Genetic structure and dispersal in plant populations

JOHAN FOGELQVIST
Dissertation presented at Uppsala University to be publicly examined in zootissalen, EBC, Norbyvägen 9, Uppsala, Friday, September 26, 2008 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in Swedish.

Abstract

This thesis focuses on the spatial structure and methods to identify spatial structure in plants.

Methods that investigate genetic structure can mainly be divided into equilibrium methods that reveal summed dispersal over many generations, and cluster methods, that reveal more recent dispersal events. Depending on the spatial level, local or global, suitable methods are different.

The thesis consists of four papers.

The first explores the spatial genetic structure in two epiphytic bryophytes that have different dispersal strategies (Orthotrichum speciosum and O. obtusifolium) using three different approaches based on pairwise kinship coefficients assessed from AFLP data. The spatial kinship structure was detected with both autocorrelation analysis and generalized additive models, but linear regression failed to detect any structure in O. speciosum.

In the second paper the spatial genetic structure in marginal populations of the forest tree Quercus robur is investigated at both local and regional scales. At the local scale, dispersal kernels as estimated using maximum likelihood parentage methods showed to be comparable to results acquired in central located populations. At the regional scale the degree of isolation at the margin of the distribution is shown.

The third paper compares a number of sibship clustering methods. It was found that the performances of the sibship reconstruction algorithms are strongly dependent on fulfilling the assumptions of the model and that using an overly simple model produced very unreliable results. The amount of information included in the model affected the results; models including all the available information outperformed the models using only a subset of the information.

In the last paper we show that the number of clusters as estimated by the software Structurama depends on sample size. At high number of subpopulations, the estimated number of clusters tends to be grossly underestimated when the number of sampled individuals per subpopulation is low.

Keywords: population genetics, dispersal, structure, plant, gene flow

Johan Fogelqvist,

© Johan Fogelqvist 2008

ISSN 1651-6214
urn:nbn:se:uu:diva-9211 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-9211)
Tree by tree by tree by tree
Karl
List of papers

This thesis is based on the following papers, which are referred to by their Roman numerals:


II Fogelqvist J, Parducci L and Lascoux M. Genetic variation and dispersal at the northern limit of the *Quercus robur* natural range (Manuscript)

III Fogelqvist J, Mortier F, Gerber S, and Lascoux M. Algorithms for reconstruction of full-sibs families from molecular markers: a comparison study (Manuscript)

IV Fogelqvist J, Niittyvuopio A, Savolainen O and Lascoux M. Cryptic population genetic structure: the number of inferred clusters depends on sample size (Manuscript)

Paper I is reprinted with kind permission from Blackwell Publishing.
Contents

Introduction.....................................................................................................9
Regional scale ..........................................................................................13
  Equilibrium methods ...........................................................................14
  Clustering methods .............................................................................15
Local scale................................................................................................17
  Parentage methods .............................................................................18
  Sibship classification ..........................................................................19
Spatial genetic structure in epiphytic bryophytes (Paper I) ......................23
  Dispersal pattern and spatial genetic structure in Quercus robur at the
  northern limit of the distribution (Paper II) ...........................................24
  Clustering individuals into sibships (Paper III) ....................................25
  Sample size dependent clustering (Paper IV) .......................................28
Conclusions...................................................................................................30
Summary in Swedish ....................................................................................31
  Spridning och rumslig genetisk struktur hos växter .........................31
Acknowledgements.......................................................................................34
References.....................................................................................................35
Introduction

Population structure is a key concept of population genetics. As an effect of restricted dispersal virtually all populations are, in some sense and to various degrees, structured. Population structure has played a central role in evolutionary biology: it is at the heart of Sewall Wright adaptive landscape and of theories of allopatric speciation. Recently population structure and its counterpart, gene flow, have received a lot of attention for practical reasons: pollen-mediated gene escape for transgenic crops (Lavigne et al. 1996; Baker & Preston 2006) is directly related to it and population structure can be a major nuisance in association studies, where it can lead to false positives (Zhao et al. 2007).

Spatial structuring of individuals, which is a consequence of limited dispersal and environmental heterogeneity, will make nearby individuals more genetically similar on average. In a homogenous landscape, limited dispersal will create a clinal pattern, but if the dispersal is restricted in some directions then clusters or subpopulations will arise. Clearly species have different dispersal abilities: some have high rates of dispersal and thus a high colonisation ability, whereas others might have a lower rate of dispersal, but instead could, for example, have higher offspring survival. These different life-history strategies leads to successions of the landscape: first an area is colonized by invasive species (that generally have a high dispersal rate), but after a while the area becomes more dominated by late succession species (that generally has lower dispersal capabilities but instead higher capacity to grow and survive in dense vegetation). Some plants that today might be considered as short dispersal species, might, in fact, have high dispersal capabilities under suitable conditions. For example oaks have been found to have a dispersal rate of several hundred meters per year (Huntley & Birks, 1983) during the colonisation following the last glaciation.

As plants may disperse their genes both by pollen and seeds, it is important to distinguish between gene flow and (seed) dispersal. An organism may have a high level of gene flow, but this does not necessarily imply that the organism has a high ability to colonize new areas. As pollen and seeds do not carry exactly the same DNA, for example in Angiosperms cpDNA is usually only transmitted through seeds, the effect of seed and pollen flow can, at least in principle, be disentangled. Usually pollen flow is much higher than seed flow (eg Ouborg et al. 1999), and dispersal curves are typically leptokurtic i.e. the majority of dispersal events being in the short distance
range, only a small proportion dispersing over longer distance. Despite being rare, these long distance dispersal events are of major importance since they can affect the species ability to colonize new areas, the probabilities of persistence in a fragmented habitat and metapopulation dynamics (Ouborg et al. 1999, Cain et al. 2000). Unfortunately, characterizing the tail of the distribution is inherently difficult and empirical studies are few (Cain et al. 2000).

Subpopulations arising from limited dispersal open up for local adaptations, making organisms more fitted to their environment, which could, in the end, lead to allopatric speciation. In association mapping, local population structure might lead to spurious correlations, thus ideally one should have samples with minimal population structure. These samples can, however, depending on the study, be very difficult to collect. Thus techniques have been developed to account for structured samples (e.g. Zhao et al. 2007). A requirement is however a correct identification of the clusters.

For various reasons people have been interested in identifying subpopulations for a long time. Most, if not all, population and quantitative genetics concepts are, by definition, defined at the population level. Thus any structure within the population has to been taken into account. For the most trivial cases where barriers to dispersal are obvious, e.g. isolated islands, geographical subpopulations are often well defined and easily identified, but in the case of a more or less continuous population the concept of subpopulations becomes more artificial. Sampled inaccurately, a population with a clinal change in allele frequencies could be identified as being constituted of a number of distinct subpopulations, which could be a misleading conclusion. The meaning of “subpopulation” will also depend on the spatial level of the study. At the global level, a globally distributed species might be separated into a rather small number of clusters, typically 4-6 or at least less than 10. Each cluster might in turn be constituted of a number of subpopulations, and so on. This is to say, there often is a hierarchical population structure that likely reflects processes acting on different time scales.

All identification of clusters is in essence based on finding individuals sharing alleles that are identical by descent (IBD). Individuals sharing many alleles IBD are assumed to be more related than individuals sharing less alleles IBD. In the presence of recurrent mutations two alleles can be identical by state, without being identical by descent. Nonetheless individuals sharing a high proportion of alleles identical in state (IIS) are likely to have a recent common ancestor. The impact of mutation on the data depends on the scale of the study/the amount of time that mutation has been acting and the type of genetic marker used. Ideally, mutation should have acted long enough to separate different clusters (as with no mutation there would be only one allele present) but not so much that the signal of IBD is lost. Identity in state is usually denoted as $Q$, and can in a subdivided population be refined into $Q_w$ denoting genes within subpopulations IIS and $Q_s$, denoting genes IIS between subpopulations. The quantity:
\[
F_{st} = \frac{Q_w - Q_b}{1 - Q_b}
\]

has since Wright (1931) been one of the most often used measure of genetic differentiation between subpopulations.

From a population genetics perspective, population structure acts by lowering the probability of identity in state of any randomly chosen alleles in the population, since there is a probability larger than zero that they come from different subpopulations. Considering first a single unstructured population, samples taken from that population can, under the assumptions of coalescent theory, be shown to coalesce at time intervals exponentially distributed depending on the number of lineages present at each time interval (Figure 1) (Kingman 1982).

![Figure 1. Coalescent tree representing a single unstructured population. The time between coalescent events depends on the number of lineages present.](image)

However if the population is structured (under the infinite island model), samples from within subpopulations will coalesce first, until all lineages are in separate demes, then the different lineages will start to coalesce at a rate depending on the migration among demes (Wakeley 2000). The genealogy of a structured population can thus be viewed as the sum of two coalescent
processes, acting on different time scales. The first and fastest process is usually called the “scattering phase”, as all lineages will be collected into separate demes. The second and slower process is called the “collecting phase”, as lineages from different demes coalesce. The slower process will behave exactly as a rescaled version of the standard coalescent process for an unstructured population, where lineages instead coalesce at time intervals depending on the number of demes present. Depending on the migration rate the phases may or may not overlap. The relative length of these phases will depend on the migration rate, the lower the migration rate the longer time will it take for the separate demes to coalesce. Thus the coalescent process will behave as if the population were larger.

![Coalescent tree representing a subdivided population.](image)

*Figure 2. Coalescent tree representing a subdivided population.*

Different models have been devised to reflect local population structure. Wrights (1943) F-statistics were developed assuming an infinite island model, where each island contribute equally to a pool of migrants that spread randomly to all islands. This was later refined to the “stepping stone” model, where subpopulations are considered to be arranged in a one- or two-
dimensional space, and each subpopulation has a certain migration rate to its neighbours. As nearby individuals (demes) are more related to each other than to more distant ones, these are called “isolation by distance” models. By setting the migration rate appropriately, various population structure patterns such as hierarchical structure and contact zones can be modelled. However, by adding parameters the complexity of the model increases so that analytic solutions are more difficult to obtain (Wakeley 2005, Rousset 2001). Even if the infinite island model is overly simplistic, it remains popular because of the ease of calculations.

![Migrant pool](image)

**Figure 3.** Graphic representation of three migration models, each circle represents a subpopulation. **A:** Island model. Each subpopulation sends and receives migrants to a common migrant pool. **B:** Two-dimensional stepping stone model. In this example migration can occur only between neighboring populations, the dark gray population in the middle can thus send and receive migrants from the light grey populations. **C:** Hierarchical structure consisting of four demes (light-gray) each constituted of four subpopulations. Migration can occur between any pair of subpopulations, but is more likely within deme than between.

**Regional scale**

Depending on the question at hand, samples may be taken differently. For example, in association studies samples are often taken at a large geographical scale, but if one is interested in directly estimating dispersal kernels samples must be taken at a local scale. Thus different techniques have been developed to deal with these different cases. One may also be interested in either (or both) average dispersal over many generations or more recent dispersal events. Estimates of the former would be the relevant ones to address.
evolutionary questions, whereas estimates of the latter would be needed for conservation or behavioural studies.

Equilibrium methods

Dispersal is often difficult to observe directly. For large organisms, such as moving animals or heavy seeds, direct observation can be done by, for example, attaching radio transmitters to the organism whose movements can thereby be traced. Using this technique, we can get very exact measures of individual dispersal distances, but the drawbacks, apart from being quite labour-intensive, is the fact that in population genetics only realized dispersal is of interest. For example, the first experiments studying pollen dispersal used fluorescent dye marked pollen at a source plant and trap plants at various distances, whereby the movements from the source to the sink could be directly observed (Campbell and Wasser 1989). However when genetic markers emerged at enough resolution and one could compare the result with realized dispersal, defined as the proportion of offspring at various distances that actually were pollinated by the source plant, the fit was poor; much of the gene flow came from outside the studied area, and the average distances were greater than estimated by direct pollen movement (Campbell 1991 – allozymes, \textit{Ipomopsis aggregata}). The most probable cause for this discrepancy is the genetic correlation of nearby individuals resulting from limited dispersal. Thus nearby individuals are likely to be siblings, mother offspring pair or related in some other way. Post pollination events, i.e. reduced stigma reception of related pollen grain or inbreeding depression might then lead to seed abortion in those cases. As population structure is closely related to dispersal, it is often much more feasible to make indirect inferences on the dispersal process by studying the population structure. For example, under the rather demanding assumptions of the Wright island model (Wright 1931, 1943) classic F-statistics is directly linking the amount of genetic structure (Fst) to the average number of migrants (Nm) each generation:

\[ F_{st} = \frac{1}{4}N_{m} + 1 \quad [2] \]

Basically this approach compares the variation within and among subpopulations, and computes the number of migrants per generation that would explain this structure under equilibrium. This means that the population genetic structure observed is the result of gene flow and migration over many generations. Given that the populations are at drift-migration equilibrium and having large enough samples, we can potentially detect very low frequency dispersal events using this indirect approach. Apart from assuming equilibrium, the F-statistics also cannot separate the effect of migration and mutation, which becomes a problem when studying situations where the mutation rates are of the same order of magnitude as the migration rates. This can be
the case, for instance, when populations are geographically very distant and have diverged long ago. Unfortunately, there is no equivalent to equation [2] for more realistic models. One can, however, still use $F_{ST}$ statistics to infer some aspects of gene flow. For instance in a 2-dimensional isolation-by-distance model, considering all subpopulation pairs, the regression of $F_{ST}/(1-F_{ST})$ on log(distance) yields a slope inversely proportional to $4D\delta^2$, where $\delta^2$ is the variance of the one dimensional dispersal distance and $D$ is the population density (Rousset 1997). The quantity $4\pi D\delta^2$ is often referred to as neighbourhood size, although the meaning of that term remains unclear (Rousset 2001). $F_{ST}$ can in this model be defined at the individual level, alleviating the need to a priori define subpopulations (Rousset 2000). Estimates obtained with this approach have been shown to be reasonably closed to direct estimates (mark-recapture) of mean dispersal distance and neighbourhood size (e.g. Sumner et al. 2001). At the local level, autocorrelation methods such as Moran Index are sometimes used to characterize population structure. This method does not estimate any dispersal parameters, even if Epperson and Li (1997) established a numerical correlation between some autocorrelation statistics and $D\delta^2$.

In practice populations will seldom be in equilibrium and the assumptions of the Wright-Fisher model are usually violated. Population size may fluctuate over time, sex-ratio may be uneven, and selection may act on adjacent loci even if neutral markers are used. Theoretically this can be accounted for by scaling the population size to a size that behaves as an equilibrium population, in practice the effective population size is however difficult to estimate. Usually the effective population size is lower than the census size, though, interestingly, population subdivision can increase the effective population size.

Recently, as an effect of the increasingly amount of genetic information available as well as the increase in computational power, alternative approaches have started to emerge and become practicable.

Clustering methods

As $F_{ST}$ and $F_{ST}$-related methods are based on equilibrium and thus measures the effective dispersal summed over many generations, these methods do not always yield relevant estimates, if one is primarily interested in recent demographic events. Furthermore $F_{ST}$-based approaches are intrinsically single-locus methods and thereby do not use linkage disequilibrium, the non random association of alleles at different loci, Yet, linkage disequilibrium is influenced by recent demographic events too and therefore constitutes a source of information on the latter. The realization of this, together with the availability of large number of highly variable loci have led to the development of a new class of methods that might be called assignment methods, or clustering methods. These individual-based methods try to assign the sam-
pled individuals into different clusters or into some kind of predefined popu-
lation structure. A criterion of how a subpopulation is constituted, i.e. Hardy-
Weinberg and linkage equilibrium, is the basis of these approaches, and then
individuals are partitioned among the clusters in a way that maximize the fit
to this criterion. The number of clusters and the assumed origin of individu-
als may or may not be included depending on the method. Using this ap-
proach recent immigrants or their descents may be identified. The estimated
structure obtained through these methods is thus reflecting more recent dis-
persal events than FST based methods.

In assignment tests, as implemented in geneclass (Cornuet et al. 1999) for
instance, all individuals are first partitioned into predefined clusters (the
location where the individuals have been sampled for instance). Thereafter
the most probable origin of each individual, one at a time, is determined.
This is done calculating the likelihood (Peatkau et al. 1995) or the marginal
probability (Rannala and Mountain 1997) of observing the genotype of the
focal individual given the allele frequencies in each subpopulation. Confi-
dence level of the analysis can be checked using simulations. Using these
approaches recent immigrants can be detected, and the amount of calcula-
tions is limited, proportional to the product of the numbers of samples and
clusters. However the assumed population structure may not be correct. And
if subpopulations are not spatially separated and therefore not easily distin-
guishable, it might be difficult to pre-assign individuals into clusters in the
first place.

To address this problem, Pritchard et al. (2000) developed a Bayesian ap-
proach where individuals could be partitioned into clusters without any prior
information of population structure. The idea behind the method is to create
clusters that are in Hardy-Weinberg and linkage equilibrium, by assigning
individuals to a predefined number of clusters, K, until the best possible
configuration is achieved. As the number of possible configurations quickly
becomes enormous, Markov chain Monte Carlo simulations must be used.
The whole procedure could then be repeated for different number of clusters,
and the most probable number of clusters can be determined using different
methods (Pritchard et al. 2000; Evanno et al. 2005). The method quickly
became hugely popular and was later extended to incorporate linkage among
markers (Falush et al. 2003). It is certainly the most widely used method for
estimating population structure. There are, however, a number of alternative
methods available, some of which are compared in Latch et al. (2006) and
Chen et al. (2007). Being Bayesian, additional information such as geo-
 graphical origin and selfing rate is relatively easy to incorporate in the
framework, which is also done in some of the methods reviewed by Latch et
al. (2006) and Chen et al. (2007). Using simulations, the approach has been
shown to perform well under a wide range of conditions (Rosenberg et al.
Latch et al. 2006, Chen et al. 2007). However as the concept of population,
used by these methods, is rather artificial, the result may sometimes be difficult to interpret (Setakis et al. 2006). Often the spatial distribution of individuals is more or less continuous. As dispersal generally is limited we will have some spatial structure, at least isolation by distance. But this structure might be gradual as well as clumped. If we were to sample a number of individuals at some random locations and analyse the output, we would probably get a number of clusters, regardless if the spatial structure is gradual or clumped. This has led to a dispute about the spatial genetic distribution of humans, some arguing that there is mainly gradual change in allele frequencies among human populations and that the discrete units implicitly suggested by STRUCTURE or similar programs are misleading artefacts (Rosenberg et al. 2005, Serre & Pääbo 2004, Lawson-Handley et al. 2007)).

Local scale

At the local scale cluster methods are often inappropriate as individuals are too related to separate into distinct clusters. Thus other methods are more suitable to study the local genetic structure.

The amount of structure at the local scale can, for instance, be investigated by pairwise comparison of point estimators such as relatedness or kinship coefficient. As these point estimators are continuous measures of the degree of relatedness between pair of individuals compared to the population mean, family structure on the local scale can be approximated. For example full sibs are expected to share 50% of their alleles, whereas half sibs share on average only 25% of their alleles. Commonly used point estimators are kinship coefficients, often defined as probabilities of identity by descent (which is not perfectly true when the estimate of the parameter is based on genetic markers) (i.e. Loiselle et al. 1995, Ritland 1996), and relatedness coefficients, which estimate the genetic similarity of two individuals relative to a reference population (i.e. Queller and Goodnight 1989). The drawback of these approaches is that if not only full sibs vs unrelated pairs are considered, several relationships may have the same expected value, for example both full sibs and parent-offspring have an expected allele sharing proportion of 50%. Furthermore the large variance of the estimate makes specific inferences on familial structure difficult unless the familial structure is really simple, i.e. the sample comprises only sibs and unrelated individuals. On the other hand, having enough genetic information we can get closer to the expected value, and if detailed knowledge about familial structure is not important, kinship and/or relatedness measure yields interesting information about local genetic structure. Article I uses this approach to identify genetic structure among epiphytic mosses in a Finnish forest.
Parentage methods

As the number of individuals inhabiting a local population is limited, it is sometimes, depending on organism, possible to sample all or at least the majority of the potential mates and offspring within an area. As genetic information is nowadays obtained at high resolution and speed, it has become feasible to genotype a set of offspring and a large fraction of the potential parents. Information on dispersal and mating can then be obtained by direct comparison of the genotypes and the fit to expectations given Mendelian rules. If the maternal genotype is known we can use the joint information of offspring and mother genotypes to estimate the paternal genetic part. If a great portion of potential fathers is also known, we can infer the paternity of offspring with great precision and thereby acquire detailed information on individual dispersal events. This is usually called parentage studies, and a number of methods have been developed to tackle this (reviewed in Jones et al. 2003). The major advantage of these methods is that very detailed estimates of dispersal traits can be obtained, given that enough potential parents have been sampled. The drawback is, however, that a fair amount of genetic information is required for the method to work properly. Often a large part of the gene flow comes from outside the studied area (i.e. Streiff et al. 1999) and thus the tail of the distribution kernel is left unknown. These methods are most suited for studying long-lived species without seed bank as offspring with no found parents are translated into gene flow from outside the stand, which might not be the case for seeds originating from a seed bank, or if the parent(s) died prior to sampling.

The first approaches involved direct comparison of the genotypes, aiming to exclude all potential males but the one to be the real father. For example, if a offspring has genotype a1a2, b1b2, c1c2…n1n2 and the mother has genotype a1a3, b2b3, c1c2, ..., n1n2, the father has to be either a2ax,b1bx,c1cx or c2cx,…,n1nx or n2nx. In practice total exclusion of all but one male is seldom achieved (Chakraborty et al. 1988). Firstly, errors inevitably enter large datasets, which is a problem that increases as more loci are added, and secondly unless the population is completely isolated a fraction of pollen will come from outside the investigated area. This problem was tackled through fractional (Devlin et al. 1988, Smouse and Meager 1994) and categorical (Meagher & Thompson 1986) likelihood based procedures. The principle is to calculate the likelihood of each parent-offspring relationship: fractional methods then split the parentage of offspring to all potential parents proportional to the likelihood whereas the categorical methods assign each offspring to a particular parent. Fractional methods were shown to have better statistical properties in at least some cases (proportion of offspring parented by each adult (Devlin et al. 1988; Smouse & Meagher 1994; Neff et al. 2001), allow comparison of the reproductive success of different categories of males (Nielsen et al. 2001; Signorovitch &
Nielsen 2002) and the incorporation of prior information about the biology of the species into the analysis (Neff et al. 2001; Nielsen et al. 2001), but suffer from being biologically unrealistic as no offspring can have shared paternity. In categorical methods, after the likelihood of each parent–offspring pair has been calculated, simulations are usually carried out to decide proper threshold values above which a father could be regarded as the true father (Marshall et al. 1998; Gerber et al. 2000). As error is inevitable in large datasets, some kind of error is usually incorporated in the model. Without errors incorporated in the model the precision usually decreases as more loci are added to the data, and there seems to be little or no negative effect of assuming errors in an error free data set (Sancristobal & Chevalet 1997). Often parentage can be determined with great precision, given that a sufficient proportion of the potential males have been sampled. However, in numerous studies, the parentage of a large proportion of the offspring can not be determined. Apart from the offspring that are left unassigned due to lack of power, these will usually be the product of the most long distance dispersal events, as the potential fathers typically are sampled around the trap trees. In studies of dispersal, we are often interested in the shape of the dispersal kernel, whether it is leptokurtic or not, a question difficult to answer in the absence of samples from the tail (Dyers 2007). An alternative approach, not relying on sampling all potential fathers, was therefore developed by Smouse et al. (2001). This method, called TWOGENER, is based on the differentiation among inferred pollen pools of a sample of females. The estimated parameter Φ_{st}, that is analogous to Fst, compares the variation within and among pollen pools and is easily translated to the average pollination distance (times adult density). The TWOGENER method is not directly comparable to parentage studies, the former putting emphasis on “effective pollen flow” suggest in general smaller number of effective pollinators and yields shorter average pollination distance (Smouse and Sork 2004).

Sibship classification

In recent years, with the parallel increase in computational power and genetic marker availability, it has become possible to reconstruct pedigrees using genetic information only. Several methods are available, all using different assumptions (see review in Blouin 2003). Sometimes information about parentage and/or parental genotypes can be included, missing data may or may not be accepted, error may or may not be included in the model and the assumed familial structure may differ in complexity. Often sibship reconstruction methods are broadly divided into pairwise methods, considering all pairs in a sample, and full partition methods, that jointly tries to estimate the most likely partition using all information simultaneously. Pairwise methods require less calculation than partition methods as the number of pairs increases only with the square number of samples, and were the first to
be explored. The reconstruction is based on using molecular data to investigate pairwise relationships between individuals, either by moment estimators such as relatedness (Lynch 1988; Queller & Goodnight 1989; Ritland 1996; Lynch & Ritland 1999) or through likelihood, modelling the probability that a given pair of individuals belong to a given set of relationship (Thompson 1975; Mousseau et al. 1998). Full partition methods are more computationally demanding, but make more efficient use of the data. By considering only pairs of individuals at a time, pairwise methods can never exclude two individuals from being full-sibs, but with three individuals such an exclusion is possible with three or more codominant alleles (ignoring mistyping/mutation) (Almudevar and Field 1999). Full partition methods usually try to find partitions of offspring into full- and half-sibships compatible with Mendelian inheritance, the likelihood of the configuration is calculated, and in the end the most likely configuration represents the estimated family configuration. As the number of configurations is enormous, even for a limited sample, a full search is not possible in a reasonable amount of time if the number of individuals exceeds about 15-20. To arrange a sample of 100 individuals into clusters of some kind, there are about $4.76 \cdot 10^{115}$ possible configurations. The most commonly used method to address this issue is so-called Markov Chain Monte Carlo (MCMC) simulations. MCMC enables us to explore a huge parameter space by sampling at different points in the parameter space, and then investigate the resulting sample. As the number of samples taken increases, the distribution of the samples taken is converging towards the stationary distribution of the Markov chain. Samples are taken by defining a chain that moves in the parameter space in a specified way, so that it does not get stuck somewhere, and is more likely to move to a point of high likelihood than to a point of low likelihood. In other words the Markov chain has to be irreducible (able to move from anywhere to anywhere) and aperiodic (it does not cycle). It can be shown that for any irreducible and aperiodic Markov chain there exists exactly one stationary distribution, and that the distribution of samples is converging to the stationary distribution as the number of samples increases to infinity (Häggström 2002). To show that a Markov chain is aperiodic and irreducible is often not very easy, it is often easier to show that the chain is reversible (equally likely to move from a to b as from b to a), in which case the chain also can be shown to have a stationary distribution (Häggström 2002).

The Metropolis algorithm is the most widely used one for updating the chain: The process basically has three components, $\pi$ (the distribution of interest), $T$ (a transition function that controls the movement between states) and the chain of states $x_1, x_2 \ldots x_n$. Now we start in state $x_1$. Then we induce a random perturbation using the transition function $x_1 \to x'$ and calculate the change using the acceptance/rejection function:
\[ \rho(x, y) = \min \left\{ 1, \frac{\pi(y)T(y, x)}{\pi(x)T(x, y)} \right\} \]  

[3]

We then accept the change with probability \( \rho(x,y) \) and let \( x_2 = x' \), otherwise let \( x_2 = x_1 \). If \( T(x', x) = T(x, x') \) the transition function is said to be symmetric and cancels out from the acceptance/rejection function. This special case of Metropolis updating was originally suggested by Metropolis \textit{et al.} (1953) and Hastings (1970). For discrete spaces, Peskun (1973) showed the Metropolis algorithm to be optimal in terms of statistical efficiency, however in terms of convergence rate this is less clear.

Another commonly used algorithm is the Gibbs sampler (Geman & Geman 1984), also a special case of Metropolis–Hastings transition. The point in Gibbs sampler is to reduce the number of dimensions by constructing a Markov chain of a sequence of conditional distributions which are chosen so that \( \pi \) is invariant with respect to each of these “conditional” moves. \( X \) is thus viewed as \( x = (x_1, x_2, \ldots, x_d) \), and at each \( t \) one chooses one \( x_i \) \( (x_1 \ldots, x_d) \) to propose a change, conditioned on \( x[-i] \). This can be made randomly or systematically. The Gibbs sampler can speed up convergence, especially if correlated variables are grouped together in the Gibbs sample. The conditionals needed in the update are also commonly available in Bayesian and likelihood computations.

To improve mixing, i.e. allow the MCMC scheme to move more freely across the state space, parallel tempering or Metropolis coupled chains (Geyer 1991) is sometimes used. The idea is to run several differently tempered chains simultaneously: in the hot chains the probability surface is flattened so that movements in the chain are more probable. So, we have a number of distributions \( \pi_i \) with different tempering, and we sample from all chains simultaneously. In each step, if \( u \sim \text{Unif}[0,1] \leq a_0 \) we do a parallel step, updating all \( x_i \) to \( x_i^{t+1} \), else we do a swapping step and randomly choose two neighbouring chains, \( i \) and \( i+1 \) and propose a swap using standard Metropolis update. This approach had been shown to be powerful and can utilize the information from multiple MCMCs. However, in Metropolis coupled chains, only samples from the cold chain are used when making inferences (Brooks 1998). The drawback is thus wasting a lot of samples, for the sake of good mixing.

In many cases, one must have a sampler that can jump across different dimensional spaces. This is problematic, but in principle this can be solved, see Green (1995) for a formalization of this reversible jump. The problem is the fact that not all moves from the higher dimension space to the lower dimension space can be reversed directly as information is lost in the transition to the lower dimension space. The solution is to come up with a “matching space” \( Z \), so that the dimensions of lower dimension space times \( Z \) is equal to the dimensions of the higher dimension space.
Now that we have a way of sampling the distribution, we only need to define our model. In the case of reconstructing a family structure, this model is often based on Mendelian inheritance, i.e. we define the probability of observing a family group as the probability of the offspring inheriting their genotypes given the parental genotypes, where the parental genotypes can be known or, most often at least partially, missing. In the case of missing parental genotypes, the probability may either be calculated integrating over all compatible parental genotypes, or the parental genotypes can be treated as an unknown variable, and estimated jointly. In the simplest scenario of only full sibs vs. unrelated, the parental genotypes are restricted as all alleles present in each sib-group originate from the two parents (ignoring mutation and mistyping). If one sex is polygamous, the space of possible genotypes is increased but sib groups remains as distinct units sharing the same mother or father depending on which sex is monogamous. If both sexes are polygamous there are no longer any distinct sib-groups and all individuals may in principle be interconnected, thus it is very demanding to integrate over all possible parental genotypes; these are better treated as a random variable in such cases. Technically, we are not limited to estimate sib-groups, the same methods could in principle be used to estimate more complex family groups, such as sibs-parents-grand parents. The genetic precision needed is however increasing with the complexity of the problem.

\[\text{ Increasing complexity of familial structure}\]

*Figure 4.* Schematic illustration of mating systems. The presence of polygamous sex increases the complexity of the familial structure. If both sexes are polygamous, as in wind-pollinated plants, all offspring may be interconnected. **A:** Monogamy, only full-sib and unrelated pairs are considered. **B:** Polygamy, one sex is multiple mating so that both full- and half-sibs are considered. **C:** Promiscuous, both sexes have multiple mating.
Spatial genetic structure in epiphytic bryophytes (Paper I)

The aim of paper I was to assess the spatial genetic structure of two epiphytic bryophytes, Orthotrichum speciosum and O. obtusifolium, that have different dispersal strategies. Both mosses are haploid, but the former is monoecious and disperses only by spores (≈20 μm) whereas the later is dioecious and disperses by spores but also by gemmae (≈100 μm), which are asexually produced outgrowth from the leaf, usually much bigger than spores. Both species grow on the bark of tree species with high bark pH and nutrition content (Kuusinen 1996).

This study was conducted in central-eastern Finland (68°98’ N, 29°14’ E) in the semi-natural nature reserve of Teeri-Losouo. Within an area of 210 ha all individuals were sampled whenever they were found within reach, in total 79 and 85 samples were found of O. speciosum and O. obtusifolium respectively. Various environmental variables were also recorded at each sampling location, to separate the effect of restricted dispersal from local adaptation. As no microsatellite markers were available for these organisms, genotyping was done using amplified fragment length polymorphisms (AFLP), a dominant marker system, which is assumed to be neutral. The spatial genetic structure was investigated using pairwise approach: first all pairwise kinship coefficients were estimated following Loiselle et al. (1995), then the spatial structure of kinship was analysed using three different approaches: linear regression models (similar to Mantel test; Manley 1997), generalized additive models (GAM), and spatial autocorrelation analysis (similar to Moran I). Spatial genetic structure was found for both species in all models except for the linear regression model in O. speciosum. In both species pairwise kinship coefficients were significantly higher than expected from random for distance up to 300-350m, and lower in the highest distance class. However there was no association between kinship coefficients and local environmental variables.

This was the first study using genetic marker to investigate dispersal pattern at the landscape level in epiphytic bryophytes. Restricted dispersal has previously been confirmed in O. speciosum (Snäll et al. 2003, Hedenås et al. 2003) but has been difficult to detect in O. obtusifolium. This illustrates the problem of drawing inferences about dispersal from species distribution pattern.
Dispersal pattern and spatial genetic structure in *Quercus robur* at the northern limit of the distribution (Paper II)

In this paper the spatial genetic structure of *Quercus robur* is investigated at both the local and regional scales. *Q. robur* is a long lived forest tree with wind dispersed pollen and heavy seeds. Because of its important economic value in central and southern Europe it has been studied thoroughly. The pattern of colonization of Europe since the last ice age is rather well known (Petit *et al.* 2002) and, at the local scale, several paternity studies have been performed revealing detailed information on dispersal distances of pollen (e.g Streiff *et al.* 1999, Valbuena-Carabánña *et al.* 2005). Several microsatellite markers are available (Steinkellner *et al.* 1997) as well as linkage maps (Barreneche *et al.* 1998). However no study had been performed at the northern limit of the distribution: The repeated founder effect during colonization after the last ice age is expected to have decreased the amount of genetic variation (Austerlitz *et al.* 2000) and populations at the margins are often more fragmented than central populations. Also, most studies done so far have been performed in more or less managed forests, in contrast we used a stand with a natural age distribution of trees, that is known to have been unmanaged for some time apart from some selective felling of conifers.

At the regional level, about 30 samples (33.18± 5.81) were taken from eight forest stands located along the river Dalälven and from four stands located at various distances from the river, and genotyped at six microsatellite loci (zag1/5, zag20, zag104, zag9, Steinkellner *et al.* 1997, msq4, msq13, Dow *et al.* 1995).

We then analysed the spatial structure using both F-statistics and a clustering algorithm as implemented in STRUCTURE 2.1. Using F-statistics, only weak isolation by distance was found over the total range of the investigated area, but looking more in detail at the populations located along the river we found a clear trend with increasing average pairwise Fst values and decreasing FIS values as one moved upstream the river. As the distribution of *Q. robur* is getting patchier as one moves upstream this pattern probably reflects the degree of isolation in these populations. Using the clustering approach we were able to separate the populations located along the river Dalälven from the southern ones, with the individuals populating the intermediate population showing shared ancestry from the two clusters. This could perhaps seem a bit contradictory to the fact that the F-statistics showed no separation of the Dalälven vs non-Dalälven populations, but instead showed variation among the Dalälven populations. However this variation was clinal, whereas the cluster algorithm searches for more distinct units.
For the local analysis, all adult individuals (442) were sampled in one population located along the river Dalälven. We also sampled acorns from 47 randomly selected trees and all 1-yr old seedlings we could find. In total 951 acorns and 152 seedlings were sampled. On 76 of the sampled seedlings it was possible to retrieve the pericarp tissue, that is of maternal origin and thus in principle could provide us with the maternal genotype of the seedling. We used two maximum likelihood methods (Marshall et al. 1998(?) and Gerber et al. 2003) to identify the most likely father among the sampled parents for all acorns. As in other similar studies performed in southern Europe we found that the majority of pollen originates from outside the stand and the resulting pollen dispersal function were also very similar to the ones obtained by Streiff et al. (1998). The maternity of sampled seedlings was determined using two methods: first we used the genotype of the maternally inherited pericarp whenever it could be found attached to the seedlings. As already mentioned this would in principle give us the maternal genotype, unfortunately, as the DNA content was degraded, the genotype could only be partially be retrieved and in only a few cases the complete genotype could be unambiguously scored. Thus we used an error tolerant maximum likelihood approach to identify the most likely mother given the partially retrieved genotype of the pericarp. Secondly we used the maximum likelihood method described by Gerber et al. (2003) to identify the most likely parent pair among the sampled adults, given the seedling genotype. The two methods gave concordant results in only a few cases. This could simply be due to a lack of power as only 36% and 8% of the seedlings with and without pericarp tissue could be assigned maternity. Surprisingly we found that very small trees contribute to a relatively large fraction of the seedlings. This illustrates the difference between actual and realized seed flow, large trees produces far more acorns, but for germination and establishment the acorn has to be dispersed into a suitable environment, which, perhaps is more likely found near smaller trees.

Clustering individuals into sibships (Paper III)

As in the previous paper, it is common in parentage studies to observe a high amount of gene flow coming from outside the study area. One way to make use of the information in the offspring with no assigned fathers is to reconstruct the sibships, thereby gaining information about the structure of long distance dispersal events. There are many sibships algorithms available, based on different methods and having different assumptions and requirements. This study compares a number of methods on both simulated datasets and real data acquired for parentage studies. We thus have a number of offspring with maternity and maternal genotype information, and we want to reconstruct the family structure of the offspring. It is a difficult problem, as
we could expect the family sizes to be very small, and both parents to be polygamous. Thus the problem does not fit the assumptions of a number of the methods, most of which assume that the families are comprised of full-sib families nested within half-sib progeny arrays. On the other hand we know the maternity and the maternal genotypes, so that the paternal part can easily be extracted in most cases.

We tested seven methods and a novel one. The performance of a method was measured as its ability to correctly estimate the proportion of paternal half-sibs and maternal full-sibs, as well as the power and false assignment rate, using data simulated under different familial conditions.

We found that the performances of the sibship reconstruction algorithms are strongly dependent on fulfilling the assumptions of the model and that using an overly simple model produced very unreliable results. The amount of information included in the model affected the results, models including all the available information outperformed the models using only a subset of the information i.e. only the offspring genotypes. Most methods could not separate paternal sib-groups from the case of no paternal family structure, i.e. all offspring having different fathers, unless fathers sired a relatively large number of offspring (about more than 4).

In the real datasets, most methods could detect a higher proportion of maternal full-sibs than would be expected assuming all offspring having separate fathers, however only a few could detect a higher proportion of paternal half-sibs.
Figure 5. Inferred full-sibs and paternal half-sibs in simulated data, full-sib estimates in the left column, paternal half-sib estimates in the right column. Results are grouped and averaged over three kinds of methods: promiscuous, half-sibs and full-sib methods. Top row: estimated proportion of maternal half-sib pairs (left) and paternal half-sib pairs (right). The black dotted line represents the true proportion. Mid row: power to detect maternal full-sibs (mFS) and paternal half-sibs (fHS). Bottom row: false assignment rates of maternal half-sib pairs (left) and paternal half-sib pairs (right).
Sample size dependent clustering (Paper IV)

In the last article, we investigated the effect of sample size on the estimated number of clusters, when the sample consists of a rather high number of subpopulations. The study was motivated by the observation that the resulting cluster numbers observed in *Arabidopsis thaliana* seemed to be dependent on the sampling strategy. In most cases, a single individual was sampled per location, the logic being that *A. thaliana* being a selfer, subpopulations would be in practice monomorphic (Kuittinen et al. 1997; Bergelson et al. 1998). However in studies where a larger number of individuals were sampled per location, this did not seem to be the case (e.g. Nordborg *et al.* 2005, Stenoien *et al.* 2005). It could also be noticed that in studies where more individuals were sampled per location, differentiation among subpopulations was found to be higher (e.g Stenoien *et al.* 2005, Bakker *et al.* 2006).

STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.* 2003) is the most widely used clustering algorithm, although there are several other methods available. It has been shown to behave well in a wide range of controlled conditions, its main limitation being the amount of computing power needed, especially when the number of clusters is large, as each MCMC simulation has to be conditioned on a specific value of K, the number of clusters. Thus we used STRUCTURAMA (Huelsenbeck & Andolfatto 2007) for the estimation of the number of clusters, which is a similar approach based on the same principles, but which jointly estimate the number of clusters.

As the population structure could be assumed to affect the resulting clustering, we simulated data assuming both an island model of migration, and a hierarchical model. In both cases we assumed 20 subpopulations. In the hierarchical migration model, the populations were grouped into five equally sized demes. We then varied the migration rates, and the amount of recombination.

To simulate different sampling strategies, we created subsets of the dataset, by randomly drawing 1-10 individuals from each subpopulation, and estimated the number of clusters using STRUCTURAMA. As STRUCTURAMA is not as widely used as STRUCTURE, we also did a limited study instead using STRUCTURE for the estimation of K, to verify that the results were similar.

We found that in almost all cases, the sampling strategy had a large effect. Sampling only a few individuals per subpopulation led to a strong underestimation of the number of clusters. The effect of sampling was weaker in the hierarchical population structure with high migration. In these cases the number of clusters found corresponded to the number of demes, thus even if only one individual was sampled per subpopulation, as there were four subpopulations in each deme the number of samples per deme never was below four. The sample size dependency on the estimated number of clusters was also shown in a real dataset comprised of *A. thaliana* sampled from mainly
Scandinavia. This suggests that variation within subpopulation is important for successful estimation of population structure even in highly selfing species such as *A. thaliana.*

![Graphs showing inferred number of clusters]

**Figure 6.** Number of clusters inferred by *Structurama* as a function of the number of individuals sampled from each subpopulation, for four cases: island model with high (top-left) and low (top-right) migration, and hierarchical structure with high (bottom-left) and low (bottom-right) migration. Mean (solid line) and 95% confidence interval (dotted line) based on the simulated data is showed. The total number of simulated subpopulations is 20, in the cases of hierarchical structure partitioned into five demes. Simulations were carried out assuming stepwise mutation model and free recombination.
Conclusions

Classical population structure descriptors such as Fst are still widely used. The theory behind is rather robust and the ease of use makes Fst probably the most used measurement of population structure. The main drawback is the lack of resolution to reflect recent dispersal and to study dispersal at the individual level. Recent development of Bayesian and likelihood based cluster and pedigree reconstruction methods have yielded several algorithms that can be shown to work, given that the assumptions of the methods are met and that we have enough genetic resolution. Conformation to the assumptions of the model seems to be of high importance, for example a method only discriminating full-sibs from unrelated will have problems in a mixture of full-and half sibs. The genetic resolution needed can be rather large, and the time needed to perform the necessary calculations can in some cases be extremely long. However, as the computational as well as the genotyping capacity is continuously growing, this might be a vanishing problem so that the complexity of clustering problems that can be solved is also increasing.
Spridning och rumslig genetisk struktur hos växter.


Metoder för att studera genetisk struktur på en regional nivå kan grovt sett delas in i jämviktsmetoder, som studerar effekten av spridning över många generationer, och klustermetoder, som försöker partitionera individer i olika subpopulationer. Jämviktsmetoder är baserade på $F_{ST}$, ett mått som jämför variationen i allelfrekvenser inom och mellan subpopulationer och är, givet Wrights $\theta$-modell (Wright 1943, 1931), direkt kopplat till antalet migranter per generation. Klustermetoder är istället baserade på att hitta den partition av individer som är mest sannolik givet de genetiska data man har. På den lokala skalan kan man även studera individuella pollineringar och spridning-
ar av frön med hjälp av vad som kallas föräldrastudier. Alla potentiella föräldrar och avkomma inom ett område karteras och genotypas och den mest troliga föräldern kan utses.

Avhandlingens första artikel studerar den rumsliga genetiska strukturen hos två epifytiska mossor, hättemossa (Orthotrichum speciosum) och trubbhättemossa (O. obtusifolium), i norra Finland. Eftersom epifyter växer på andra träd är de även beroende av värdträdens spridning och utbredning, de måste hinna med hela sin livscykel innan värdträdet faller. Begränsad spridning har tidigare konstaterats hos Orthotrichum speciosum/O. obtusifolium med hjälp av ekologiska metoder, då man studerat utbredningsmönster. Denna studie kunde verifiera detta med genetiska metoder.


Den tredje artikeln studerar olika metoder att rekonstruera familjegrupper utifrån genetiska data, som t ex. de ekollen som ej tilldelats någon far i artikel två. De enklaste metoderna, som enbart letar efter helsyskon, var otillräckliga i situationer med mer komplexa familjestrukturer. De mer avancerade metoderna som även beaktar halvsysskon och promiskuösa föräldrar gav bättre resultat men de datorbaserade analyserna kunde vara mycket tidskrävande. Eftersom vindpollinerande växter som ekar genererar en väldigt komplicerad familjestruktur, måste avancerade och tidskrävande metoder tillämpas, tillika krävs en hög genetisk precision (många genetiska markörer) för att säkert bestämma familjegrupper.

Avhandlingens fjärde och sista del studerar hur insamlingsstrategier kan påverka resultatet vid studier som söker dela in ett urval av självbefruktande organismer i olika subpopulationer.
Då t.ex. backtrav (*Arabidopsis thaliana*) är självbefruktad har man länge tänkt sig att det räcker med ett prov från varje lokal, eftersom alla individer inom en lokal borde vara genetiskt identiska. I studier som analyserat flera individer från samma lokal visade det sig dock att det fanns genetisk variation inom lokalen. Vi visar, både med hjälp av simulerade data och med ett riktigt dataset bestående av *A. thaliana* huvudsakligen från Skandinavien, att ett för litet antal analyserade individer per lokal leder till en underskattning av antalet subpopulationer.

Acknowledgements

First of all I want to thank my main supervisor Martin Lascoux for supporting me in life, music and science during these years. My second supervisor, Sophie Gerber, has also supported me, mainly in science, and given me valuable advises.

Thanks to all the people at the department that have helped me and made the time funnier be a during the years, Niclas for always having a solution for whatever problem, Hanna for lot of work during my parental leave, for being like an extra mum, and also giving me your room in the last year, Laura for helping me with the pericarp lab work, Tord for the collaboration with the bryophytes, Cilla for helping me out of the tricky Gävle-incident, Anna P for helping me start with some kind of Betula project that in the end I was unable to finish (I forgot why), A-C for being funny, Martin C for being weird, SÅ for having funny sleeping times, Tomas for being from Östersund, Kerstin for being from Skåne, Tanja for convincing me to learn Pearl (that turned out to be a cornerstone of the thesis) and generally for having funny discussions with, Karl for the best drunk conversations, Harald for general support recently (like setting up a deadline, otherwise I guess this would never be finished), Lena C for introducing me to the teaching that turned out to be one of the funnier elements of the Phd-period, my dual-xeon computer for doing most of the work, the cellar room in house 5, and RM for giving me salary.

Special thanks to Per, for discussions about all kinds of meaningless stuff that made the time meaningful, also, sometimes math question and other scientific things.

Thanks to all people that have helped me with the field work, especially Jonas Casslén for interesting political discussions and Johan Wallén for being a good worker. In the lab, I have been helped and received valuable advises from Niclas, Kerstin, Cilla, Vladimir and others, I guess.

Thanks to all other friends and family, special to Kalle and the Sallnäs family, kompost-Lasse, and Bastins. My mother and father, my full sibs Simon, Petter & Jonte, most cousins and the Laudon family. And last, but foremost, Jännu, together with Jerk, Frej & Tua. Without them it would probably have been easier to finish the thesis, but much more boring.
References


Lavigne C; Godelle B; Reboud X; Gouyon PH (1996). A Method to determine the mean pollen dispersal of individual plants growing within a large pollen source. *Theoretical and Applied Genetics*, 93, 1319-1326.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se

urn:nbn:se:uu:diva-9211

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2008