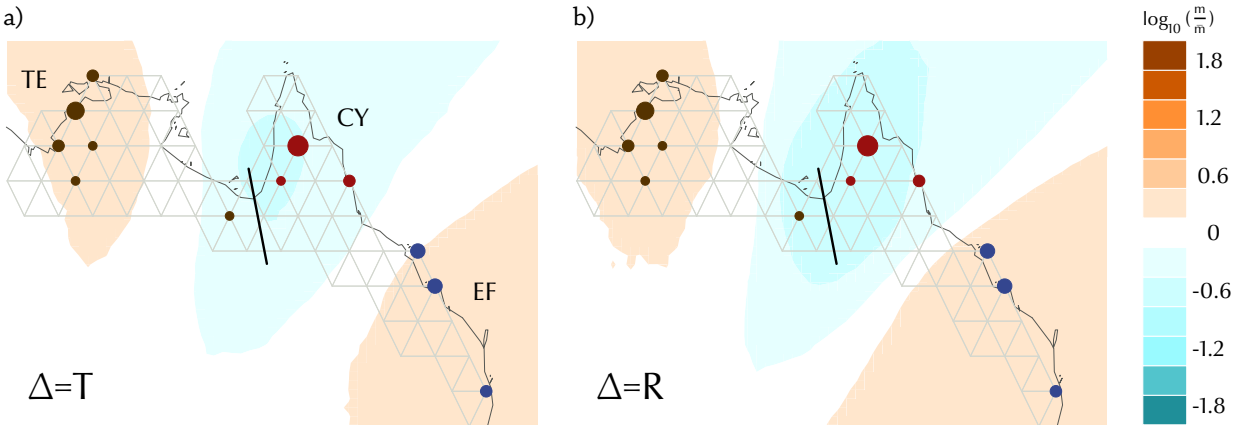


Figure 6.5: Estimated relative rates of effective migration for the RBFW dataset using two distance metric on the graph: a) expected coalescence time $\underline{T} = (T_{\alpha\beta})$; b) effective resistance $\underline{R} = (R_{\alpha\beta})$.

6.1.1 *What is the effect of the Carpentarian barrier on genetic differentiation?*

Since we plot relative migration rates, a completely white migration surface would correspond to uniform migration; the colors indicate deviations from the expectation under uniform migration.

For the RBFW dataset, the most interesting feature is the area of lower effective migration that roughly covers the Cape York (CY) biogeographic region and the Carpentarian barrier. This result is consistent with the hypothesis that the Carpentarian barrier affects genetic differentiation. It is also consistent with the hypothesis that the CY group has a slightly lower effective population size (similar to the simulations in Section 5.2). Furthermore, CY is relatively less similar to TE than it is to EF as CY and TE are separated by longer effective distance [i.e., darker blue color]. Although this can also be inferred from the PCA and STRUCTURE analysis, the effective migration plot combines information about genetic dissimilarities and geographic distances and thus is an intuitive representation of spatial patterns in genetic variation.

Finally, we show draws from the posterior distribution of effective migration [Figure 6.6]. Although in most instances the region of the Carpentarian barrier falls inside a tile of lower effective migration [relative to the rest of the habitat], there is a lot of variability in the location, shape and rate of the "barrier". One possible explanation is that the Carpentarian barrier does not have a strong effect on the genetic structure of this species. However, the RBFW dataset is small and it is also possible that our method

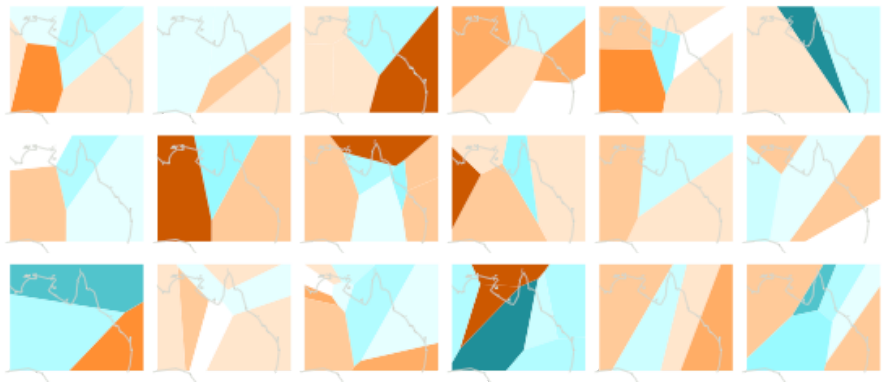


Figure 6.6: Draws from the posterior distribution of effective migration in red-backed fairywrens.

cannot detect a strong barrier effect with certainty.

6.2 *Forest and savanna elephants in Africa*

Here we present results for a dataset of African elephants sampled throughout the range of the species in Sub-Saharan Africa. The sample includes both forest elephants (*Loxodonta africana cyclotis*) and savanna elephants (*Loxodonta africana africana*). Both subspecies are under threat, partly from poaching, and the data were collected to help assign contraband tusks to their location of origin [Wasser et al., 2004, 2007].

There is observational and genetic evidence that forest and savanna elephants hybridize in the areas where their ranges meet [Wasser et al., 2004]. Therefore, we remove putative hybrids so that the dataset we analyze consists of 223 forest elephants and 896 savanna elephants genotyped at 16 microsatellites. These genetic markers were chosen in part because they can be isolated and amplified in samples of low quality and thus microsatellite DNA can be extracted from a small piece of tusk [Wasser et al., 2004].

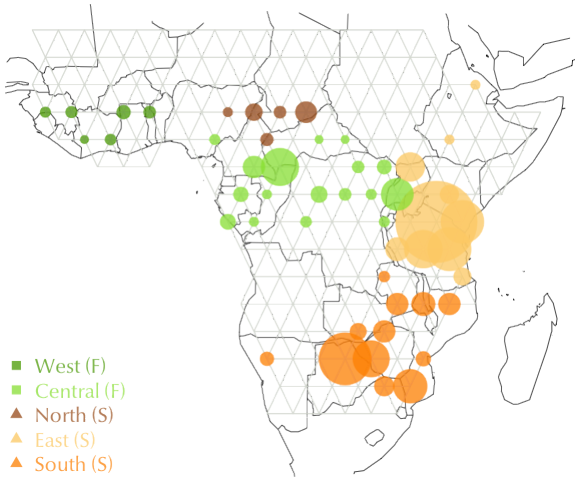


Figure 6.7: Irregular triangular grid (V, E) spanning the habitat of the African elephant. The map shows five regions as identified in [Wasser et al., 2004]. The west and central regions comprise the range of the forest elephant. The north, east and south regions comprise the range of the savanna elephant. Samples are assigned to the closest deme in the grid.

[Wasser et al., 2004] show that forest and savanna elephants can be accurately discriminated. This is also evident in the PC scatterplot of the sample covariance matrix [Figure 6.8] where the leading principal component separates forest (West, Central) and savanna (North, East, South) and explains 29% of the observed genetic variation. PCA analysis also indicates that there is more genetic diversity in forest than in savanna elephants and suggests no further population structure within the two subspecies. However, the sample configuration is very uneven with about 4 times savanna than forest elephants, so the PCA results might be biased [Section 4.5].

We applied our method to the data provided by [Wasser et al., 2004]. The results confirm that forest and savanna elephants are genetically differentiated enough to distinguish between the two subspecies. In Figure 6.9 we observe a prominent barrier in effective migration that curves through the habitat to separate the west and central regions (the range of forest elephants) from the north, east and south regions (the range of savanna) elephants.

Our model estimates migration rates to explain the overall sample structure. However, each genotyped site has its own genealogy and thus observed genetic distances vary across sites. With microsatellites, mutation rates are higher, and since more mutations mean more information about relative branch lengths, we can also fit the model at each microsatellite separately [Figure 6.10]. There is great variability in effective mi-

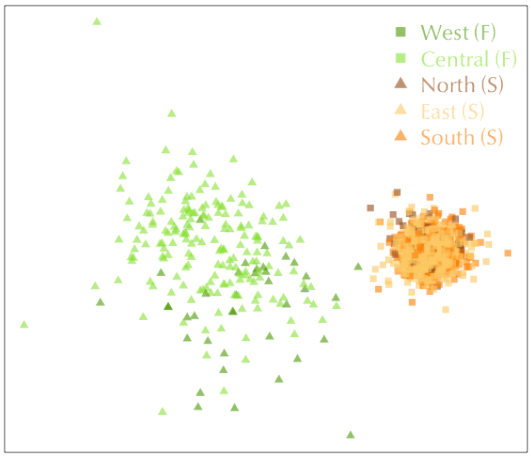


Figure 6.8: PCA analysis of the forest and savanna elephants (FS) data.

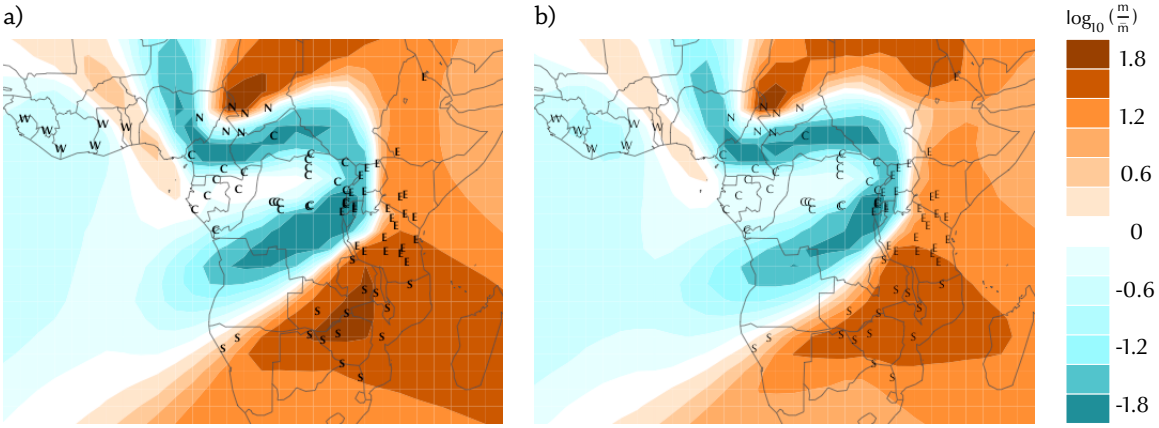


Figure 6.9: Effective migration rates for forest and savanna elephants (a) using all 16 microsatellites and (b) excluding the most variable locus.

gration across microsatellites. And the pattern of effective migration at the sixth locus produces most of the overall pattern, except for the relationship between the west and central regions — essentially, the relationship among forest elephants. This suggests that elephants can be categorized with high accuracy as forest or savanna based on just this one microsatellite.

We also split the sample into only forest and only savanna elephants to explore subtler population structure in each subspecies. The genetic variation in forest elephants is consistent with isolation by resistance with a very small bridge of higher effective migration between the west and central regions. The genetic variation in savanna elephants deviates from isolation by distance considerably: the central region is separated from the rest with a barrier of low effective migration while the south and east regions are more genetically similar than the large area they span would suggest.

6.2.1 STRUCTURE and GENELAND results

As a clustering method, GENELAND [Guillot et al., 2005] looks for distinct clusters and therefore sharp boundaries between them. Removing putative hybrids makes differences in allele frequencies between biogeographic regions easy to detect [Figure 6.12]. However, GENELAND does not explain the relationship between the regions — they are all distinct from each other.

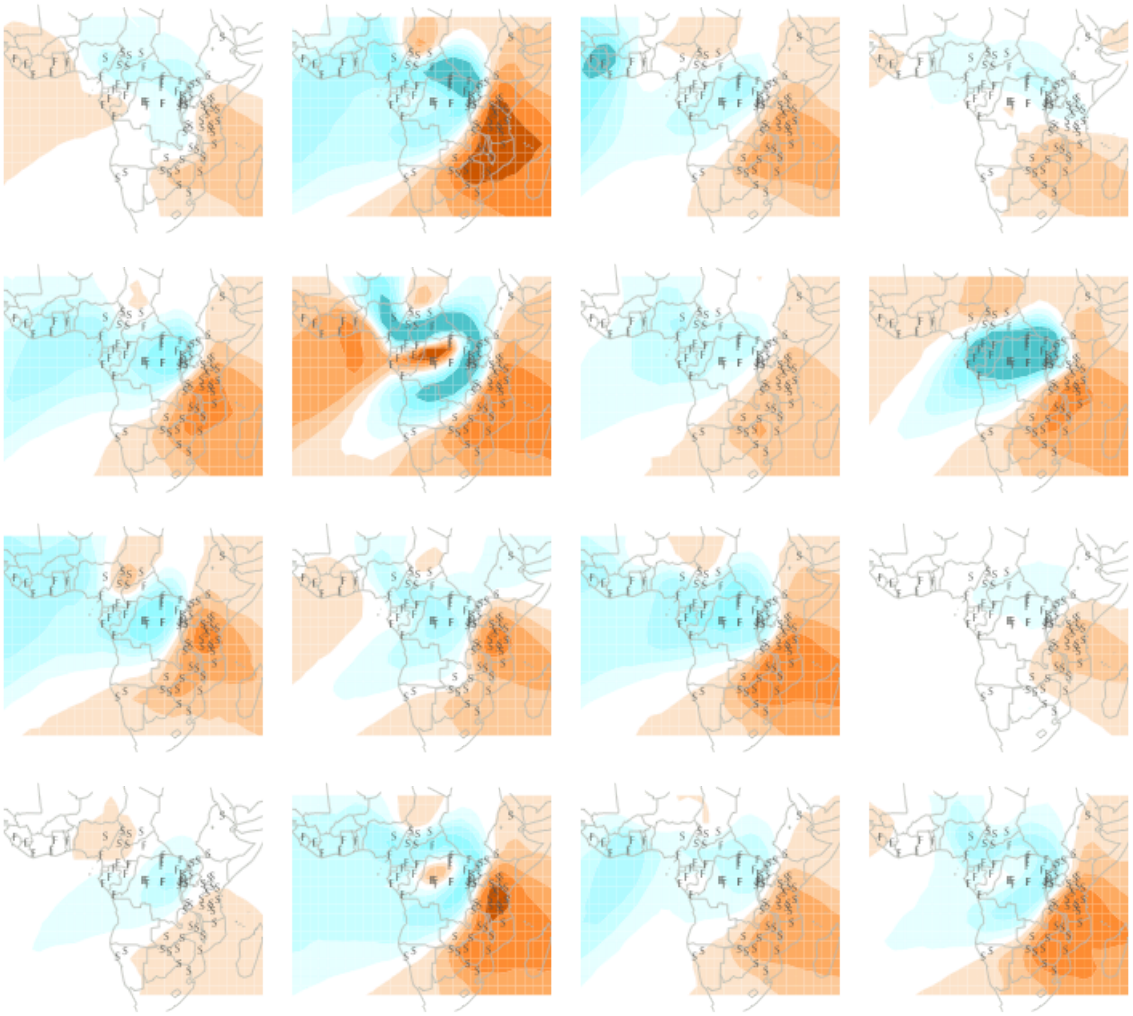


Figure 6.10: Effective migration rates at each of sixteen microsatellites. The 6th locus is most variable, presumably due to highest mutation rate.

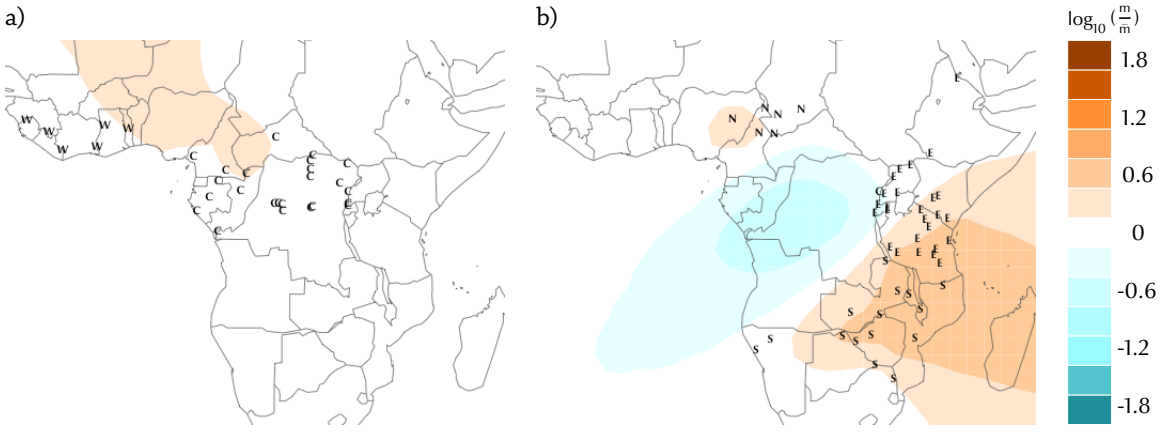


Figure 6.11: Effective migration rates for (a) only forest elephants and (b) only savanna elephants, using the same triangular grid as in Figure 6.7 and all 16 microsatellites.

On the other hand, STRUCTURE [Pritchard et al., 2000] with sampling location prior [Hubisz et al., 2009] provides intuition for the relationship between the five biogeographic regions [Figure 6.13]. It clearly detects the difference between forest elephants (west, central) and savanna elephants (north, east, south) as they fall into difference clusters. Furthermore, STRUCTURE shows some evidence for isolation by distance, particularly in savanna elephants. Most of these individuals are represented as weighted mixtures of four clusters that do not correspond to distinct geographic areas.

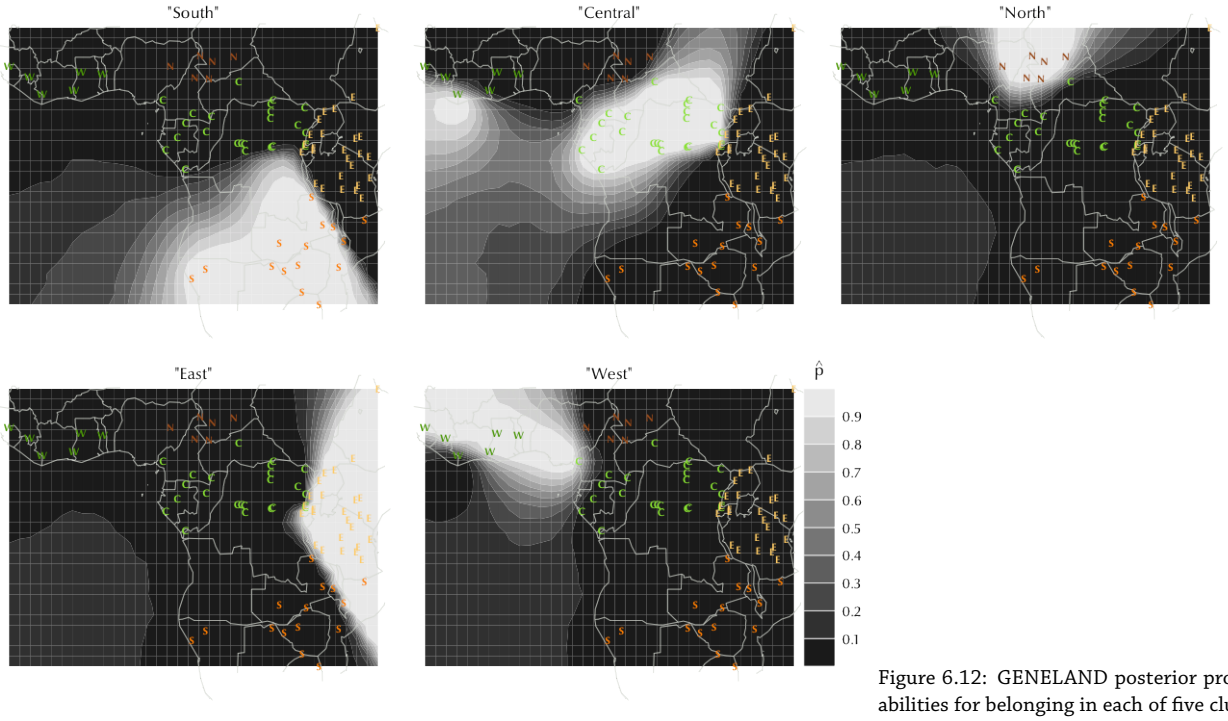


Figure 6.12: GENELAND posterior probabilities for belonging in each of five clusters, which correspond directly to the five biogeographic regions.

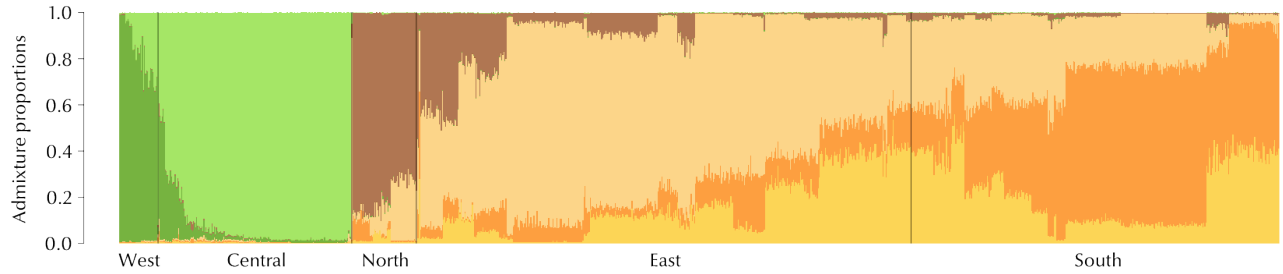


Figure 6.13: STRUCTURE membership proportions in six clusters when using sampling locations (not regions) to provide prior information for cluster assignments.

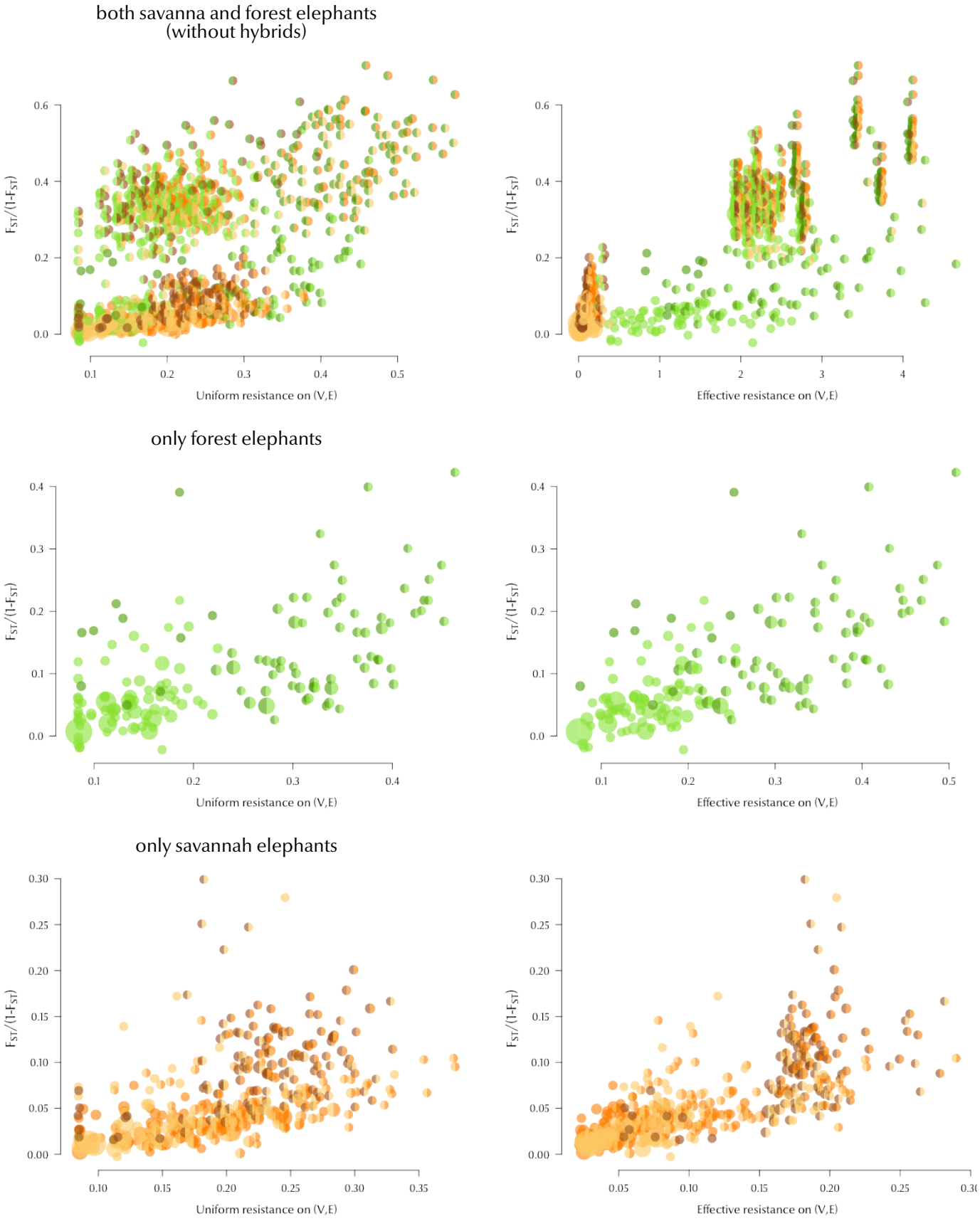


Figure 6.14: Scatterplots of genetic dissimilarities versus distances on the population graph with a) uniform migration and b) estimated effective migration.

6.3 *Human populations from Europe and Africa*

The genetic structure of human populations has been extensively studied since [Menozzi et al., 1978] first used PCA to summarize human genetic variation across continents. Here we analyze two large-scale genome-wide datasets. The European dataset is part of the POPRES collection [Nelson et al., 2008] and consists of 1387 individuals genotyped at 197,146 autosomal SNPs. Most samples were collected in Western Europe, so we analyze a subset of 1208 individuals from 15 countries. The Sub-Saharan African dataset consists of 314 individuals from 21 ethnic groups genotyped at 27,922 autosomal SNPs.

[Novembre et al., 2008, Lao et al., 2008] use PCA analysis to characterize the spatial structure of genetic variation within Europe and find a close correspondence between genetic and geographic distances, and hence, evidence for isolation by distance. In fact, [Novembre et al., 2008] shows that the two leading principal components are strongly correlated with latitude and longitude, respectively. [Wang et al., 2012] use their Procrustes method to analyze the population structure within Africa [as well as within Europe, Asia and world-wide] and also observe similarity between genetic and geographic maps, after excluding hunter-gatherer populations.

The PCA plots reveal that the human population structure in Europe and Africa is continuous: while individuals from the same group tend to cluster together, the overall arrangement qualitatively resembles the configuration of sampling locations [Figure 6.15]. The correspondence between the PCA projections and the geographic map can be improved with a rotation transformation such as Procrustes [Wang et al., 2010] but this cannot improve the PCA analysis — for example, by correcting for biased sampling.

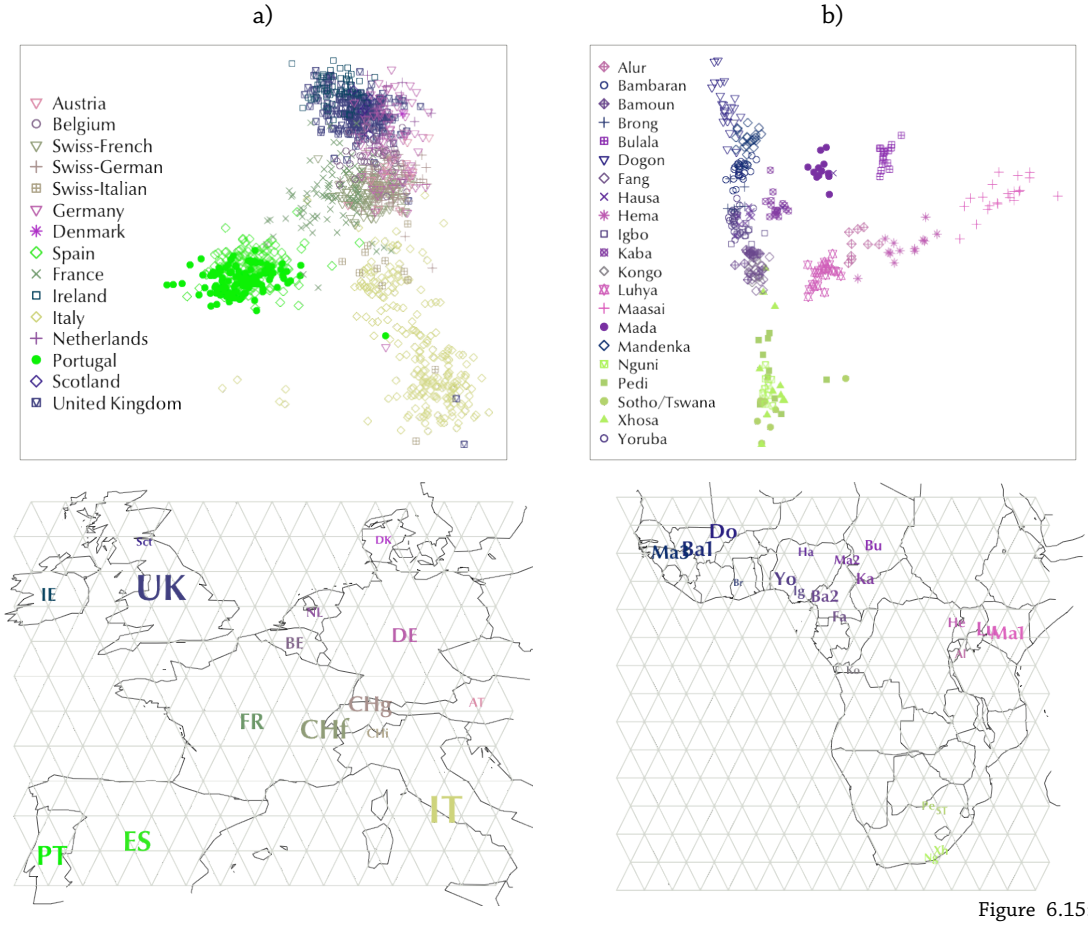


Figure 6.15: Sample configuration and PCA analysis of the European and African datasets. In the bottom row font size indicates relative sample size.

Figure 6.15 also illustrates the limitations of the sampling scheme. In both cases it is biased but, more importantly, geographic locations are implied as individuals from the same populations are automatically assigned to the same coordinates. [For the European dataset, population membership is determined based on grandparents' country of origin or self-repored country of birth [Novembre et al., 2008].] This is clearly not representative of human spatial distribution in either Europe or Africa and, furthermore, the geographic information might be too coarse to detect substructure within populations. [On the other hand, the stepping-stone model is discrete. Since observations are assigned to the nearest deme, the grid itself implies a limit on how much geographic resolution our method can represent.]

As [Novembre et al., 2008, Wang et al., 2012] have shown, the spatial structure of human genetic variation in both Europe and Africa exhibits broad isolation by distance as genetic differentiation tends to increase gradually with geographic distance [Figure 6.16; top row]. However, while geography explains some patterns of genetic differentiation, a homogeneous habitat (i.e., uniform migration) might not provide the best explanation for the observed data.

We applied our method to estimate the effective human migration in both Europe and Africa and plotted genetic differentiation against the inferred effective distances [Figure 6.16; bottom row]. The linear relationship between genetic dissimilarity and effective distance is stronger [r^2 increases from 33% to 85% for the European data, and from 24% to 91% for the African data]. However, since there are so many pairwise comparisons in the scatterplots, it is more instructive to analyze the inferred effective migration surface [Figure 6.17].

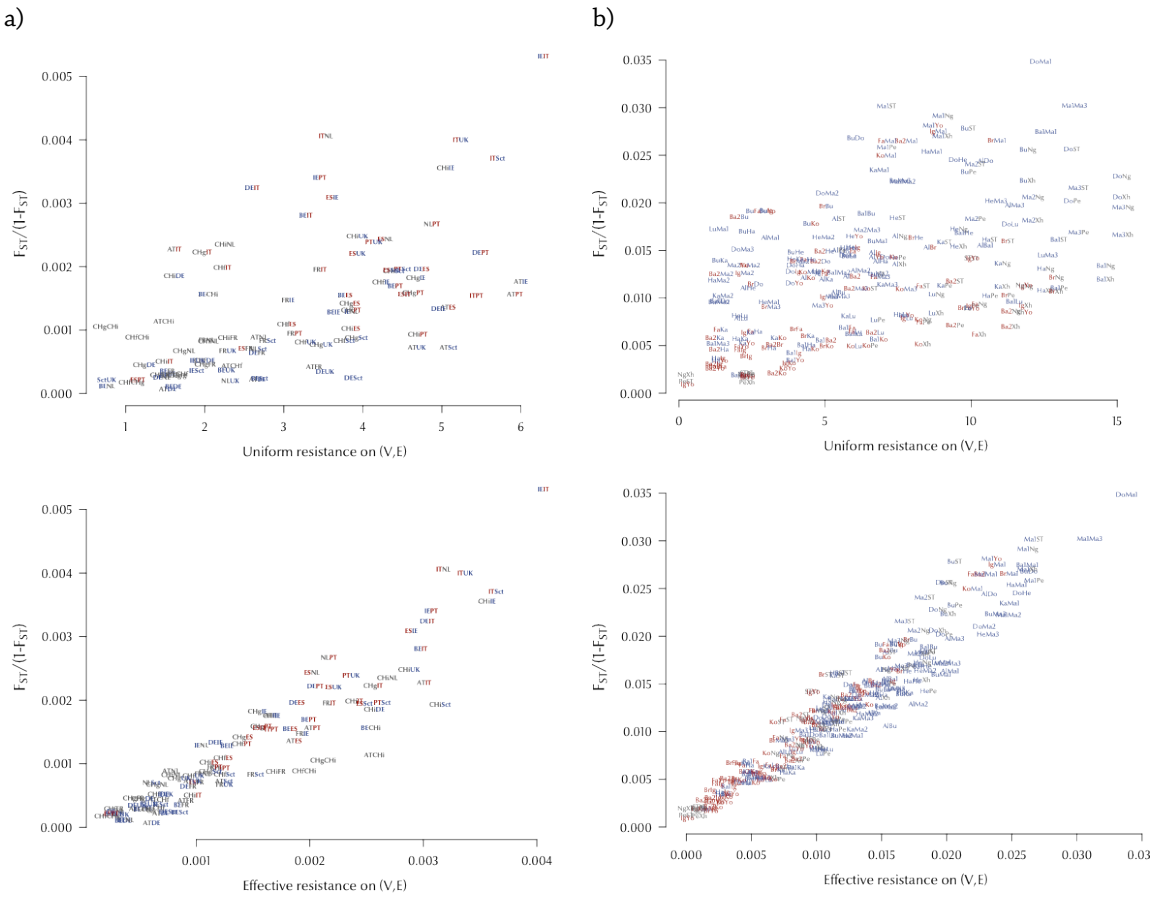


Figure 6.16: Genetic differentiation (linearized F_{ST}) versus resistance distance ($R_{\alpha\beta}$) with either uniform migration rates or estimated effective migration rates on the population graph (V, E). The colors, which match those in Figure 6.17, are chosen to emphasize the difference between the populations in red and those in blue.

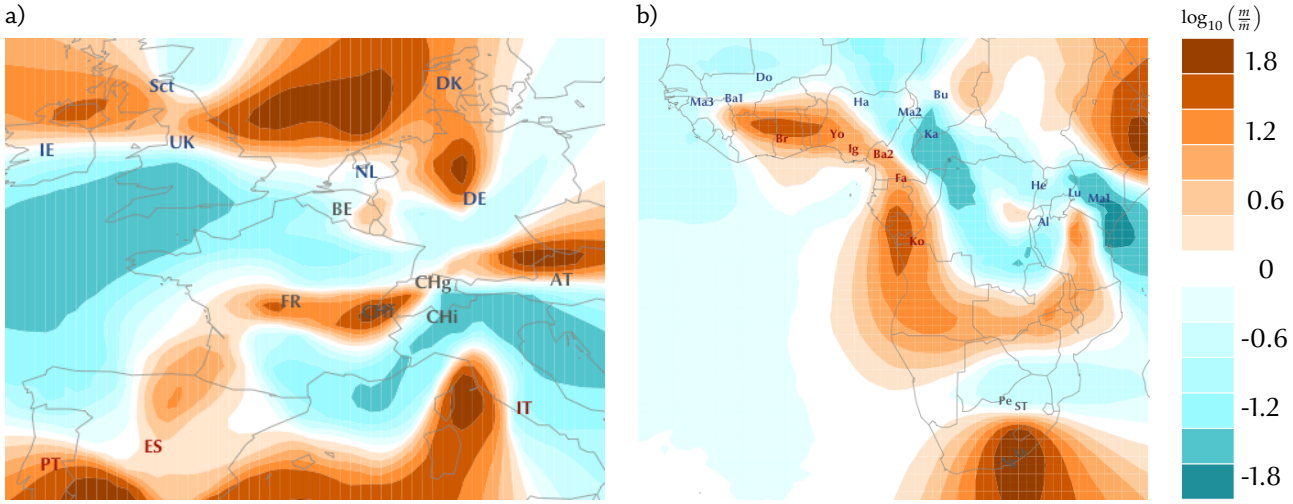


Figure 6.17: Inferred effective migration surfaces for the West European and Sub-Saharan African datasets.

From the inferred effective migration surface in Europe [Figure 6.17 a)] we can make several observations that are difficult to make from the PCA plot or the distance scatterplot. For example, the northern countries (Ireland, the UK, Scotland, Denmark, the Netherlands, Germany — in blue) are more genetically similar that we would expect based on the geographic distance alone. The same is true for the three southern countries (Portugal, Spain and Italy — in red). On the other hand, a barrier to effective migration separates the British Isles and the Iberian peninsula, and another barrier separates Italy and France [roughly where the Alps are]. However, we cannot conclude that the effect is due only to lower migration rates across bodies of water or mountain ranges. The observed patterns can also be influenced by differences in effective population size and other evolutionary processes. Finally, the inferred migration surface also suggests that there is more differentiation in the north/south direction rather than the east/west direction as the north and the south are separated by two areas of lower effective migration. This results is consistent with the hypothesis that a north/south cline is a distinguishing feature of population structure within Europe [Tian et al., 2008].

We can also make interesting observations from the inferred effective migration surface in Africa [Figure 6.17 b)]. There is higher effective migration along the Atlantic coast than in the interior of the continent, and therefore, inland populations (in blue) are more genetically dissimilar than coastal populations (in red). Consequently, there is more differentiation in the east/west direction than in the north/ south direction. This pattern can be observed in the PCA plot, as noted by [Wang et al., 2012], where populations along the coast cluster closer together, inland populations form more isolated clusters and the E/W-associated principal component explains twice as much variation as the N/S-associated one. On the other hand, the four Bantu speaking groups at the southern tip cluster together in the PCA plot but not in the effective migration surface. However, this might be the result of lower geographic resolution in that region: Pedi and Sotho/Tswana are assigned to the same deme, and similarly, Nguni and Xhosa are assigned to another deme. The first pair has lower genetic differentiation than the second [$F_{ST}(Pe, ST) = 0.0012, F_{ST}(Ng, Xh) = 0.0019$].

We use the program GENEPOP [Rousset, 2008] to compute pairwise F_{ST}s.

These patterns are present, to some extent, in the PCA plot and the distance scatterplot. But they are easy to observe only if we categorize the locations and color the points

appropriately. In contrast, the pattern is clear after the analysis of effective migration.

6.4 *Arabidopsis thaliana in Europe and North America*

Arabidopsis thaliana is a small flowering plant and a commonly studied model organism in population genetics. It has a broad natural range — Europe, Asia and North Africa — and now grows in North America as well. Although *A. thaliana* is a selfing plant with low gene flow, its genetic variation has significant spatial structure [Nordborg et al., 2005, Platt et al., 2010]. On the continental scale, in Europe there is broad isolation by distance with east-west gradient that has been interpreted as evidence for post-glaciation colonization [Nordborg et al., 2005]. On the other hand, in North America there is genome-wide linkage disequilibrium and haplotype sharing that have been interpreted as evidence for recent human introduction from Europe [Nordborg et al., 2005].

A large geographically referenced dataset is available from the Regional Mapping (RegMap) project [Horton et al., 2012]. We split the full dataset (1193 accessions, ~ 220, 000 SNPs) into two subsets — North America (180 plants) and Europe (823 plants) — which we analyze both separately and together. [We exclude plants from Asia because the continent is very sparsely sampled.]

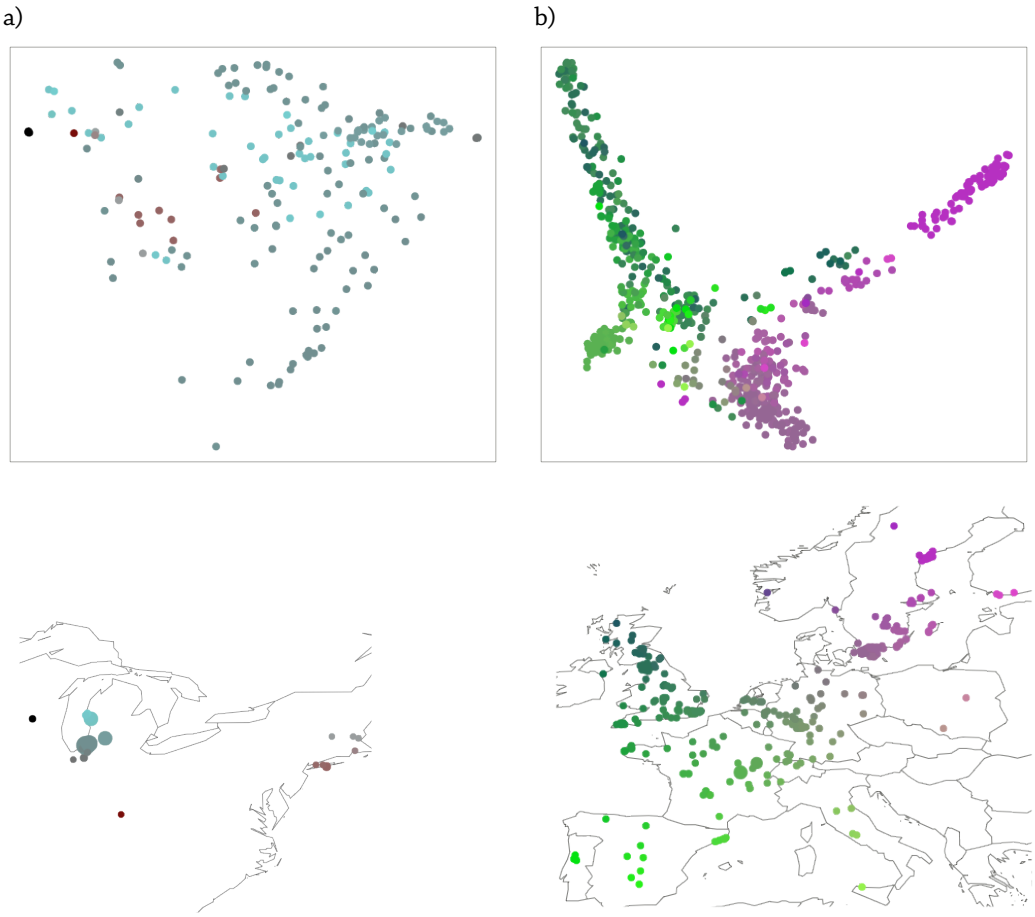


Figure 6.18: Sample configuration and PCA analysis of *Arabidopsis thaliana* data from the RegMap project; a) North America, b) Europe.

First we perform principal components analysis [Figure 6.18]. As we would expect if *A. thaliana* has different history in Europe and North America, there are differences in genetic variation on the two continents [Nordborg et al., 2005]. There is little pop-

ulation structure in the North American subset: the samples are separated to some extent in a north/south direction, with no obvious separation between samples from around Lake Michigan and those from the Atlantic coast despite the geographic distance. On the other hand, the population structure in the European subset is continuous, with some correspondence between genetic variation and geographic distribution, as we would expect under isolation by distance [Platt et al., 2010].

Next we apply our method to estimate the effective migration for *A. thaliana* in North America and Europe [Figure 6.19]. In North America, the two sampled regions — Lake Michigan and the Atlantic coast — are connected by a strip of high effective migration. This indicates that the regions are similar genetically even though they are far apart in space. [This is the opposite of what we expect under isolation by distance.] There is an area of higher effective migration at the south tip of Lake Michigan where most of the North American samples are collected. Therefore, our results are consistent with the observation that there is extensive haplotype sharing (which indicates identity by descent) not only within but also between sampling locations [Nordborg et al., 2005].

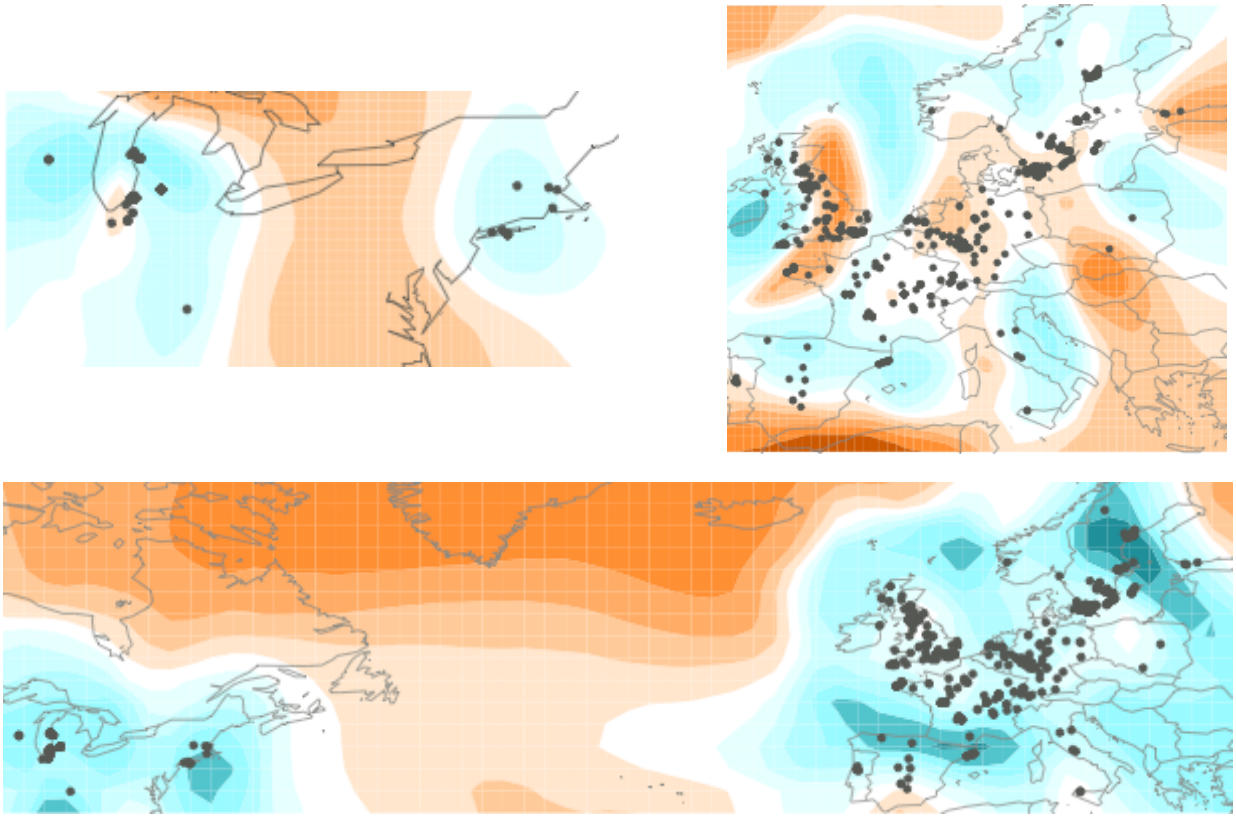


Figure 6.19: Inferred effective migration surfaces for *Arabidopsis thaliana* from two RegMap subsets; a) North America; b) Europe; c) North America and Europe combined.

In continental Europe, the overall pattern in broad isolation by distance, with small variability in effective migration rates. On the other hand, the north of the British Isles is separated from the rest of Britain which in turn is connected to northern France by an area of high migration. Our results are consistent with previous studies of the population structure of *A. thaliana*. [Platt et al., 2010] find that in Eurasia there is a strong trend of isolation by distance (at three distance scales) while in North America there is no relationship between geographic distance and allelic similarity (except at fine distance scale). And [Horton et al., 2012] observe that in the PCA plot most accessions from the British Isles are projected closest to France but some plants from Britain cluster with lines from Sweden. Our method summarizes and visualizes these patterns.

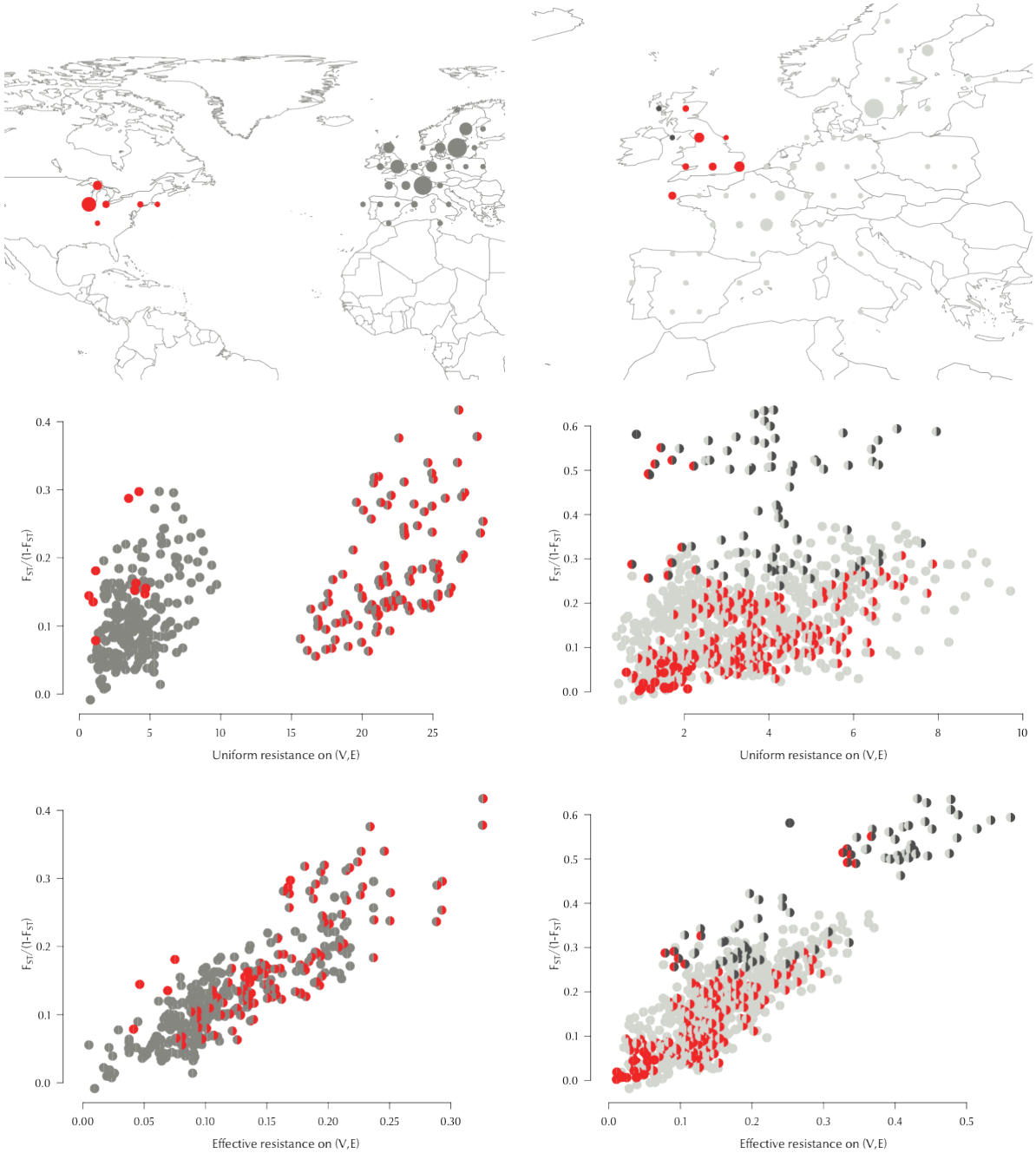


Figure 6.20: Genetic differentiation (linearized F_{ST}) versus resistance distance ($R_{\alpha\beta}$) with either uniform migration rates or estimated effective migration rates on the population graph (V, E). The samples are assigned to a regular triangular grid [because the designation of nations as populations is not relevant] and the colors are chosen to emphasize interesting patterns [that correspond to regions with strong deviation from isolation by distance].

In the combined analysis, the overall pattern is a prominent corridor of high effective migration across the Atlantic ocean. While effective migration rates are symmetric, this supports the hypothesis that recent directed migration introduced *A. thaliana* from Europe to the New World. On this larger scale, effective migration rates within North America and Europe are low [as we would expect in a plant species with low gene flow].

[Platt et al., 2010] considers two models of the continuous population structure of *A. thaliana* — a model with constant uniform migration and another with uniform migration and a single shift in dispersal rate — and concludes that neither version is a good fit for the observed patterns of genetic dissimilarities. That is, even though the overall pattern is consistent with isolation by distance, there might be deviations from uniform migration (or too much noise). We can observe such complex details in the effective migration surface, e.g., the British Isles in Figure 6.19. When we combine the two samples the strongest signal in the data is the genetic similarity between the two continents even though they are separated by the expanse of the Atlantic ocean. This illustrates how we can observe finer details at smaller scales because effective migration rates are parametrized relative to the overall mean log rate.

6.5 Conclusion

Genetic variation in natural populations often exhibits spatial structure as genetic similarity tends to decay with geographic distance. However, this relationship is often not homogeneous and the distribution of similarities across the habitat contains information about the evolutionary and ecological history of the population.

Visualization is an important tool for detecting and understanding patterns of population structure. We have developed a model-based method for estimating and visualizing effective migration to explain observed deviations from homogeneous dispersal and isolation by distance. It represents the population as a triangular grid of discrete components and its effective migration as a colored partition of a two-dimensional map.

Our method is particularly useful for characterizing continuous population structure (even though the underlying stepping-stone model is discrete) because it models a spatially distributed population instead of a collection of distinct and isolated subpopulations. Neither is it necessary to categorize samples into regional groups to compute measures of differentiation such as F_{ST}. [Nevertheless, assigning colors and names to groups of genetically similar individuals, in areas of high effective migration, can be helpful for subsequent analysis.] If spatial structure is continuous, clustering samples into biogeographic regions might not be a well-defined problem. In this scenario cluster-based methods such as STRUCTURE and GENELAND may be inappropriate.

Our method also offers some advantages over PCA analysis. PCA produces two-dimensional visual summaries of observed genetic variation and can capture both continuous and discrete population structure at the sample level. In contrast, our method produces a visual representation of geographic and genetic information at the population level. Consequently, it is easier to make qualitative comparisons between populations or between geographic regions, in terms of both geographic and genetic distances. And our method can detect deviations from uniform migration (and hence, isolation by distance) that are not clearly evident in PC projections because PCA is strongly affected by sampling bias and does not estimate relevant demographic parameters.

7
Appendices

7.1 Properties of the stepping-stone model of population subdivision

The goal of this section is to derive the system of linear equations (2.15) for the expected pairwise coalescence times $T = (T_{\alpha\beta})$ as a function of the migration rates $M = (m_{\alpha\beta})$ and the coalescence rates $q = (q_\alpha)$ in the stepping-stone model.

7.1.1 Probabilities of identity by descent

In population genetics, the probability of identity is a measure of relatedness due to shared ancestry. The concept of identity can be defined as either the event that the lineages have the same ancestor in a reference population at a specified time in the past or the event that no mutations have occurred since the lineages diverged from their most recent common ancestor. We use the second definition of identity known as *identity by state* [versus *identity by descent*].

Let $\phi_{\alpha\beta}(\theta)$ be the probability of identity by descent for two distinct lineages drawn at random from demes α and β. The parameter $\theta = 2N_0u$ is the mutation rate per $2N_0$ generations for a single lineage, or equivalently, the total mutation rate per N_0 generations for a pair of lineages.

To derive expressions for $\phi_{\alpha\beta}(\theta)$ for every pair (α, β), consider the history of a sample of size 2 backwards in time. Let $x(t) = \{x_\alpha^{(t)}\}$ be the state of the ancestral process t generations ago when the sample has $x_\alpha^{(t)}$ ancestors in deme α. It is convenient to consider time in units of N_0 generations. On this timescale and under certain assumptions about reproduction and migration, the discrete-time ancestral process $\{x(t) : t = 0, 1, \dots\}$ converges to a continuous-time ancestral process $\{x(t) : t \geq 0\}$, called the *structured coalescent* [Notohara, 1990, 1993]. Mutations are generated by a Poisson process with intensity θ such that in t units of time a lineage accummulates $K \sim \text{Po}(\theta t)$ mutations.

To derive the probability of identity for the pair (α, β), consider the first event that results in a change of state. The initial state is $x(0) = \{x_\alpha^{(0)} = 1, x_\beta^{(0)} = 1, x_\gamma^{(0)} = 0 : \gamma \neq \alpha, \gamma \neq \beta\}$. If the two lineages are drawn from the same deme, i.e., $\alpha = \beta$, the first event can be a coalescence with rate q_α, a mutation with rate θ, or a migration to deme γ with rate $2m_{\alpha\gamma}$. If a mutation occurs, the lineages are no longer identical by descent. Therefore, under equilibrium,

Since the process starts with two lineages in α and the migration rate from α to γ is $m_{\alpha\gamma}$ for a single lineage, the total rate of movement is $2m_{\alpha\gamma}$. Similarly, the combined mutation rate is θ.

$$\phi_{\alpha\alpha}(\theta) = \frac{q_\alpha}{\theta + q_\alpha + 2m_\alpha} + \sum_{\gamma:\gamma\neq\alpha} \frac{2m_{\alpha\gamma}}{\theta + q_\alpha + 2m_\alpha} \phi_{\alpha\gamma}(\theta), \tag{7.1}$$

where $m_\alpha = \sum_{\gamma:\gamma\neq\alpha} m_{\alpha\gamma}$ is the total rate of migration out of α. More precisely, since the coalescent proceeds backwards in time, $m_{\alpha\gamma}$ is the rate at which offspring in α have

parents from γ and m_α is the total rate of "outside-deme" parentage.

When the two lineages are drawn from two different demes, i.e., $\alpha \neq \beta$, they cannot coalesce in a single step. In this case, the first event can be a mutation with rate θ, a migration from deme α to deme γ with rate $m_{\alpha\gamma}$, or a migration from deme β to deme γ with rate $m_{\beta\gamma}$. Under equilibrium,

$$\phi_{\alpha\beta}(\theta) = \sum_{\gamma:\gamma\neq\alpha} \frac{m_{\alpha\gamma}}{\theta + m_\alpha + m_\beta}\phi_{\gamma\beta}(\theta) + \sum_{\gamma:\gamma\neq\beta} \frac{m_{\beta\gamma}}{\theta + m_\alpha + m_\beta}\phi_{\alpha\gamma}(\theta). \tag{7.2}$$

Equations (7.1) and (7.2) represent a system of linear equations for the probabilities of identity by descent in terms of the mutation rate θ, the coalescence rates q_α and the migration rates $m_{\alpha\beta}$. In matrix notation,

$$\operatorname{diag}\{q\}[\operatorname{diag}\{\Phi\} - I] = [M - (\theta/2)I]\Phi + \Phi[M - (\theta/2)I]'. \tag{7.3}$$

Here $M = (m_{\alpha\beta})$ is the infinitesimal generator of the migration process with diagonal entries $-m_\alpha = -\sum_{\gamma:\gamma\neq\alpha} m_{\alpha\gamma}$, $\Phi \equiv \Phi(\theta) = (\phi_{\alpha\beta}(\theta))$ is the matrix of probabilities of identity at fixed mutation rate θ, and $q = (q_\alpha)$ is the vector of coalescence rates.

7.1.2 *Expected pairwise coalescence times*

A linear system for the expected pairwise coalescence times can be derived correspondingly. Since by definition ϕ is the probability that no mutation occurs in either lineage before coalescence at time t,

$$\phi(\theta) = \mathrm{P}\{K = 0\} = \mathrm{E}\left\{e^{-\theta t}\right\}. \tag{7.4}$$

That is, the probability of identity ϕ is the Laplace transform of the coalescence time t [Hudson, 1990]. Therefore,

$$\mathrm{E}\{t\} = -\phi'(0) \text{ where } \phi' = \frac{\partial}{\partial\theta}\phi. \tag{7.5}$$

To obtain a system for the expected coalescence times, differentiate equations (7.1) and (7.2) with respect to the mutation rate θ and evaluate at $\theta = 0$. The result is

$$1 = (q_\alpha + 2m_\alpha)T_{\alpha\alpha} - \sum_{\gamma:\gamma\neq\alpha} 2m_{\alpha\gamma}T_{\alpha\gamma} \quad \text{and} \tag{7.6a}$$

$$1 = (m_\alpha + m_\beta)T_{\alpha\beta} - \sum_{\gamma:\gamma\neq\alpha} m_{\alpha\gamma}T_{\gamma\beta} - \sum_{\gamma:\gamma\neq\beta} m_{\beta\gamma}T_{\alpha\gamma}. \tag{7.6b}$$

Equivalently, in matrix notation,

$$\operatorname{diag}\{q\}\operatorname{diag}\{T\} - MT - TM' = 11'. \tag{7.7}$$

[Here $11'$ is a $d \times d$ matrix of 1s.] This method for deriving equations (7.6a) and (7.6b) is developed in [Bahlo and Griffiths, 2001]. Alternatively, [Hey, 1991] constructs a Markov chain with $d(d+1)/2$ non-absorbing states for each unique pair (α, β). The set of states includes d homoallelic states, when the two lineages are in the same deme, and $d(d-1)/2$ heteroallelic states, when two lineages are in different demes. There is also an absorbing state, which corresponds to coalescence. Transition probabilities between all these states reflect the migration rates $m_{\alpha\beta}$ and coalescence rates q_α.

Furthermore, since the population evolves under equilibrium, migration is conservative and $M'q^{-1} = 0$ by definition. If we multiply equation (7.7) by $(q^{-1})'$ on the left

and by q^{-1} on the right, we obtain

$$1' \operatorname{diag}\{T\} q^{-1} = (1' q^{-1})^2 \quad \Leftrightarrow \quad \sum_\alpha T_{\alpha\alpha}(N_\alpha/N_0) = \left(\sum_\alpha N_\alpha/N_0\right)^2$$
$$T_0 \equiv \sum_\alpha T_{\alpha\alpha}(N_\alpha/N_T) = N_T/N_0 = d \qquad (7.8)$$

where T_0 is the [weighted] average within-deme coalescence time, $N_T = \sum_\alpha N_\alpha$ is the total population size and $N_0 = N_T/d$ is the coalescent timescale. Therefore, under conservative migration, the *average* within-deme coalescence time does not depend on the exact pattern and rates of migration [Strobeck, 1987]. If migration is isotropic — a much stronger assumption — the within-deme coalescence times $T_{\alpha\alpha}$ for all demes α do not depend on the migration process.

7.2 Distance matrices

Here we discuss distance matrices, also called dissimilarity matrices, and review some relevant properties. Two examples of a distance matrix are the matrix of expected pairwise coalescence times, $\underline{T}$, and the matrix of effective resistance distances, $\underline{R}$.

First we state two equivalent definitions of a distance matrix.

Definition 7.1 *The matrix $D = (d_{ij}^2)$ is a distance matrix if there exist squared lengths $\ell = (\ell_i^2) \in \mathbb{R}_+^n$ such that*

$$\ell 1' + 1\ell' - D \succcurlyeq 0. \qquad (7.9)$$

$\mathbb{S}^n$ is the set of symmetric $n \times n$ matrices; $\mathbb{S}_+^n$ is the set of symmetric $n \times n$ matrices with nonnegative elements.

Definition 7.2 *The matrix $D = (d_{ij}^2)$ is a distance matrix if there exists pairwise similarities $S = (S_{ij}) \in \mathbb{S}_+^n$ such that*

$$d_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}. \qquad (7.10)$$

Let $X = (x_1, \dots, x_n)' \in \mathbb{R}^{n\times p}$ represent n points in p-dimensional inner product space. For example, in the setting of analyzing population structure, x_i is a genotype vector of p polymorphic sites. Then the squared distance between points i and j is given by

$$d_{ij}^2 = \langle x_i - x_j, x_i - x_j\rangle = \langle x_i, x_i\rangle + \langle x_j, x_j\rangle - 2\langle x_i, x_j\rangle \equiv \ell_i^2 + \ell_j^2 - 2S_{ij}, \qquad (7.11)$$

where $S_{ij} = \langle x_i, x_j\rangle$ is the inner product between two vectors in $\mathbb{R}^p$, and $S = XX'$ is positive definite as a Gram matrix. In matrix notation,

$$D = \operatorname{diag}\{S\}1 + 1\operatorname{diag}\{S\}' - 2S. \qquad (7.12)$$

D is nonnegative with 0s on the main diagonal because the dissimilarity of a point with itself is trivially 0.

Clearly the similarity matrix S contains more information about X than the dissimilarity matrix D: $\operatorname{diag}\{S\} = \ell$ while $\operatorname{diag}\{D\} = 0$. That is, S captures the *absolute* position of each point in the space (the length ℓ_i is the distance to the center O) while D reflects only the *relative* difference for each pair of points.

Theorem 7.1 *The matrix $D \in \mathbb{D}^n$ is a distance matrix if and only if it is conditionally negative definite.*

Sketch of proof.

- Suppose that D is a distance matrix. For every vector $\alpha \in \mathbb{R}^n$ such that $1'\alpha = 0$ (that is, α is a contrast)

$$0 \le 2\alpha' D\alpha = \alpha'(\ell 1' + 1\ell' - D)\alpha \qquad (7.13a)$$
$$= \alpha'\ell(1'\alpha) + (\alpha' 1)\ell'\alpha - \alpha' D\alpha = -\alpha' D\alpha. \qquad (7.13b)$$

  Therefore, D is conditionally negative definite.

- Suppose that D is conditionally negative definite. Choose a vector $w \in \mathbb{R}^n$ such that $1'w = 1$ and define

$$P = I - 1w', \tag{7.14a}$$

$$S = -\frac{1}{2}PDP' = -\frac{1}{2}(I - 1w')D(I - w1'). \tag{7.14b}$$

Then P is a centering matrix such that

$$PP = I - 1w' - 1w' + 1(w'1)w' = I - 1w' = P, \tag{7.15a}$$

$$w'Px = w'x - (w'1)w'x = w'x - w'x = 0 \tag{7.15b}$$

for every $x \in \mathbb{R}^n$. That is, P is an orthogonal projection onto the hyperplane $\{w\}^\perp$. Furthermore, $(I - w1')x$ is a contrast and

That is, $1'(I - w1')x = 0$.

$$x'(I - 1w')D(I - w1')x = -2x'Sx \leq 0 \tag{7.16}$$

since D is conditionally negative definite. Therefore, S is a positive definite matrix and it has a decomposition $S = YY'$. It is straightforward to show that

$$D = \operatorname{diag}\Big\{-\frac{1}{2}PDP'\Big\}1' + 1\operatorname{diag}\Big\{-\frac{1}{2}PDP'\Big\}' + PDP'. \tag{7.17}$$

That is, the vectors $Y = (y_1, \dots, y_n)'$ generate the distance matrix D. However, note that the similarity matrix S depends on the choice of w. It is not surprising that D does not determine S (nor Y) uniquely since it contains information only about relative differences.

Using equation (7.12) the (i,j)-th element is

$$\begin{aligned}
&-\tfrac{1}{2}e_i(PDP')e_i - \tfrac{1}{2}e_j(PDP')e_j + e_i(PDP')e_j \\
&= -\tfrac{1}{2}[w'Dw - w'De_i - e_i'Dw] \\
&\quad -\tfrac{1}{2}[w'Dw - w'De_j - e_j'Dw] \\
&\quad + w'Dw - w'De_i - e_i'Dw + D_{ij} = D_{ij}
\end{aligned}$$

where $D_{ii} = 0$ and e_i is the i-th standard basis vector.

The vector w determines the position of the origin O. For example, $w = 1/n$ corresponds to placing the origin at the centroid (the center of mass) $1'Y/n = \bar{y}$ and $w = e_i$ —at the ith point $e_i'Y = y_i$. Different decompositions $S = YY'$ give different orientations about the origin O.

The condition $1'w = 1$ implies that w is a vector of weights.

□

Now we consider the special case when the lengths ℓ_i are all equal to r for some $r > 0$ and thus the points x_i are the same distance from the center O. Geometrically, the points lie on the circumference of a sphere with radius r in $\mathbb{R}^p$ [Gower, 1985] and so D is called a *spherical* distance metric. This puts a constraint on the choice of w. In general,

A covariance matrix is a circumhypersphere with radius σ^2 and a correlation matrix – with radius 1.

$$\operatorname{diag}\{S\} = \frac{1}{2}\operatorname{diag}\big\{1w'D + Dw1' - 1w'Dw1' - D\big\} \tag{7.18a}$$

$$= Dw - \frac{1}{2}(w'Dw)1. \tag{7.18b}$$

$\operatorname{diag}\{D\} = 0$ and $\operatorname{diag}\{w1'\} = w$.

If $\ell = r^2 1$, then $\operatorname{diag}\{S\} = \ell = r^2 1$. Therefore,

$$Dw - \frac{1}{2}(w'Dw)1 = r^2 1. \tag{7.19}$$

If D is nonsingular, D^{-1} exists and we can right-multiply by D^{-1}. Then

$$w = \Big(\tfrac{1}{2}w'Dw + r^2\Big)D^{-1}1. \tag{7.20}$$

Recall that w satisfies $w'1 = 1$, so that

$$1'w = \Big(\tfrac{1}{2}w'Dw + r^2\Big)1'D^{-1}1 = 1 \tag{7.21}$$

This implies $1'D^{-1}1 \neq 0$ and

$$w = \frac{D^{-1}1}{1'D^{-1}1}, \qquad r^2 = \frac{1/2}{1'D^{-1}1}. \tag{7.22}$$

In this case $P = I - 11'D^{-1}/1'D^{-1}1$ is the orthogonal projection onto the hyperplane $\{D^{-1}1\}^\perp$.

Thus we have proved the following

Corollary 7.1 *Suppose that $D \in \mathbb{D}^n$ is a distance matrix such that* $\det\{D\} \neq 0$. *Then* $1'D^{-1}1 > 0$.

7.3 Conditional definite matrices

Here we derive a sufficient condition for positive definiteness of the covariance matrix $\Sigma = 11' - \lambda\Delta$ where $\Delta \in \mathbb{D}^n$ is a distance matrix [or more generally, a conditionally negative definite matrix] and λ is a positive constant. The derivation is based on [Bapat and Raghavan, 1997] and uses the Spectral Theorem which states that $\Sigma \succ 0$ if and only if all its eigenvalues are positive.

Definition 7.3 *A matrix $\Delta \in \mathbb{S}^n$ is conditionally negative definite if*

$$\alpha'\Delta\alpha \leq 0 \tag{7.23}$$

for all $\alpha \in \mathbb{R}^n$ such that $\sum_i \alpha_i = \alpha' 1 = 0$. Thus, conditional negative definiteness is equivalent to negative definiteness on the subspace $\{1\}^\perp$.

Theorem 7.2 *A conditionally negative definite (c.n.d) matrix $\Delta \in \mathbb{S}^n$ has at most one positive eigenvalue.*

Sketch of proof. We consider the c.n.d. case. Suppose to the contrary that Δ has two positive eigenvalues $u_1 > 0$ and $u_2 > 0$ with corresponding eigenvectors x and y. Without loss of generality, we can assume that the eigenvectors are normalized so that $\sum_i x_i = \sum_i y_i \Leftrightarrow \sum_i (x_i - y_i) = 0$. That is, $(x-y)'1 = 0$ and $x - y$ is a contrast. Furthermore,

$$\begin{aligned}
(x-y)'\Delta(x-y) &= x'\Delta x + y'\Delta y - y'\Delta x - x'\Delta y && \text{(7.24a)}\\
&= u_1 x'x + u_2 y'y - u_1 y'x - u_2 x'y && \text{(7.24b)}\\
&= u_1 x'x + u_2 y'y > 0 && \text{(7.24c)}
\end{aligned}$$

$x \perp y \Leftrightarrow x'y = 0$

since $u_1, u_2 > 0$. This contradicts the definition of conditionally negative definite matrices. □

A distance matrix Δ is nonnegative, with main diagonal of 0s, and is also conditionally negative definite by Theorem 7.1.

Corollary 7.2 *Suppose that Δ is a nonnegative, nonzero symmetric matrix. Then Δ has at least one positive eigenvalue.*

Sketch of proof. Since Δ is symmetric, by the Spectral Theorem it has real eigenvalues $u = \{u_i\}$. Furthermore, $\operatorname{tr}\{\Delta\} = \sum_{i=1}^n u_i \geq 0$. The trace of Δ is nonnegative because Δ is nonnegative; its eigenvalues are not all zero because Δ is nonzero. Since $\sum_i u_i \geq 0$, at least one of the eigenvalues is positive. □

Corollary 7.3 *Suppose that Δ is a nonnegative, nonzero, conditionally negative definite matrix. Then Δ has exactly one positive eigenvalue.*

Sketch of proof. By Theorem 7.2 Δ has at most one positive eigenvalue while by Corollary 7.3 it has at least one positive eigenvalue. Therefore, it has exactly one positive eigenvalue. If Δ is strictly c.n.d, its other $n-1$ eigenvalues are negative. □

So far we know that Δ is both conditionally negative definite and nonnegative, and therefore, it has exactly one positive eigenvalue. On the other hand, $\Sigma = 11' - \lambda\Delta$ is conditionally positive definite: for all α such that $\alpha' 1 = 0$, we have

$$\alpha'\Sigma\alpha = \alpha'(11' - \lambda\Delta)\alpha = (\alpha' 1)^2 - \lambda(\alpha'\Delta\alpha) \geq 0. \tag{7.25}$$

Therefore, Σ has at most one negative eigenvalue. Finally, by the matrix-determinant lemma for a rank-one update,

$$\prod_{i=1}^{n} u_i^* = \det\{\Sigma\} = \left(1 - \frac{1'\Delta^{-1}1}{\lambda}\right)\det\{-\lambda\Delta\} = \left(1 - \frac{1'\Delta^{-1}1}{\lambda}\right)\prod_{i=1}^{n}(-\lambda)u_i, \tag{7.26}$$

where $u = \{u_i\}$ are the eigenvalues of Δ, $u^* = \{u_i^*\}$ are the eigenvalues of Σ.

To ensure that the u_i^*s are positive, we use the fact that the product on the left in equation (7.26) has at most one negative term while the product on the right has exactly one negative term. Therefore, a necessary and sufficient condition for $\Sigma \succcurlyeq 0$ is that λ satisfies

$$1 - \frac{1'\Delta^{-1}1}{\lambda} \leq 0. \tag{7.27}$$

7.4 *Restricted maximum likelihood (REML) in the general case*

Consider the model $Y \sim \mathrm{N}(X\beta, \Sigma)$ with design matrix X and covariance matrix Σ. Let $K \in \mathbb{R}^{n\times p}$ be a basis for the mean space and $L \in \mathbb{R}^{(n-p)\times n}$ be a basis for the residual space. For example, $K = X$ if the design matrix has full rank, or otherwise, K is p linearly independent columns of X. By construction $LK = 0$ and $\ker\{L\} = \operatorname{span}\{K\}$. Also let $Q[\Sigma]$ be the unique orthogonal projection with kernel K given by

$$Q = I - K(K'\Sigma^{-1}K)^{-1}K'\Sigma^{-1}. \tag{7.28}$$

[McCullagh, 2009] shows that regardless of the choice for L is, Q has an equivalent characterization given by

$$Q = \Sigma L'(L\Sigma L')^{-1}L. \tag{7.29}$$

To prove this, it is sufficient to show that

- $\Sigma L'(L\Sigma L')^{-1}L$ is a projection:
  $QQ = \left(\Sigma L'(L\Sigma L')^{-1}L\right)\left(\Sigma L'(L\Sigma L')^{-1}L\right) = \Sigma L'(L\Sigma L')^{-1}L = Q$
- $\Sigma L'(L\Sigma L')^{-1}L$ is self-adjoint with respect to the inner product $\langle u, v\rangle = u\Sigma^{-1}v$:
  $Q'\Sigma^{-1} = \left(\Sigma L'(L\Sigma L')^{-1}L\right)'\Sigma^{-1} = L'(L\Sigma L')^{-1}L = \Sigma^{-1}\left(\Sigma L'(L\Sigma L')^{-1}L\right) = \Sigma^{-1}Q$
- $\ker\left\{\Sigma L'(L\Sigma L')^{-1}L\right\} = \{K\}$: $QK = \left(\Sigma L'(L\Sigma L')^{-1}L\right)K = \Sigma L'(L\Sigma L')^{-1}L(LK) = 0$

The orthogonal projection with kernel K (i.e., the orthogonal projection onto the residual space) is unique, so (7.28) = (7.29).

To rewrite the Wishart log-likelihood in equation (4.13), we derive the following expressions involving Σ, Q, L and K.

$$\begin{aligned}
\det\{L\Sigma L'\}^{-1}\det\{LL'\} &= \det\left\{(L\Sigma L')^{-1}(LL')\right\} \\
&= \operatorname{Det}\left\{L'(L\Sigma L')^{-1}L\right\} \\
&= \operatorname{Det}\left\{\Sigma^{-1}\Sigma L'(L\Sigma L')^{-1}L\right\} \\
&= \operatorname{Det}\left\{\Sigma^{-1}Q\right\}
\end{aligned} \tag{7.30}$$

where the standard determinant, denoted by det, is the product of all eigenvalues and the generalized determinant, denoted by Det, is the product of the nonzero eigenvalues. The first equality holds because $(L\Sigma L')^{-1}LL'$ is $(n-p)\times(n-p)$ with $n-p$ nonzero eigenvalues and $L'(L\Sigma L')^{-1}L$ is $n\times n$ with $n-p$ nonzero eigenvalues and the two matrices have the same nonzero eigenvalues:

$$\text{If} \quad (L\Sigma L')^{-1}LL' = \lambda u, \quad \text{then} \qquad L'(L\Sigma L')^{-1}L(L'u) = \lambda(L'u).$$

$$\begin{aligned}\text{If} \quad L'(L\Sigma L')^{-1}Lv = \lambda y, \quad \text{then} \qquad LL'(L\Sigma L')^{-1}Lv &= \lambda Ly,\\ (L\Sigma L')^{-1}Lv &= (LL')^{-1}Ly,\\ (L\Sigma L')^{-1}(LL')(LL')^{-1}Ly &= (LL')^{-1}Ly.\end{aligned}$$

Similarly,

$$\det\{K'\Sigma^{-1}K\}^{-1}\det\{K'K\} = \operatorname{Det}\left\{\Sigma(I-Q)'\right\}. \tag{7.31}$$

Following [Verbyla, 1990], let $A = [K, L']$. Using both characterizations of the projection Q and the formula for the determinant of a block matrix,

$$\begin{aligned}\det\{A'\Sigma A\} = \det\begin{pmatrix} K'\Sigma K & K'\Sigma L' \\ L\Sigma K & L\Sigma L'\end{pmatrix} &= \det\{L\Sigma L'\}\det\left\{K'\Sigma K - K'\Sigma L'(L\Sigma L')^{-1}L\Sigma K\right\}\\ &= \det\{L\Sigma L'\}\det\left\{K'\left[I - L'(L\Sigma L')^{-1}L\Sigma\right]\Sigma K\right\}\\ &= \det\{L\Sigma L'\}\det\left\{K'\left[I-Q\right]\Sigma K\right\}\\ &= \det\{L\Sigma L'\}\det\left\{K'\left[K(K'\Sigma^{-1}K)^{-1}K'\Sigma^{-1}\right]\Sigma K\right\}\\ &= \det\{L\Sigma L'\}\det\left\{K'K(K'\Sigma^{-1}K)^{-1}K'K\right\}\\ &= \det\{L\Sigma L'\}\det\{K'K\}\det\{K'\Sigma^{-1}K\}^{-1}\det\{K'K\};\end{aligned}$$

$$\det\{A'A\} = \det\begin{pmatrix} K'K & K'L' \\ LK & LL'\end{pmatrix} = \det\begin{pmatrix} K'K & 0 \\ 0 & LL'\end{pmatrix} = \det\{K'K\}\det\{LL'\}.$$

Since A is full-rank by construction,

$$\det\{\Sigma\} = \frac{\det\{A'\Sigma A\}}{\det\{A'A\}} = \frac{\det\{K'K\}\det\{L\Sigma L'\}}{\det\{LL'\}\det\{K'\Sigma^{-1}K\}}. \tag{7.32}$$

Finally, by applying first (7.30) and then (7.31),

$$\det\{\Sigma\} = \det\{K'K\}\left[\det\{K'\Sigma^{-1}K\}\operatorname{Det}\{\Sigma^{-1}Q\}\right]^{-1} \tag{7.33a}$$

$$= \det\{LL'\}^{-1}\det\{L\Sigma L'\}\operatorname{Det}\left\{\Sigma(I-Q)'\right\}. \tag{7.33b}$$

7.4.1 *Restricted maximum likelihood (REML) in a special case*

Rather than a general covariance matrix Σ, our model for population structure in terms of distances on a population graph specifies

$$\Sigma = 11' - \lambda\Delta, \tag{7.34}$$

where Δ is a conditionally negative definite matrix such that $1'\Delta^{-1}1 = 1$. The normalization simplifies notation and defines equivalence classes $\left\{\Delta_* : (1'\Delta_*^{-1}1)\Delta_* = \Delta\right\}$. It is

also convenient because under this parametrization $\lambda \in (0,1)$ is a sufficient condition for $\Sigma \succ 0$, as we show in Appendix 7.3.

Because the covariance matrix has the form (7.34) we can avoid explicitly constructing Σ and instead work with Δ. Using the Sherman-Morrison formula for the inverse of a rank-one update,

$$\Sigma^{-1} = -\frac{1}{\lambda}\Big(\Delta^{-1} - \frac{\Delta^{-1}11'\Delta^{-1}}{1-\lambda}\Big) \tag{7.35a}$$

$$\Sigma^{-1}1 = -\frac{1}{\lambda}\Big(\Delta^{-1}1 - \frac{\Delta^{-1}1}{1-\lambda}\Big) = \frac{1}{1-\lambda}\Delta^{-1}1 \tag{7.35b}$$

$$1'\Sigma^{-1}1 = \frac{1}{1-\lambda} \tag{7.35c}$$

The orthogonal projection $Q = Q[\Sigma]$ with kernel $K = 1$ is given by

$$Q = I - \frac{11'\Sigma^{-1}}{1'\Sigma^{-1}1} = I - 11'\Delta^{-1} \tag{7.36a}$$

$$\Sigma^{-1}Q = -\frac{1}{\lambda}\Delta^{-1}\Big(I - \frac{1}{1-\lambda}11'\Delta^{-1}\Big)Q = -\frac{1}{\lambda}\Delta^{-1}Q \tag{7.36b}$$

The projection matrix Q is not symmetric in general but for every Q,

$$Q'\Sigma^{-1} = Q'\Sigma^{-1}Q = \Sigma^{-1}Q \quad \text{and} \quad Q'\Delta^{-1} = Q'\Delta^{-1}Q = \Delta^{-1}Q. \tag{7.37}$$

That is, $\Sigma^{-1}Q$ and $\Delta^{-1}Q$ are symmetric.

Now let's express $\mathrm{Det}\{\Sigma^{-1}Q\}$ as a function of $\mathrm{Det}\{\Delta^{-1}Q\}$ where the generalized determinant Det is the product of the nonzero eigenvalues. Since

$$\Delta^{-1}Qv = \alpha v \qquad \Leftrightarrow \qquad \Sigma^{-1}Q = -\frac{1}{\lambda}(\alpha v), \tag{7.38}$$

the two generalized eigenvalue problems are equivalent up to a proportionality constant and

$$\mathrm{Det}\{\Sigma^{-1}Q\} = \Big(\frac{1}{\lambda}\Big)^{n-1}\mathrm{Det}\{-\Delta^{-1}Q\} = \mathrm{Det}\{-\Delta^{-1}Q/\lambda\}. \tag{7.39}$$

where $\mathrm{rank}\{\Sigma^{-1}Q\} = \mathrm{rank}\{-\Delta^{-1}Q\} = n-1$. Finally, we apply equation (7.32) with $\Sigma = S$ and $K = 1$ to obtain

$$\det\{LSL'\} = \det\{S\}\det\big\{1'S^{-1}1\big\}\frac{\det\{LL'\}}{\det\{1'1\}}, \tag{7.40}$$

and equation (7.30) with $\Sigma = 11' - \lambda\Delta$ and $L\Sigma L' = -\lambda(L\Delta L')$ to obtain

$$\begin{aligned}\det\big\{-(L\Delta L')^{-1}/\sigma^*\big\} &= \mathrm{Det}\big\{\Sigma^{-1}Q/\sigma^2\big\}/\det\{LL'\} \\ &= \mathrm{Det}\big\{-\Delta^{-1}Q/\sigma^*\big\}/\det\{LL'\} \qquad (7.41a)\\ \mathrm{tr}\big\{-(L\Delta L')^{-1}LSL'/\sigma^*\big\} &= \mathrm{tr}\big\{-\big(L'(L\Delta L')^{-1}L\big)S/\sigma^*\big\} \\ &= \mathrm{tr}\big\{-\Delta^{-1}\Delta\big(L'(L\Delta L')^{-1}L\big)S/\sigma^*\big\} \\ &= \mathrm{tr}\big\{-(\Delta^{-1}Q)S/\sigma^*\big\}. \qquad (7.41b)\end{aligned}$$

7.5 *Efficient computation*

Let $R = (R_{\alpha\beta})$ be the matrix of effective resistances between pairs of observed demes (α,β) in the population graph $G = (V,E,M)$. From [McRae, 2006] we know that $T_{\alpha\alpha} \approx$

d and $T_{\alpha\beta} \approx d(1 + R_{\alpha\beta}/4)$ where d is the number of demes in the population graph and o is the number of observed demes. With this motivation, let

$$(\Delta_{\alpha\beta}) = d\Big(1_o 1_o' + (R_{\alpha\beta})/4 - I_o\Big) \tag{7.42}$$

be the matrix of (expected) pairwise distances between observed demes.

If individuals are exchangeable within demes, we can model distances between individuals in terms of distances between demes. For a pair $(i \in \alpha, j \in \beta)$,

$$\Delta_{ij} = d\Big(1 + R_{\alpha\beta}/4 - \mathbb{1}_{\{i=j\}}\Big). \tag{7.43}$$

Equivalently, in matrix notation,

$$(\Delta_{ij}) = d\Big(1_n 1_n' + JRJ'/4 - I_n\Big) \tag{7.44}$$

where $J = (J_{i\alpha}) \in \mathbb{Z}^{n\times o}$ is an indicator matrix such that

$$J_{i\alpha} = \begin{cases} 1 & \text{if } i \in \alpha \\ 0 & \text{if } i \notin \alpha \end{cases}. \tag{7.45}$$

To simplify the notation, we will drop the subscripts and write plainly 1 for the vector of ones and I for the identity matrix. The dimension will be clear from the context if we keep in mind that $R = (R_{\alpha\beta})$ is an $o \times o$ matrix and $\Delta = (\Delta_{ij})$ is an $n \times n$ matrix.

To evaluate the Wishart log-likelihood in equation (4.13) we need to compute the terms $\mathrm{tr}\{\Delta^{-1}QS\}$ and $\mathrm{Det}\{-\Delta^{-1}Q\}$ where

$$Q = I - \frac{11'\Delta^{-1}}{1'\Delta^{-1}1} \tag{7.46}$$

is a projection matrix, which removes the common mean, and S is the observed similarity matrix. We also standardize the distance matrix Δ so that $1'\Delta^{-1}1 = 1$. With this normalization, multiplying Δ by a (positive) constant has no effect on the product $\Delta^{-1}Q$, so we can ignore the scale d in equation (7.44).

The distance matrix $\Delta = 11' + JRJ'/4 - I$ has an "almost-block" structure, except for the diagonal of zeros: specifically, $\Delta = JBJ' - I$ where $B = R/4 + 11'$ is a known $o \times o$ matrix. [B is a function of the migration rates.] The inverse Δ^{-1} is also an almost-block matrix:

$$\Delta^{-1} = JXJ' - I, \tag{7.47}$$

where X is an unknown $o \times o$ matrix. Since $\Delta\Delta^{-1} = I$, the solution X must satisfy

$$JBCXJ' - JBJ' - JXJ' + I = I, \tag{7.48a}$$

$$J'(BC - I)XJ' = JBJ', \tag{7.48b}$$

Where $C = JJ' = \mathrm{diag}\{n_\alpha\}$ is the diagonal matrix of sample counts.

Since every term in equation (7.48b) has an exact block structure which depends on the sample configuration through J, it is sufficient to solve the lower-dimensional problem

$$(BC - I)X = B \quad \Leftrightarrow \quad (C - B^{-1})X = I. \tag{7.49}$$

This is a system of linear equations for the unknown X in terms of the effective resistances R and the counts C, and therefore, it can be solved efficiently without matrix

inversions. The diagonal matrix C is invertible because here we consider only observed demes, i.e., locations with at least one observation; the auxiliary matrix $B = R/4 + 11'$ is invertible because the matrix of effective resistances R is invertible.

Once we solve for X, we could explicitly construct Δ^{-1} from X according to equation (7.47). However, this is not necessary because we only need to compute $\operatorname{Det}\{-\Delta^{-1}Q\}$ and $\operatorname{tr}\{\Delta^{-1}QS\}$ where S is the (average) observed similarity. Using the definition of the orthogonal projection Q (equation 7.46) and the properties of the trace,

$$\operatorname{tr}\left\{\Delta^{-1}QS\right\} = \operatorname{tr}\left\{\Delta^{-1}S\right\} - \frac{1}{1'\Delta^{-1}1}\operatorname{tr}\left\{11'\Delta^{-1}S\Delta^{-1}\right\}. \tag{7.50}$$

We consider each of these terms in turn:

$$1'\Delta^{-1}1 = 1'\big(JXJ' - I\big)1 = \operatorname{tr}\left\{X\big(J'11'J\big)\right\} - n, \tag{7.51a}$$

$$\begin{aligned}\operatorname{tr}\left\{\Delta^{-1}S\right\} &= \operatorname{tr}\left\{\big(JXJ' - I\big)S\right\} \\ &= \operatorname{tr}\left\{X\big(J'SJ\big)\right\} - \operatorname{tr}\{S\},\end{aligned} \tag{7.51b}$$

$$\begin{aligned}\operatorname{tr}\left\{11'\Delta^{-1}S\Delta^{-1}\right\} &= 1'CX\big(J'SJ\big)XC1 + 1'S1 \\ &\quad - 2\operatorname{tr}\left\{X\big(J'S11'J\big)\right\}.\end{aligned} \tag{7.51c}$$

$\operatorname{tr}\{YT\} = \sum_{\alpha,\beta} Y_{\alpha\beta}T_{\alpha\beta}$. So the trace can be computed as `sum(sum(Y.*T))`.

All the terms in red are constants and can be precomputed and stored for easy access. The point is that there is no need to construct the $n \times n$ matrix Δ^{-1} in order to compute $\operatorname{tr}\{\Delta^{-1}QS\}$; we can work with the $o \times o$ matrix X instead.

Next we show how to compute efficiently the generalized determinant $\operatorname{Det}\{-\Delta^{-1}Q\}$. Since $\Delta \in \mathbb{D}^n$ is conditionally negative definite (and nonnegative),

$$\operatorname{Det}\left\{-\Delta^{-1}Q\right\} = \frac{(1'1)/(1'\Delta^{-1}1)}{-\det\{-\Delta\}}. \tag{7.52}$$

$$\operatorname{Det}\left\{\Sigma^{-1}Q\right\} = \frac{(1'1')/(1'\Sigma^{-1}1)}{\det\{\Sigma\}}$$

and

$$\det\{\Sigma\} = \left(1 - 1'\Delta^{-1}1/\lambda\right)\det\{-\lambda\Delta\}$$

$$1'\Sigma^{-1}1 = \left(1 - \lambda/1'\Delta^{-1}1\right)^{-1}$$

Furthermore, Δ has one positive eigenvalue and $n-1$ negative eigenvalues, as we show in Appendix 7.3. Therefore, $-\det\{-\Delta\}$ is guaranteed to be positive and it is sufficient to compute $|\det\{\Delta\}|$, or equivalently, find the eigenvalues of Δ. Since $\Delta = JBJ' - I$,

$$\operatorname{eig}\{\Delta\} = \operatorname{eig}\{JBJ'\} - 1, \tag{7.53}$$

where JBJ' is a block matrix and thus it has o nontrivial eigenvalues besides 0, which has multiplicity $n - o$. Furthermore, for any vector $v \in \mathbb{R}^o$,

$$\Delta(Jv) = JBCv - Jv = J(BC - I)Cv. \tag{7.54}$$

Therefore, if (v, λ) is an eigenpair for $BC - I$, then (Jv, λ) is an eigenpair for Δ. That is, the o nontrivial eigenvalues of Δ are equal to the eigenvalues of $BC - I$.

7.6 *Markov chain Monte Carlo*

Expected coalescence times in a stepping-stone model are determined by the migration rates between demes and the coalescence rates within demes according to equation (2.15). Throughout, we assume that the coalescence rate is the same for all demes and migration is symmetric. The approximation in terms of effective resistances on an undirected graph given by equation (2.28) makes the symmetry assumption explicitly and the equal size assumption implicitly. To use either expression for computing effective distances, we need to specify a migration rate for each undirected edge (α, β) in the grid (V, E). We assume that the migration rates are piecewise constant and we model them

in terms of a colored Voronoi tiling $\mathcal{V}$ of the habitat $\mathcal{H}$. Under the tessellation $\mathcal{V}$, each tile has its own migration log rate and all edges within a tile share this parameter.

Since the spatial structure of the population is unknown, an appropriate Voronoi tessellation of the habitat must be estimated given the data. We use a version of the method based on colored Voronoi tessellations implemented in GENELAND [Guillot et al., 2005]. The main difference is that the "colors" in GENELAND are cluster indices; in our framework the "colors" are log (base 10) migration rates as edges within the same tile share a common rate to encourage locally smooth migration surfaces.

7.6.1 *Updating the number of tiles T with birth/death moves*

Unlike the log rates and locations of tiles, the number of tiles present a transdimensional inference problem because adding or removing a tile changes the dimensionality of the parameter space. For such a problem we can use the birth-death Markov chain Monte Carlo algorithm (BD-MCMC) which has been applied to other variable dimension problems such as a mixture model with unknown number of components [Stephens, 2000, van Lieshout, 2000].

Assume that the Markov chain is currently in state (t, Θ_t) with t Voronoi tiles and parameters Θ_t, and that there are two options for the next move: with probability $a(t)$ the proposed move is $(t+1, \Theta_{t+1})$, i.e., the birth of a tile; with probability $1 - a(t)$ the proposed move is $(t-1, \Theta_{t-1})$, i.e., the death of a tile. Since we consider only these two moves, we assume that they occur with equal probability: $a(1) = 1$ and $a(t) = 1 - a(t) = \frac{1}{2}$ for $t > 1$. For a given number of tiles t, the model parameters include the migration log rates $\{\ell m_1, \dots, \ell m_t\}$ and locations $\{u_1, \dots, u_t\}$, as well as common parameters θ that do not depend on the tessellation and are not updated during a birth/death move. Let

A death event is impossible with only one tile.

$$\Theta_t = (\ell m_1, \dots, \ell m_t, u_1, \dots, u_t, \theta). \tag{7.55}$$

A full Bayesian model for the Voronoi tiling is specified by the likelihood on pairwise distances (4.4) together with the following prior distributions for the number of tiles and tile-specific parameters:

$$T \mid \nu \sim \mathrm{Po}(\nu), \tag{7.56a}$$

$$\underline{u} \mid T \overset{\text{iid}}{\sim} \mathrm{U}(\mathcal{H}), \tag{7.56b}$$

$$\ell\underline{m} \mid \omega, T \overset{\text{iid}}{\sim} \mathrm{N}(\ell\bar{m}, \sigma_m^2). \tag{7.56c}$$

where T is the number of tiles, $(\underline{u}, \ell\underline{m})$ are the tile centers and log rates, respectively, and $\omega = (\ell\bar{m}, \sigma_m^2)$ are hyperparameters: the mean log rate $\ell\bar{m}$ and the variance σ_m^2. The intensity (Poisson rate) ν controls the spatial organization. This prior specification implies that rates and locations are *a priori* independent.

It is convenient to denote the component parameters (location and log rate) by $\phi_t = (u_t, \ell m_t)$. Since the tiles are not ordered,

$$\pi(T, \underline{u}, \ell\underline{m} \mid \nu, \omega) \equiv \pi(T, \phi_1, \dots, \phi_T \mid \nu, \omega) \tag{7.57a}$$

$$= \pi(T \mid \nu) \times T! \times \widetilde{\pi}(\phi_1 \mid \omega) \cdots \widetilde{\pi}(\phi_T \mid \omega) \tag{7.57b}$$

That is, conditional on the number of tiles T, the ϕ_ts are independent and identically distributed from a product distribution with density

$$\widetilde{\pi}(\phi \mid \omega) \propto \mathbb{1}\{\phi(1) \in \mathcal{H}\} \cdot \mathrm{N}(\phi(2)\,;\, \ell\bar{m}, \sigma_m^2). \tag{7.58}$$

Note that the prior is invariant under relabeling of the tiles, i.e.,

$$\pi(T, \phi_1, \dots, \phi_T \,|\, \nu, \omega) = \pi(T, \phi_{\sigma(1)}, \dots, \phi_{\sigma(T)} \,|\, \nu, \omega) \tag{7.59}$$

for every permutation σ of the indices $1, \dots, T$. That is, the tile parameters are exchangeable.

Next we construct a birth-death MCMC that allows only two types of moves: the birth of a new tile and the death of an existing tile (when $T > 1$). Suppose that the current state is $(t, \phi_1, \dots, \phi_t)$. If the proposal is a birth, we add a new tile with log rate $\ell m_{t+1} \sim \mathrm{N}(\ell\bar{m}, \sigma_m^2)$ and location $u_{t+1} \sim \mathrm{U}(\mathcal{H})$. We denote the birth density by $b(t) = \widetilde{\pi}(\phi_{t+1} \,|\, \omega)$. If the proposal is a death, we select a tile to remove uniformly at random, i.e., with probability $d(t) = \frac{1}{t}$.

To guarantee that the birth-death chain is reversible and the stationary distribution is the posterior $\pi(T, \phi_1, \dots, \phi_T \,|\, z, \nu, \omega, \theta)$ given observed data z, we choose the acceptance probabilities $\alpha(\cdot, \cdot)$ so that they satisfy the detailed balance condition:

$$\begin{aligned} &a(t)b(t)\pi(t, \phi_1, \dots, \phi_t \,|\, z, \nu, \omega, \theta)\alpha(t, t+1) \\ &\quad = [1 - a(t+1)]d(t+1)\pi(t+1, \phi_1, \dots, \phi_{t+1} \,|\, z, \nu, \omega, \theta)\alpha(t+1, t). \end{aligned} \tag{7.60}$$

Since $a(t) = a(t+1) = \frac{1}{2}$,

$$r(t) = \frac{\alpha(t, t+1)}{\alpha(t+1, t)} = \frac{d(t+1)}{b(t)} \frac{\pi(t+1, \phi_1, \dots, \phi_{t+1} \,|\, z, \nu, \omega, \theta)}{\pi(t, \phi_1, \dots, \phi_t \,|\, z, \nu, \omega, \theta)} \tag{7.61a}$$

$$= \frac{d(t+1)}{b(t)} \frac{\pi(t+1 \,|\, \nu)}{\pi(t \,|\, \nu)} \frac{\pi(\phi_1, \dots, \phi_{t+1} \,|\, t+1, \omega)}{\pi(\phi_1, \dots, \phi_t \,|\, t, \omega)} \frac{f_{t+1}(z; \boldsymbol{\phi}_{t+1}, \theta)}{f_t(z; \boldsymbol{\phi}_t, \theta)} \tag{7.61b}$$

$$= \frac{\nu}{t+1} \frac{f_{t+1}(z; \boldsymbol{\phi}_{t+1}, \theta)}{f_t(z; \boldsymbol{\phi}_t, \theta)} \tag{7.61c}$$

Apply equation (7.57) with $\pi(t \,|\, \nu) = \nu^t e^{-\nu}/t!$

and $\alpha(t, t+1) = \min\{r(t), 1\}$. Therefore, the following algorithm simulates a Markov chain with stationary distribution $\pi(T, \boldsymbol{\phi}_T \,|\, z, \nu, \omega, \theta)$ where, for simplicity of notation, we write $\boldsymbol{\phi}_t = (\phi_1, \dots, \phi_t)$.

1. Choose between a birth event and a death event, with equal probability.

2. If a birth is proposed, its location, migration log rate and coalescence log rate are sampled from the priors, and the acceptance probability is

$$\alpha(T, T+1) = \min\left\{\frac{\lambda}{T+1} \frac{f_{T+1}(z; \Theta_{T+1})}{f_T(z; \Theta_T)}, 1\right\} \tag{7.62}$$

3. If a death is proposed, a tile to be removed is selected uniformly at random, and the acceptance probability is

$$\alpha(T+1, T) = \min\left\{\frac{T+1}{\lambda} \frac{f_T(z; \Theta_T)}{f_{T+1}(z; \Theta_{T+1})}, 1\right\} \tag{7.63}$$

because a deletion move is the reverse of an addition move [Byers and Raftery, 2002].

7.6.2 *Updating the Voronoi centers (for a fixed number of tiles T)*

This is a Metropolis-Hastings symmetric random-walk update. Sequentially, for each tile t, we propose a new center u_t^*. The proposal distribution is bivariate normal centered at the current value $u_t^c = (x_t, y_t)$ [with correlation 0]. The proposal is accepted with probability

$$\alpha = \min\left\{\frac{\pi(u_t^* \,|\, Z, \Theta_{\setminus u_t})}{\pi(u_t^c \,|\, Z, \Theta_{\setminus u_t})}, 1\right\} = \min\left\{\frac{f(Z\,;\, \Theta^*)\pi(\underline{u}^*)}{f(Z\,;\, \Theta^c)\pi(\underline{u}^c)}, 1\right\}. \tag{7.64}$$

The prior distribution of the center locations $\underline{u} = (u_t : t = 1, \dots, T)$ is uniform over the domain $\mathcal{H}$,

$$\pi(\underline{u}) \propto \mathbb{1}\{u_t \in \mathcal{H} : t = 1, \dots, T\}. \tag{7.65}$$

On the log scale, $\log(\alpha) = -\infty$ if at least one component of $\underline{u}^*$ falls outside of the domain $\mathcal{H}$. Otherwise,

$$\log(\alpha) = \min\Big\{\ell(\Theta^* \,|\, Z) - \ell(\Theta^c \,|\, Z), 0\Big\}. \tag{7.66}$$

7.6.3 *Updating the log-transformed migration rates $\ell\underline{m}$*

We assume that the migration log (base 10) rates are normally distributed with common mean $\ell\bar{m}$ and variance σ_m^2,

$$\ell m_t \,|\, \ell\bar{m}, \sigma_m^2 \overset{\text{iid}}{\sim} \mathrm{N}(\ell\bar{m}, \sigma_m^2), \tag{7.67}$$

or in an equivalent parametrization,

$$\ell m_t = \ell\bar{m} + e_t, e_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_m^2). \tag{7.68}$$

where $\ell\bar{m}$ is the mean log rate and e_t is the effect of tile t, relative to the mean. The second parametrization is more convenient because it allows scaling all migration rates simultaneously by adjusting $\ell\bar{m}$.

We choose a vague prior for the hyperparameters $\ell\bar{m}, \sigma_m^2$ assuming prior independence of location and scale,

$$\ell\bar{m} \sim \mathrm{U}(lob, upb), \tag{7.69}$$

$$\sigma_m^2 \sim \text{Inv-G}(\tfrac{a}{2}, \tfrac{b}{2}). \tag{7.70}$$

That is, the hyperprior on $(\ell\bar{m}, \sigma_m^2)$ is is semi-conjugate.

To simulate a Markov chain with stationary distribution $\pi(T, \underline{u}, \ell\underline{m} \,|\, z, \nu, \omega)$,

1. Update each error in turn (or all errors at once) with a Metropolis-Hastings step and a random-walk proposal. That is, we draw a new migration log rate parameter $\ell m_t^* \sim \mathrm{N}(\ell m_t^c,)$ for each tile in the current Voronoi decomposition and accept the proposal $\ell\underline{m}^* = \{\ell m_t^* : t = 1, \dots, T\}$ with probability

$$\alpha = \min\left\{\frac{f(Z\,;\, \Theta_{\setminus \ell\underline{m}}, \ell\underline{m}^*)\pi(\ell\underline{m}^* \,|\, \ell\bar{m}, \sigma_m^2)}{f(Z\,;\, \Theta_{\setminus \ell\underline{m}}, \ell\underline{m}^c)\pi(\ell\underline{m}^c \,|\, \ell\bar{m}, \sigma_m^2)}, 1\right\}. \tag{7.71}$$

2. Update the mean migration log rate $\ell\bar{m}$ with a Metropolis-Hastings step and a random-walk proposal.

3. Update the common log rate variance σ_m^2 with a Gibbs step by sampling from its full conditional distribution:

$$\pi(\sigma_m^2 \,|\, Z, \Theta) \propto \text{Inv-G}\Big(\tfrac{a}{2}, \tfrac{b}{2}\Big) \prod_{t=1}^{T} \mathrm{N}\Big(e_t\,;\, 0, \sigma_m^2\Big) \tag{7.72a}$$

$$\propto \left\{\frac{1}{\sigma_m^2}\right\}^{a/2+1} \exp\left\{-\frac{b}{2\sigma_m^2}\right\} \times \prod_{t=1}^{T} \left\{\frac{1}{\sigma_m^2}\right\}^{1/2} \exp\left\{-\frac{e_t^2}{2\sigma_m^2}\right\} \tag{7.72b}$$

$$\propto \left\{\frac{1}{\sigma_m^2}\right\}^{a/2+T/2+1} \exp\left\{-\frac{1}{2\sigma_m^2}(b + s_m^2)\right\}, \tag{7.72c}$$

where $s_e^2 = \sum_{t=1}^{T} e_t^2$ is the sum of squares for the relative tile effects on the log scale. Because we conveniently choose the conjugate inverse-gamma prior for σ_m^2, we can update this parameter by drawing

$$\sigma_m^2 \sim \text{Inv-G}\Big((a+T)/2, (b+s_e^2/2)\Big). \tag{7.73}$$

7.6.4 *Updating the degrees of freedom k*

Here we consider updating the degrees of freedom k. The proposal distribution is

$$k^* \sim \text{N}(k^c, v_{new}) \tag{7.74}$$

where k^c is the current value and v is the proposal variance.

Since the Wishart degrees of freedom for a $n \times n$ matrix is a real number ν that satisfies $\nu > n-1$, the support of this parameter is (n, p). If the proposed value k^* is not valid,

$$\log\left\{\frac{\pi(k^*)}{\pi(k^c)}\right\} = -\infty \tag{7.75}$$

and the proposal is rejected. Otherwise, it is accepted with probability

$$\alpha = \min\left\{\frac{\pi(k^*) f(Z;\, \Theta_{\setminus k}, k^*)}{\pi(k^c) f(Z;\, \Theta_{\setminus k}, k^c)}, 1\right\} \tag{7.76}$$

Here $f(Z;\, \Theta_{\setminus k}, k^*)$ is the likelihood for the given value of k with the rest of the parameters $\Theta_{\setminus k}$ fixed to their current values. The prior on the degrees of freedom is uniform on the log scale, i.e.,

$$\pi(k) \propto \frac{1}{k}. \tag{7.77}$$

Since k is bounded, the prior is proper with normalizing constant $\log(p) - \log(n)$.

7.6.5 *Updating the scale nuisance parameter σ**

The nuisance parameter $\sigma^* = \lambda\sigma^2$ can be efficiently updated with a Gibbs step if we choose the conjugate prior distribution, Inv-G$(c/2, d/2)$. Then the full conditional is also Inverse Gamma with shape and scale parameters given by

$$c^* = c + k(n-1), \tag{7.78a}$$

$$d^* = d + k\,\text{tr}\left\{\Delta^{-1} QS\right\}. \tag{7.78b}$$

For microsatellites, $\pi(\sigma_1^*, \ldots, \sigma_p^* \,|\, Z, \Theta)$ factorizes into the full conditional of each site-specific scale parameter σ_s^*, so there is no loss of efficiency to estimate a small number of microsatellites.

7.7 *MATLAB implementation*

7.7.1 *Triangular (isometric) grid*

Suppose that the genotypes individuals are sampled within a rectangular region $\mathcal{H}$ bounded by (x_0, y_0) on the bottom right and (x_1, y_1) on the top right.

By convention, x denotes longitudes and y latitudes.

To initialize the program, we specify the dimensions $\ell_x \times \ell_y$ of a triangular grid (V, E) to tile the habitat $\mathcal{H}$. The resulting grid is regular but not strictly isometric, unless ℓ_x and ℓ_y are chosen to match the size of the habitat.

By definition, a triangular grid is formed by dividing the plane regularly into equilateral triangles.

7.7.2 *Data structures*

Here I describe the MATLAB implementation and data structures. The problem is specified in terms of

- $\ell_x \times \ell_y$ triangular grid (V, E) which spans the habitat $\mathcal{H}$;
- $(\ell_x \ell_y) \times (\ell_x \ell_y)$ symmetric matrix M of migration rates.

The order of (V, E) is $|V| = \ell_x \ell_y$ and the size is $n_e \equiv |E| = (\ell_x - 1)\ell_y + (2\ell_x - 1)(\ell_y - 1)$. Both the grid (V, E) and the migration matrix M are very sparse because each vertex $v \in V$ has at most six neighbors and

$m_{\alpha\beta} = 2N_0 \dot{m}_{\alpha\beta}$ where N_0 is the coalescent timescale.

$$M = (M_{\alpha\beta}) = \left\{ \begin{array}{cc} m_{\alpha\beta} & \text{if } (\alpha, \beta) \in E \\ 0 & \text{otherwise.} \end{array} \right\} \tag{7.79}$$

That is, (V, E) and M together describe a weighted matrix $G = (V, E, M)$. It is not required that M be symmetric; the linear system for Δ is valid as long as (V, E) is connected: If all demes communicate, the sample will eventually coalesce, i.e., the distance between lineages is finite. This guarantees that

$$\Delta = (d^2_{\alpha\beta}) = \left\{ d^2_{\alpha\beta} < \infty \text{ for } (\alpha, \beta) \in V \times V. \right\} \tag{7.80}$$

Although the two matrices have the same size, M is sparse but Δ is full and hence might be expensive to compute. With a denser grid (V, E), few of the demes are sampled from and computing the entire distance matrix Δ is not necessary. To compute the likelihood of the data, we need only the sample distance matrix $\underline{\Delta}$.

In the rest of this section, let $n_v = \ell_x \ell_y$ be the number of demes and $n_p = \binom{n_v}{2} = n_v(n_v - 1)/2$ be the number of unique pairs of demes. The number of unknowns is $n_v + n_p$, the number of within-deme coalescence times plus the number of between-demes coalescence times.

Vertex set representation: The vertices V are stored in a $n_v \times 2$ matrix `Vcoord`, with the x (longitude) coordinates in the first column and the y (latitude) coordinates in the second column. The locations of the Voronoi sites S are stored similarly in `Scoord`. The triangular grid (V, E) is fixed, so `Vcoord` does not change. On the other hand, the Voronoi decomposition of $\mathcal{H}$ is updated regularly, which is reflected by (row) changes in `Scoord`.

The two matrices are used to update the Voronoi tessellation whenever a tile moves its location. Recall that by definition the Voronoi tile (cell) $T(s)$ consists of the points closer to s than to any other site.

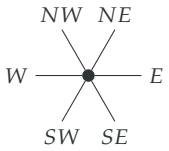
Compute all distances between the demes in `Vcoord` and the sites in `Scoord`.
For each deme $v \in V$, find the closest Voronoi site $s \in S$.

```
euDist = rdist(Vcoord,Scoord);

[temp,Colors] = min(euDist,[],2);
```

The vector `Colors` indicates which tile each deme falls into.

Edge set representation The edges E are stored in a $n_v \times 6$ matrix `Edges`. There is one row for each vertex (deme) and the columns are its six adjacent vertices, in the order W, NW, NE, E, SE, SW (clockwise). Vertices are identified by their row index in `Edges`. If the deme does not have a neighbor in some positions, the corresponding entries of `Edges` are set to 0. The number of nonzero entries is twice the number of edges $2n_e$.



Rate parameters representation The backward migration matrix is stored in a $n_v \times n_v$ sparse matrix `Mrates` with $2n_e$ nonzero elements.

7.7.3 *Computing coalescence distances*

Our MCMC implementation requires repeatedly solving a system of linear equations $Ax = b$. The matrix $A = [A_1; A_2]$ is large, sparse, nearly symmetric and positive definite. The regularity of the grid G gives A its structure and sparseness.

A_1 represents the n_v within-deme equations

$$(q_\alpha + m_\alpha)T_{\alpha\alpha} - \sum_{\gamma \in Nei(\alpha)} m_{\alpha\gamma} T_{\alpha\gamma} = 1, \tag{7.81}$$

and A_2 represents the n_p between-demes equations

$$(m_\alpha + m_\beta)T_{\alpha\beta} - \sum_{\gamma \in Nei(\alpha)} m_{\alpha\gamma} T_{\beta\gamma} - \sum_{\gamma \in Nei(\beta)} m_{\beta\gamma} T_{\alpha\gamma} = 2. \tag{7.82}$$

Here $Nei(\alpha) = \{\gamma \in V : (\alpha, \gamma) \in E\}$ is the set of vertices adjacent to α and $m_\alpha = \sum_{\gamma \in Nei(\alpha)} m_{\alpha\gamma}$ is the rate of migration into α. The equations also shows that $b = [1_{n_v}; 1_{n_p}]$.

The matrix A is positive definite because

$$A_2 = \mathcal{L}_2(\{m_{\alpha\beta}\}), \tag{7.83}$$

$$A_1 = \operatorname{diag}\{q\} + \mathcal{L}_1(\{m_{\alpha\beta}\}). \tag{7.84}$$

The Laplacian matrices $\mathcal{L}_1$, $\mathcal{L}_2$ are functions of only the migration rates and $\mathcal{L} = [\mathcal{L}_1; \mathcal{L}_2]$ is also a Laplacian matrix, and therefore, it is positive definite. We note that the matrix $Q = -\mathcal{L}$ is the infinitesimal generator the migration process where the lineages move from deme to deme according to M. For a continuous-time stochastic process, the infinitesimal generator is the matrix $Q = (q_{x,y})$ with entries

$$q_{x,y} = \begin{cases} -\lambda_x & \text{if } x = y, \\ \lambda_x a_{x,y} & \text{otherwise} \end{cases} \tag{7.85}$$

where λ_x is the holding rate for state x and $\mathcal{A} = (a_{x,y})$ is the transition probability matrix of the embedded jump chain. In this case, the transition probabilities are

$$\frac{m_{\alpha\beta}}{\sum_{\gamma \in Nei(\alpha)} m_{\alpha\gamma}} = \frac{N_0 \dot{m}_{\alpha\beta}}{\sum_{\gamma \in Nei(\alpha)} N_0 \dot{m}_{\alpha\gamma}} = \frac{\dot{m}_{\alpha\beta}}{1 - \dot{m}_{\alpha\alpha}}. \tag{7.86}$$

Solving $Ax = b$, and thus finding all coalescence times at once, has the advantage of reducing numerical errors. Because we use an iterative procedure (preconditioned conjugate gradient), we control how close the approximate solution x is to the true solution x. If we first solve for x_2 and then substitute to find x_1, numerical errors in x_2 are propagated in x_1.

7.7.4 *Computing resistance distances*

Consider again the matrix of migration rates between neighboring demes, M. Let L be its Laplacian matrix,

$$L = \operatorname{diag}\{M1\} - M \tag{7.87}$$

The effective resistance $R_{\alpha\beta}$ between a pair of demes (α, β) is equal to the βth element of the vector x given by

$$L_{-\alpha} x = e_\beta \tag{7.88}$$

where $L_{-\alpha}$ is the Laplacian L with the αth row and column removed, and e_β is the standard basis vector with a 1 at the βth coordinate and 0s elsewhere. This method can be optimized by solving $L_{-\alpha}X = E$ where $E = (e_\beta)$, so that we compute the effective resistance between α and all other demes with a single matrix operation.

There are other methods to compute $\underline{R}$, e.g., with a single matrix inversion [Babić et al., 2002]

$$H = (L + n_v^{-1}J)^{-1} \tag{7.89}$$

$$R_{\alpha\beta} = H_{\alpha\alpha} + H_{\beta\beta} - 2H_{\alpha\beta} \tag{7.90}$$

This method is more efficient for sparser grids where most demes are sampled from.

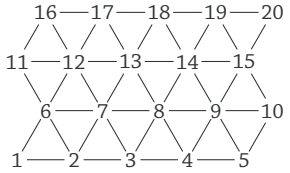


Figure 7.1: This is a 5×4 regular triangular grid. If line weight is proportional to migration rate, this pattern corresponds to uniform migration with equal deme sizes.

7.8 *Simulations with* `ms`

Here we describe how we produce samples from a spatially distributed population that evolves under Kimura's stepping-stone model using the program `ms` [Hudson, 2002]. For all simulations, we first construct a regular triangular grid (V, E) of $\ell_x \times \ell_y$ demes (vertices) with coordinates (x_α, y_α). The spatial information is not explicitly used by `ms` and instead we set $m_{\alpha\beta} = 0 = m_{\beta\alpha}$ for all pairs of demes such that $(\alpha, \beta) \notin E$. We also specify the size of each deme, N_α, and the migration rates across each edge in the two opposite directions, $m_{\alpha\beta}$ and $m_{\beta\alpha}$. Migration is not necessarily symmetric but is conservative [i.e., it preserves deme sizes.] Both the deme sizes $N_\alpha = N_0 p_\alpha$ and the migration rates $m_{\alpha\beta} = 4N_0\dot{m}_{\alpha\beta}$ are relative to the coalescent timescale N_0. [That is, p_α is the relative size of deme α and $\dot{m}_{\alpha\beta}$ is the backward migration fraction from α to β per generation.] The input arguments are p_α and $m_{\alpha\beta}$, respectively.

7.8.1 *Spatial structure due to constant migration*

Here edges in the middle of the habitat have migration rate that is a factor of magnitude lower than the rate of the edges on either side. [The rates range from 0.3 to 3.] This patterns is a barrier to actual migration and it results in a barrier to effective migration.

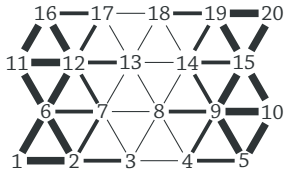


Figure 7.2: Barrier to migration. If line weights are proportional to migration rates, this patterns corresponds to a barrier across the middle of the habitat.

```
ms 20 1 -s 1 -I 20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
-m 1 2 3.0 -m 2 1 3.0 -m 1 6 3.0 -m 6 1 3.0 -m 2 3 1.7 -m 3 2 1.7
-m 2 6 3.0 -m 6 2 3.0 -m 2 7 1.7 -m 7 2 1.7 -m 3 4 0.3 -m 4 3 0.3
-m 3 7 0.3 -m 7 3 0.3 -m 3 8 0.3 -m 8 3 0.3 -m 4 5 1.7 -m 5 4 1.7
-m 4 8 0.3 -m 8 4 0.3 -m 4 9 1.7 -m 9 4 1.7 -m 5 9 3.0 -m 9 5 3.0
-m 5 10 3.0 -m 10 5 3.0 -m 6 7 1.7 -m 7 6 1.7 -m 6 11 3.0 -m 11 6 3.0
-m 6 12 3.0 -m 12 6 3.0 -m 7 8 0.3 -m 8 7 0.3 -m 7 12 1.7 -m 12 7 1.7
-m 7 13 0.3 -m 13 7 0.3 -m 8 9 1.7 -m 9 8 1.7 -m 8 13 0.3 -m 13 8 0.3
-m 8 14 0.3 -m 14 8 0.3 -m 9 10 3.0 -m 10 9 3.0 -m 9 14 1.7
-m 14 9 1.7 -m 9 15 3.0 -m 15 9 3.0 -m 10 15 3.0 -m 15 10 3.0
-m 11 12 3.0 -m 12 11 3.0 -m 11 16 3.0 -m 16 11 3.0 -m 12 13 1.7
-m 13 12 1.7 -m 12 16 3.0 -m 16 12 3.0 -m 12 17 1.7 -m 17 12 1.7
-m 13 14 0.3 -m 14 13 0.3 -m 13 17 0.3 -m 17 13 0.3 -m 13 18 0.3
-m 18 13 0.3 -m 14 15 1.7 -m 15 14 1.7 -m 14 18 0.3 -m 18 14 0.3
-m 14 19 1.7 -m 19 14 1.7 -m 15 19 3.0 -m 19 15 3.0 -m 15 20 3.0
-m 20 15 3.0 -m 16 17 1.7 -m 17 16 1.7 -m 17 18 0.3 -m 18 17 0.3
-m 18 19 1.7 -m 19 18 1.7 -m 19 20 3.0 -m 20 19 3.0
```

7.8.2 *Spatial structure due to variation in diversity*

Here some demes have bigger size and thus lower coalescence rate and higher genetic diversity. In the first version, migration rates are constant but there are differences in effective population size. Since demes in the "east" and "west" of the habitat are 5 times bigger than those in the middle, the effect is a barrier to effective migration that is qualitatively very similar to the true barrier in the previous simulation.

[A few edges are directed, with rate $m_{\alpha\beta} = 0.2$ from a big deme to a small deme and rate $m_{\beta\alpha} = 1$ in the other direction. These edges cross the "boundary" between the areas of high and low diversity and their rates are assigned so that migration is conservative: the same number of migrants are exchanged between α and β because $N_\alpha m_{\alpha\beta} = N_\beta m_{\beta\alpha}$.]

```
ms 20 1 -s 1 -I 20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
-n 1 5.0 -n 2 5.0 -n 3 1.0 -n 4 1.0 -n 5 1.0 -n 6 5.0 -n 7 1.0
-n 8 1.0 -n 9 1.0 -n 10 5.0 -n 11 5.0 -n 12 1.0 -n 13 1.0 -n 14 1.0
-n 15 5.0 -n 16 1.0 -n 17 1.0 -n 18 1.0 -n 19 5.0 -n 20 5.0
-m 1 2 1.0 -m 2 1 1.0 -m 1 6 1.0 -m 6 1 1.0 -m 2 3 0.2 -m 3 2 1.0
-m 2 6 1.0 -m 6 2 1.0 -m 2 7 0.2 -m 7 2 1.0 -m 3 4 1.0 -m 4 3 1.0
-m 3 7 1.0 -m 7 3 1.0 -m 3 8 1.0 -m 8 3 1.0 -m 4 5 1.0 -m 5 4 1.0
-m 4 8 1.0 -m 8 4 1.0 -m 4 9 1.0 -m 9 4 1.0 -m 5 9 1.0 -m 9 5 1.0
-m 5 10 1.0 -m 10 5 0.2 -m 6 7 0.2 -m 7 6 1.0 -m 6 11 1.0 -m 11 6 1.0
-m 6 12 0.2 -m 12 6 1.0 -m 7 8 1.0 -m 8 7 1.0 -m 7 12 1.0 -m 12 7 1.0
-m 7 13 1.0 -m 13 7 1.0 -m 8 9 1.0 -m 9 8 1.0 -m 8 13 1.0 -m 13 8 1.0
-m 8 14 1.0 -m 14 8 1.0 -m 9 10 1.0 -m 10 9 0.2 -m 9 14 1.0 -m 14 9 1.0
-m 9 15 1.0 -m 15 9 0.2 -m 10 15 1.0 -m 15 10 1.0 -m 11 12 0.2
-m 12 11 1.0 -m 11 16 0.2 -m 16 11 1.0 -m 12 13 1.0 -m 13 12 1.0
-m 12 16 1.0 -m 16 12 1.0 -m 12 17 1.0 -m 17 12 1.0 -m 13 14 1.0
-m 14 13 1.0 -m 13 17 1.0 -m 17 13 1.0 -m 13 18 1.0 -m 18 13 1.0
-m 14 15 1.0 -m 15 14 0.2 -m 14 18 1.0 -m 18 14 1.0 -m 14 19 1.0
-m 19 14 0.2 -m 15 19 1.0 -m 19 15 1.0 -m 15 20 1.0 -m 20 15 1.0
-m 16 17 1.0 -m 17 16 1.0 -m 17 18 1.0 -m 18 17 1.0 -m 18 19 1.0
-m 19 18 0.2 -m 19 20 1.0 -m 20 19 1.0
```

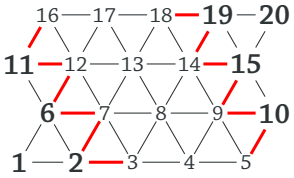


Figure 7.3: Barrier to effective migration due to differences in effective population size. The demes in bold are 5 times bigger; the edges in red are *directed* — this is necessary to preserve equilibrium in time.

In the second version, differences in migration rates compensate for differences in deme size because $N_\gamma m_{\gamma\omega} = N_\omega m_{\omega\gamma}$ for *all* edges $(\gamma, \omega) \in E$. The result is no variation in effective migration although both the deme sizes and the migration rates vary across the habitat.

```
ms 20 1 -s 1 -I 20 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0
-n 1 5.0 -n 2 5.0 -n 3 1.0 -n 4 1.0 -n 5 1.0 -n 6 5.0 -n 7 1.0
-n 8 1.0 -n 9 1.0 -n 10 5.0 -n 11 5.0 -n 12 1.0 -n 13 1.0 -n 14 1.0
-n 15 5.0 -n 16 1.0 -n 17 1.0 -n 18 1.0 -n 19 5.0 -n 20 5.0
-m 1 2 0.2 -m 2 1 0.2 -m 1 6 0.2 -m 6 1 0.2 -m 2 3 0.2 -m 3 2 1.0
-m 2 6 0.2 -m 6 2 0.2 -m 2 7 0.2 -m 7 2 1.0 -m 3 4 1.0 -m 4 3 1.0
-m 3 7 1.0 -m 7 3 1.0 -m 3 8 1.0 -m 8 3 1.0 -m 4 5 1.0 -m 5 4 1.0
-m 4 8 1.0 -m 8 4 1.0 -m 4 9 1.0 -m 9 4 1.0 -m 5 9 1.0 -m 9 5 1.0
-m 5 10 1.0 -m 10 5 0.2 -m 6 7 0.2 -m 7 6 1.0 -m 6 11 0.2 -m 11 6 0.2
-m 6 12 0.2 -m 12 6 1.0 -m 7 8 1.0 -m 8 7 1.0 -m 7 12 1.0 -m 12 7 1.0
-m 7 13 1.0 -m 13 7 1.0 -m 8 9 1.0 -m 9 8 1.0 -m 8 13 1.0 -m 13 8 1.0
-m 8 14 1.0 -m 14 8 1.0 -m 9 10 1.0 -m 10 9 0.2 -m 9 14 1.0 -m 14 9 1.0
```

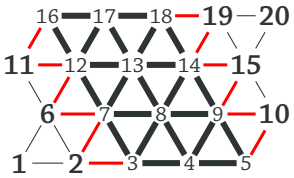


Figure 7.4: Uniform effective migration even though there are differences in both population size and in migration rates. The demes in bold are 4 times bigger; the edges in red are *directed* — this is necessary to preserve equilibrium in time.

```
-m 9 15 1.0 -m 15 9 0.2 -m 10 15 0.2 -m 15 10 0.2 -m 11 12 0.2
-m 12 11 1.0 -m 11 16 0.2 -m 16 11 1.0 -m 12 13 1.0 -m 13 12 1.0
-m 12 16 1.0 -m 16 12 1.0 -m 12 17 1.0 -m 17 12 1.0 -m 13 14 1.0
-m 14 13 1.0 -m 13 17 1.0 -m 17 13 1.0 -m 13 18 1.0 -m 18 13 1.0
-m 14 15 1.0 -m 15 14 0.2 -m 14 18 1.0 -m 18 14 1.0 -m 14 19 1.0
-m 19 14 0.2 -m 15 19 0.2 -m 19 15 0.2 -m 15 20 0.2 -m 20 15 0.2
-m 16 17 1.0 -m 17 16 1.0 -m 17 18 1.0 -m 18 17 1.0 -m 18 19 1.0
-m 19 18 0.2 -m 19 20 0.2 -m 20 19 0.2
```

7.8.3 *Spatial structure due to a split event*

Here the effect of a barrier to effective migration is produced by a past event that zeroes out some migration rates and thus disconnects the "east" and "west" regions of the habitat. The split is instantaneous and occurs $3N_0$ generations back in the past. This creates a barrier in time that is detected as a barrier to effective migration.

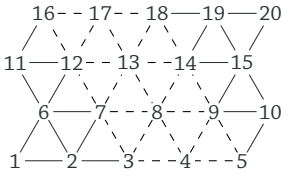


Figure 7.5: Barrier to effective migration due to a split in time and otherwise uniform migration rates. The dashed edges are disconnected at the same time in the past.

```
ms 20 1 -s 1 -I 20 4 3 0 0 0 3 0 0 0 0 0 0 0 0 3 0 0 0 3 4 0
-m 1 2 1.0 -m 2 1 1.0 -m 1 6 1.0 -m 6 1 1.0 -m 2 3 1.0 -m 3 2 1.0
-m 2 6 1.0 -m 6 2 1.0 -m 2 7 1.0 -m 7 2 1.0 -m 5 10 1.0 -m 10 5 1.0
-m 6 7 1.0 -m 7 6 1.0 -m 6 11 1.0 -m 11 6 1.0 -m 6 12 1.0 -m 12 6 1.0
-m 9 10 1.0 -m 10 9 1.0 -m 9 15 1.0 -m 15 9 1.0 -m 10 15 1.0
-m 15 10 1.0 -m 11 12 1.0 -m 12 11 1.0 -m 11 16 1.0 -m 16 11 1.0
-m 14 15 1.0 -m 15 14 1.0 -m 14 19 1.0 -m 19 14 1.0 -m 15 19 1.0
-m 19 15 1.0 -m 15 20 1.0 -m 20 15 1.0 -m 18 19 1.0 -m 19 18 1.0
-m 19 20 1.0 -m 20 19 1.0
-em 3.0 3 7 1.0 -em 3.0 3 8 1.0 -em 3.0 3 4 1.0 -em 3.0 4 3 1.0
-em 3.0 4 8 1.0 -em 3.0 4 9 1.0 -em 3.0 4 5 1.0 -em 3.0 5 4 1.0
-em 3.0 5 9 1.0 -em 3.0 7 12 1.0 -em 3.0 7 13 1.0 -em 3.0 7 8 1.0
-em 3.0 7 3 1.0 -em 3.0 8 7 1.0 -em 3.0 8 13 1.0 -em 3.0 8 14 1.0
-em 3.0 8 9 1.0 -em 3.0 8 4 1.0 -em 3.0 8 3 1.0 -em 3.0 9 8 1.0
-em 3.0 9 14 1.0 -em 3.0 9 5 1.0 -em 3.0 9 4 1.0 -em 3.0 12 16 1.0
-em 3.0 12 17 1.0 -em 3.0 12 13 1.0 -em 3.0 12 7 1.0 -em 3.0 13 12 1.0
-em 3.0 13 17 1.0 -em 3.0 13 18 1.0 -em 3.0 13 14 1.0 -em 3.0 13 8 1.0
-em 3.0 13 7 1.0 -em 3.0 14 13 1.0 -em 3.0 14 18 1.0 -em 3.0 14 9 1.0
-em 3.0 14 8 1.0 -em 3.0 16 17 1.0 -em 3.0 16 12 1.0 -em 3.0 17 16 1.0
-em 3.0 17 18 1.0 -em 3.0 17 13 1.0 -em 3.0 17 12 1.0 -em 3.0 18 17 1.0
-em 3.0 18 14 1.0 -em 3.0 18 13 1.0
```

8
Bibliography

D. Babić, D. J. Klein, I. Lukovits, S. Nikolić, and N. Trinajstić. Resistance-distance matrix: A computational algorithm and its application. *International Journal of Quantum Chemistry*, 90(1):166–176, 2002.

M. Bahlo and R. C. Griffiths. Coalescence time for two genes from a subdivided population. *Journal of Mathematical Biology*, 43(5):397–410, 2001.

R. B. Bapat. Resistance matrix of a weighted graph. *MATCH: Communications in Mathematical and in Computer Chemistry*, 50:73–82, 2004.

R. B. Bapat and T. E. S. Raghavan. *Nonnegative matrices and applications*. Cambridge University Press, 1997.

P. Beerli and J. Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences (PNAS)*, 98(8):4563–4568, 2001.

C. A. Brewer, G. W. Hatchard, and M. A. Harrower. ColorBrewer in print: a catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32, 2003.

S. D. Byers and A. E. Raftery. Bayesian estimation and segmentation of spatial point processes using Voronoi tilings. In Andrew B. Lawson and David G.T. Denison, editors, *Spatial Cluster Modeling*, page 109–121. Chapman&Hall, 2002.

L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The history and geography of human genes*. Princeton University Press, 1994.

A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4):312–340, 1996.

A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research*, 15 (11):1496–1502, 2005.

C. C. Cockerham. Variance of gene frequencies. *Evolution*, 23(1):72–84, 1969.

J. Felsenstein. A pain in the torus: Some difficulties with models of isolation by distance. *The American Naturalist*, 109(967):359–368, 1975.

J. C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67(1):81–97, 1985.

G. Guillot, A. Estoup, F. Mortier, and J. F. Cosson. A spatial statistical model for landscape genetics. *Genetics*, 170(3):1261–1280, 2005.

E. M. Hanks and M. B. Hooten. Circuit theory and model-based inference for landscape connectivity. *Journal of the American Statistical Association*, 108(501):22–33, 2013.

J. Hey. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoretical Population Biology*, 39(1):30–48, 1991.

M W. Horton, A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, and et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics*, 44(2):212–216, 2012.

M. J. Hubisz, D. Falush, M. Stephens, and J. K. Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9(5):1322–1332, 2009.

R. R. Hudson. Gene genealogies and the coalescent process. In Douglas Futuyma and Janis Antonovics, editors, *Oxford surveys in evolutionary biology*, volume 7, pages 1--44. Oxford University Press, 1990.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.

M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561–576, 1964.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43, 1982a.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248, 1982b.

O. Lao, T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf, A. Caliebe, and et al. Correlation between genetic and geographic structure in Europe. *Current Biology*, 18(16): 1241–1248, 2008.

D. J. Lawson and D. Falush. Population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13:337–361, 2012.

J. Y. Lee and S. V. Edwards. Divergence across Australia's Carpentarian barrier: Statistical phylogeography of the red-backed fairy wren Malurus melanocephalus. *Evolution*, 62(12):3117–3134, 2008.

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2008.

P. McCullagh. Marginal likelihood for distance matrices. *Statistica Sinica*, 19:631–649, 2009.

B. H. McRae. Isolation by resistance. *Evolution*, 60(8):1551–1561, 2006.

B. H. McRae, B. G. Dickson, T. H. Keitt, and V. B. Shah. Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, 89(10):2712–2742, 2008.

G. McVean. A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686, 2009.

P. Menozzi, A. Piazza, and L. L. Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358):786–792, 1978.

T. Nagylaki. The strong-migration limit in geographically structured populations. *Journal of Mathematical Biology*, 9(2):101–114, 1980.

M. Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences (PNAS)*, 70(12):3321–3323, 1973.

M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, and *et al.* The population reference sample, POPRES: A resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008.

R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 2000.

M. Nordborg, T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, and et al. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, 3(7):e196, 2005.

M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1):59–75, 1990.

M. Notohara. The strong-migration limit for the genealogical process in geographically structured populations. *Journal of Mathematical Biology*, 31(2):115–122, 1993.

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, and C. D. Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial tessellations : concepts and applications of Voronoi diagrams*. Wiley Series in Probability and Statistics. Wiley, 2000.

A. Platt, M. Horton, Y. S. Huang, Y. Li, A. E. Anastasio, and et al. The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, 6(2):e1000843, 2010.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*, 1(6):e70, 2005.

F. Rousset. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145(4):1219–1228, 1997.

F. Rousset. *Genetic structure and selection in subdivided populations*. Princeton University Press, 2004.

F. Rousset. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, 8:103–106, 2008.

D. Serre and S. Pääbo. Evidence for gradients of human genetic diversity within and among continents. *Genome Research*, 14(9):1679–1685, 2004.

M. Slatkin. Inbreeding coefficients and coalescence times. *Genetical Research*, 58(2): 167–175, 1991.

M. Stephens. Bayesian analysis of mixture models with an unknown number of components —an alternative to reversible jump methods. *The Annals of Statistics*, 28(1): 40–74, 2000.

C. Strobeck. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics*, 117(1):149–153, 1987.

C. Tian, R. M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, and et al. Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genetics*, 4(1):e4, 2008.

M. N. M. van Lieshout. *Markov point processes and their applications*. Imperial College Press, 2000.

A. P. Verbyla. A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, 32(2):227–230, 1990.

C. Wang, Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 13, 2010.

C. Wang, S. Zöllner, and N. A. Rosenberg. A quantitative comparison of the similarity between genes and geography in worldwide human populations. *PLoS Genetics*, 8(8): e1002886, 2012.

S. K. Wasser, A. M. Shedlock, K. Comstock, E. A. Ostrander, B. Mutayoba, and M. Stephens. Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences (PNAS)*, 10 (41):14847–14852, 2004.

S. K. Wasser, C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, B. Clark, and M. Stephens. Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences (PNAS)*, 104(10):4228–4233, 2007.

G. H. Weiss and M. Kimura. A mathematical analysis of the stepping stone model of genetic correlation. *Journal of Applied Probability*, 2(1):129–149, 1965.

S. Wright. Isolation by distance. *Genetics*, 28(2):114–138, 1943.