GENETIC MARKER STUDIES OF THE *LARIX GMELINII* COMPLEX AND THE DEVELOPMENT OF GENETIC MARKER THEORY FOR PLANT POPULATIONS

XIN-SHENG HU

A thesis submitted to the University of Edinburgh

for

the degree of Doctor of Philosophy

Institute of Ecology and Resource Management, University of Edinburgh

January 1998

Acknowledgements

The co-operative project between China and United Kingdom (ODA), " Early Establishment and Tree Improvement of *Larix* in China", provided me with a valuable opportunity to study in the University of Edinburgh, Scotland. The old, calm, handsome city with modern library is an excellent place to dig the earth of knowledge.

My supervisors Drs. Richard A. Ennos and Amanda C.M. Gillies have shown me extreme patience to guide my study throughout all stages. Dr. Ennos published an important paper that made me understand a new area in the population genetics of the plant kingdom. He used his specific method to bring me gradually into the door of population genetics. I always benefit from his stimulating discussion during the entire struggle. Dr. Gillies gave me many encouragements with careful correction and suggestions.

I deeply appreciate Dr. Nick H. Barton for helpful comments on some theoretical results.

The molecular ecology lab provides me with a friendly environment to learn new technology. Mr.J. Morman gave me much help and guidance. Dr. Eric Easton offered me and my family much help in study and life in the UK. Drs. Billy Sinclar, Anthony Langdon, Ashley Robertson and Theodore R. Allnutt were generous to let me try their methods. Miss Nicola Preston's successful help in DNA sequencing analysis brought my experiments to their conclusion.

I greatly appreciate Dr. Malcolm, Dr. Rook and Mr. W. Hogg for their frequent concern with my study. I am grateful to Profs. Shi-Ji Wang, Xiao-Shan Wang and Yi-Fan Han for much advice and encouragement. It is rather unfortunate that Prof. Shi-Ji Wang has had no chance to comment on my thesis. I am greatly indebted to him for my teaching in China. Prof. Ben-Li Pan and Prof. Wan-Chun Gu are appreciated for providing part of the seeds for the experiments.

To all those who are not mentioned above but gave me help in some or another, I express my sincere gratitude.

Last but not least, I thank my parents, my wife Fei Xiao and my son Yang Hu for immeasurable love and support. To them I would like to dedicate this thesis: my parents, Fei Xiao and Yang Hu.

Xin-Sheng Hu
January 1998, Edinburgh

I declare that all the work in this thesis was done purely by myself except where identified.

XIN-SHENG HU

January 1998, Edinburgh

Abstract

The thesis is composed of two parts that are related by the theme of genetic markers. The first part involves the application of genetic markers to investigate the mating system, population genetic structure, and evolutionary relationship of the three Chinese larch taxa: *Larix gmelinii*, *L. olgensis* and *L. principis-rupprechtii*. The second part of the thesis explores the development of populations genetic theory that is relevant to plant populations and follows the behaviour of uniparentally as well as biparentally inherited markers.

Seventeen populations of the *Larix* taxa listed above were analysed using eight polymorphic allozyme markers. Results indicated that mating system was variable among taxa and among natural populations within taxa. Outcrossing rates were $t_m = 0.986 \pm 0.081$ for one population of *L.gmelinii*, $t_m = 0.684 \pm 0.107 \sim 1.203 \pm 0.371$ for six populations of *L. olgensis*, and $t_m = 0.792 \pm 0.169 \sim 0.930 \pm 0.149$ for two populations of *L. principis-rupprechtii*. Population differentiation of each taxa was very small, showing that less than 2% of total genetic variation occurred among populations. Spatial distribution of genetic variation of *L. gmelinii* was random, but a weak pattern of isolation by distance was detected in *L. olgensis*.

The genetic relationship among the three taxa elucidated by allozyme markers indicated that the genetic distances were very low between them. Nei's arithmetic genetic distance was about 0.01 between taxa and 0.002 among populations within taxa. *L. gmelinii* was more closely related to *L. olgensis* than to *L. principis-rupprechtii*. Analyses of PCR-RFLP and sequencing of three non-coding regions of cpDNA from genes *trn*L(UGU) to *trn* F(GAA) showed no differences at all between the three taxa. Thus it may be concluded that divergence among the three taxa has occurred within recent history. Based on morphological traits and the results obtained by allozyme and cpDNA sequence markers, it is reasonable to consider *L. olgensis* and *L. principis-rupprechtii* to be two varieties of *L. gmelinii* rather than two separate *Larix* species.

In the second part of this thesis, theories of plant population genetic structure were developed to incorporate biparentally, paternally, and maternally inherited genes into a variety of models. Population differentiation for each of the three plant genomes was

formulated in the island, stepping stone and isolation by distance models of population structure. The results showed that maternally inherited organelle genes maintain larger differentiation than paternally inherited organelle genes, which in turn maintain larger differentiation than biparentally inherited nuclear genes. In the stepping-stone model, differences in genetic correlation with distance among the differently inherited genomes were conditional on the values of long and short distance migration for pollen and seeds. The relative contribution to migration of seed and pollen flow can be estimated in terms of gene frequency data or DNA sequence data. This can be carried out using Wright's F-statistics, Nei's genetic distance, and the number of segregating nucleotide sites.

When genes located on haploid genomes are under selection in a cline, results show that reparametrization may render previous cline theory suitable for plant organelle genes. One important results is that both the ratio of pollen to seed flow, and the ratio of fitnesses between paternally and maternally inherited genes play a critical role in determining cline displacement of these two types of genetic markers.

Integration of the theoretical results with practical work suggests that investigation of population structure in the *L. gmelinii* complex using maternally inherited markers (mtDNA markers) in addition to the nuclear markers already scored, is likely to yield the most interesting results concerning their biology, ecology and history.

CONTENT

CHAPTER 1

Introduction to the *Larix gmelinii* complex in China

1.1. Introduction

A genetic marker is usually defined as any allele used as an experimental probe to mark a nucleus, chromosome or gene (Riger, *et al.*, 1991). Here, the meaning of genetic marker is broadened to include any trait or character controlled by genes, or any DNA sequence itself used for the same purpose. This is because the allele itself may be difficult to identify in practical work.

Genetic markers are key tools used for addressing many problems in population genetics and ecology for a number of reasons. The first is that genetic markers behave according to simple rules of segregation every generation. Changes from one generation to next generation can therefore be simply modelled.

The second reason for the usefulness is that genetic markers can provide a record of historical events because they may alter through mutation and selection. These events may change the state of genetic markers and their frequencies in the population. The influence of these events may differ for different genetic markers. For example, mutation rate may be different among regions of DNA sequence, such as between coding and non-coding regions in some species (Gielly and Taberlet, 1994). Larger mutation rates are expected to occur in non-coding regions of the genome.

The third attribute of genetic markers that may prove useful is that they can display different modes of inheritance. They may be associated with biparentally inherited nuclear genomes, paternally inherited chloroplast genomes in conifers, and maternally inherited chloroplast and mitochondrial genomes in angiosperms (see review by Mogensen, 1996). Comparison between the behaviours of these controlling markers may provide information about the biology and ecology of the species concerned.

The fourth important characteristic of genetic markers is that they differ in their degree of resolution of genetic difference. DNA sequence data provides the ultimate in resolution of genetic differences between individuals.

These properties enable genetic markers, especially molecular genetic markers, to be powerful and flexible tools for measuring genetic variation within and between species.

Analysis of such variation can yield information on biology, ecology and history (Avise, 1994). Using relevant theory on the behaviour of such markers, many important historical events can be inferred. The following are some examples where molecular genetic markers have been applied in population genetics.

Use of codominant genetic marker, such as allozyme, provides a convenient way to score the mating system of natural populations of plants, especially conifers (Mitton, 1983, 1992). With the help of the mixed mating model (Ritland and Jain, 1981), or the neighbourhood model (Adams, 1992), outcrossing rate can be estimated.

An important event related to population structure is migration. Estimates of the number of migrants can be made using selectively neutral marker under several theoretical models (Slatkin and Barton, 1989). For example, then average number of migrants can be indirectly estimated by $(1/F_{st}-1)/4$ according to Wright's island model (1951), or by the private allele method (Barton and Slatkin, 1986). Estimates of interpopulation gene flow are now available in many studies, for example in fourteen gymnosperm and seven angiosperm forest species (Govindaraju, 1989).

The use of biparentally and maternally inherited genetic markers for inferring the history regarding post-glacial migration have been reported in recent years. For example, Jøhnk and Iegismund (1997) used the variation of allozyme and chloroplast DNA markers to infer post-glacial migration routes of *Quercus robur* and *Q. petraea* in Denmark. The underlying theory for this inference is that different vectors of migration are used by biparentally and maternally inherited genes. For biparentally inherited genes, migration can be mediated by either seed flow or pollen flow, while only seed flow contributes to the migration of maternal genes. The expected consequences for selectively neutral markers is then that population differentiation is larger for biparental genes than that for maternal genes (Ennos, 1994; Petit, *et al*, 1993b). Using this theory and the relationship between allele frequency and geographical distance, the migration routes can be inferred.

Taxonomic problems have been extensively studied using a variety of genetic markers, from morphological traits to allozyme to DNA sequence markers (Quicke, 1996). The underlying theoretical foundation for taxonomic studies is that taxa are assumed to be initially derived from a common ancestor. They diverged at different times due to the influences of mutation,

selection, recombination, etc.. A key hypothesis for the use of genetic marker for this purpose is that these events have been preserved and can be detected in genetic markers. Thus, the evolutionary relationship between taxa can be inferred using genetic similarity or distance, which is obtained by investigating variation of genetic markers. Several theories have been developed for estimating such genetic distances, including Nei's genetic distance (Nei, 1972) using frequency data, and Jukes and Cantor's one-parameter model (Jukes and Cantor, 1969) and Kimura's two-parameter model using DNA sequence data (Kimura, 1980). The absolute divergence time between taxa studied can be inferred under the molecular clock hypothesis.

Two clear characteristics can be seen from the above examples. On the one hand, variation of genetic markers provides much information regarding genetic phenomena, evolutionary processes, and influences of external factors, such as ecological factors. On the other hand, this information can only be inferred with the support of relevant theories. Theory provides us with a foundation for using the variation of genetic markers to infer important events involved within and between species.

The first objective of this thesis is thus to use genetic markers to answer important basic questions about an economically important taxa of *Larix* in China, i.e. to survey genetic variation in natural populations of native Chinese *Larix* taxa so as to provide background information for further genetic improvement. These are applications of mostly biparentally inherited nuclear markers, for which theory has already been developed. In addition, chloroplast DNA marker are used to elucidate the evolutionary relationship among the three Chinese *Larix* taxa.

In recent years, many more genetic markers have been developed, such as PCR (polymerase chain reaction; Mullis, *et al.*, 1986;Williams, *et al.*, 1990) based genetic markers on each of the three plant genomes. Organelle markers with different properties are now available (Taberlet, *et al.*,1991; Demesure, *et al.*, 1995). Use of these markers to survey plant population genetic structure may greatly broaden our knowledge of population genetics, especially the role of seed and pollen flow in gene migration.

Insight into the role that seed and pollen flow play in plant population genetic structure can only be gained through the application of relevant theory. However, analysis of population

genetic structure for *plants* has suffered neglect in terms of theory. Most theories have been developed for animal populations and are not necessarily appropriate for plants.

Therefore, the second part of the thesis explores population genetic theories suitable for plant species. We are especifically interested in hermaphrodite plant species, and the focus is on the impacts of seed and pollen flow on population genetic structure of the three plant genomes possessing different modes of inheritance in a variety of situations. These theories will be essential to allow genetic markers to be used to infer the role of seed and pollen flow in plants. This part will begin in Chapter 5.

Before embarking on the first part of the thesis (the application of genetic markers to *Larix* species) the situation with respect to larch species in general will be introduced. We will focus upon existing taxonomic studies and also those of population structure within three main Chinese larch taxa. In order to understand the position of Chinese larch on a world-wide scale, the distribution of *Larix* is considered. Then, the three Chinese larch taxa are reviewed separately, with particular emphasis on their native distribution in China and ecological requirements. The work that has already been done on population genetic structure and taxonomy of the species, using different genetic markers, will be then reviewed, emphasising the gaps that exist in our knowledge. Finally, the objectives of this study of Chinese *Larix* are outlined.

1. 2 Geographical distribution of *Larix*

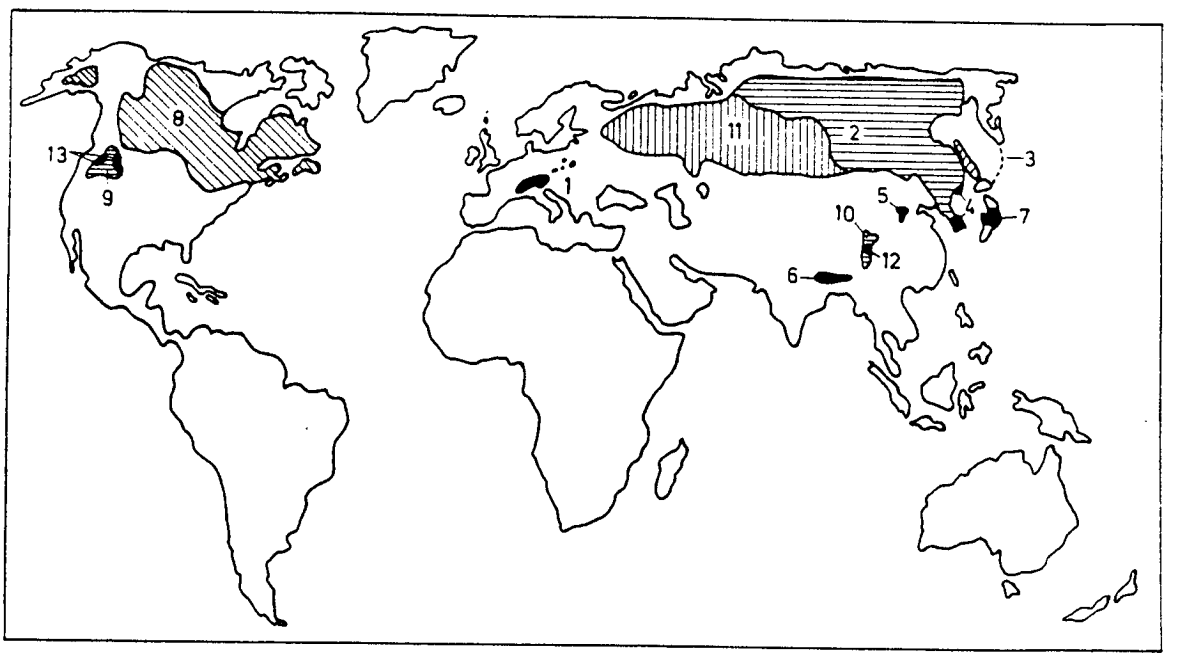
1.2.1. Distribution

Larix Miller is one of largest genera in the family Pinaceae and occurs in the temperate and cold temperate regions of the northern hemisphere. It is an economically important conifer species and is used for timber and in paper making. Larch species grow faster at an earlier stage in their development than do other members of the Pinaceae (Yang, 1995). Therefore, larch plantations are becoming increasingly commonplace and studies concerning population genetic improvement in the species are being emphasised.

Larix differs from all other genera in the family Pinaceae in that the species within it are deciduous, and needles are born on dwarf shoots (Ostenfeld and Larsen, 1930). Following the description made by Ostenfeld and Larson (1930), *Larix* can be divided into two

Fig.1.1 Distribution of *Larix* species, cited from Tang *et al.* (1995). The species and varieties indicated are:

1. *L. decidua*
2. *L. gmelinii*
3. *L. gmelinii* var. *japonica*
4. *L. gmelinii* var. *olgensis*
5. *L. principis-rupprechtii*
6. *L. griffithiana*
7. *L. kaempferi*
8. *L. laricina*
9. *L. occidentalis*
10. *L. potaninii*
11. *L. sibirica*
12. *L. mastersiana*
13. *L. lyallii*



subgenera, containing ten species and three varieties. One subgenus, Sect. *Multiseriales*, is characterised by species possessing bracts on the cone that are longer than the cone scales. It includes five species: Himalayan larch (*L. griffithiana* [Lindley & Gordon] Carrière), masters larch (*L. mastersiana* Rehder & Wilson), Chinese larch (*L. potanini* Batalin), western larch (*L. occidentalis* Nuttall), and Alpine larch (*L. lyallii* Parlatore). The other subgenus, Sect. *Larix*, is characterised by individuals possessing bracts that are shorter than the cone scales. It includes five species and three additional varieties: Japanese larch (*L. kaempferi* [Lambert] Sargent), Dahurian larch (*L gmelinii* Turczaninow), Siberian larch (*L. sibirica* Ledebour), European larch (*L. decidua* Miller), tamarack (*L. laricina* [DuRoi]K.Koch), Polish larch (*L. decidua* var. *polonica*[Raciborski] Ostenfeld & Syrach Larsen), Kurile larch (*L. gmelinii* var. *olgensis* [Mayr] Ostenfeld & Syrach Larsen) and Prince Rupprecht larch (*L. gmelinii* var. *principis-ruppechlii* [Mayr]Ostenfeld & Syrach L.).

The geographic distribution of *Larix* species is illustrated by Fig. 1.1. Species in Sect. *Multiseriales* are restricted to small disjunct areas, mostly located in mountainous regions. *L. griffithiana*, for example, occurs within the Himalayas at a height of from 1800m to 2900m above sea level, while *L. potaninii* and *L. mastersiana* are situated within the regions between the Da Xue Shan and the Ming Shan Mountains (Zhang, *et al.* 1992). *L. occidentalis* occurs in Western Canada, the most north-easterly part of the state of Washington, extreme west of Montana, and the northern parts of Idaho (Ostenfeld and Larsen, 1930). *L. lyallii* occurs in two separate regions: one towards the east of the United States in the Rockies, and the another to the west in the Cascade Mountains.

Species in Sect. *Larix* occupy larger geographic regions than do those in Sect. *Multiseriales* (Fig.1.1). One exception, *L. kaempferi*, occurs naturally in the interior of Hondo, Japan. *L. gmelinii* is a very common tree throughout the entire forest-clad regions of Eastern Siberia, especially in the north, where it alone forms the tree line (Fig.1.1). *L. gmelinii* var. *olgensis* is mainly located in the Chang Bai Shan Mountains in China (Zhang, *et al* 1992). *L. gmelini* var. *principis-rupprechtii* is situated in the mountains of Sanxi and Hebei province in China, while *L. sibirica* occurs in western Siberica and north-eastern Russia (Fig.1.1). The distribution of *L. sibirica* extends from Lake Baikal in the east to the White Sea, and terminates in the west near to Lake Onega. The northern edge of its range reaches Jenisej, and the Altai Mountains in the south. *L. decidua* occurs most commonly in the region stretching from Dauphine and Provence northwards

and eastwards through the Alps to a point 40-50 km south-west of Vienna, then it extends southwards to latitute 46° N in the north-west former Yugoslavia and the north-east corner of Italy. *L. decidua* var *polonica* occurs mainly in Poland.

According to Ostenfeld and Larsen (1930), four species and two varieties of *Larix*, are located in China (Fig.1.1). These taxa, particularly *L. gmelinii*, *L. gmelinii* var. *olgensis* and *L. gmelinii* var. *principis-rupprechtii*, play an important role in timber production in China (Yang, 1995) and are widely used for mine shaft supports, sleeper supports, wire poles, building, bridge link, trailer frame and furniture (Zheng, 1983).

1.2.2. Three Chinese larch taxa and their ecological requirements

In this section, a brief introduction to three particular larch taxa, namely *L. gmelinii, L. olgensis* and *L. principis-rupprechtii*, will be given. All of these taxa are of considerable importance to Chinese forestry and provenance trials that are at least 10 years old have been planted for all of them (Yang, *et al.*, 1991; Ma,1992).

1.2.2.1 *Larix gmelinii*

Chinese *L. gmelinii* covers part of the southern extension of its entire species range. Only a small part of its entire distribution, however, is in China (Fig. 1.1), and it mainly occurs in the Greater Xingan Mountains, generally in areas below 1200m above sea level. A small part of its distribution occurs in the Lesser Xingan Mountains

L.gmelinii grows in the cold temperate zone in China and is a cold tolerant species. In the Greater Xingan Mountains, the annual growth period for *L.gmelinii* is short, from 100 to 120 days, since for more than seven to eight months of the year, the temperature is below 0°C and the minimum temperature is -51°C. Annual precipitation is 300~600mm and the soil has a permanent frost layer 1m beneath the surface. Even under these conditions, *L. gmelinii* still flourishes (Yang, 1995).

L. gmelinii readily hybridises with other larch species in areas where they overlap. For example, *L. amurensis* B. Kolesn is the hybrid formed between *L. gmelinii* and *L. olgensis*,

while *L. ochotensis* B. Kolesen is the hybrid formed between *L. gmelinii* and *L. gmelinii* var. *japonica* (Yang, 1995).

1.2.2.2 *Larix olgensis*

The distribution of *L. gmelinii* var. *olgensis* in China is centered in the Chang Bei Shan Mountains. The southern end of its distribution extends into northern Korea, while it continues northwards up to 45° 20'N and westward to 125°E, in the Kuan Dian county. It grows between 500m and 1800m above sea level.

In contrast to *L. gmelinii*, *L. olgensis* grows in the wet temperate zone where the annual average temperature is 2.0~6.4 °C and temperatures reach -13 ~ - 21°C in January, while in July they are as high as 18 ~ 24 °C. The annual precipitation in this zone is 540 - 1200 mm. *L. olgensis* is both cold tolerant and light sensitive and has no strong requirement for particular soil conditions. Therefore, it can grow in poor soil or wet land (Yang, 1995).

L. olgensis undergoes natural hybridisation with *L. gmelinii* and *L. principis-rupprechtii*. For example, *L. lubarsikii* Suk is the hybrid between *L. olgensis* and *L. principis-rupprechtii* (Yang, 1995).

1.2.2.3 *Larix principis-rupprechtii*

L. principis-rupprechtii, historically, occurred in large areas of northern China within the range from 36°30' to 43°30'N and from 111° to 120°E (Ma, 1992). The eastern range almost extends to that of *L. olgensis*, while the northern range may overlap with *L. gmelinii*. Probably due to over logging and some characteristics of the species, for example its light sensitive character, its distribution is mainly now limited to the mountains, including the Tai Yue Shan, Guan Di Shan, Guan Qian Shan, Wu Ta Shan and Han Shan Mountains. There are also some remnant populations occuring in other mountains, giving a fragmented appearance in the distribution of the species (Ma and Tao, 1992; Wang, 1995).

According to Ma and Wang (1992), ecological requirement for *L. principis-rupprechtii* differ from those of *L. gmelinii* and *L. olgensis* in that it requires warmer climate conditions.

In its distribution region, the period in which the temperature is higher than 0°C is from April to October (Ma and Wang, 1992).

In summary, while the geographical distribution of *L. gmelinii* and *L. olgensis* overlap in places, they are both isolated from *L. principis-rupprechtii*. The ecological requirements are not the same between them. However, all three taxa can freely interbreed and produce hybrids.

1.3 Taxonomic problems

1.3.1. Taxonomy

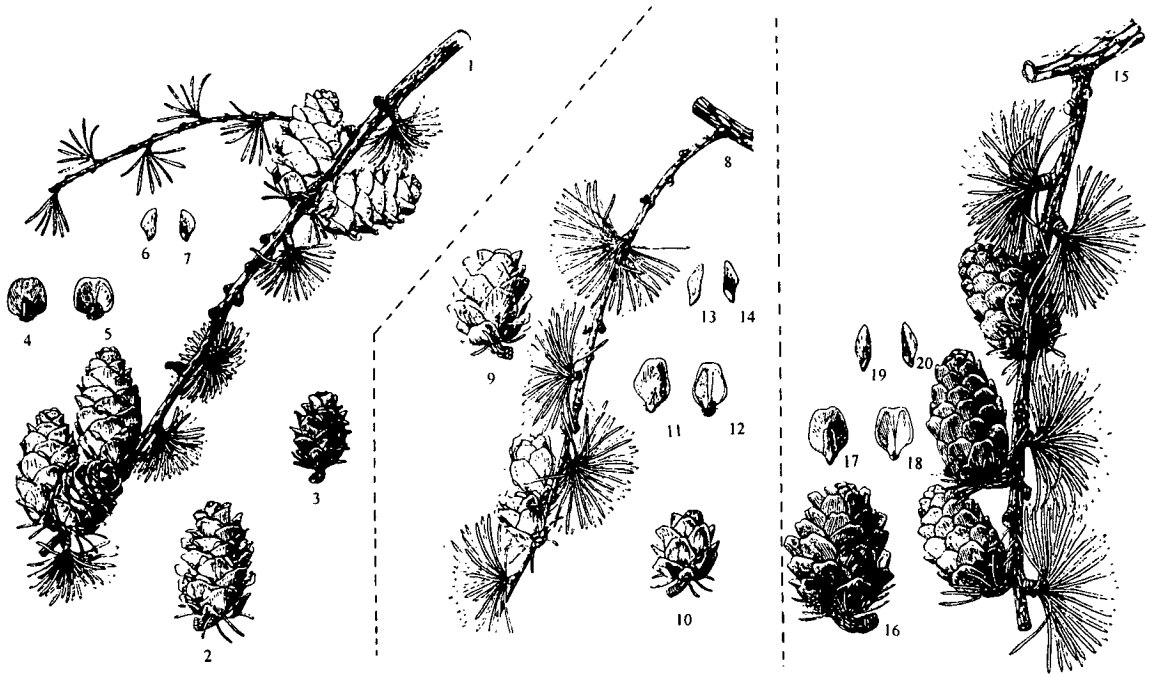
The characters used in plant taxonomy are turning from traditional morphological and anatomical features to the use of protein and DNA markers (Quicke, 1996). Traditional taxonomic studies have used morphological characters and anatomical structure, such as leaves, stems, perianth, bracts fruit, etc. (Heywood, 1967; Jones and Luchsinger, 1979). For example, floral characteristics have been widely used for classification purpose in the angiosperms (Jones and Luchsinger, 1979). The results so formed are not reliable and may even be flawed due to strong modification of morphological traits caused by environmental factors. Currently, molecular markers, especially DNA sequence data, are being used for plant systematics or phylogeny, rendering the results more accurate. Such sequence data are not subject to environmental modification and hence providing reliable results from which to infer the historical events involved in the phylogeny (Avise, 1994; Quicke, 1996).

Classification of *Larix* species is fraught with confusion because different authors employ different morphological traits in their taxonomy. Ostenfeld and Larsen (1930) classified *Larix* species into 10 species and 3 varieties according to their cone traits. They considered *L. olgensis* and *L. principis-rupprechtii* to be two varieties of *L. gmelinii*. However, according to Zhang *et al* (1992), there are ten species and five varieties of *Larix* in China. Chinese scientists consider *L. olgensis* and *L. principis-rupprechtii* to be different species (Zheng, 1983), though the number of taxa in China is still open to question (Zheng, 1983; Zhang, *et al.*, 1992). A key based on morphological characters is given in Appendix I.

As indicated by the key, the three larch taxa, *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii*, can be distinguished based on morphological and cone characters (Fig. 1.2).

Fig. 1.2. Morphological characters of the three Chinese larch taxa (after Zheng, *et al.* 1983): *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii.*

Cone-beared shoot:	1, 8, 15.
Cone:	2, 3, 9, 10, 16.
Cone-scales and bract-scales:	4, 5, 11, 12, 17, 18.
Seed:	6, 7, 13, 14, 19, 20.



L. olgensis **Henry** *L. gmelinii* **Rupr.** *L. principis-rupprechtii* **Mayr**

For example, the cone of *L. gmelinii* varies from cup form to ellipse, and its length is 1.5~2.0 (2.5) cm, with average number of cone-scales being 20, seldom 30. These characters can be used to distinguish it from *L. principis-rupprechtii* whose cone shape varies from reniform to widely reniform and the length is 2.0 ~ 2.7 cm, with cone-scales average more than 30, seldom less than 30. Both taxa are characterised by smooth and shining cone-scales, which can be used to distinguish *L. olgensis* which exhibits pilose cone-scales (Appendix I).

According to Zhang *et al* (1992), there is one variety of *L. principis-rupprechtii*, i.e. *L. principis-rupprechtii* var. *wulingshanensis* and three varieties of *L. olgensis*, i.e. *L. olgensis* var. *changpaiensis*, *L. olgensis* var. *heilingensis* and *L. olgensis* var. *koreana*. These varieties can be distinguished by certain morphological traits. For example, according to cone length, *L. olgensis* var. *koreana* (1.4~3.0cm) can be distinguished from *L. olgensis* var. *changpaiensis* (>3.0cm; see Appendix I).

These taxonomic treatments are summarised in Figure1.3. There are three varieties of *L. olgensis* and one variety of *L. principis-rupprechtii*, which were not mentioned in Ostenfeld and Larsen (1930). These varieties were not named at all in 1983 (Zheng, 1983). At the present time, nomination of these varieties is still arguable in China, which could indicate the unreliability of using morphological traits in taxonomy.

1.3.2. Genetic relationship of the three taxa

A number of studies, using a range of different techniques, have been conducted to assess the classification and genetic relationship of *Larix* species. Although the delimitation of *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii* can be made using morphological characters, such as cone traits, the genetic relationships among the three species, still remain unclear. According to Ostenfeld and Larsen (1930), *L. olgensis* and *L. principis-rupprechtii* are considered to be varieties of *L. gmelinii*. Thus these three taxa are members of the same species and therefore very closely related.

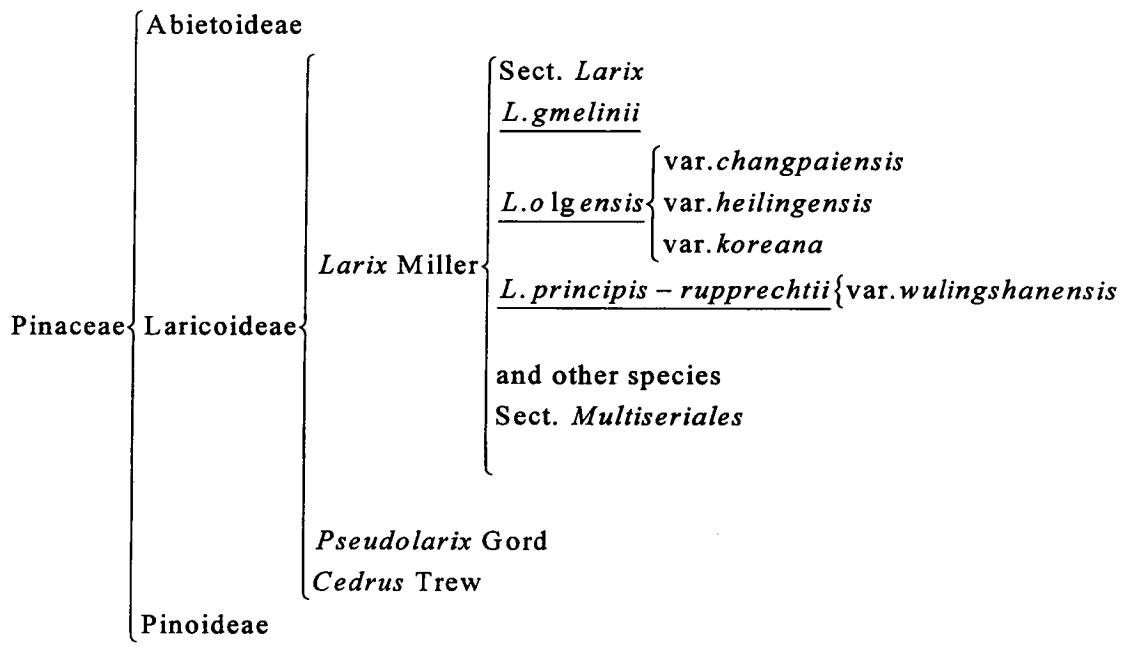


Fig. 1.3. The position of *L. gmelinii* complex in the family Pinaceae. Varieties of *L. olgensis* and *L. principis-rupprechtii* are classified based on work by Zhang *et al.*(1992) using morphological traits.

According to Ma (1992), who based his conclusions on results obtained by former Soviet Union scientists, *L. gmelinii* is a larch species that was present in the northeast region of Siberia overlapping with the distribution of *L. sibirica* during the Pleistocence. The current distribution of *L. gmelinii* is a result of its invasion into the distribution of *L. sibirica*, a process which has not stopped since the Pleistocene. The western and southern extension of *L. gmelinii* is parallel to the expansion of severe climate conditions and the area where soil is permanently frozen. *L. principis-rupprechtii* is not a relic species, but was formed more recently in the warmer climate during the southwards extension of *L. gmelinii.* The recent formation of *L. olgensis* was brought about by the southern migration of *L. gmelinii* that penetrated into the species range of *L. sibirica.* The isolated part of *L.gmelinii* so formed within the distribution of *L. sibirica* underwent speciation to form *L. olgensis* after the rising of the Chang Bei Shan Mountains (Ma, 1992).

Zhou (1962) studied the evolution of Chinese larch species by investigating woody structure characters. He inferred that, in the Sect. *Larix, L. sibirica* and *L. gmelinii* were relatively young species. However, he did not believe that *L. olgensis* occurred in China, because the woody anatomy structure of *L. olgensis* is quite similar to *L. gmelinii* rather than those of *L. olgensis* from pre-Soviet Union scientists (Ma, 1992). Thus, it may be seen that a great deal of confusion exists over the classification of the *L. olgensis* and *L.gmelinii* and that they are clearly closely related.

Zhang *et al.*(1985) studied evolutionary relationships among five *Larix* species: *L. gmelinii, L. olgensis, L. sibirica, L. kaempferi,* and *L. principis-ruppechtii*, using chromosomes characters. They found that chromosome structure for three of the species (*L. gmelinii, L.olgensis* and *L. kaempferi*) was 2n = 2x = 24 = 12m(4sc) + 10cm + 2st; while *L. principis-rupprechtii* and *L.sibirica* were shown to be 2n = 2x = 24 = 12cm (4sc) + 12sm. According to Stebbin's karyotype classification, which was devised to classify different types of symmetry according to the combinations between the ratio of largest to smallest chromosome and the proportion of chromosomes with arm ration smaller than 2.0 (Stebbin, 1958), all of them are type 2B (2:1~ 4:1 vs. 1~50%). Exceptions are *L. principis-rupprechtii* and *L. sibirica* that are type 2A (<2:1 vs. 1~50%). It can be inferred from Stebbin's classification that type 2B exhibits more asymmetry in chromosomes than type 2A. It is commonly accepted that karyotype evolution is a consequence of chromosomal structural changes (inversions, translocations, centric fusion, etc.), hence resulting in progressive reduction of the basic chromosome number and an increased asymmetry (Stebbin, 1958; John and Lewis, 1968). Therefore, Zhang *et al.* (1985) inferred that the evolutionary trend of the species was from *L. sibirica* to *L. principis-ruppechtii* and then to the other larch species. This means that Chinese *L. gmelinii* and *L. olgensis* may have evolved from *L. principis-rupprechtii.*

Using DNA markers to elucidate relationship between the three Chinese species may provide additional insight into the evolution of *Larix.* Indeed, Tang *et al* (1995) investigated evolutionary relationships between nine species and three varieties of larch using RFLP analysis of chloroplast DNA (cpDNA), which has been commonly used as a markers for reconstrcting plant phylogenies in more recent years (reviewed by Rieseberg and Soltis, 1991; Clegg and Zurawski, 1992; see Chapter 4). Using six restriction enzymes (*Bam*HI, *Bc*III, *Dra* I, *Hind* III and *Kpn* I) and eleven non-overlapping probes, Tang *et al.* found low

nucleotide divergence among taxa and divided them into three groups: The first group was comprised of just one species, *L. griffithiana*; while the second included *L. sibirica, L. laricina and L. occidentalis*; and the third group consisted of *L. gmelinii, L. potaninii, L. kaempferi,* and *L. decidua.* This means that the three Chinese larch species are quite closely related. Since the mutation rate (point mutation) in cpDNA is very small (Wolfe *et al*, 1987) and the molecule is predominantly uniparently inherited without recombination (Szmidt, 1991; Harris and Ingram, 1991), the close genetic relationship between the three Chinese larch species was demonstrated.

Using random amplified polymorphic DNA (RAPD) markers, Shiraishi *et al.* (1995) studied evolutionary relationship between five larch species: *L. kaempfei, L. gmelinii, L. gmelinii* var. *japonica, L. olgensis* and *L. decidua.* Their results indicated that *L. gmelinii* and *L. olgensis* were sister taxa, hereby providing evidence in support of a close relationship between *L. gmelinii* and *L. olgensis*. Unfortunately, they did not include *L. principis-rupprechtii* in their study.

In summary, the relationship between these three Chinese larch species is still very much confused. For example, Chinese scientists believe that *L.olgensis* and *L. principis-rupprechtii* are separate species, while others believe them to be varieties of *L. gmelinii* (Ostenfeld and Larsen, 1930). Moreover, three varieties of *L. olgensis* and one variety of *L. principis-rupprechtii* are considered to exist (Fig.1.2), resulting in still more complicated evolutionary relationship between the taxa. Clearly, there is a need to resolve the evolutionary relationship between these three taxa.

1.4. Mating system and population structure

Knowledge of population structure is very important for assisting in decision making for tree improvement and genetic conservation programmes (Hamrick, 1994). An understanding of natural population genetic structure may also provide background information that is important for future genetic improvement of a species.

The genus *Larix* contains a number of important timber species. Correspondingly, there have been extensive studies of breeding systems and genetics in the genus (Martinsson, 1995a; Schmidt and McDonald, 1995). However, the mating systems and population genetic

structure for the three Chinese larch species have not yet been studied using molecular markers. Undoubtedly, the results of such a study may influence our understanding of genetic structure of natural populations and future tree improvement in China. In this section, the achievements in studies of population genetic structure in selected larch species will be discussed, and then compared with information obtained for the three Chinese larch species.

1.4.1. Achievements outside China

1.4.1.1 Mating system

Mating system influences the mode of transmission of genes from one generation to the next (Brown, 1990), and thus determines the distribution of genotypes within populations. It also influences the degree of population differentiation. Its importance in population genetics has long been appreciated (Wright, 1931, 1969). Generally speaking, outcrossing promotes gene flow, and brings genotype distribution toward Hardy-Weinberg equilibrium. Selfing reduces gene flow, and leads to reduction of the proportion of heterozygous in a population. Thus, knowledge of the mating system of a species may also allow greater understanding of strategies for maintaining genetic diversity, and assist in the formulation of optimal strategies for hybridisation and tree genetic improvement, especially for those using wind-pollinated seeds for reforestation. For example, inbreeding depression may result in reduction in survival and growth of seedling progeny of some conifers (Sorensen and Miles, 1974, 1982).

Historically, research on the plant mating systems falls into three distinct periods: survey period before 1960, exact model analysis, and use of allozyme markers from 1970 (Brown, 1990). Currently, allozyme markers are still widely used to score plant mating system because allozyme polymorphism is common and readily detectable (see Chapter 2), even if DNA markers have been employed for this purpose more recently (Gitzendanner, *et al.*, 1996).

According to Brown's classification (Brown, 1990), plant species conveniently fall into five classes of mating systems: predominant selfing (outcrossing rate, $t<0.10$), predominant outcrossing ($t > 0.95$), mixed selfing and outcrossing, apomictic, and haploid selfing (Brown, 1990). Many conifers fall into the mixed mating category (Mitton, 1992). Examples

from the family Pinaceae are given in Table 1.1. The mixed mating system means that some proportion of seed (s) is produced by selfing, and the complementary proportion (t=(1-s)) is produced by outcrossing.

Table 1.1. Outcrossing rate of some conifer species in the family Pinaceae detected by allozyme markers

Species & References	Outcrossing rate (t_s or t_m)†	Mating system (Pure selfing, outcrossing,or mixed)
Balsam fir (*Abies balsamea*) Neale & Adams, 1985a	0.78~ 0.99, mean 0.89	mixed
White spruce (*Picea glauca*) King, *et al*, 1984	0.75~0.99, mean 0.90	mixed
Jack pine (*Pinus banksiana*) Cheliak, *et al*, 1985	0.88 ±0.047	mixed
Logepole pine (*Pinus contorta* ssp. *latifolia*) Epperson & Allard, 1984	1.03±0.04	mixed
Jeffrey pine(*Pinus jeffreyi*) Furnier & Adams, 1986	0.881~0.971	mixed
Ponderosa pine (*Pinus ponderosa*) Farris & Mitton, 1984	0.81±0.054 (low density) 0.96±0.046 (high density)	mixed
Douglas fir (*Pseudotsuga menziesii*) Shaw & Allard, 1982 Neale & Adam, 1985b	0.90(t_m) 0.94~1.00(t_m)	mixed
Tamarack (***Larix laricina***) Knowles, *et al*, 1987	0.316~0.897 (t_s) 0.729 (low density) (t_m) 0.908 (high density) (t_m)	mixed
European larch (***Larix decidua***) Gomory & Paule, 1992	0.64~1.0 (t_s) 0.852 ±0.007 (t_m)	mixed

† : t_s : outcrossing rate estimated by single locus; t_m: outcrossing rate estimated by multilocus methods.

Outcrossing rates can be variable from species to species. For example, the multilocus outcrossing rate is 0.89 in balsam fir (*Abies balsamea*; Neale and Adams, 1985) and 1.03 in the lodgepole pine (*Pinus contorta* ssp. *latifolia*; Epperson and Allard, 1984). Many factors, such as population density and age structure, can influence the mating systems between different populations of a species (see review by Mitton, 1992). For example the outcrossing rate of ponderosa pine (*Pinus jeffreyi*; Furnier and Mitton, 1986) was estimated to be 0.81±0.054 in low density population and 0.96±0.046 in high density population. Thus, the mating system of a species usually varies in space and time (see review by Mitton, 1992).

Two species of *Larix*, *L. laricina* (Knowles, 1987) and *L. decidua* (Gömöry and Paule, 1992;Furnier and Paule, 1992), have been investigated with regard to their mating systems. The mean multilocus outcrossing rate over five allozyme markers was found to be 0.729 in five natural populations of *L. laricina* in Ontario (Knowles, *et al*, 1987), which is lower than estimates that have been reported for most other conifers. *L. decidua* has also been shown to exhibit significant levels of selfing in a seed orchard that was designed for maximum outcrossing (Gömöry and Paule, 1992). The multilocus outcrossing rate of *L.decidua* is 0.852 (Gömöry and Paule, 1992). In an old stand of Polish larch (*L. decidua* var. *polonica*), Lewandowski *et al.* (1991) revealed that the multilocus outcrossing rate was $t_m = 0.943 \pm 0.055$, while single locus estimates varied from 0.80 to 1.13. Results from these two larch species would suggest that outcrossing rates in the genus are quite variable and some population may exhibit significant levels of selfing, implying that, possibly, a different mating system exists between larch and other conifers possessing predominant outcrossing.

1.4.1.2 Population structure

Extensive studies have been carried out on population (provenance) structure and geographical variation in many larch species around the world (Schmidt and McDonald, 1995; Martinsson, 1995a). The importance of these studies can be reflected from the two international cooperative provenance trials of European larch species, the first in 1944 and the second in 1958-1959; and one international provenance trial of Japanese larch (*L. kaempferi*) in 1956. Since the 1960s provenance trials of different *Larix* species have been carried out in many countries (reviewed by Yang, 1995). Martinsson (1995b) recently proposed that another international research project should be initiated: Systematics and

differentiation in the genus *Larix* in Eurasia. One objective in the proposed project involves phytogeographical analysis and the analysis of the genetic structure and polymorphism within and between populations, ecotypes and species of the genus *Larix* in Russia and Europ (Martinsson ,1995b). These activities indicate the importance of genetic improvement in *Larix* at the population level.

However, most of these studies have focused on genetic structure for quantitative trait structure, especially field growth performance and improvement (provenance trials), and this is appropriate for selection of the best provenance suitable for growth in a particular location.

Another important aspect regarding tree improvement is to investigate population structure using selectively neutral markers such as allozymes, which may provide more information not about adaptive variation, but about the history of populations and the extent of gene flow among them (Crawford, 1983). It has been found using allozyme markers, that most outcrossing plant species maintain a large proportion of genetic variation within populations and a small proportion of total genetic variation among populations (Hamrick, *et al.*, 1991). For example, in a recent review, Hamrick *et al.* (1991) observed that only 9.5% of total genetic variation occurred between populations of selected 426 woody plant taxa, with 9.9% in temperate woody species (231 taxa) and 13.5% in tropical woody species (124 taxa). The reason for this is that in woody plant taxa gene flow among population tends to be extensive. At equilibrium between drift and migration, very little genetic differentiation is found among populations.

Some *Larix* species exhibit a similar distribution of genetic variation in natural populations. For example, Liu and Knowles (1991) used allozymes to investigate genetic structure of 44 populations of *L. laricina* from northern Ontario, and found that approximately 2% of the total genetic variation was evident among populations. Using four methods of cluster analysis, based on genetic distance, and a discriminant function analysis of genotypic structure of the populations, they revealed that the variations patterns were not related to geographic location, although one of the clustering procedures did show a weak east-west pattern. They, therefore, concluded that the distribution of the variation provided little evidence in support of the two routes proposed for post glacial reinvasion that meet west of Lake Superior (Liu and Knowles, 1991).

Ying and Morgenstern (1991) found similar results for eight natural populations of *L. laricina* in New Brunswick. The F_{st} estimated using 13 polymorphic allozyme loci indicated that 3.8% of total genetic variation resided between populations. A correlation between Nei's genetic distance and altitude was found to be stronger than that between Nei's distance and linear geographic distance, implying marked non-random patterns for the genetic variation between populations relative to their altitude.

Cheliak *et al.* (1988) used 19 allozyme loci to investigate 36 populations of *L. laricina* that represented the natural range of the species in North America. They found that each was in Hardy-Weinberg equilibrium, and about 5% of total genetic variation occurred between populations.Using discriminant analysis of genotypic structure of the populations investigated, they observed a general east versus west pattern in the natural range, with populations in the Great Lakes basin being further differentiated. They concluded that the present-day population distribution, population density and reinvasion routes after the last glaciation could account for these observed patterns of genetic variation.

Only one report is available regarding population genetic structure in the three Chinese species considered in this study, using allozyme markers. Potenko and Razumov (1996) investigated six stands of *L. gmelinii* in two areas of the Russian Far East, with altitudes ranging from 100 to 1000m, and found that 96% of genetic variation occurred within stands, which is similar to that in *L. laricina* (Cheliak, *et al.*, 1988).

1.4.2. Achievements inside China

Extensive studies have been carried out in China on each of the three *Larix* species considered in this study, due to the important role that they play in timber production. Genetic studies in each of the three taxa have also been utilised, for selecting seed sources for provenances trials, hybrids, seed orchard and vegetative propagation. The following are some results related to population structure, mainly based on Yang *et al.* (1990a, b) and Ma (1992). It should be noted again that these results indicated patterns of adaptive variation that are different from patterns of neutral variation due to different mechanisms involving in their formation (see review by Barton and Turelli, 1989).

1.4.2.1 *Larix gmelinii*

Studies concerning provenance selection in *L. gmelinii* began in 1980 (Yang, *et al.*, 1990a,b). Sixteen seed sources were planted in 13 separate locations, and traits such as growth, phenology and resistance to disease, were utilised to evaluate the performance of these seed sources and their geographical variation.

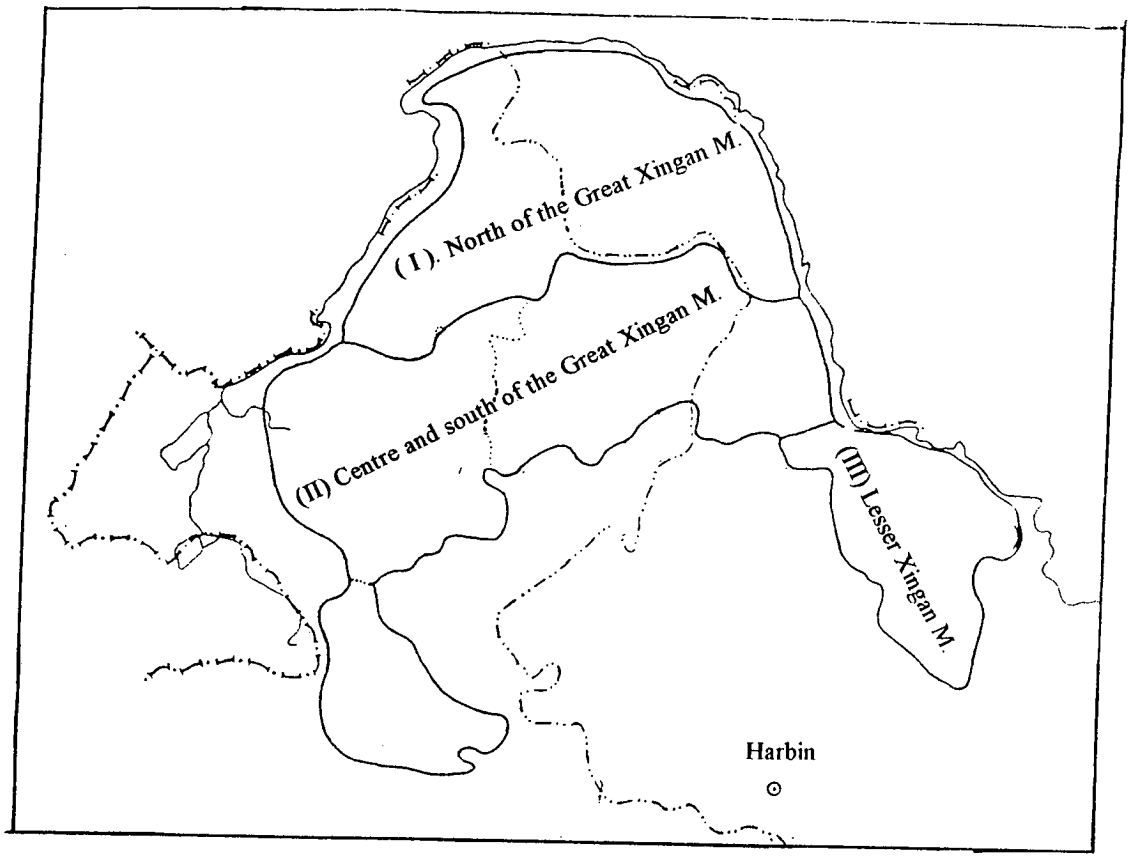
One interesting result from the Chinese provenance trials is that the patterns of geographical variation were more marked with a change in longitude than latitude with regard to the growth traits (Yang, *et al*, 1990a,b). In this way, the entire species range in China can be grouped into three seed zones: (I) north part of the Great Xingan Mountains; (II) central and southern parts of the Great Xingan Mountains; (III) Lesser Xingan Mountains (Fig1.4).

Li *et al.* (1991) used an allozyme marker, peroxide (PER), to investigate geographical variation of 14 provenances. They found that there were significant difference between provenances in terms of total band number. According to the coefficients of variation (CV%) of the number of bands, they classified the 14 provenances into four seed zones that are not consistent with seed zones obtained according to field growth performance. However, it should be noted here that the method they used did not involve analysis in terms of locus and allele frequency, but only comparison of total number of bands between populations. It is, therefore, difficult to extract more information from the study regarding population structure, such as the distribution of genetic variation between and within populations.

1.4.2.2 *Larix olgensis*

The first Chinese provenances trials for *L. olgensis* began in 1980 (Yang, *et al*, 1991). Results show significant geographical variation in terms of growth traits, e.g. height. Based on these results, Yang *et al.* (1991) concluded that: ① The basic pattern of variation in terms of growth traits was more marked with a change in altitude than in latitude. ② The Xiaobeihu provenance, growing at a lower elevation and lower equivalent latitude, was a good source of material in terms of rapid growth, good stability and timber quality. ③ The compounded effect of precipitation and temperature was an important factor in influencing current genetic variation. ④ Growth traits including height and diameter were the most

Fig. 1.4. Seed zones of *L. gmelinii* (I). North part of the Great Xingan Mountain; (II) Central and southern parts of the Great Xingan Mountain; (III) Lesser Xingan Mountains.



important characters in provenance delimination. ⑤ Better field performance with respect to growth traits can be obtained by transferring seeds from low equivalent latitude to the northern region for afforestation.

L. olgensis provenances could be divided into four seed zones based on these trials (Fig. 1.5), which would indicate a significant difference between zones, but no such difference between provenances within each zone.

Population structure within *L. olgensis* has not yet been investigated using selectively neutral markers. Therefore, the distribution of genetic variation within and among populations is not known for selectively neutral markers.

1.4.2.3 *L. principis-rupprechtii*

Provenance trials for *L. principis-rupprechtii* have been established for at least ten years (Ma, 1992). Results from the trials show that significant difference were observed among provenances in terms of growth traits, with the exception of Fengning, Hebei Province and in Lusi, Henan Province. The interaction between provenance and site (provenance × site) is also significant. Ma and Tao (1992) concluded that the island distribution of the remnants of *L. principis-rupprechtii* populations may be the main cause for such differentiation in the species.

Based on the results of these provenance trials, the species range of *L. principis-rupprechtii* can be divided into three zones: northern, central and southern (Fig. 1.6). No obvious geographical correlation was found for the patterns of growth traits compounded with latitude and longitude. The correlation coefficients between growth traits and latitude or longitude are not significant (Ma and Tao, 1992).

In summary, the extensive provenance trials that have been planted for each of the three *Larix* species have provided a lot of useful information that will guide future genetic activities in the genus. However, more basic background information, such as the mating system and genetic variation of natural populations for selectively neutral markers, is still not available, which is likely to be quite different from those revealed by provenance experiments.

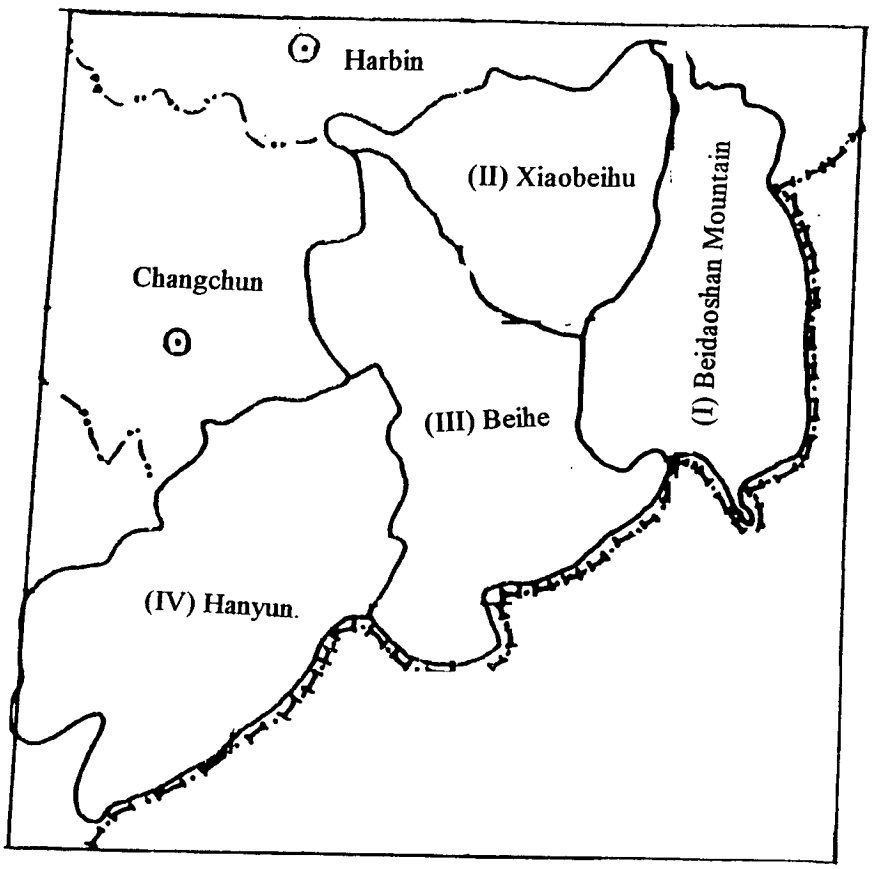


Fig.1.5. Seed zones of *L. olgensis*: (I) Beidaoshan Mountain; (II) Xiaobeihu; (III) Beihe and (IV) Hanyun.

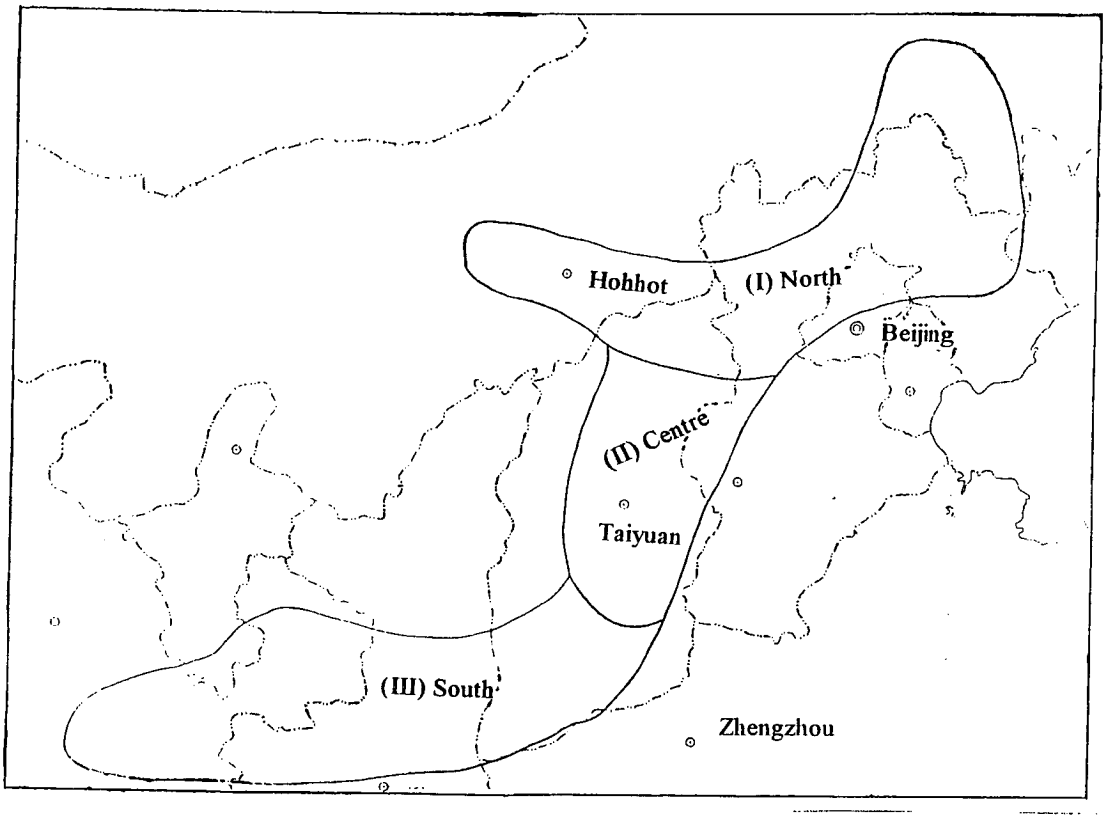


Fig.1.6. Seed zones of *L. principis-rupprechtii*: (I) North; (II) Central and (III) South.

1.5. Objectives of this study

This study uses molecular markers to address three important problems concerning future genetic improvement of *Larix*. The first concerns the mating system of this species, the second is taxomy of the three species, and the third is related to population structure. Addressing the first issue is important in tree improvement to determine the possible occurrence of inbreeding and loss of fitness in wild collected seed. The second question must be resolved to improve our understanding of the genetic relationship between three Chinese larch taxa and therefore to elucidate past events that have occurred in their history. In addition to resolving taxonomic problem within the genus, the third issue, i.e. population genetic structure, may help us to determine the distribution of genetic variation within and between populations of the three species. The expected results may tell us more about the history of the species and its population delineation of evolutionary units, and, also help us with decision making regarding how to best improve and conserve these species.

Thus, the aim of the first part of the thesis is to develop allozyme and DNA markers that may provide information to increase our understanding of genetic diversity in *Larix*. Studies are concerned with:

- **Population genetic structure** within the three taxa, the *L. gmelinii* complex. The distribution of genetic variation will be investigated using allozyme markers. Spatial patterns of these structure are tested using an isolation by distance model (see Chapter 2).

- **Mating system** of natural populations in each of the three taxa. Outcrossing rates will be estimated using single- and multi-locus models and allozyme markers (see Chapter 3).

- **Taxonomy of the three taxa** Both allozyme and DNA markers will be employed to elucidate evolutionary relationship among the three taxa and their populations (see Chapter 2 and 4).

CHAPTER 2

Use of Allozymes to Assess Population Structure and Evolutionary Relationships within the *L. gmelinii* Complex

2.1. Introduction

The distribution of genetic markers can yield useful information that is of practical value to foresters or others who are managing and utilising *Larix* in northern China. First, it can be used to tell us the levels of gene diversity or polymorphism that indicates whether populations have been through a bottleneck or whether they retain genetic variation, and hence tell us the potential adaptability of populations studied to variable environments. For example, effects of bottleneck or founder events may reduce gene diversity. Measurement of such diversity can be carried out in terms of percentage of polymorphic loci, such as P(99%), the percentage of loci whose common allele frequency is less than 0.99, or in terms of He, the gene diversity. The levels of gene diversity or polymorphism within natural populations of the three Chinese larch taxa have not been reported, and this forms part of this chapter.

Second, the arrangement of genetic diversity at single loci may tell us about the distribution of genotypes within the populations investigated. Are these alleles randomly associated or not? Measurement of these deviations from Hardy-Weinberg equilibrium can indicate the extent of inbreeding. Two kinds of reasons may be responsible for the inbreeding, matings between related individuals including selfing, and population subdivision (Wright, 1943). The former in this case can be measured by Wright's inbreeding coefficient, F_{is} .

Third, arrangement of genetic diversity at different loci may tell us whether alleles at different loci are randomly associated or not. Measurement of this relationship can be performed using a linkage disequilibrium test. Very tight linkage disequilibrium may be related to inbreeding or to small population size.

Fourth, the arrangement of genetic diversity within and among populations provides guidance for many aspects of genetic activities, such as genetic conservation. Measurement of population differentiation can be obtained via Wright's F_{st}. Several factors may influence the distribution of genetic variation within and between populations, such as drift, migration and selection. The usefulness of this analysis of population differentiation for selectively neutral markers is that it can be used to estimate the average number of migrants between populations within each of the three *Larix* taxa. The relationship between genetic difference and geographic distance can also be inferred.

Fifth, genetic distance between populations within and among the three Chinese *Larix* taxa, can also be addressed using genetic markers. Larger genetic distance should exist between taxa than within taxa. This information may also be used to reconstruct phylogenies of the three *Larix* taxa. Thus, are the three *Larix* taxa genetically distinct, and more different than populations within taxa? What is the absolute value of genetic distance? Does it correspond to differences at species or subspecies level ? The use of genetic markers can answer these questions to some extent. Information of this type is important for tree genetic improvement of the three *Larix* taxa in China.

It can be seen from the above that many important questions, which have not been addressed in the three *Larix* taxa, can be investigated using genetic markers. Thus, choice of the most appropriate genetic marker is critical in the present study. However, many factors may influence this decision including the aim of a study, the genetic properties of markers and the economic cost involved. For historic and technical reasons, the markers chosen for genetic studies range from morphological and physiological traits, to chromosome karyotypes, allozymes and, most fundamentally, to the level of DNA variation (Mallet, 1996).

As far back as 1966, Hubby and Lewontin (1966) pointed out the following criteria that an ideal molecular technique must satisfy. These requirements still hold today: ① Allelic expression should be distinguishable in individuals; ② The effect of each allelic substitution should be locus-specific and distinguishable from substitutions at other loci; ③ All base substitutions should be detectable;④ Loci should be sampled at random, irrespective of their function or likely level of polymorphism.

These criteria are met to different extents by using different genetic markers, from morphological to DNA markers. Volumes of books concerning the use of allozymes in plant genetics and breeding are available (Tanksley and Orton, 1983). Basing on the following reviews on the work that has been done mostly in recent years, allozyme markers were chosen for the present study. In the following the principle of allozyme analysis is briefly introduced. Comments on the allozyme marker are then given, and the aims of present study are finally presented.

2.1.1. Use of allozyme marker in this study

2.1.1.1 Principle

Allozyme marker variation is detected at the level of the protein that is obtained from translation of mRNA, which in turn is transcribed from DNA. If mutation occurs, including substitution and deletion or addition causing the change of at least one amino acid in a polypeptide coded by genes, it may result in a change in the net electrostatic charge on the polypeptide. This change will in turn change the net charge on the enzyme or other protein of which such a polypeptide is a constituent. Usually an enzyme is in a state of three dimensional (Tertiary) structure, and consists of several polypeptides that are made up from one or more structural genes. Thus the electrophoretic difference in enzyme proteins will segregate as single mendelian genes (Hubby and Lewontin, 1966). If a large number of enzymes and samples of individuals are surveyed, the electrophoretic mobility of some enzymes may be different in individuals, and thus can be used as marker to investigate genetic variation.

2.1.1.2 Advantage and disadvantage of allozyme analysis

Traditionally, studies have utilised morphological traits and secondary compounds such as flavonids and terpenoids whose variation is controlled by several or many genes. One important characteristic of all these traits is possible phenotypic plasticity induced by changes in environmental factors, such as temperature and precipitation. Thus, the primary difficulty of using these data is to screen an appropriate marker that provides an accurate prediction of genotype from phenotype (Crawford, 1983).

Compared with morphological traits and secondary compounds, enzyme electrophoresis provides data that has many advantages, as described below (Brown and Weir, 1983; Crawford, 1983; Adams, 1983; Mitton, 1983): ① Allozyme expression is usually codominant and exhibits additive effects. ② Enzyme specificity allows interpretation of banding patterns as alleles at different loci and the comparability of loci in different populations or species. ③Each allelic difference is detected as a mobility difference and is independent of its functional role. ④ A large numbers of different loci can be assayed conveniently on one individual, using small amounts of material with minimal preparation

and expense. ⑤Allozymes are especially useful for studying conifers due to the haploid megagametophyte and diploid embryo which result in ease of detection of heterozygotes and, hence, to investigation of mating systems.

However, several disadvantages associated with allozymes may restrict their application in studies of plant genetics (Brown and Weir, 1983): ① It can be difficult to decipher the genetic basis of the observed protein banding pattern. ② The information that allozymes supply regarding levels of variation detected is underestimated. This is because only one quarter of base substitutions result in amino acid replacements that alter the net charge on the protein detectable by routine electrophoresis. ③ Substitution in non-coding region of DNA cannot be detected using allozymes. This may limit their potential as fingerprints for distinguishing individuals.

Although there are many disadvantages, allozyme markers have been extensively used in plant population structure (Mitton, 1983; Schaal, *et al.*, 1991), in tree breeding (Adams, 1983; Mitton, 1983), for example, the identification of parents and clones, and plant systematics (Crawford, 1983). Thus, in the following, studies using allozymes markers in plant population structure, mating system, and phylogenies are briefly commented upon.

2.1.1.3 Use of allozymes in studies of plant population structure

As was emphasised in Chapter 1, distribution of genetic variation within and among populations is an important indicator in practice for developing strategies made for genetic conservation (Ennos, 1996). Allozyme markers have been extensively employed in studies for this purpose in more recent years. For example, a review by Hamrick and Godt (1989) has summarised that population genetic structure of about 468 plant species were investigated by allozyme markers, showing that 11.3% of genetic variation occurred between populations. Use of allozyme marker to quantitatively estimate the population structure was emphasised by Hamrick *et al.* (1991). Hamrick and Godt (1996) concluded that " Generalisation from the plant allozyme literature can be used to predict the levels and distribution of genetic diversity in unstudied species, but the accuracy of such prediction is low...". The use of allozyme marker to investigate population structure of previously unstudied species is still reported today.

Populations of *Larix* species have also been investigated using allozyme markers. For example, about 5% of total genetic variation was detected to occur between populations in *L. laricina* (Cheliak *et al* ,1988), which is less than that found in *L. occidentalis* where 8.6% of genetic variation was due to that between populations (Fins and Seeb, 1986). An even smaller genetic differentiation (2%) was found between natural populations of *L. laricina* from northern Ontario (Liu and Knowles, 1991), and in this study, no significant correlation was found between genetic variation and geographic locations, implying existence of a random spatial pattern.

One important point that influences the choice of allozyme marker in this study is that comparable level of population structure assessed by allozyme and by DNA markers can be obtained in some cases studies described below. It is likely that the capability for fingerprinting individuals may increase from allozyme data to DNA sequence data. This is because many introns and silent point mutation may occur in DNA sequence and degeneration occurs for protein synthesis from gene to transcription to translation. We do not know how many point mutations or how large other kinds of mutation, such as insertion/deletion, need to be to lead to presence of a new allozyme allele detectable on a gel. But it is conservative to accept that accumulation of at least one mutation may cause a new allozyme allele (band(s)) to appear.

Results of several studies provide evidence in support of this point. For example, Spooner *et al.* (1996) used different molecular markers, for measuring relationships among the wild relatives of *Solanum* section *Etuberosum*, and showed that the capability for detecting interspecific sampling variance is different. A gradation exists from allozyme (low) to RAPD to RFLP (nuclear DNA) (high), and the contrasting capability for detecting intraspecific variation, grades from RFLPs (low) to RAPDs (high). Similar results were found in two aspen species, trembling aspen (*Populus tremuloides*) and bigtooth aspen (*P. grandidentata*) (Liu and Furnier, 1993). Using nuclear DNA RFLP and RAPDs and allozyme markers, Liu and Furnier (1993) found that the changes for the polymorphism were from RAPD (100%), to allozyme (77%), to RFLP (71%) in populations of trembling aspen, and from RAPD (88%), to RFLP(65%), to allozyme (29%) in populations of bigtooth aspen. This case indicates that RFLP and allozymes revealed comparable patterns of genetic variation in populations of trembling aspen not bigtooth aspen.

Several studies have indicated that allozymes and RAPDs both detect comparable levels of genetic variation within and among populations. For example, Szmidt *et al.*(1996) investigated two populations of *Pinus sylvestris* (L.) from northern Sweden using 20 allozyme and 22 RAPD loci, and found that, when complete genotype information was obtained, RAPD analysis provided genetic information similar to that revealed by analysis of allozyme variation. Similarly, in *Buchloe dactyloides*, a plant species that is widely distributed throughout the Great Plains of North America, Peakall *et al.* (1995) surveyed two diploid populations in both Mexico and Gulf Texas regions, using twelve allozyme loci and 98 RAPD polymorphic bands. Their results indicated that RAPD bands revealed greater variation among regions (54% of total variance) than allozymes (45.2%), but less variation among individuals within populations (31.9% for RAPD vs. 45.2% for allozymes); the proportion of genetic variance among populations within regions was similar (9.7% for RAPDs vs. 9.6% for allozymes).

As we know, variation detected by RAPD marker probably reflects the genetic variation of whole plant genomes for an array of given primers because the regions amplified by these primers are based on the whole genomes information. Variation detected by allozyme markers in contrast reflects the genetic variation of protein coding regions of whole genome with respect to an array of given enzymes. Theoretically, RAPD markers may be more powerful for detecting variation than allozyme because potential variation is reduced from DNA sequence to protein data. However, levels of population structure assessed by RAPD and allozyme markers are expected to be comparable since population structure parameters depend upon the distribution of variation, not its absolute value.

2.1.1.4 Use of allozymes to study plant phylogeny

The use of allozymes for reconstructing phylogenies and making ystematic inferences was reviewed by Crawford (1983). Crawford (1983) pointed out that allozyme makers could provide a reasonably precise and quantitative measure of genetic divergence between populations, subspecies and species. One important conclusion from Crawford's review is that the majority of studies demonstrated high genetic identity between conspecific populations and between subspecies. The genetic identity ranged from 0.87 to 1.00 for conspecific populations (Table 1 of Crawford, 1983) and from 0.75 to 0.99 for subspecies (Table 2 of Crawford, 1983).

For example, Zanetto *et al.* (1994) used allozymes markers to elucidate the interspecific differentiation of two white oaks, *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. They surveyed 14 populations of these two species, using ten polymorphic allozyme loci in *Q. petraea* and nine in *Q. robur*. Variation among populations within species was low for both species, 2.4% for *Q. robur* and 3.2% for *Q. petraea*. Differentiation between species was low, 3.3%, which was equivalent to that between populations. Comparison of local interspecific genetic distances indicated no clear geographic pattern of inter-specific differentiation among seven different regions.

Similar to the analyses of Crawford (1983), two possible explanations for these quite close relationships between conspecific populations or between subspecies elucidated by allozyme markers are: (i) The subdivision between populations or the divergence between subspecies occurred recently, and thus changes have not occurred at allozyme loci due to the time factor. (ii) Possible hybridisation between subspecies prevents divergence. The extensive gene flow between conspecific populations or between subspecies may significantly reduce the genetic divergence between them.

If the above conclusion obtained by using allozyme markers is quite general to conspecific populations or to subspecies diverged within a short period of time, these results may be extended to other similar case studies. Thus it is expected that a similar situation may occur for the three larch taxa in this study because the time of divergence leading to their formation is not known but was shown to be quite close to each other in terms of Nei's genetic distance (Tang, *et al.*, 1995).

Plant phylogeny elucidated by allozyme markers are often, but not always concordant with those obtained by morphological characters because allozyme markers are usually selective neutral. For example, Vickey (1990) used 11 loci with 30 alleles to investigate 2000 plants belonging to 85 populations of the *Mimulus glabratus* complex (Scrophulariaceae). He found that allozyme results were consistent with a tentative phylogeny of the complex developed from cytological and biogeographic data, showing clear differences between almost all 17 of the races. The groups delimited by allozymes correspond remarkably well with the geographic races. However, in a separate study of the genus *Wolffia*, Crawford and Landolt (1995) used 14 allozyme loci to score a total of 133 clones representing 10 of the 11 recognised species. The genetic identities among most pairs of species were zero, with non-

zero values ranging from 0.14 to 0.40. Crawford and Landolt concluded that enzyme electrophoresis provided limited resolution of species relationships in the genus *Wolffia* because of lack of shared alleles between the majority of species pairs in this genus.

Similar arguments can be extended to the comparison of plant phylogeny elucidated by allozymes and DNA markers. It is still difficult to make a clear judgement on this comparison. Mummenhoff *et al.* (1995) used RFLP analysis of cpDNA to examine the phylogeny among *Lepidium* taxa which is usually classified into sections; *Lepia*, *Lepiocardamon* and *Cardamon*. By using 15 restriction endonucleases, filter hybridisation experiments, and comparative mapping procedures, a total of 119 variable restriction sites were detected. Of these, 56 were phylogenetically informative and were used in cladistic analysis. The resulting phylogenetic tree agrees with results derived from morphology, allozyme electrophoresis and the analysis of glucosinolates. In a separate study, Sharma *et al.* (1995) used RAPD markers to distinguish between six different *Lens* taxa representing cultivated lentil and its wild relatives. Twenty-four arbitrary sequence 10 primers were identified, which generated a total of 88 polymorphic bands in 54 accessions. The total variation was partitioned into variation within and among *Lens* taxa. The relationships among the six taxa elucidated by RAPD markers corresponded well with previous isozyme and RFLP studies. These two cases also indicate that use of allozymes to elucidate plant phylogeny is consistent with that elucidated by DNA markers. However, a case study for the discordance for the phylogeny elucidated by DNA markers and allozymes has not been found.

It is very important to understand the relationship between phylogeny construction and the type of marker employed. Since the markers used for this purpose underly quite variable evolutionary mechanisms such as mutation, selection, recombination, inheritance mode etc., the phylogeny structure inferred by these different markers are likely to be different. However, the true phylogeny that we infer using different markers is only one. Which marker is best to infer phylogeny is still arguable due to influence of many factors involved in the markers used. Although allozyme markers mainly reflect variation of protein coding regions of DNA, it is after all at level of protein not the variation at DNA sequence level. However, if a sufficient number of allozyme markers are used, it is likely that the phylogeny elucidated by allozymes may approach the same one elucidated by coding regions for the same genomes, but may not approach the one elucidated by non-coding regions.

In summary, allozymes are still a useful marker even to date and widely employed in studying population structure and phylogeny.

2.1.2. Aims of present study

The aims of this chapter are to use allozymes: (i) to investigate the distribution of genetic variation within and between natural populations of the three larch taxa; (ii) to explore the possibility of elucidating evolutionary relationships among the three larch taxa..

2.2. Materials

Open pollinated seeds were collected from natural populations of the three larch taxa: eight in *L. gmelinii*, six in *L. olgensis* and two in *L. principis-rupprechtii*. In addition, one population of *L. olgensis* was sampled from a seed orchard in Liaoning Province. Samples of *L. gmelinii* and *L. olgensis* cover most of their distribution in China, but the two populations of *L. principis-rupprechtii* represent only the North seed zone (Fig. 1.5 in Chapter 1). Locations of the sampled populations are shown in Fig.2.1.

Table 2.1 lists the geographic location and sample size of these populations studied. Some of the sampled populations were collected as mixtures of half-sibs, and others as separate half-sib families. Mixed samples shown in Table 2.1 for *L. gmelinii* were those used for provenance trials, provided by Prof. Ban-li Pan, Forestry Academy of Heirongjiang, China. The number of half-sib families are not clear.

2.3. Methodology

2.3.1 Seed preparation and enzyme extraction

Mature seeds were surface-sterilised using H_2O_2 for about 20 minutes, and then allowed to germinate for three or four days prior to analysis. The seed coat was then excised, and the macrogametophyte tissue and embryo were isolated. Both of these tissues were then separately homogenised by hand grinding on ice in 25 µl and 30 µl of seed extraction buffer (0.013M Tris; 0.0043M citric acid; 0.50 mg/ml NADP; 0.50 mg/ml NAD; 0.18mg/ml

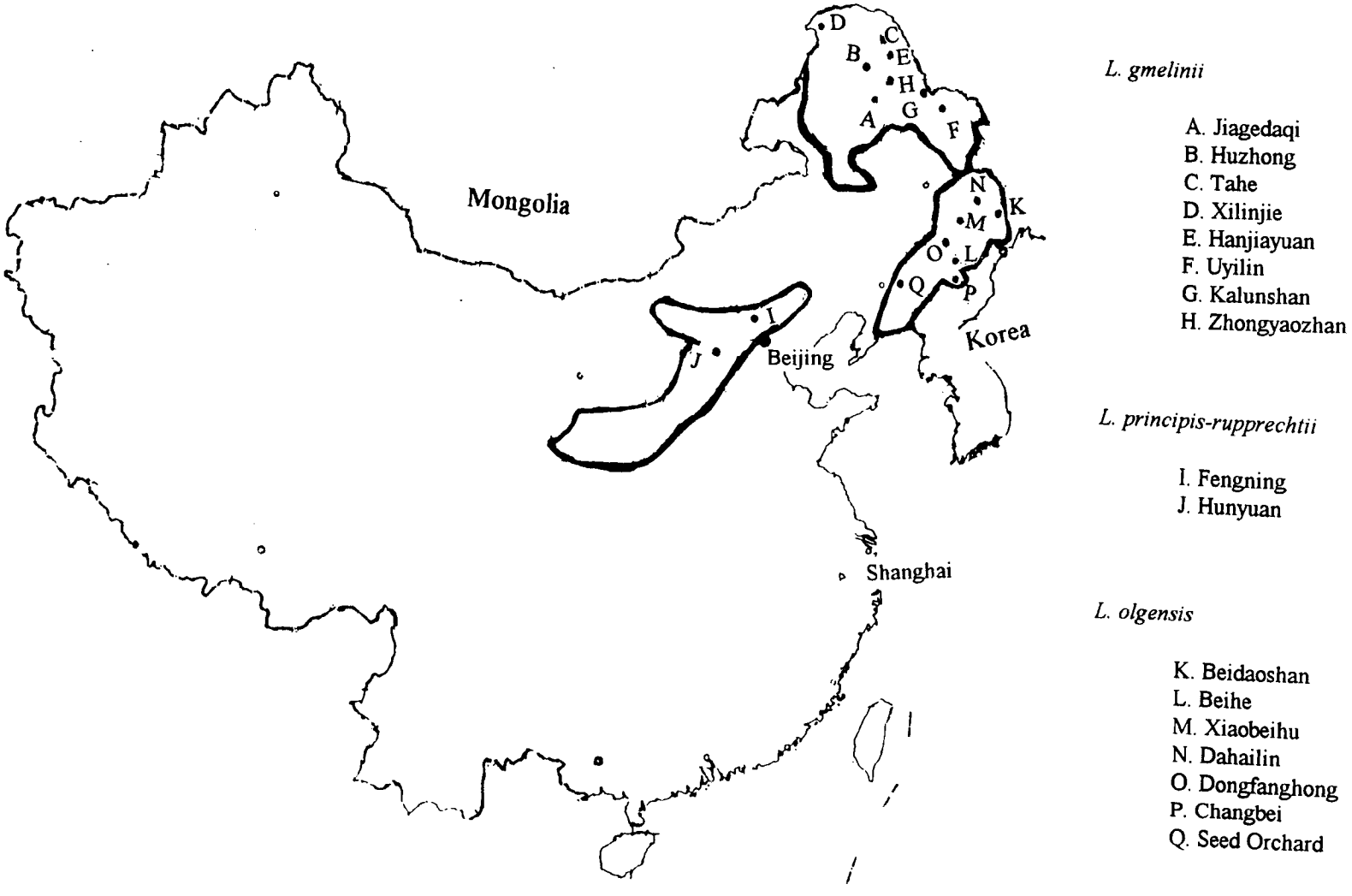


Fig.2.1. Natural distribution of the three *Larix* taxa in this study and locations of populations sampled for allozyme analysis.

ascorbic acid; 0.34 mg/ml EDTA; 0.10 mg/ml bovine serum albumin; 0.15% (v/v) β-mercaptoethanol; pH7.5; see Cheliak and Pitel, 1984), respectively.

2.3.2 Buffer systems and starch gel preparation

Buffer systems Two different buffer systems were used: system I and II. For system I, the starch gel buffer was composed of 12.1mg/ml tris-base-HCl (pH 8.5), while the electrode buffer contained: 2 mg/ml NaOH, 18.5 mg/ml boric acid pH 8.0. For system II (Chelik and Pitel, 1984), the starch gel buffer contained 2.62 mg/ml histidine-HCl, 0.13 mg/ml EDTA pH7.0, and the electrode buffer 15.1375 mg/ml Tris pH7.0.

Table 3.1. Location and sample size of the 17 *Larix* populations investigated using allozyme analysis

Species/Population	Latitude(N)	Longitude(E)	Half-sibs	Seeds	Seeds/half-sib
L. gmelinii					
Huzhong	51°56'	123°42'	mixed	90	
Tahe	52°30'	124°45'	mixed	90	
Xilinjie	53°20'	122°10'	mixed	90	
Hanjiayuan	52°15'	125°45'	mixed	90	
Uyilin	48°30'	129°26'	mixed	90	
Kalunshan	49°58'	127°30'	mixed	54	
Zhongyaozhan	50°45'	125°07'	mixed	56	
Jiagedaqi	50°24'	124°07'	21	126	6
L. principis-rupprechtii					
Fengning	41°12'	116°32'	20	121	> 6
Hunyuan	39°32'	113°41'	9	75	> 6
L. olgensis					
Beidaoshan	44°00'	131°07'	20	120	6
Beihe	42°25'	128°08'	29	174	6
Xiaobeihu	44°01'	128°50'	33	198	6
Dahailin	44°28'	129°48'	25	288	> 6
Dongfanghong	42°39'	128°06'	20	139	> 6
Changbei	41°26'	128°11'	21	209	> 6
Seed orchard	41°54'	124°06'	mixed	188	

Starch gel preparation 27.5 g of hydrolysed starch was mixed with 250 ml of the required starch gel buffer (11% w/v) in a side-arm flask and heated. The solution was evacuated, then poured into a plastic mould, and immediately covered with a plastic plate, avoiding air bubbles. The gel was allowed to cool for 30 minutes, then covered with cling-film and left overnight at room temperature, prior to analysis.

2.3.3 Electrophoresis

Electrophoresis was conducted according to Cheliak and Pitel (1984). Filter paper wicks about 3.0 mm in width were soaked in enzyme homogenate and placed in a slit cut through the gel approximately 2.5 cm from its cathodal end. Homogenate from both macrogametophyte and embryo tissue were loaded into the same gel next to each other, so that heterozygous individuals could be scored easily. Bromocresol green ($C_{21}H_{14}Br_4O_5S$, pH5.4) was used as a tracker dye and loaded into three wells: two on both ends and one in the middle of the gel.

The gel was run at 30 mA for 30 min and then at 60 mA for about 4 hours, until the buffer front marked by the tracker dye had migrated a sufficient distance, 2-3 cm to the frontier edge. The gel was sliced and stained with different enzyme recipes, after Cheliak and Pitel (1984; see Appendix II).

2.3.4 Scoring of gels

The common allele at each locus was designated as allele 1. The migration distance of the common allele within each locus was measured and compared to the total migration distance of the tracker dye (Cheliak and Pitel, 1984a), to measure the R_f value. Any variants (alleles) observed at a particular locus were then measured relative to its common allele, to obtain the R_m of each allele. This was obtained using the following :

$$R_m = \frac{\text{migration distance of variant (mm)}}{\text{migration distance of standard(mm)}}$$

2.4. Data Analysis

Data analyses involved in this study are: (i) genetic diversity within populations; (ii) single-

and multi-locus structure within populations; (iii) genetic structure among populations; (iv) genetic distances within and between larch taxa.

Allele frequencies within each population were calculated for each locus. Average observed (Ho) and expected (He) heterozygosity (under Hardy-Weinberg equilibrium) within each population were calculated using Fstat (version 1.2) package (Goudet, 1995).

Test of Hardy-Weinberg equilibrium within each population was conducted for each locus. The null hypothesis H_0 is that there is random combination of gametes. Fisher's exact test using contingency tables was used. Haldane's (1954) exact probability (P-value) was used to calculate the probability of the occurrence of the observed sample under the null hypothesis (also see Weir, 1991, p78-79). Two alternative effective hypotheses were assumed to calculate the probability of excess or deficiency. The test introduced by Rousset and Raymond (1995) was employed. For the hypothesis of heterozygote deficiency, the P-value is the sum of probabilities of samples in which heterozygotes are less than that observed. For the hypothesis of heterozygote excess, the P-value is the sum of probabilities of samples in which heterozygotes are greater than that observed. In fact, they are two one-tailed tests.

Allele polymorphism was measured as the fraction of polymorphic loci in which the commonest allele frequency was less than 99%, denoted by P(99%), i.e. the ratio of the number of loci whose common allele frequency was smaller than 0.99, to the total number of loci analysed. The mean number of alleles per locus within each population was calculated.

Linkage disequilibrium between any pair of loci within each population or over populations was tested using the software package Genepop (version 2.0; Raymond and Rousset, 1995). Fisher's exact test (P-value) was conducted using contingency tables.

Population genetic structure was analysed using the Fstat package (version 1.2; Goudet, 1995). Wright's *F*-statistics within each of the three larch taxa were calculated for each locus according to the method used by Weir and Cockerham (1984). The significance of population differentiation was carried out for each locus and over loci using Fisher's probability test (P-value). However, the variance of each *F*-statistic was estimated by jack-knife methods over loci.

The spatial pattern for population genetic structure was tested for evidence of isolation by distance using the method introduced by Slatkin (1993). Using simulation under a variety of models, Slatkin (1993) proved that an approximate log-log linear relationship existed between the number of migrants (Nm) and geographic distance (D), i.e. Log(Nm)=a+bLog (D), where a and b are regression parameters. If the b value is smaller than zero, the number of migrants will decrease with geographic distance, and *vice versa*. The number of migrants (Nm) can be estimated according to Wright's formula, i.e. $Nm = (1 / F_{st} - 1) / 4$. The spatial heterogeneity in terms of Fst was also tested using Mantel's test (Mantel, 1967), which uses a permutation procedure to produce the distribution of the test statistics, *Z*, and then calculate the exact probability (P-value) for the observed sample under the null hypothesis (also see Pigliucci and Barbujani, 1993). Estimate of migrants from data on private alleles was tested using Genepop (version 2.0; Raymond and Rousset, 1995). Number of migrants (Nm) was also estimated using the method introduced by Barton and Slatkin (1986).

Phylogeny reconstruction among the three larch taxa was carried out with Biosys-1 (Swofford and Selander, 1981) according to Nei's genetic distance (Nei, 1972), using the unweighted pair-group method with arithmetic averaging (UPGMA).

2.5. Results

2.5.1 Primary screening of polymorphic markers

Eleven enzyme systems were screened for useful markers in each of the three larch taxa. These enzymes were aspartate aminotransferase (AAT; E.C.2.6.1.1); Glucose-6-phosphate dehydro-genase (G6PD; E.C.1.1.1.49); Glutamate dehydrogenase (GDH; E.C.1.4.1.3); Isocitrate dehydrogenase (IDH;E.C.1.1.1.42); Leucine amino-peptidase (LAP; E.C.3.4.11.1), malate dehydrogenase (MDH; E.C.1.1.1.37); 6-phosphogluconate dehydrogenase (6PGD; E.C.1.1.1.44); phosphoglucose isomerase (PGI; E.C.5.3.1.9); phosphoglucomutase (PGM; E.C.2.7.5.1); shikimic acid dehydrogenase (SDH; E.C.1.1.1.25) and superoxide dismutase (SOD; E.C.1.15.1.1.). Among these 11 enzyme systems scored, six were found to be polymorphic in at least one of populations studied; while others were monomorphic or difficult to resolve. The six polymorphic enzyme systems were PGI, MDH, 6PGD, AAT, PGM and SDH. These six polymorphic enzymes resolved a total of eight loci..

2.5.2 Interpretation of banding pattern

Phosphoglucose isomerase (PGI)

One zone of PGI activity was clearly observed, which is the same as that in *L. laricina* (Cheliak and Pitel, 1985). It expressed a total of three different bands (phenotype) in haploid tissue. If the diploid tissue was heterozygous for this enzyme system, then three bands were apparent (Fig. 2.2a). Thus, this was interpreted as representing one locus with three alleles, i.e. alleles Pgi^1, Pgi^2 and Pgi^3.

6-Phosphogluconate dehydrogenase (6PGD)

Two zones of 6PGD activity were observed, which is the same as *L. laricina* (Cheliak and Pitel, 1985). 6PGD-I (Fig.2.2a) exhibited monomorphic, but 6PGD-II exhibited one locus with two different band variants in haploid tissue, i.e. alleles $6Pgd^1$ and $6Pgd^2$. If the diploid tissue was heterozygous, three clear bands were apparent (Fig. 2.2a).

Malate dehydrogenase (MDH)

Four zones of MDH activity were observed (Fig.2.2b) which is the same as in *L. laricina* (Cheliak and Pitel, 1985) and *L decidua* (Lewandowski and Meinartowicz, 1988). MDH-I exhibited two alleles in haploid tissue, i.e. $Mdh\text{-}I^1$ and $Mdh\text{-}I^2$, and heterozygous individuals possessed three bands in diploid tissue. MDH-II and -III were monomorphic. MDH-IV exhibited two bands in haploid tissue, representing two alleles. However, the bands are too weak to be scored in diploid tissue, and were not used for analysis. Therefore, for MDH, only MDH-I was used for analysis (Fig. 2.2b).

Aspartate aminotransferase (AAT)

Three zones of AAT activity were evident (Fig.2.2b), which is the same as in *L. laricina* (Cheliak and Pitel, 1985). They are AAT-I, AAT-II and AAT-III (Fig.2.2b). Both AAT-I and AAT-II locus exhibited two alleles in haploid tissue, and heterozygous individual possessed three bands in diploid tissue. Since the remaining three bands always occurred together and were able to be clearly scored only in haploid tissue, thus these three bands were treated as

one locus with four alleles, i.e. alleles *Aat-III*^{1}, *Aat-III* ^{2}, *Aat-III* ^{3} and *Aat-III*^{4} (Fig. 2.2b). However, alleles of AAT-III locus were too weak to be scored in diploid tissue. Therefore, for the AAT-III locus, only haploid data were used for analysis.

Phosphoglucomutase (PGM)

One zone of PGM activity was observed, which is the same as in *L. laricina* (Cheliak and Pitel, 1985). Four alleles in haploid tissue were observed, i.e. alleles *Pgm*^{1}, *Pgm*^{2}, *Pgm*^{3} and *Pgm*^{4}, and heterozygous individuals possessed two bands (Fig.2.2c).

Shikimate dehydrogenase (SDH)

One zone of SDH activity was evident (Fig. 2.2c), which is the same as in *L. decidua* (Lewandoeski and Mejnartowicz, 1990) and in Japanese and European larch (Ennos and Tang, 1995). Four alleles in haploid tissue were observed, i.e. alleles *Sdh*^{1}, *Sdh*^{2}, *Sdh*^{3} and *Sdh*^{4}, and heterozygous individuals possessed two bands in diploid tissue (Fig. 2.2c).

2.5.3 Allele frequency and polymorphism

The allele frequencies at all loci were calculated for all populations investigated, and were shown to be variable among populations (Table 2.2).

PGI expressed two alleles in all populations of *L.gmelinii*, but three alleles could be found in populations Xiaobeihu, Dongfanghong and Changbei of *L. olgensis*, and populations Fengning and Hunyuan of *L. principis-rupprechtii* (Table 2.2). The frequency of the most common allele (*Pgi*^{1}) was greater than 0.9 in all populations of *L. gmelinii* and *L. olgensis*, but was less than 0.8 in the two populations of *L. principis-rupprechtii*, i.e. Fengning and Hunyuan,.

MDH-I exhibited a similar structure to PGI in all populations investigated. The frequency of the most common allele (*Mdh-I*^{1}) was greater than 0.9, but that of *Mdh-I*^{2} allele was more variable, ranging from 0.00 to 0.039. Three populations of *L. gmelinii*, i.e. Tahe, Kalunshan and Zhongyaozhan, were fixed for *Mdh-I*^{1}, but this was not the case in any populations of *L.olgensis* and *L. principis-rupprechtii* (Table 2.2).

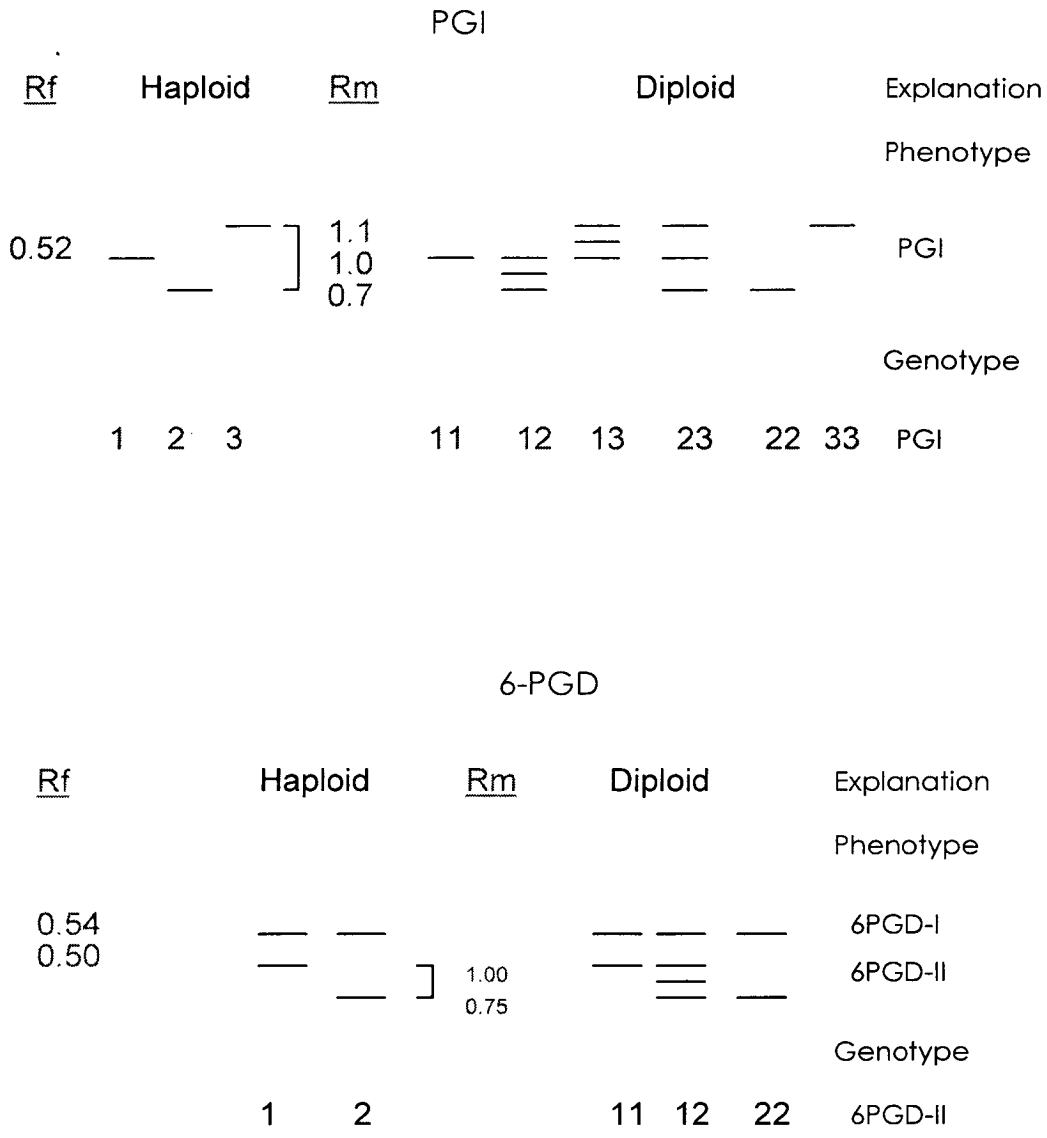


Fig. 2.2a. Zymogrammes representing isozyme banding pattern (PGI and 6PGD) within the three Chinese *Larix* taxa.

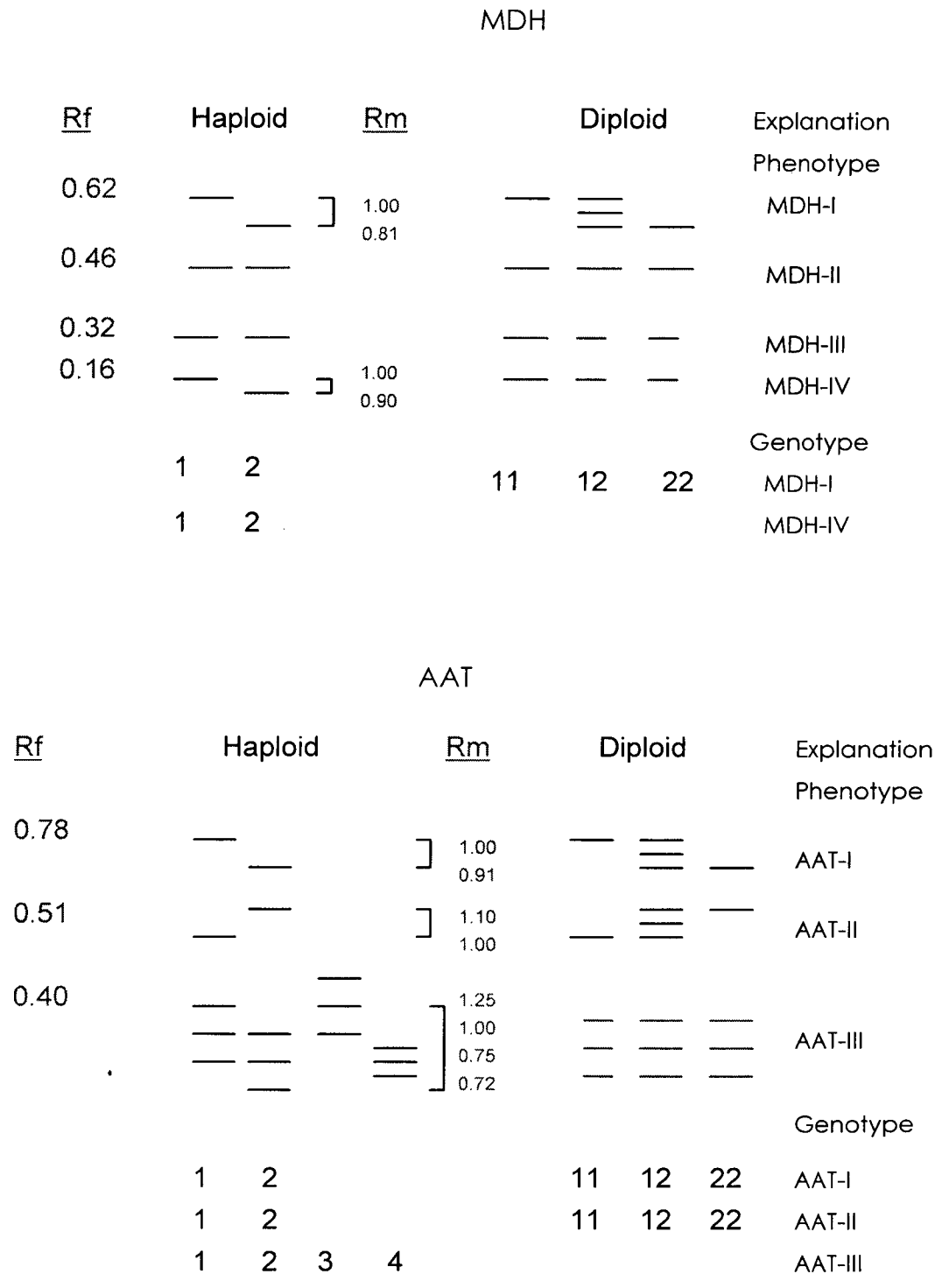


Fig. 2.2b Zymogrammes representing isozyme banding pattern (MDH and AAT) within the three Chinese *Larix* taxa.

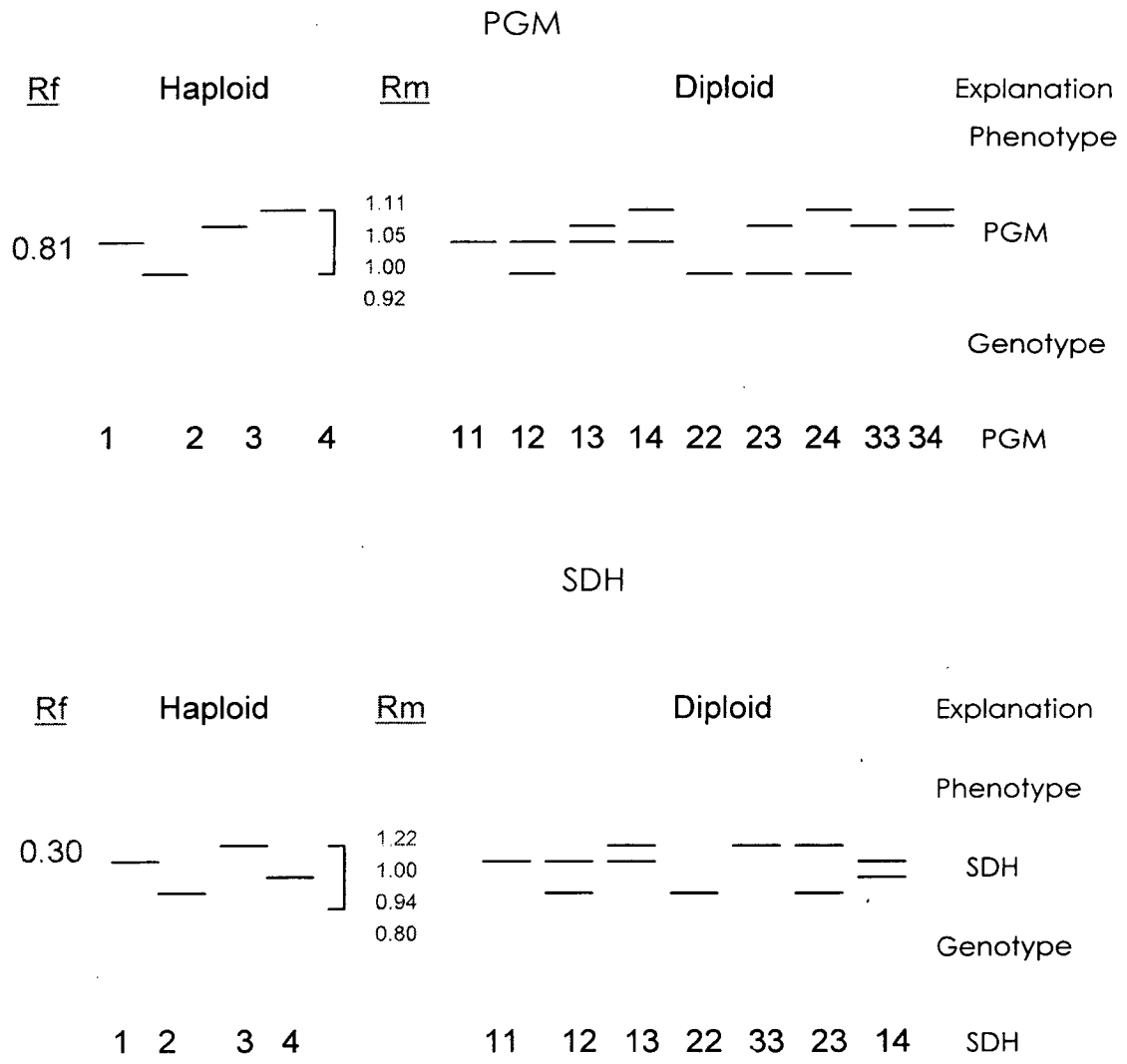


Fig. 2.2c Zymogrammes representing isozyme banding pattern (PGM and SDH) within the three Chinese *Larix* taxa.

6PGD was fixed for allele *6Pgd*^{1} in all the populations investigated except *L. gmelinii* in Xilinjie and Kalunshan. The two populations of *L. principis-rupprechtii* were fixed for allele *6Pgd*^{1} as well. However, all populations of *L. olgensis* except Dahailin were polymorphic at this locus (Table 2.2).

AAT also exhibited difference in the populations investigated. For the AAT-I locus, populations Jiagedaqi, Hanjiayuan and Uyilin, were fixed for allele *Aat-I*^{1}, but the remaining populations of *L. gmelinii* were polymorphic. The two populations of *L. principis-rupprechtii* were slightly different from each other. Population Fengning exhibited nearly complete fixation for allele *Aat-I*^{1}, with frequency being 0.996. Population Hunyuan , however , exhibited *Aat-I*^{1} at a frequency of 0.920 and *Aat-I*^{2} at 0.080. Most populations of *L. olgensis* were fixed, or nearly fixed, for allele *Aat-I*^{1}, except for populations Beidaoshan and Changbei in which frequencies of *Aat-I*^{1}were 0.975 and 0.971, respectively (Table 2.2).

For AAT-II, all populations investigated were polymorphic except for population Tahe in *L. gmelinii* that was monomorphic for *Aat-II*^{1}. AAT-III exhibited differences among the three larch taxa. For example, most populations of *L. gmelinii* exhibited two alleles at this locus with the exception of populations Tahe and Uyilin that has three alleles. All populations of *L.olgensis* and *L. principis-rupprechtii* , however, were fixed for *Aat-III*^{1} (Table 2.2).

PGM was highly polymorphic in all populations of each of the three larch species. The frequencies of the most common allele *Pgm*^{1}, ranged from 0.627 to 0.778 in populations of *L. gmelinii*, from 0.711 to 0.867 in populations of *L. olgensis*, and from 0.717 to 0.719 in the two populations of *L.principis-rupprechtiii* (Table 2.2).

SDH was expressed differently in the three larch taxa. In *L. gmelinii*, for example, four alleles were detected and high levels of polymorphism were found in each population. Frequencies of the most common allele *Sdh*^{1} ranged from 0.873 to 0.928. In *L. principis-rupprechtii*, however, both populations investigated were monomorphic for allele *Sdh*^{1}. In *L. olgensis*, SDH was polymorphic and variable among populations. Frequency of the common allele SDH-1 ranged from 0.830 in Dahailin to 1.00 in Changbei (Table 2.2).

The levels of polymorphism of resolved loci in all studied populations of the three species

have been estimated and are summarised in Table 2.3. It can be seen that the number of alleles per locus was comparable between populations within each species, ranging from 2.00 to 2.37 in populations of *L. gmelinii*, from 1.87 to 2.00 in two populations of *L. principis-rupprechtii*, and from 1.87 to 2.28 in populations of *L. olgensis*. The differences between taxa were not the same. The average number of alleles per locus was 2.20 in *L. gmelinii*, larger than 2.11 in *L. olgensis*, larger than 1.93 in *L. principis-rupprechtii* (Table 2.3).

The percentage of polymorphic loci, P(99%) value, revealed a similar relationship between species as did the average number of alleles per locus, *L. gmelinii* exhibited the largest percentage of polymorphic loci, with P(99%) value being 68%, while *L. principis-rupprechtii* exhibited the lowest polymorphism, with P(99%) value being 49%, and that for *L. olgensis* was 66% .

The level of observed heterozygotes over loci in *L. gmelinii* ranged from 0.075 (Hanjiayuan) to 0.126 (Kalunshan), with a mean of 0.097, while in *L. olgensis*, values ranged from 0.067 (Xiaobeihu) to 0.141(seed orchard), with a mean of 0.090. The two populations of *L. principis- rupprechtii* showed slightly higher observed heterozygotes than those in the other two species, 0.103 in Fengning and 0.104 in Hunyuan, with a mean of 0.103. In most populations expected heterozygote frequencies were slightly larger than those of observed heterozygotes with the exceptions of populations such as in Kalunshan, 0.126 (Ho) vs 0.115 (He), and in the seed orchard (0.141vs 0.135). Mean of observed and expected heterozygotes of *L. gmeliniii* 0.097 (Ho) and 0.100 (He), was comparable to that of *L. olgensis*, 0.090 (Ho) and 0.097 (He), but both were slightly smaller than that of *L. principis-rupprechtii* , 0.103 (Ho) and 0.126 (He).

Table 2.2. Allele frequencies estimated in each population investigated (' ------' means that data were not obtained for various technical reasons)

		L. gmelinii							
	Allele	Jiagedaqi	Huzhong	Tahe	Xilinjie	Hanjiayuan	Uyilin	Kalunshan	Zhongyaozhan
Pgi	-1	0.996	0.983	0.994	0.944	0.978	0.950	0.907	0.964
	-2	0.004	0.017	0.006	0.056	0.022	0.050	0.093	0.036
Mdh-I	-1	0.996	0.961	1.000	0.983	0.994	0.989	1.000	1.000
	-2	0.004	0.039	0.000	0.017	0.006	0.011	0.000	0.000
6Pgd	-1	1.000	1.000	1.000	0.994	1.000	1.000	0.991	1.000
	-2	0.000	0.000	0.000	0.006	0.000	0.000	0.009	0.000
Aat-I	-1	1.000	0.983	0.972	0.978	1.000	1.000	0.991	0.973
	-2	0.000	0.017	0.028	0.022	0.000	0.000	0.009	0.027
Aat-II	-1	0.952	0.972	1.000	0.972	0.972	0.980	0.981	0.964
	-2	0.048	0.028	0.000	0.028	0.028	0.020	0.019	0.036
Aat-III	-1	0.746	0.856	0.811	0.867	0.867	0.900	0.889	0.821
	-2	0.254	0.133	0.156	0.133	0.133	0.080	0.111	0.179
	-3	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.000
	-4	0.000	0.000	0.033	0.000	0.000	0.000	0.000	0.000
Pgm	-1	0.627	0.739	0.639	0.728	0.778	0.761	0.759	0.696
	-2	0.151	0.106	0.206	0.100	0.094	0.067	0.111	0.071
	-3	0.044	0.078	0.067	0.117	0.078	0.078	0.074	0.036
	-4	0.179	0.078	0.089	0.056	0.050	0.094	0.056	0.196
Sdh	-1	0.873	0.889	0.878	0.900	0.928	0.917	0.852	0.893
	-2	0.119	0.094	0.083	0.089	0.072	0.083	0.111	0.098
	-3	0.008	0.017	0.028	0.011	0.000	0.000	0.028	0.009
	-4	0.000	0.000	0.011	0.000	0.000	0.000	0.009	0.000

Table 2.2 (*Continued*)

		L principis-rupprechtii		*L. olgensis*						
	Allele	Fengning	Hunyuan	Beidaoshan	Beihe	Xiaobeihu	Dahailin	Dongfanghong	Changbei	Seed Orchard
Pgi	-1	0.760	0.715	0.979	0.937	0.937	0.910	0.929	0.947	0.953
	-2	0.236	0.215	0.021	0.063	0.058	0.090	0.046	0.032	0.047
	-3	0.004	0.069	0.000	0.000	0.005	0.000	0.025	0.021	0.000
Mdh-I	-1	0.967	0.993	0.967	0.980	0.997	0.976	0.971	0.931	0.914
	-2	0.033	0.007	0.033	0.020	0.003	0.024	0.029	0.069	0.086
6Pgd	-1	1.000	1.000	0.975	0.974	0.997	1.000	0.971	0.960	0.930
	-2	0.000	0.000	0.025	0.026	0.003	0.000	0.029	0.040	0.070
Aat-I	-1	0.996	0.920	0.975	0.991	1.000	1.000	0.996	0.971	0.992
	-2	0.004	0.080	0.025	0.009	0.000	0.000	0.004	0.029	0.008
Aat-II	-1	0.996	0.973	0.962	0.960	0.934	0.913	0.983	0.939	0.977
	-2	0.004	0.027	0.038	0.040	0.066	0.087	0.017	0.061	0.023
Aat-III	-1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Pgm	-1	0.719	0.717	0.867	0.862	0.838	0.840	0.842	0.838	0.711
	-2	0.004	0.022	0.046	0.023	0.051	0.049	0.025	0.053	0.031
	-3	0.050	0.109	0.025	0.032	0.051	0.056	0.004	0.011	0.055
	-4	0.227	0.152	0.063	0.083	0.061	0.056	0.129	0.098	0.203
Sdh	-1	1.000	1.000	0.996	0.994	0.978	0.830	-------	1.000	-------
	-2	0.000	0.000	0.004	0.006	0.022	0.066		0.000	
	-3	0.000	0.000	0.000	0.000	0.000	0.101		0.000	
	-4	0.000	0.000	0.000	0.000	0.000	0.003		0.000	

Table 2.3 Measures of genetic variability in 17 populations of *Larix**

Species	Population	A	P(99%)	Ho	He
L. gmelinii					
	Jiagedaqi	2.12	50	0.106	0.102
	Huzhong	2.25	87	0.100	0.102
	Tahe	2.25	50	0.081	0.094
	Xilinjie	2.37	87	0.096	0.108
	Hanjiayuan	2.00	62	0.075	0.078
	Uyilin	2.12	75	0.085	0.089
	Kalunshan	2.37	62	0.126	0.115
	Zhongyaozhan	2.12	75	0.104	0.107
Mean		2.20	68	0.097	0.100
L. principis-rupprechtii					
	Fengning	1.87	37	0.103	0.117
	Hunyuan	2.00	62	0.104	0.134
Mean		1.93	49	0.103	0.126
L. olgensis					
	Beidaoshan	2.12	75	0.072	0.071
	Beihe	2.12	62	0.076	0.076
	Xiaobeihu	2.12	50	0.067	0.077
	Dahailin	1.87	62	0.101	0.125
	Dongfanghong	2.28	71	0.091	0.090
	Changbei	2.12	75	0.083	0.105
	Seed orchard	2.14	71	0.141	0.135
Mean		2.11	66	0.090	0.097

*: A: average number of alleles per locus: P(99%), percentage of polymorphic loci where the frequency of the most common allele is <0.99.; Ho and He, observed and expected heterozygote, respectively.

2.5.4 Hardy-Weinberg equilibrium

Results of the tests of Hardy-Weinberg equilibrium within each population of the three larch taxa are summarised in Table 2.4. Inbreeding coefficient (F_{is}) and the probabilities (P-value) for occurrences of three types of departure from Hardy-Weinberg equilibrium (general, deficiency and excess) are listed in Table 2.4.

In *L. gmelinii*, AAT-II was found not to be in Hardy-Weinberg equilibrium in populations Jiagedaqi, Huzhong and Uyilin, while PGM and SDH were not in Hardy-Weinberg equilibrium in populations Xilinjie and Tahe respectively. All other populations tested were in Hardy-Weinberg equilibrium.

In *L. principis-rupprechtii*, PGI was not in Hardy-Weinberg equilibrium in population Hunyuan. In *L. olgensis*, the following populations did not show Hardy-Weinberg equilibrium in different enzymes: Xiaobeihu in PGI, AAT-II and PGM; Changbei in MDH, AAT-II and PGM; Beidaoshan and Seed orchard in AAT-II; Dahailin in GOT-II, PGM and SDH.

The main reason for this departure from Hardy-Weinberg equilibrium in most of these cases was due to a deficit of heterozygotes, rather than to excess of heterozygotes (Table 2.4), with the exceptions of populations Jagedaqi for AAT-II, Tahe for SDH and Changbei for PGM. By chance some values will differ from Hardy-Weinberg equilibrium when so many tests are carried out.

2.5.5 Linkage disequilibrium

The null hypothesis that any pair of loci are independent of each other is accepted from the tests of linkage disequilibrium in each population of each of the three species. For simplicity, the results of Fisher's global tests for each pair of loci over all populations in each of the three larch taxa are summarised in Table 2.5. It can be concluded that these seven loci are independent of each other.

Table 2.4. Test of Hardy-Weinberg equilibrium for all loci. Type-I error probabilities were listed for rejecting null hypothesis for all possible reasons(General) or for only heterozygote deficit (Deficit) or excess (Excess). Bold characters indicate significant values (P<0.05). Symbol '—' stands for monomorphic loci in this sample of embryos.

		L. gmelinii							
		Jiagedaqi	Huzhong	Tahe	Xilinjie	Hanjiayuan	Uyilin	Kalunshan	Zhongyaozhan
PGI	Fis	—	- 0.011	—	- 0.053	- 0.017	- 0.043	- 0.093	- 0.028
	General		1		1	1	1	1	1
	Deficit		1		1	1	1	1	1
	Excess		0.983		0.768	0.966	0.900	0.633	0.946
MDH-1	Fis	—	- 0.035	—	- 0.011	—	- 0.006	—	—
	General		1		1		1		
	Deficit		1		1		1		
	Excess		0.886		0.983		0.994		
6PGD	Fis	—	—	—	—	—	—	—	—
	General								
	Deficit								
	Excess								
AAT-I	Fis	—	- 0.011	- 0.023	- 0.017	—	—	—	- 0.019
	General		1	1	1				1
	Deficit		1	1	1				1
	Excess		0.983	0.944	0.966				0.973
AAT-II	Fis	0.129	0.796	—	0.388	- 0.023	1	- 0.010	0.488
	General	**0.022**	**0.000**		0.055	1	**0.010**	1	0.053
	Deficit	0.240	**0.000**		0.055	1	**0.010**	1	0.053
	Excess	0.978	1		0.999	0.944	0.989	0.991	0.946
PGM	Fis	- 0.063	- 0.000	0.075	0.229	0.005	0.036	- 0.141	- 0.131
	General	0.131	0.297	0.512	**0.003**	0.319	0.499	0.885	0.877
	Deficit	0.731	0.350	0.135	**0.000**	0.424	0.206	0.982	0.924
	Excess	0.269	0.649	0.864	0.999	0.575	0.801	0.086	0.117
SDH	Fis	0.010	- 0.102	0.203	0.029	0.094	- 0.079	- 0.054	0.268
	General	0.760	1	**0.041**	0.610	0.376	1	0.222	0.102
	Deficit	0.545	1	0.079	0.521	0.376	1	0.728	0.071
	Excess	0.690	0.305	0.921	0.799	0.941	0.762	0.449	0.990

Table 2.4. (*Continued*)

		L.principis-rupprechtii				*L. olgensis*				
		Fengning	Hunyuan	Beidaoshan	Beihe	Xiaobeihu	Dahailin	Dongfanghong	Changbei Seed orchard	
PGI	Fis	0.124	0.373	- 0.017	0.129	0.025	- 0.096	0.009	- 0.039	- 0.039
	General	0.060	**0.000**	1	0.135	**0.003**	0.605	0.456	1	1
	Deficit	0.095	**0.010**	1	0.135	**0.002**	1	0.456	1	1
	Excess	0.925	0.989	0.958	0.979	0.998	0.289	0.716	0.675	0.884
MDH-I	Fis	- 0.030	—	- 0.030	- 0.018	—	- 0.021	- 0.026	0.746	- 0.086
	General	1		1	1		1	1	**0.000**	1
	Deficit	1		1	1		1	1	**0.000**	1
	Excess	0.887		0.886	0.940		0.928	0.914	1	0.625
6PGD	Fis	—	—	- 0.021	- 0.021	—	—	- 0.026	- 0.039	- 0.068
	General			1	1			1	1	1
	Deficit			1	1			1	1	1
	Excess					0.938	0.899	0.914	0.761	0.739
AAT-I	Fis	—	0.282	- 0.021	- 0.006	—	—	—	- 0.027	—
	General		0.059	1	1				1	
	Deficit		0.059	1	1				1	
	Excess		0.996	0.938	0.991				0.860	
AAT-II	Fis	—	-0.021	0.426	- 0.039	0.508	0.433	- 0.013	0.215	0.663
	General		1	**0.006**	1	**0.000**	**0.000**	1	**0.022**	**0.023**
	Deficit		1	**0.006**	1	**0.000**	**0.000**	1	**0.022**	**0.023**
	Excess		0.959	1	0.761	1	1	0.975	0.997	1
PGM	Fis	0.098	-0.056	- 0.097	- 0.038	0.057	0.273	- 0.029	0.102	- 0.105
	General	0.269	0.854	1	0.337	**0.042**	**0.000**	0.157	**0.047**	0.801
	Deficit	0.294	0.836	1	0.288	**0.012**	**0.000**	0.059	0.057	0.913
	Excess	0.714	0.200	0.092	0.722	0.987	1	0.948	0.944	0.126
SDH	Fis	—	—	—	- 0.003	- 0.018	0.161	—	—	—
	General				1	1	**0.021**			
	Deficit				1	1	**0.007**			
	Excess				0.997	0.956	0.992			

Table 2.5 Fisher's global tests for linkage disequilibrium of any pair of loci within the three larch taxa. The probabilities of exact tests (P-value) are listed.

Locus pair		*L. gmelinii*	*L. olgensis*	*L. principis-rupprechtii*
PGI	MDH-I	0.9918	0.7810	0.5065
PGI	6PGD	1.0000	0.3434	Not possible*
MDH-I	6PGD	1.0000	0.9999	Not possible
PGI	AAT-I	0.9706	0.4487	0.4775
MDH-I	AAT-I	1.0000	0.8616	0.4176
6PGD	AAT-I	1.0000	0.9996	Not possible
PGI	AAT-II	0.0786	0.8043	0.4574
MDH-I	AAT-II	1.0000	0.8353	1.0000
6PGD	AAT-II	1.0000	0.9987	Not possible
AAT-I	AAT-II	1.0000	0.1399	0.5265
PGI	PGM	0.6287	0.2422	0.2884
MDH-I	PGM	0.2386	0.9945	Not possible
6PGD	PGM	0.3613	0.5196	0.3637
AAT-I	PGM	0.1968	0.9330	1.0000
AAT-II	PGM	0.3191	0.1994	Not possible
PGI	SDH	0.7042	0.9747	Not possible
MDH-I	SDH	0.9947	1.0000	Not possible
6PGD	SDH	1.0000	1.0000	Not possible
AAT-I	SDH	0.9964	1.0000	Not possible
AAT-II	SDH	0.8369	0.6147	Not possible
PGM	SDH	0.2091	0.7704	Not possible

*: The impossibility of calculating the P-value is due to the existence of one monomorphic locus in the pair of loci tested.

2.5.6 Population differentiation

Unbiased estimates of F_{st} for each of the three taxa are summarised in Table 2.6. F_{st} estimates for single loci were variable in *L. gmelinii*, ranging from -0.002 (SDH) to 0.022 (PGI). Population differentiation was significant for the single loci PGI, MDH-I, AAT-I and PGM, but not for 6PGD, AAT-II and SDH. Estimates of F_{st} for AAT-III, which was polymorphic only in *L. gmelinii*, using haploid data, was also not significant (0.013 ± 0.016).

Population differentiation of *L. principis-rupprechtii* was significant for single loci AAT-I (F_{st} = 0.079**) and AAT-II (F_{st} = 0.015 **), but not for PGI, MDH-I and PGM. In *L. olgensis*, population differentiation was significant for all single loci, with F_{st} ranging from 0.005 to 0.164.

Table 2.6 Estimates of F-statistics for 7 polymorphic loci over all populations of the three larch taxa. —: monomorphic locus ;*: P<5%;**: P<1%

	L. gmelinii			*L. principis-rupprechtii*			*L. olgensis*		
Locus	F_{IT}	F_{ST}	F_{IS}	F_{IT}	F_{ST}	F_{IS}	F_{IT}	F_{ST}	F_{IS}
PGI	-0.029±0.011	0.022±0.013**	-0.053±0.015	0.223	0.001	0.222	0.008±0.035	0.005±0.005**	0.003±0.038
MDH-I	-0.010±0.004	0.015±0.011**	-0.025±0.014	-0.016	0.009	-0.026	0.287±0.253	0.019±0.011**	0.273±0.256
6PGD	-0.002±0.001	0.000±0.003	-0.002±0.002	—	—	—	- 0.025±0.009	0.015±0.011**	- 0.040±0.022
AAT-I	-0.013±0.006	0.005±0.006**	-0.018±0.003	0.316	0.079**	0.257	- 0.011±0.006	0.012±0.005**	- 0.023±0.006
AAT-II	0.303±0.137	0.001±0.007	0.303±0.139	-0.002	0.015**	-0.018	0.338±0.089	0.006±0.005**	0.334±0.090
PGM	0.026±0.038	0.016±0.016**	0.010±0.041	0.045	0.005	0.041	0.049±0.044	0.008±0.008**	0.042±0.047
SDH	0.043±0.044	- 0.002±0.002	0.045±0.045	—	—	—	0.382±0.186	0.164±0.076**	0.239±0.123
Over loci	0.033±0.028	0.012±0.007**	0.021±0.034	0.132	0.009**	0.124	0.102±0.057	0.019±0.013**	0.084±0.053

Multilocus F_{st} estimates showed a significant departure from zero, although they were very small, less than 2%; indicating that most of the total genetic variation was maintained within populations in each of the three larch taxa.

A comparison of population differentiation between taxa indicated that *L. olgensis* presents larger F_{st} (=0.019) than that of *L. gmelinii* (F_{st} = 0.012), which in turn is larger than that of *L. principis-rupprechtii* (F_{st} =0.009). It should be noted that only two populations that were sampled from the northern seed zone of *L. principis-rupprechtii* have been included in the analysis.

2.5.7 Isolation by distance

In order to detect the relationship between the log(Nm) and the Log(D), i.e. Log(Nm) = a+bLog (D), the number of migrants (Nm), calculated from unbiased F_{ST}'s (Weir and Cockerham, 1984), were employed to regress to corresponding geographic distances. Estimates of the a and b constants are summarised in Table 2.7.

The log-log linear correlations were not significant for either single- or multi-locus estimates in *L. gmelinii*, even through population differentiation was shown to be significant (Table 2.7). For most of the estimates, b is larger than zero; implying no existence of a pattern of isolation by distance in *L. gmelinii* (Fig.2.3a).

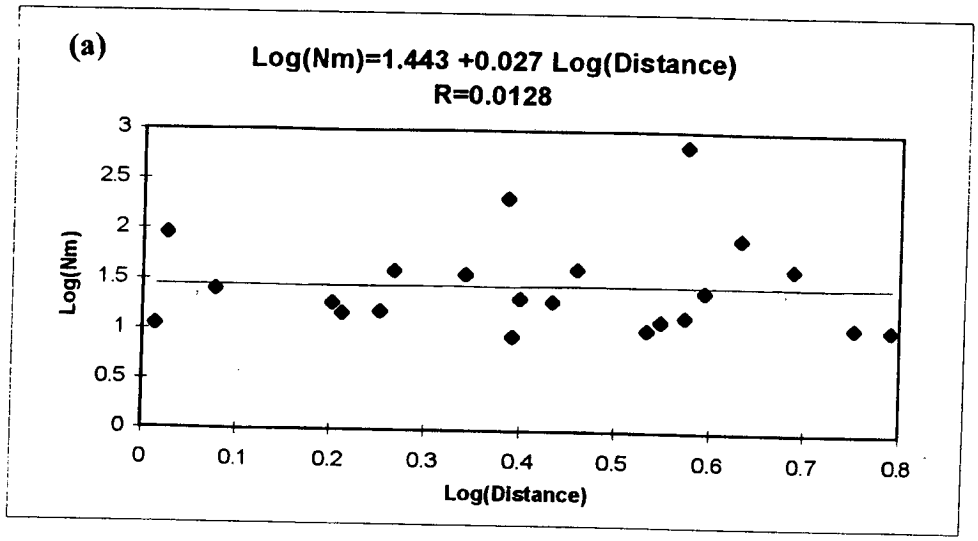
In *L. olgensis*, however, for most of the estimates, b is negative, and the linear correlations between Log(Nm) and Log(Distance) were shown to be significant in the single loci 6PGD (r = - 0.605), PGM (r = - 0.604), and in the multilocus analysis as well (r = - 0.637). The deterministic coefficients (r^2) were still small, less than 40%, therefore it may be concluded that a weak pattern of isolation by distance exists between natural populations of *L. olgensis* (Fig.2.3b).

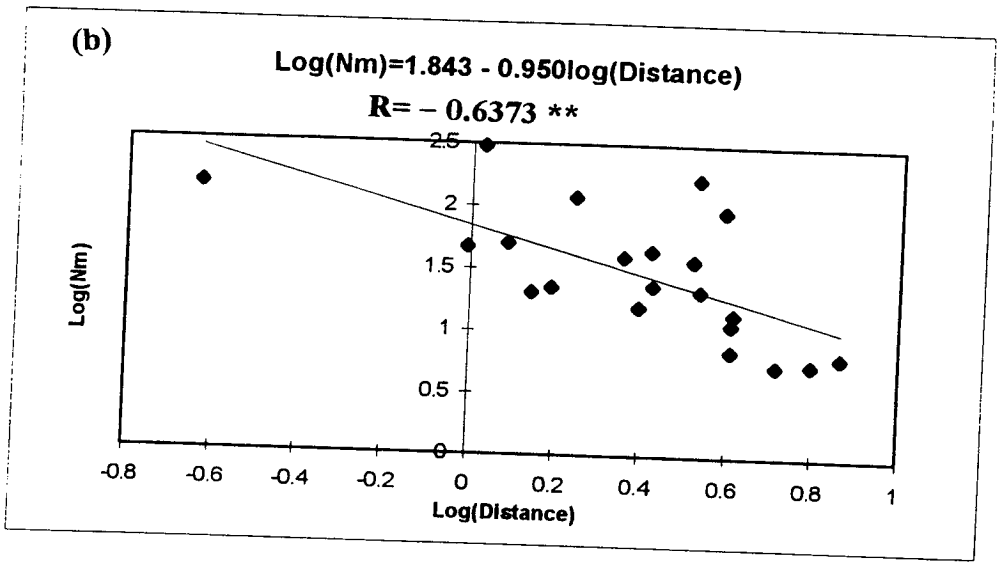
The above result was further proven using Mantel's exact test of heterogeneity in space (Mantel, 1967). The probability for the observed sample to take place under the null hypothesis, i.e. the hypothesis that the Fst values between populations are independent of geographic distances, is P= 0.792 in *L. gmelinii* and P=0.033 (<5%) in *L. olgensis*; showing significant relationship in *L. olgensis* but not in *L. gmelinii*.

Table 2.7 Estimates of parameters, a and b in Log(Nm)=a+bLog(Distance) and the correlation coefficient, r, between Log(Nm) and Log(Distance) **: P<1%

	L. gmelinii			*L. olgensis*		
Locus	a	b	r	a	b	r
PGI	1.496	-0.625	-0.249	1.796	-0.605	-0.511
MDH-I	1.537	0.023	0.011	1.219	-0.585	-0.494
6PGD	1.908	0.430	0.136	1.681	-1.347	-0.605**
AAT-I	1.394	-0.033	-0.020	1.309	0.262	0.154
AAT-II	1.375	0.419	0.135	1.625	-0.276	-0.204
PGM	1.026	0.404	0.243	2.169	-1.385	-0.604**
SDH	1.560	0.503	0.297	—	—	—
Over Loci	1.443	0.027	0.013	1.843	-0.950	-0.637**

Fig.2.3 $Log_{10}(Nm)$ plotted against Log_{10}(Distance) for (a) *L. gmelinii* and (b) *L. olgensis*. A significant pattern of isolation by distance was detected in *L. olgensis*, but not in *L. gmelinii*.





Since only two populations of *L. principis-rupprechtii* were analysed, detection of the isolation by distance was not conducted.

2.5.8 Genetic relationship among three taxa

Two types of Nei's genetic distances between any pair of populations investigated were calculated and summarised in Table 2.8. If the mutation rates vary with different loci, Nei's distance using geometric mean is used. If the mutation rates are assumed to be the same between loci, distance using arithemetric mean over loci is used, which is slightly larger than the former. It can be seen from Table2.8 that the genetic distances between populations, or between *Larix* taxa, are very small.

According to the distance using arithmetric mean, distances within each taxa are 0.00256±0.00183 for *L. gmelinii*, 0.0020 for *L. principis-rupprechtii*, and 0.00216±0.00164 for *L. olgensis*. However, distances between taxa are slightly larger than those within, 0.01435 ± 0.00449 between *L. gmelinii* and *L. principis-rupprechtii*, 0.00752 ±0.00406 between *L. gmelinii* and *L. olgensis*, 0.00886 ± 0.00133 between *L. olgensis* and *L. principis-rupprechtii*.

According to the distance using geometric mean, distances within each taxa are 0.0027±0.00197 for *L. gmelinii*, 0.0020 for *L. principis-rupprechtii*, and 0.00162±0.00134 for *L. olgensis*. However, distances between taxa are slightly larger than those within, 0.0109 ± 0.00352 between *L. gmelinii* and *L. principis-rupprechtii*, 0.00547 ±0.00338 between *L. gmelinii* and *L. olgensis*, 0.00697 ± 0.00094 between *L. olgensis* and *L. principis-rupprechtii*.

A UPGMA dendrogram showing relationships between fifteen populations of the three larch taxa was devised using eight loci and Nei's genetic distance (arithmetric mean; Fig 2.4). Two populations, the Seed orchard and Dongfanghong, were not included in the analysis due to a lack of SDH data.

The three larch taxa could be grouped into three distinct clusters using these eight allozyme markers, even though the distance between them is quite small. It is important to note that *L. gmelinii* appears more closely related to *L. olgensis* than *L. principis-rupprechtii*.

Table 2.8 Nei's genetic distance among all 17 populations investigated. Top diagonal distances are calculated using geometric mean over loci. Bottom diagonal distances are calculated using arithmetic mean over loci (Nei, 1972).

	L. gmelinii								*L. principis-rupprechtii*		*L. olgensis*						
	Jiagedaqi	Huzhong	Tahe	Xilinjie	Hanjiayuan	Uyilin	Kalunshan	ZhongYaozhan	Fengning	Hunyuan	Beidaoshan	Beihe	Xiaobeihu	Dahailin	Dongfanghong*	Changbei	Seed orchard*
Jiagedaqi		0.0045	0.0033	0.0058	0.0058	0.0068	0.0063	0.0020	0.0170	0.0175	0.0127	0.0129	0.0122	0.0131	0.0124	0.0121	0.0115
Huzhong	0.0041		0.0022	0.0004	0.0003	0.0007	0.0007	0.0024	0.0111	0.0094	0.0031	0.0035	0.0029	0.0035	0.0037	0.0034	0.0053
Tahe	0.0025	0.0021		0.0028	0.0030	0.0042	0.0030	0.0044	0.0163	0.0143	0.0079	0.0088	0.0076	0.0084	0.0094	0.0081	0.0105
Xilinjie	0.0052	0.0005	0.0026		0.0004	0.0008	0.0005	0.0035	0.0107	0.0080	0.0036	0.0037	0.0029	0.0035	0.0047	0.0042	0.0064
Hanjiayuan	0.0053	0.0004	0.0030	0.0005		0.0006	0.0008	0.0033	0.0114	0.0094	0.0025	0.0028	0.0022	0.0033	0.0037	0.0032	0.0064
Uyilin	0.0067	0.0008	0.0040	0.0007	0.0006		0.0008	0.0027	0.0076	0.0065	0.0017	0.0016	0.0013	0.0023	0.0017	0.0021	0.0034
Kalunshan	0.0069	0.0014	0.0038	0.0008	0.0013	0.0010		0.0036	0.0093	0.0074	0.0034	0.0033	0.0026	0.0025	0.0034	0.0039	0.0060
Zhongyaozhan	0.0020	0.0019	0.0030	0.0025	0.0025	0.0024	0.0033		0.0088	0.0097	0.0064	0.0060	0.0062	0.0071	0.0051	0.0059	0.0047
Fengning	0.0220	0.0137	0.0187	0.0117	0.0131	0.0090	0.0104	0.0126		0.0020	0.0088	0.0064	0.0076	0.0082	0.0067	0.0077	0.0054
Hunyuan	0.0245	0.0145	0.0193	0.0115	0.0135	0.0099	0.0107	0.0146	0.0020		0.0076	0.0057	0.0061	0.0067	0.0068	0.0069	0.0070
Beidaoshan	0.0160	0.0050	0.0109	0.0052	0.0040	0.0032	0.0055	0.0085	0.0100	0.0103		0.0002	0.0004	0.0020	0.0007	0.0003	0.0037
Beihe	0.0163	0.0055	0.0117	0.0052	0.0043	0.0029	0.0050	0.0083	0.0072	0.0078	0.0004		0.0003	0.0017	0.0004	0.0005	0.0031
Xiaobeihu	0.0149	0.0047	0.0103	0.0042	0.0035	0.0023	0.0040	0.0077	0.0079	0.0079	0.0007	0.0004		0.0013	0.0011	0.0008	0.0042
Dahailin	0.0167	0.0059	0.0120	0.0056	0.0055	0.0039	0.0038	0.0095	0.0104	0.0108	0.0037	0.0031	0.0023		0.0016	0.0023	0.0045
Dongfanghong	0.0161	0.0053	0.0118	0.0056	0.0049	0.0027	0.0041	0.0076	0.0072	0.0086	0.0009	0.0004	0.0012	0.0018		0.0006	0.0020
Changbei	0.0166	0.0059	0.0120	0.0062	0.0052	0.0039	0.0066	0.0089	0.0087	0.0094	0.0004	0.0006	0.0012	0.0043	0.0007		0.0025
Seed orchard	0.0149	0.0063	0.0115	0.0069	0.0073	0.0043	0.0067	0.0076	0.0074	0.0105	0.0041	0.0037	0.0047	0.0053	0.0026	0.0029	

*: The distances between the marked population and other populations were calculated using 7 loci excluding SDH.

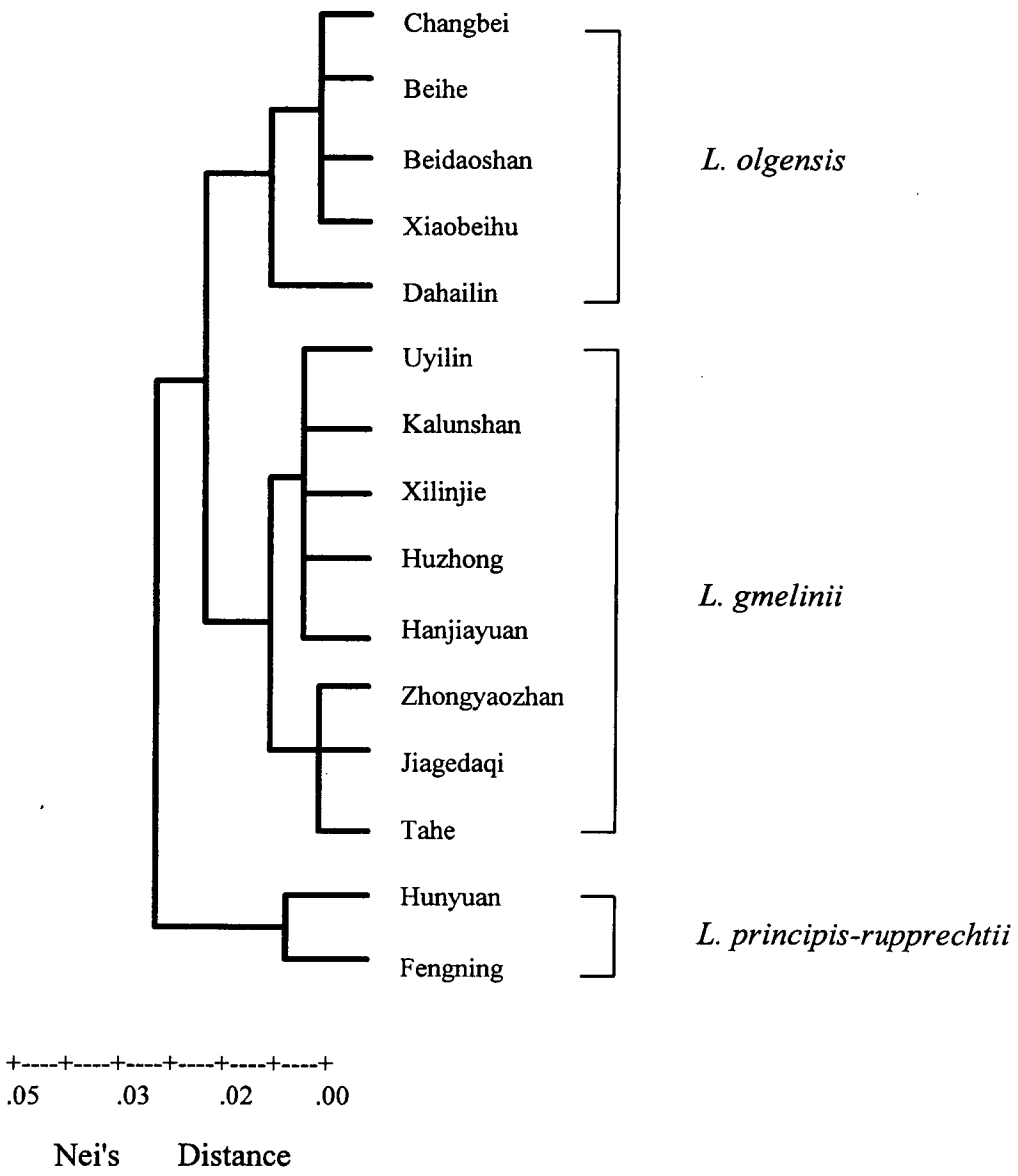


Fig. 2.4 A UPGMA dendrogram calculated according to Nei's (1972) genetic distance measure using Biosys 1.0 (Swofford and Selander, 1981). Only those populations for which data from all eight polymorphic loci are available are included.

2.6 Discussion

2.6.1 Allozyme markers

Six enzymes with a total of eight polymorphic loci were expressed in at least one population of the three larch taxa in this study. These loci were independent of one another according to linkage disequilibrium tests, implying that they maybe located on different chromosomes, or linked, but no linkage disequilibrium were observed. However, four pairs of genes have already been found to be significantly linked in *Larix laricina*: AAT and SOD, SOD and ACO, 6PGD and PGI, and AAT and PGI (Cheliak and Pitel, 1984b).

Allele frequencies were quite variable among populations and species (Table 2.2), with PGI, MDH-I, AAT-I, AAT-II and PGM being polymorphic in all species. 6PGD was monomorphic in *L. principis-rupprechtii*, with low levels of polymorphism in *L. gmelinii*, but it was highly polymorphic in *L. olgensis*. AAT-III was polymorphic in *L. gmelinii*, but monomorphic in the other two species. SDH, however, was polymorphic in both *L. gmelinii* and *L. olgensis*, but not in the *L. principis-rupprechtii*. These differences may be used for distinguishing the three species, but this will be difficult in practice since a species-specific marker was not found and hence large number of seeds would be required for analysis.

2.6.2 Population structure and genetic conservation

One common phenomenon in woody species is that most genetic variation occurs within populations and only a small proportion between populations in plant species (Hamrick, 1994). Similar results have been found in this study of the three *Larix* taxa, and most genetic variation is maintained within populations. However, population differentiation, less than 2%, is smaller than that found in some other conifers. For example, 5% of total genetic variation was found between populations of *L. laricina* (Cheliak, *et al.*, 1988). A comparable level of population differentiation to the three larch taxa in this study, was found in natural populations of *L. laricina* from northern Ontario (Liu and Knowles, 1991), about 2% of total genetic variation occurring between populations, and from New Brunswick (Ying and Morgenstern, 1991), about 3.8% of total genetic variation occurring between populations.

One possible explanation for the low level of population differentiation is the high amount of gene flow that prevents large population differentiation caused by genetic drift. The

postulated mean number of migrants, Nm, is 20.58 in *L. gmelinii*, 12.9 in *L. olgensis*, and 27.5 in *L. principis-rupprechtii* according to Nm = (1/Fst-1)/4 obtained in the island model (Wright, 1951). These values are much greater than 1.0 and, hence, lead to a low level of population differentiation (<< 20%). Furthermore, no private allele was detected in populations of *L. gmelinii*; indicating that migration takes place extensively (Barton and Slaktin, 1986). However, private alleles Sdh^3 and Sdh^4 were found in Dahailin of *L. olgensis*, with mean number of migrants (Nm) being 0.87 (<1) estimated using the formula introduced by Barton and Slaktin (1986). This indirectly indicates that migration in *L. olgensis* was not as extensive as that in *L. gmelinii*, resulting in a relatively higher level of population differentiation compared with *L.gmelinii*.

Environmental factors, such as climate, are also important in influencing current population structure revealed in this study. Although the markers used are considered to be selectively neutral, the indirect effect through 'hitch-hiking', caused by linkage between selectively unneutral and neutral genes, may influence population structure in terms of selectively neutral markers. This effect is difficult to test in this study. However, the following analysis for the limited distribution of the three larch taxa may likely indicate that this effect is small, or is comparable among populations. *L. principis-rupprechti* is restricted mainly to semi-arid areas in south temperate zone, while *L. olgensis* grows in wet areas in mid temperature zone where the climate is influenced by the Japanese sea. *L. gmelinii* mainly occurs in the north temperature zone where the climate is influenced by the continental climate from north of the region (Fig. 2.1). Therefore, if there are effects of the hitch-hiking, these effects are likely to be among the three taxa rather than within these three larch taxa. The result will be that current population genetic structure among populations within taxa assessed by allozymes in this study is dominated by migration and genetic drift.

Although there is significant population differentiation in *L. gmelinii*, it is lacking in geographic pattern, implying random distribution of genetic variation in space. However, this is not the case in populations of *L. olgensis* where there is weak geographical pattern of the genetic distribution, implying non-random distribution caused by isolation by distance. The possible reason for this weak pattern may be due to mountains barriers which block extensive gene flow.

Results of provenance trials conducted by Ma *et al.* (1992) showed that no significant

relationship was observed between growth traits and geographical latitudes in *L. gmelinii* and *L. principis-rupprechtii*, even though there were significant differences among provenances. However, a weak relationship was found between growth traits and geographical latitudes in *L. olgensis*, showing that the performance of the populations from the northern region of the distribution of *L. olgensis*, such as Xiaobeihu, grew better than those from the southern region (Ma, *et al*, 1992; Yang, *et al*, 1991). Their results are similar to those obtained in this study, i.e. existence of weak pattern between genetic differentiation and geographic distance in *L. olgensis* but not in other two larch taxa. This concordance is likely to be the result of an accumulation of mutations at allozyme loci and morphological divergence through time.

However, distribution of genetic variation between and within populations may be quite different for growth traits and allozyme markers. For instance, in the 22 population of white spruce (Furnier, *et al.* 1991), the height growth exhibits strong difference among populations, with 48.0% and 54.1% (broad sense heritability) of total genetic variation at age 9 and 19. However, using 6 polymorphic allozyme markers, only 3.8% of total genetic variation are due to between populations (Furnier, *et al.*, 1991). A similar case to white spruce takes place in two larch taxa of this study. Results of provenance trials of *L. gmelinii* indicated that more than 70% of total genetic variation (broad sense heritability) occurred among provenances in terms of height growth at age of 8 years (Yang, *et al*, 1990a). More than 65% of total variation occurred among provenances of *L. olgensis* in terms of height growth at age of 10 years old (Yang *et al*, 1991). Distribution of genetic variation among provenances of *L. principis-rupprechtii* was not reported. The results obtained using allozymes indicated that less than 2% of variation occurs among populations in the three larch taxa (Chapter 2), indicating disconcordance between growth traits and some allozyme markers, in terms of distribution of genetic variation within and between populations.

Hamrick *et al.* (1991) pointed out that the distribution of genetic variation within and between populations provides the prerequired information for establishment of effective and efficient conservation practices. The current distribution of genetic variation in the three larch taxa indicates that the conservation of the three taxa should focus on within populations rather than between, because few samples of populations can be used to represent the entire distribution of the larch taxa without losing much genetic variation, which was also proposed for *L. laricina* by Ying and Morgenstern (1991). However, such

conclusion clearly hold only for conservation of selectively neutral variation, such as the allozyme variation measured here.

2.6.3 Genetic relationship among three taxa

Larix olgensis and *L. principis-rupprechtii* are considered by Ostenfeld and Larsen (1930) to be two varieties of *L. gmelinii*, and by Wang (1992), Zheng (1983) and others as two species in their own right. Results obtained by Zhang *et al* (1985), using chromosomes characters, and by Tang *et al* (1995), using cpDNA RFLP markers, imply a close genetic relationship among the three larch taxa. Zhang (1985) further inferred that the evolutionary trend among the three larch taxa was from *L. principis-rupprechtii* to *L. olgensis* and *L. gmelinii.* Elucidation of evolutionary relationships among the three larch taxa may greatly contribute to our knowledge of the classification of the Chinese *Larix* taxa.

Generally, the dendrogram derived in this study using allozyme markers is in support of the results obtained by Tang *et al.* (1995), i.e. a close genetic relationship among the three taxa. The Nei's genetic distance, observed by allozyme markers, is very small and ranges from 0.005 to 0.015 (Fig.2.4). The important new result is that *L. gmelinii* is more closely related to *L. olgensis* than to *L. principis-rupprechtii.* This is consistence with results obtained by Zhang (1985) who found that both *L. gmelinii* and *L. olgensis* have Stebbin's type 2B chromosomes but *L. principis-rupprechtii* has not. As was mentioned in the introduction of this chapter, Crawford (1983) analysed that two possible explanations were likely responsible for this: One is that divergence between populations or between subspecies occurred within recent history, and hence accumulation of mutations was not large enough to produce a distant relationship. The second is that a high amount of gene flow has occurred among them. These reasons could be useful for explaining the results obtained in this study.

The morphological traits used to distinguish the three larch taxa (Appendix I) are not reliable, due to the phenotypic plasticity caused by interaction between environmental factors and genotypes. Hence, the capability is limited for using morphological traits to elucidate evolution history. However, use of selectively neutral markers, such as allozymes, can resolve this problem to some extent, since environmental modification is avoided if effects of the hitch-hiking effect and recombination are ignored. Thus population genetic diversity and polymorphism are controlled by migration, drift, and mutation.

If the time scale is so short that the effect of mutation is small enough to be ignored, the genetic variation in source populations is maintained at a higher level than that in their derived populations established via colonisation. This is because of founder effects and bottlenecks that may be involved in the colonising process, resulting in loss of genetic variation. Thus, mean number of alleles per loci or the percentage of polymorphism may decline because of stochastic migration that may result in loss of rare allele with larger probability than common allele in the process of colonisation. Tajima (1990) showed in theory that marginal population maintained lower DNA polymorphism than central populations if the migration rate in the marginal populations is lower than that of the central populations, indicating that low rate of migration may reduce genetic variation in recipient population. Thus the above analysis may provide a clue to infer the formation of populations. However, it should be remembered that the above inference will hold under the influences of migration and drift only.

Therefore, under this hypothesis, a more detailed insight into the evolutionary history of Chinese larch taxa may be inferred from integrating the results of the mean number of alleles per locus, the fraction of polymorphic loci and the geographic distribution of the three taxa. The allozyme markers used in this study have been shown to be in linkage equilibrium, and are considered to be selectively neutral. Thus, if the divergence among the three taxa occurred within recent history, the effect of mutation can be ignored. Only migration and drift dominate the genetic variation of these allozymes markers. LePage and Basinger (1995) argued that three distinct patterns of displacement were in support of current distribution of larches. One of these patterns of displacement, thought by LePage and Basinger (1990), is from " northeastern Russia along the eastern coast of Asia and into central China " Assume that the three larch taxa are the sample species intially. Experiments have shown that the mean number of alleles per locus varies from 2.20 to 2.11 and 1.93, and percentage polymorphism from 68%, to 66% and 49%, for *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii*, respectively; supporting the direction of evolution trend from *L. gmelinii* to *L. olgensis* and *L. principis-rupprechtii*. It is more likely that *L. principis-rupprechtii* is mainly influenced by *L. olgensis* via migration and colonisation rather than by *L. gmelinii*; due to shorter geographical distance to *L. olgensis* than to *L. gmelinii* (Fig.2.1). This results in further reduction in mean number of alleles per locus from 2.11 to 1.93 due to effects of bottleneck and founder effects involving in the process of colonisation.

A critical assumption for the above inference is that the divergence of the three larch taxa occurred within recent history so that influence of mutation is ignored. Moreover, the above inference contrasts with the evolutionary trend proposed by Zhang *et al.* (1985) who stated that evolutionary direction is from *L. principis-rupprechtii* to *L. olgensis* and *L. gmelinii*. Therefore, these points highlight the need to further elucidate the evolutionary relationship between the three Chinese larch taxa using DNA markers (see Chapter 4).

2.7 Summary

The genetic variation within seventeen populations, eight in *L. gmelinii*, seven in *L. olgensis* and two in *L. principis-rupprechtii*, representing three Chinese larch taxa was quantified and studied using eight polymorphic allozyme loci. Seven allozyme loci were found to be free from association in each taxa. Most of populations were found to be in Hardy-Weinberg equilibrium for these allozymes, with the exceptions of a few populations of *L. olgensis* and *L. gmelinii* due to heterozygosity deficiency, probably caused by self-fertilisation (see next Chapter).

Less then 2% of total allozyme variation occurred between populations investigated in each of the three taxa. Analyses of spatial patterns indicated that the distribution of the allozyme variation did not correlate with geographic pattern in *L. gmelinii*, but a weak correlation was found in *L. olgensis*. This may have been caused by isolation by distance, resulting in that a higher level of population differentiation being present in *L. gmelinii*, compared with *L. gmelinii*.

The values of Nei's genetic distances within each larch taxa were very small, about 0.002, while distances between taxa were larger than within, about 0.01, five times the distance within taxa. A dendrogram was reconstructed to elucidate evolutionary relationship between the three larch taxa, using these eight polymorphic enzyme loci and this indicated that *L. gmelinii* was more closely related to *L. olgensis* than to *L. principis-rupprechtii*.

CHAPTER 3

Use of Allozymes to Investigate the Mating System of Taxa within the *L. gmelinii* Complex

3.1. Introduction

3.1.1 Significance of mating system

Measurement of the mating system provides critical information relevant for many aspects of forest tree genetic improvement programmes, such as genetic conservation and predicting genetic gain. It describes how plants transmit their genetic material from the current generation to the next generation.

Knowledge of mating systems may help us to infer the genotypic composition of the population. For example, if a plant species possesses predominant selfing, the proportion of heterozygotes will quickly decline. Populations of the species will become genetically homozygous, and differentiation among populations will be large. On the other hand, if a plant species possesses predominant outcrossing, the genotypic composition within populations will tend to Hardy-Weinberg equilibrium. Low vigour of selfed seeds in an outcrossing population is expected due to inbreeding depression. Mating systems also influence the extent of linkage disequilibrium within a population and hence multilocus genotypic structure. Inbreeding populations are characterised by high levels of linkage disequilibrium, while in outcrossing populations linkage disequilibrium is only expected between tightly linked loci (Epperson and Allard, 1984).

Knowledge of the mating system may help us to judge the quality of seeds from a seed orchard. For example, a pattern of predominant outcrossing may indicate a higher proportion of good seeds as long as pollen contamination is controlled. Predominant outcrossing also predicts the quality of open-pollinated seeds from natural stands for tree breeding. This is because inbreeding depression is avoided when selfing rates are low.

Mating system provides information required for establishing seed orchard or other stands for producing seeds for plantation. Strict isolation control is required for predominant outcrossing species so as to obtain good quality of seeds. Serious pollen contamination in some established seed orchards has been reported. For example, Adams *et al.* (1997) investigated a mature Douglas-fir seed orchard in western Oregon. Pollen contamination rate from the natural stand was estimated as 0.255 ±0.096 in 1980, 0.519±0.276 in 1985, 0.389±0.237 in 1987 and 1989, and 0.259 ±0.193 in 1990. Shaw and Allard (1982) scored

the mating system of eight natural populations and one seed orchard of Douglas-fir, and showed that outcrossing rate was about 0.90 for natural stands and a seed orchard. The high proportion of outcrossing means that there is a large potential for pollen contamination.

Mating system can also be used for predicting population structure to some extent. For any mother tree, the pollen pool which it samples consists of three parts (Fig. 3.1). It may be derived from itself or genetic relatives (S); from genetic unrelated neighbouring individuals (N) within populations; and from other populations via pollen flow (M). The proportion for outcrossing is composed of parts of M and N. However, only the proportion M can be used for predicting population structure. Predominant outcrossing species possess a larger proportion of M+N than selfing plant species (Fig.3.1). There is therefore a higher probability of receiving migrated pollen and this will result in a smaller degree of population differentiation.

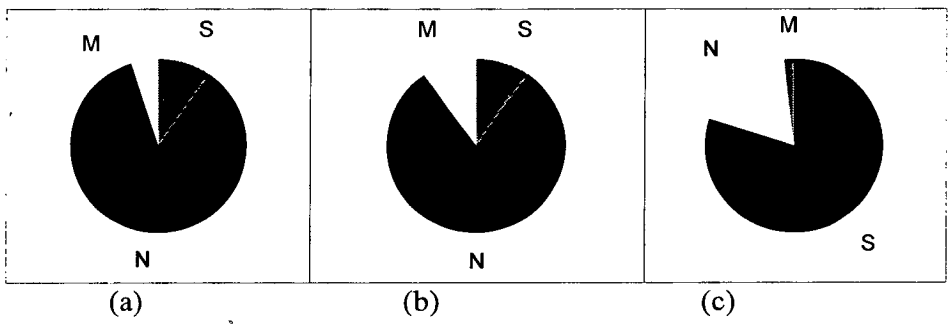


Fig.3.1. A pie chart representing the relationship between migration rate (M) and mating system. (a) and (b) are predominant outcrossing species with different migration rates. (c) is a predominantly selfing species.

In summary, quantification of mating systems may provide important information for plant breeding programme.

3.1.2 Scoring of mating system

Traditionally, mating systems were scored using polymorphic morphological marker genes. Shaw *et al.* (1981) pointed out that as early as 1916, Jones used the progeny of homozygous dwarf tomato to estimate the outcrossing rate using progenies where any outcrossing was

with pollen of normal plants. At this time the scoring of mating systems was confined to experimental populations.

This method was then developed by Fyfe and Bailey (1951) who proposed a statistical method for estimating outcrossing rate t and p, the frequency of dominant alleles in the population investigated. However, the method introduced by them is based on an assumption of equilibrium between selfing and outcrossing. Outcrossing rate can be estimated as $1-\hat{f}/1+\hat{f}$ where $\hat{f}$ is Wright's inbreeding coefficient, by the fitting observed proportions of genotypes descended from a known maternal genotype. A similar method was developed by Nei and Syakudo (1958). Shaw *et al.* (1981) pointed out that these studies extended the "... use of the method from experimental populations to natural or other populations in which allelic frequencies are unknown." However, one limitation is that maternal genotypes must be known to make use of these methods.

Two components must be required in estimating outcrossing rate: *allele frequency in pollen pool* and *maternal genotypes*. Use of codominant marker genes has widened the scoring of mating system to many different species. A single locus mixed model, wherein a proportion of zygotes are derived from selfing and the remaining zygotes are derived from outcrossing, was presented by Brown and Allard (1970) and Clegg *et al.* (1978) for estimating mating system. The genotype of the maternal parent can be inferred using a maximum likelihood method (Clegg, *et al*, 1978; Ritland, 1983). According to Shaw *et al.* (1981), a series of assumptions are involved in the single locus mixed mating model:(i) allele frequencies in pollen pool are distributed uniformly over the population of maternal plants; (ii) the probability of outcrossing is independent of maternal genotype; (iii) selection does not intervene in the time between pollination and the time that seeds or seedlings are sampled. Violations of these assumptions have been extensively reviewed by Mitton (1992). Deviation from these assumptions cause biases in the estimates of outcrossing rates. For example, heterogeneity of gene frequency between subpopulations may severely bias downward estimation of outcrossing rate (Ennos and Clegg, 1982). As a consequence, more accurate multilocus estimates of outcrossing rate have been developed.

The use of multiple loci can obviously increase the potential for scoring outcrossed individuals (Shaw *et al.*, 1981). This is because if one locus does not detect a outcross with certainty, then another locus might detect it. Moreover, Shaw *et al.* (1981) pointed out that

multilocus estimates have lower variances than the mean of all single locus estimates. Several multilocus mixed models have been developed (Brown, *et al.*, 1978; Shaw, *et al.*, 1981; Ritland and Jain, 1981). The model introduced by Brown *et al.* (1978) is only suitable for predominantly inbred species. This is because their multilocus model makes use of only multilocus homozygous genotypes using a maximum likelihood method.

Shaw *et al.* (1981) showed that a multilocus outcrossing rate can be estimated by $n / N(1-\alpha)$ where n is the number of discernible outcrossed progeny in a sample of size N, and α is the probability that an outcross will not be discerned. The expected α depends on maternal genotypic frequencies, allele frequencies in the pollen pool and the number of loci used. A larger number of loci may reduce the α value. The α can be estimated after single locus analysis to estimate maternal genotypic frequencies and allele frequencies in the pollen pool, using the model introduced by Clegg *et al.* (1978). Thus, in order to use the model introduced by Shaw *et al.* (1981) to estimate multilocus outcrossing rate, a two step procedure is used. First, a single locus analysis is conduced using the method of Clegg *et al.* (1978), and then a multilocus analysis is performed. This method is simple, but does not make efficient use of multilocus data to estimate outcrossing rate.

The deficiencies of the method proposed by Shaw *et al.* (1981) were overcome by Ritland and Jain (1981). Ritland and Jain (1981) proposed a model that can simultaneously employ multilocus data. However, the basic procedure is similar. First, maternal genotypic frequencies (if the maternal genotypes are not available) are inferred. Then, the estimates of the maternal genotypic frequencies are used for further estimating outcrossing rates. The second step is performed by an iteration equation so as to obtain maximum multilocus log likelihood (Ritland and Jain, 1981; Ritland, 1983). To date, this model has been widely used in many plant species.

Additional information about inbreeding other than selfing can be inferred by comparing estimates of single locus and multilocus outcrossing rate. Single locus estimates of outcrossing are expected to be biased downward by any biparental inbreeding in addition to selfing. Since multilocus analysis possesses higher resolution for detecting hybridised individuals than single locus analysis (Shaw, *et al.*, 1981), it is expected that mean single locus estimate will be lower than that estimated using multiple loci when mating among relatives occurs.

3.1.3 Use of allozymes to study plant mating system

Despite the fact that many PCR based markers have recently been developed, allozyme markers remain some of the best for estimating plant mating system. This is because of their codominant inheritance and cheap technical cost. They have been widely employed to score mating system, especially in conifers that contain haploid megagametophyte and diploid embryos (Adams, 1983; Mitton, 1992). In *Larix*, for example, the mating system has been scored using allozyme markers in two larch species. Natural populations of *L. laricina* (Knowles, *et al*, 1987) and seed orchard of *L. decidua* (Gömöry and Paule, 1992) have been investigated. Both studies showed a lower outcrossing rate in *Larix* than for other reported conifers, such as logepole pine (*Pinus contorta* ssp. *latifolia*; Epperson and Allard, 1984) and Jeffrey pine (*Pinus jeffreyi*; Furnier and Adams, 1986).

Studies using DNA markers to score plant mating system have already been conducted. For example, Dow and Ashely (1996) explored the use of microsatellite markers to analyse seed dispersal and parentage of 62 adult bur oak, *Quercus macrocarpa*, and 100 saplings in a stand established in northern Illinois, USA. Their study demonstrated the utility of microsatellite analysis for studying mating systems. Since RFLP markers are also codominant, this marker can be used for scoring mating systems. For example, Milgroom *et al.* (1993) used six unlinked RFLP loci to examine the outcrossing rate in a natural population of chestnut blight fungus, *Cryphonectria* (*Endothia*) *parasitica*. They found that the multilocus estimate of the outcrossing rate was 0.74, including a mixed mating system for this fungus.

Few studies have compared DNA based and allozyme based estimates of outcrossing rates. However, one study on a forest pathogen, *Cronartium ribicola* (Gitzendanner *et al.*, 1996), indicate that comparable estimates of mating system parameters were obtained using these different markers. Both RFLP and allozymes indicated that genotype frequencies were in Hardy-Weinberg equilibrium and that the rust shows random mating (Gitzendanner *et al.*,1996).

These data confirm the legitimacy of using allozyme markers for mating system estimates, and they will be used in this study of the three larch taxa.

3.1.4 Aims of the chapter

It can be seen from the above that the mating system provides important information in practice. However, mating system has not been scored in the three larch taxa. The central questions asked are: What type of mating system do the three larch taxa belong to ? Do the three taxa differ in outcrossing rate? Is there variation within taxa? If there is a deviation from random mating, how does this occur?

According to the results obtained in other conifers, most of which possess a mixed type of mating system, one expectation is that there will be little difference between the three Chinese larch taxa and other conifers because they are wind-pollinated conifers.

Thus, the aims of this chapter are: (i) to measure and compare mating systems of the three larch taxa among populations and taxa; (ii) to compare the results with those of other studies on conifers; and (iii) to estimate amount of true selfing and extent of biparental inbreeding within populations.

3.2. Materials

Open pollinated seeds were collected by family from natural populations of the three larch taxa. One population was available in *L. gmelinii*, six in *L. olgensis* and two in *L. principis-rupprechtii*. Locations sampled and the number of half-sib families involved are listed in Table 3.1. The number of half-sib families scored ranged from 9 to 33 (mean 22). At least six seeds were available from each half sib family, enabling maternal genotypes to be inferred with reasonable certainty.

3.3. Methodology

3.3.1 Seed preparation and enzyme extraction

Preparation of material and enzyme extraction were the same as in Chapter 2.

3.3.2 Buffer systems and starch gel preparation

Buffer systems and starch gel preparation were the same as in Chapter 2

Table 3.1. Location and sample size of the 9 *Larix* populations investigated for mating system using allozyme analysis

Species Population	Latitude(N)	Longitude(E)	Half-sibs	Seeds	Seeds/Halfsibs
L. gmelinii					
Jiagedaqi	50°24'	124°07'	21	126	6
L. principis-rupprechtii					
Fengning	41°12'	116°32'	20	121	> 6
Hunyuan	39°32'	113°41'	9	75	> 6
L. olgensis					
Beidaoshan	44°00'	131°07'	20	120	6
Beihe	42°25'	128°08'	29	174	6
Xiaobeihu	44°01'	128°50'	33	198	6
Dahailin	44°28'	129°48'	25	288	> 6
Dongfanghong	42°39'	128°06'	20	139	> 6
Changbei	41°26'	128°11'	21	209	> 6

3.3.3 Electrophoresis

Conduct of electrophoresis was the same as in Chapter 2.

3.3.4 Scoring of gels

Scoring of gels was the same as in Chapter 2.

3.4 Data Analysis

Putative genotypes for each embryo and gametophyte were recored for each family to build up family genotype data within each population. Data from the gametophyte scored was used to infer the maternal genotype of each halfsib family. Using these data, quantitative analysis of the mating system within each larch taxa was analysed based on Ritland's mixed model (Ritland, 1983). Ritland's MLT programme (Ritland, 1990) was employed to calculate both single locus outcrossing rate (t_s) and multilocus (t_m), using maximum

likelihood methods. Standard errors of these estimates were calculated by conducting 200 bootstraps between families (half-sibs).

3.5. Results

3.5.1 Primary screening of polymorphic markers

The screening of polymorphic markers provided a total of six polymorphic enzyme systems, i.e. PGI, MDH, 6PGD, AAT, PGM and SDH. Detailed banding patterns and genetic explanation were shown in Fig.2.2 in Chapter 2.

3.5.2 Mating system

The mating system was analysed in each of the three taxa and results are summarised in Table 3.2.

L. gmelinii

Only one populations was available. Three polymorphic loci were used for estimating outcrossing rates, i.e. AAT-II, PGM, and SDH. Nearly complete outcrossing was demonstrated in a *L. gmelinii* population Jiagedaqi because both the estimated single locus (mean $\overline{t_s}$ =0.977) and multilocus outcrossing rates (t_m=0.986) were close to 1.00. The difference between t_m and $\overline{t_s}$ is not significant (Table 3.2).

L. principis-rupprechtii

Three polymorphic loci (PGI, AAT-I and PGM) and one locus (AAT-II) showing a small amount of polymorphism were used for analysis of mating system in population Hunyuan. The single locus outcrossing rate of PGI was t_s = 0.470, which is significantly smaller than 1.0. The mean single locus value was $\overline{t_s}$ = 0.730, which is significantly smaller than 1.0, while the multilocus outcrosssing rate was not significantly less than 1.0. The difference between $\overline{t_s}$ and t_m was very small and not significantly different from zero, indicating that

selfing was responsible for the low outcrossing rate rather than inbreeding (Shaw, *et al.*, 1981).

Two polymorphic loci (PGI and PGM) and one slightly polymorphic locus (MDH-I) were used for analysis of mating system of population Fengning (Table 3.2). Both single locus outcrossing rates and multilocus outcrossing rate were not significantly different from 1.0, indicating that random mating occurred in this population. The difference between $\overline{t_s}$ and t_m was not significantly different from zero, indicating that selfing was responsible for the low outcrossing rate rather than inbreeding.

In *L. principis-rupprechtii*, different outcrossing rates were found but these were not significantly different in the two populations due to large errors. Thus, based on multilocus estimates, it can concluded that the mating systems in these two populations Hunyuan and Fengning were predominately outcrossing.

L. olgensis

Polymorphic loci with variable degrees of diversity were used for analysis in six populations of *L. olgensis* (Table 3.2). Both mean single locus and multilocus outcrossing rates were significantly smaller than 1.00 in Xiaobeihu, Daihailin and Changbei. Moreover, differences between t_m and $\overline{t_s}$ was very small in both Xiaobeihu and Changbei, indicating the presence of significant levels of selfing rather than inbreeding in these populations. However, this was not the case in population Daihailin, wherein $\overline{t_s}$ was significant smaller than t_m. This suggests the existence of inbreeding in addition to selfing (Shaw, *et al.*, 1981).

Other populations of *L. olgensis*, including Beidaoshan, Beihe and Dongfanghong, were shown to be completely outcrossing both for mean single locus and multilocus estimates, with t_m ranging from 0.847 ± 0.427 to 1.203 ± 0.371.

Outcrossing rates in *L. olgensis* were variable between populations, but difference was not significant due to large errors, using the Student t test

It may be concluded that the three larch taxa generally exhibit predominant outcrossing. However, outcrossing rates are variable between populations and allozyme markers, with some populations expressing significant levels of selfing. Biparental inbreeding was implicated in only one population. In general variation in outcrossing rate within taxa is as large or larger than variation in outcrossing rate between taxa.

3.6 Discussion

The three larch taxa generally exhibit predominantly outcrossing, with some differences existing between populations within taxa. In conifers, abortion of selfed individuals (embryos) is responsible for ensuring a high rate of outcrossed seeds (Sorensen, 1982). Nearly complete outcrossing was detected in populations Jiagedaqi of *L. gmelinii*, in Dongfanghong, Beidaoshan, and Beihe of *L. olgensis*, in Fengning and Hunyuan of *L. principis-rupprechtii*. However, outcrossing rates were low in Xiaobeihu, Dahailin, and Changbei of *L. olgensis*, with multi-locus outcrossing rates ranging from 0.684 to 0.705.

These results are similar to results already reported in studies of other conifers (Table 3.3), exhibiting mixed types of mating system. In *Larix*, lower outcrossing rates with significant selfing were observed as well. For example, in some natural populations of *L. laricina*, multilocus estimates were $t_m = 0.729 \sim 0.908$ (Knowles, *et al*, 1987) and in a seed orchard of *L. decidua* Mill., $t_m = 0.852 \pm 0.07$, (Gömöry and Paule, 1992). These results are comparable with results found in some populations of *L. olgensis*.

The variable outcrossing rates among populations may be due to a variety of causes (Mitton, 1992). The presence of more than one embryo in a single ovule (polyembryony) in *Larix* and other conifers was reported (Sorensen, 1982). Many factors, such as a variety of seasons and low pollen production, contribute to low availability of outcross pollen. Rate of outcrossing is expected to increase with stand density. In low density stands, the probability for allele composition in the pollen pool of each mother tree to be diluted by pollen of neighbours is smaller than that in high density stands, thus leading to lower outcrossing rate. One typical example is ponderosa pine at a forest-grassland ecotone in eastern Colorado (Farris and Mitton, 1984). Average of outcrossing rate was estimated to be 0.96 in a stand with normal density, using allozymes (Mitton, *et al.*, 1981), while a low rate of outcrossing was observed, 0.80, in a low density stand.

Table 3.2 Estimates of single-locus (t_s) and multilocus (t_m) outcrossing rate for populations of the three larch species (standard errors in parentheses). The symbol '—' in the table stands for the locus being monomorphic. *: P<5%;**:P<1%

	L. gmelinii	*L.principis-rupprechtii*		*L. olgensis*					
	Jiagedaqi	Fengning	Hunyuan	Beidaoshan	Beihe	Xiaobeihu	Dahailin	Dongfanghong	Changbei
PGI	—	0.680(0.341)	**0.470(0.191)****	—	0.715(0.512)	0.856(0.518)	1.999(0.000)	1.999(0.002)	1.999(0.233)
MDH-I	—	1.999(0.586)	—	1.999(0.066)	—	—	1.999(0.690)	1.999(0.584)	**0.196(0.129)****
6PGD	—	—	—	—	1.999(0.479)	—	—	1.999(0.625)	1.999(0.006)
AAT-I	—	—	0.745(0.425)	1.999(0.334)	1.999(0.761)	—	—	—	—
AAT-II	0.733(0.570)	—	1.999(0.497)	—	1.999(0.382)	**0.602(0.120)****	**0.362(0.191)****	—	**0.728(0.104)****
PGM	1.029(0.111)	0.940(0.154)	1.057(0.217)	1.158(0.380)	0.950(0.549)	0.954(0.076)	0.667(0.302)	0.841(0.257)	0.913(0.051)
SDH	0.972(0.135)	—	—	—	—	—	**0.601(0.140)****	—	—
$\overline{t_s}$	0.977(0.077)	0.873(0.097)	**0.733(0.127)***	0.847(0.427)	0.971(0.358)	**0.720(0.066)****	**0.655(0.101)****	0.987(0.367)	**0.704(0.116)***
t_m	0.986(0.081)	0.930(0.149)	0.792(0.169)	0.847(0.427)	1.203(0.371)	**0.704(0.070)****	**0.684(0.107)****	0.996(0.368)	**0.705(0.116)***
$\overline{t_s} - t_m$	-0.009(0.028)	-0.056(0.092)	-0.059(0.082)	0.000(0.031)	-0.231(0.357)	0.016(0.020)	**-0.029(0.016)****	-0.008(0.047)	0.000(0.025)

Table 3.3 Outcrossing rate of some conifer species in the family Pinaceae detected by allozyme markers

Species & References	Outcrossing rate (t_s or t_m)†	Mating system (Pure selfing, outcrossing,or mixed)
Balsam fir (*Abies balsamea*) Neale & Adams, 1985	0.78~ 0.99, mean 0.89	mixed
White spruce (*Picea glauca*) King, *et al*, 1984	0.75~0.99, mean 0.90	mixed
Jack pine (*Pinus banksiana*) Cheliak, *et al*, 1985	0.88 ±0.047	mixed
Logepole pine (*Pinus contorta* ssp. *latifolia*) Epperson & Allard, 1984	1.03±0.04	mixed
Jeffrey pine(*Pinus jeffreyi*) Furnier & Adams, 1986	0.881~0.971	mixed
Ponderosa pine (*Pinus ponderosa*) Farris & Mitton, 1984	0.81±0.054(low density) 0.96±0.046(high density)	mixed
Douglas fir (*Pseudotsuga menziesii*) Shaw & Allard, 1982	0.90	mixed
Tamarack (***Larix laricina***) Knowles, *et al*, 1987	0.316~0.897(t_s) 0.729 (low density)(t_m) 0.908 (normal density)(t_m)	mixed
European larch (***Larix decidua***) Gomory & Paule, 1992	0.64~1.0 (t_s) 0.852 ±0.007 (t_m)	mixed
L. gmelinii (This study)	0.986±0.077 (t_m)	mixed
L. principis-rupprechtii (This study)	0.792~0.930 (t_m)	mixed
L. olgensis (This study)	0.684 ~1.203(t_m)	mixed

† : t_s : outcrossing rate estimated by single locus; t_m : outcrossing rate estimated by multilocus methods.

Larix is a wind-pollinated conifer. The wind conditions at the time of pollen release will influence dispersal distance of pollen grains, which in turn may influence pollen pool composition of a mother tree and hence rate of outcrossing. Wind reduction may lead to low outcrossing rates.

Variation of environmental factors may also contribute to variation of the outcrossing rate between populations. In addition to the effect of genotype, the floral phenology is usually associated with environmental factors, such as the accumulation of temperature. The difference in the timing of pollen release and female receptivity may influence the success of fertilisation. For example, Erickson and Adams (1989, 1990) found large clone to clone variation in the timing of pollen release and female receptivity in a seed orchard of Douglas-fir.

There are many other reasons for this population to population variation in outcrossing rate. However, this study suffers from lack of data on these important factors, which limit further analysis. This information is required in the future study of mating system.

Population Dahailin of *L. olgensis* might present different characters to other populations of *L. olgensis* because biparental inbreeding in addition to selfing is apparently occuring. One possible explanation is that family clumping may contribute to inbreeding in this population (Ennos and Clegg, 1982). Thus, it is necessary to use other information, such as field morphological traits, to confirm this hypothesis.

It is a common phenomenon that the single locus outcrossing rate is quite variable between different loci within populations and species (Brown and Allard, 1970). This is also the case in the present study (Table 3.2). Genetic markers with low degree of polymorphism lead to inexact estimates of outcrossing rate due to larger standard errors (Table 3.2). However, this shortcoming is avoided using multilocus outcrossing estimates. This is because multilocus analysis provides higher resolution for detecting hybridised individuals than single locus analysis (Shaw, *et al.*, 1981). Moreover, multilocus analysis is also more robust to model violations than single locus analysis (Shaw, *et al.*, 1981; Ritland and Jain, 1981). Thus, multilocus estimates of outcrossing rate are more reliable.

Several implications can be obtained from the above results. First, the variable mean single locus outcrossing rates between populations of *L. olgensis*, or between populations of *L. principis-rupprechtii*, indicate that population genetic composition is different, detected by single locus analysis.

Second, where selfing causes inbreeding depression and leads to a lack of heterozygotes in populations. This may be associated with a lower proportion of full seed and a lower germination rate (Gömöry and Paule, 1992). A high level of inbreeding or selfing in some particular populations, such as Dahailin of *L. olgensis* and Hunyuan of *L. principis-rupprechtii*, indicates that attention should be paid when seeds are collected from natural populations or even from seed orchard for afforestation. The different extents of outcrossing rates between populations of *L. olgensis* implies that this attention is necessary when using open-pollinated seeds for afforestation, or using seeds produced by a seed orchard.

Third, the general predominant outcrossing of the three *Larix* taxa may indicate that gene flow may be extensive between populations if there is no obvious barrier. There is no evidence to show that taxa are very different from one another with respect to breeding system. Thus, strict isolation control is necessary when seed orchard or seed stands are established.

Finally, the finding that the mating system is predominantly outcrossing in the three larch taxa means that field progeny tests which assume that families are composed of outcrossed sibs, will give reasonable estimates of genetic parameters of quantitative traits, such as heritability.

3.7 Summary

Mating systems of the three Chinese *Larix* taxa were scored using allozyme markers. Population Jiagedaqi of *L. gmelinii* exhibited nearly total outcrossing (t_m= 0.986 ± 0.081). Two populations of *L. principis-rupprechtii*, Fengning and Hunyuan, shared no significant difference from random outcrossing (t_m=0.847± 0.427 ~ 0.792 ± 0.169). However, mating system was variable among populations of *L. olgensis*. Two populations of *L. olgensis*, Xiaobeihu and Changbei, exhibited significant levels of selfing (t_m ≈ 0.705). One population, Dahailin, exhibited inbreeding in addition to selfing, with t_m being 0.684 ±0.107. However, the other three populations of *L. olgensis*, Beihe, Beidaoshan and Dongfanghong, exhibited predominantly outcrossing, t_m = 0.847 ± 0.427 ~1.203 ± 0.371. These results are comparable to those found in other conifers including *L. laricina* and *L. decidua*.

CHAPTER 4

Use of chloroplast DNA to infer genetic relationships between the three *Larix* taxa *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii*

4.1 Introduction

The presence of DNA in chloroplasts (cpDNA) in *Chlamydomonas moewusii* was first demonstrated by Ris and Plaut (1962), using electron microscopic and cytochemical methods. Many studies have since been carried out concerning its structure and genetics (Downie and Palmer, 1992; Clegg and Zurawski, 1992). The circular DNA molecule contains coding regions for some specific ribosomal and transfer RNA genes, and many of the protein-coding genes necessary for the function of photosynthetic apparatus, such as *psb*C, *psb*D-H and *atp*A genes (Grierson, *et al.*, 1988; Ohyama, *et al.*, 1986). Many of the characteristics of the molecule, described below, reveal its suitability for use in studies of plant evolution, either in macroevolution or microevolution.

In this introduction a brief description will be first given of the structure and genetics of the chloroplast genome, and the features that influence the choice of this marker for elucidation of genetic relationships of the three larch taxa in this study. The aim of the present study is then presented.

4.1.1 Use of cpDNA in studies of plant evolution

4.1.1.1 Sequence organisation

Chloroplast DNA is a double-stranded circular molecule, ranging in size from 120 to 217 kbp (Downie and Palmer, 1992). It usually contains two duplicate regions in reverse orientation, the inverted repeat (IR), which are separated by large single-copy (LSC) and small single-copy (SSC) regions (Fig. 4.1). To date, the cpDNA of four species, *Nicotiana tabacum*, *Oryza sativa*, *Marchantia polymorpha*, and *Pinus thunbergii* (Sugiura, 1993 submitted to Genbank databases), has been completely sequenced and the genes mapped, providing a reference for other studies. For example, the three non-coding regions in black pine cpDNA from *trn*F(GAA) to *trn*T(UGU) are located within SSC (Fig.4.1; Fig 4.2) rather than within LSC, as is the case in *Nicotiana tabacum*. Moreover, the IR length in *Pinus thunbergii*, 495bp, is considerably shorter than in *Nicotiana tabacum*, 25.3kbp, and *O. sativa*, 20.8kbp.

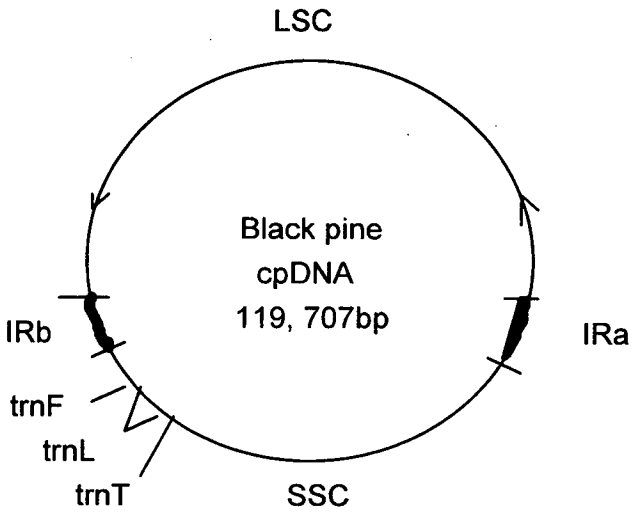
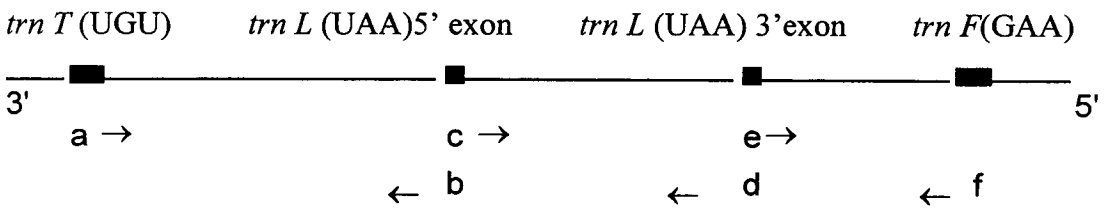


Fig.4.1. Structure of the chloroplast DNA molecule of black pine (*Pinus thunbergii*), drawn according to its complete sequence submitted by Sugiura(1993) to Genbank databases (http: //www2. ncbi. nlm. nih.gov/cgi-bin/genbank): IR= Invert repeat (495bp); LSC= large single copy region (65696bp); SSC = small single copy region (54011bp); and the three genes, *trn*F, *trn*L and *trn*T located in the SSC region, which will be investigated in the present study.



Primers:

a: 5′-CATTACAAATGCGATGCTCT-3′
b: 5′-TCTACCGATTTCGCCATATC-3′
c: 5′-CGAAATCGGTAGACGCTACG-3′
d: 5′-GGGGATAGAGGGACTTGAAC-3′
e: 5′-GGTTCAAGTCCCTCTATCCC-3′
f: 5′-ATTTGAACTGGTGACACGAG-3′

Fig. 4.2 Positions, directions and sequences of three pairs of universal primers. The three fragments (ab, cd and ef) are of similar size in some species, about 500bp (Taberlet *et al.*, 1991).

In some conifers, such as some species in the family Pinaceae: i.e. Douglas fir (*Pseudotsuga menziesii* (Mirb.) Franco), radiata pine (*Pinus radiata* D.Don; Strauss, *et al.*, 1988), and *Larix laricina* (Raubeson and Jansen, 1992), the large (20-25kbp) inverted repeat is missing. Many other species in other families also lack the invert repeat, e.g. Taxaceae, Taxodiaceae, Podocarpaceae, Cupressaceae and Araucariaceae (Raubeson and Jansen, 1992). However, later sequencing has shown that in the case of black pine a short invert repeat is present (Genbank database; Sugiura,1993). Because the inverted repeat is very short, and does not contain the coding region of 23S ribosomal RNA gene in which there are two highly conserved recognition sites for the restriction endonucleases *Kpn*I, located 800 base pairs (bp) apart, the strategy used by Raubeson and Jansen (1992) cannot be used for detecting the existence of inverted repeats. Thus, there is some risk in saying that there is no inverted repeat in *Larix* and other conifers.

4.1.1.2 Features of cpDNA suitable for analysis of macroevolution

Macroevolution here means the evolution at the level of species or at a higher level. Many features of the cpDNA molecule are considered to make it a valuable tool in phylogenetic and taxonomic analyses at the species level (Clegg and Zurawski, 1991; Palmer, *et al.* 1988). Based on the reviews by Gillies (1994; Ph.D. thesis) and Clegg and Zurawski (1991), these features are the following.

① A low mutation rate, which indicates that little intraspecific variation might be expected. Mutation in cpDNA is of two types, nucleotide substitutions (point mutation) and rearrangements. In the case of substitution, the mutation rate for structural genes is, on average, fivefold slower than for plant nuclear genes (Wolfe, *et al.*, 1987). Rearrangements, including inversions, insertions or deletions of genes and introns, and loss of one copy of the IR occur rarely in land plants. The processes that influence the formation of these rearrangements is not clear. Downie and Palmer (1991) argued that the occurrence of rearrangement may involve some major alteration of IR (its loss and expansion). Downie and Palmer (1991) also pointed out that once cpDNA rearrangements are found, these mutations should be considered to be more powerful characters than individual nucleotide substitutions and have the potential to resolve with confidence a particular branching point in a phylogeny. This can be easily understood because the rare event of the rearrangement may indicate its important role in phylogeny reconstruction. Thus the conservative nature of cpDNA in terms

of point mutation and rearrangements implies that it is a powerful indicator for macroevolution, and it has already been recommended for phylogeny reconstruction in plants, such as in some conifer species (Szmidt, 1991;Wang and Szmidt,1993).

② The predominantly uniparental inheritance of the molecule presents a clear record of historical events. The chloroplast genome is usually maternally inherited in angiosperms (Morgensen, 1996), but in most conifer species, such as *Larix* (Szmidt, 1990; Szmidt, *et al.*, 1987), it is paternally transmitted (Mogensen, 1996). Uniparent inheritance results in the absence of recombination during meiosis. Thus the effect of recombination on phylogeny is omitted.

③ The large amount of DNA present in the chloroplast makes its evaluation relatively easy. A large number of copies of the molecule, between 20 and 200, in each mature chloroplast facilitates its extraction, detection and analysis (Clegg and Zurawski, 1991).

④ As is stated before, the publication of the complete sequence of cpDNA from four species, provides a good opportunity for designing universal primers. For example, a set of universal cpDNA primers have been designed by Taberlet *et al.*(1991) and Dumolin-Lapegue *et al.*(1997). Use of these primers may facilitate phylogeny construction using DNA sequence data from some specific regions of the molecule.

It can be seen that these features, especially the slow rate of evolution and uniparental inheritance, make cpDNA a valuable tool in phylogeny studies at the species level. For example, the *rbc*L gene, which encodes the large subunit of ribulose-1,5-bisphosphate carboxylase / oxygenase (RUBISCO), has been widely sequenced from numerous plant taxa, such as the phylogeny construction in some monocots and dicots plant species (Wolfe, *et al.*1989; see review by Clegg and Zurawki, 1991, and Clegg, 1993) and also in three closely related genera *Hordeum*, *Triticum* and *Aegilops* (Gielly and Taberlet, 1994).

These features outlined above also suggest that cpDNA markers may be useful tools to elucidate the genetic relationship among the three taxa of Chinese larch species, because we do not know when the divergence among the three taxa occur. If there is a difference detected among them, the divergence probably occurred a long time ago because the mutation rate is so low in cpDNA, judged from other plant species (Clegg and Zurawki,

1991). If there is no difference in terms of cpDNA sequence data, divergence among these three larch taxa may have occurred recently. Since the point mutation rate may be smaller compared with those of nuclear DNA (Wolfe, *et al.*, 1987), use of cpDNA sequence data will provide inferences for a longer time scale than use of nuclear DNA sequence data.

However, we should also consider other features of cpDNA markers in choosing appropriate regions for elucidating the genetic relationship among the three Chinese larch taxa. These features are described below, and may help us to decide which specific regions of cpDNA are likely to be useful to detect the divergence among the three Chinese larch taxa.

4.1.1.3 Use of cpDNA in microevolution

The other feature of cpDNA is the variation within species, the intraspecific variation. Intraspecific variation in cpDNA has recently been found to be greater than was originally thought. After reviewing many studies using cpDNA in plant biosystematics, Harris and Ingram (1991) concluded that far from being rare, intraspecific cpDNA variation is relative common. Therefore, cpDNA markers have also been extensively employed to survey population genetic structure.

For example, Mason-Gamer *et al.*(1995) investigated variation within and between populations of *Coreopis grandiflora* (Asteraceae), using RFLP analysis of cpDNA. They detected sufficient cpDNA variation for analysis of intraspecific and inter-population genetic structure to provide evidence for gene flow occurring among populations. Similarly, Furnier and Stine (1995) found RFLP variation among populations of white spruce (*Picea glauca*) and estimated population differentiation to be $F_{st} = 0.147$. In four species of European oaks (Petit, *et al.*, 1993a), highly significant genetic variation has been found using cpDNA ($G_{st} = 0.895$). While Powell *et al.* (1995) have explored the use of PCR-based SSR (simple sequence repeats) analysis of cpDNA (cpSSR) in population structure of *Pinus leucodermis* Ant.. They surveyed 305 individuals from seven populations of this species, and revealed the presence of four variants with intrapopulational diversities ranging from 0.000 to 0.629, while population differentiation based on cpSSR was estimated to be $G_{st} = 0.22$. This diversity was not detected using RFLP analysis of cpDNA in the same populations, and Powell *et al.* (1995) anticipate that analysis of SSR loci within the

chloroplast genome should provide a highly informative assay for examining the genetic structure of plant populations.

It seems therefore that the sequence of cpDNA exhibits two distinct features. On the one hand it exhibits quite conservative feature due to very low mutation rate compared with other genomes; on the other hand, it also exhibits intraspecific variation in some species. These two properties may vary, depending upon the species studied, and its population history, and also upon the specific DNA sequence region analysed such as coding and non-coding regions. These double properties tell us that great care should be taken when cpDNA markers are chosen to investigate the three larch taxa. The reasons are analysed below.

Earlier workers have already investigated variation in larch species using cpDNA markers. Tang *et al.*(1995), using RFLP analysis, have already studied the genetic relationship between the three Chinese larch taxa, and found Nei's genetic distance to be zero between any pair of the three larch taxa. These results indicated that the divergence among the three taxa may have occurred recently. Based on these results, the assumption of low levels of intraspecific cpDNA variation for the three Chinese larch taxa is still used and will be tested in the present study.

The method of RFLP analysis is to use restriction enzymes to digest the genome and hence produces a population of fragments with discrete size. Then variation between individuals is probed by hybridising these fragments in each individual with a series of given radio-labelled fragments (Fig.4.3). The length variation detected by probes is caused by single base change and insertion/ deletion. RFLP analysis is based upon shared restriction fragments or not. However, several points need to be made concerning RFLP analysis.

First, hybridisation can occur if nucleotide sequence in fragments are not completely homologous to a given probe. For example, if the major part of a nucleotide sequence is homologous to a probe, this may still lead to successful DNA-DNA hybridisation. Thus the sequence difference between fragments with the same size on a gel that can hybridise with the same probe cannot be detected.

Second, the number of sites analysed by RFLP method is quite limited. For example, a 6bp recognition enzyme would be expected to cut once every 4096bp if all four kinds of bases are randomly arranged in a genome.

Third, one problem for using RFLP to reconstruct phylogeny is that it is difficult to make a clear judgement on the difference between probability of a restriction site gain and a restriction site loss. How to assign different weights to RFLP character data needs to be explored.

Fourth, use of restriction data for phylogenetic reconstruction depends on the method for scoring data. Bremer (1991) compared four different scoring methods for phylogeny reconstruction, yielding various results. Different hypotheses are involved in the scoring methods. For example, the fragment measurements are made directly on the autoradiograms from the different hybridisation. Fragments of the same size and position are scored as characters (present or absent) for phylogeny analysis. This method was called by Bremer (1990) the fragment occurrence analysis (FOA). Although the risk that two fragments of equal length come from different parts of the same genome can be ignored, the fragments used for phylogeny analysis are not evolutionarily independent. Thus, this introduces bias into the assessment of genetic relationships because the same length mutation may be scored on more than one occasion, and be given more weight than is appropriate. However, the above problem is avoided by the method of site occurrence analysis (SOA) in which only site mutation is scored rather length mutation as characters for phylogeny analysis (Bremer, 1991). But, the length mutation is omitted using SOA for phylogeny analysis. Bremer (1991) proposed that " the choice of method is dependent on a trade-off between accuracy and resource (time)."

One other point is that hybridisation to probes will not detect small length difference in the DNA due to the resolving power of the method being low. Small changes in fragment length cannot be detected.

Finally, RFLP analysis fails to tell us detailed information in terms of specific regions although it can provide general information regarding specific restriction enzyme such as mapping. In a word, RFLP analysis provides limited detection regarding site mutation and does not provide a full comparison of DNA sequences.

In published work, the divergence among twelve larch taxa was detected by RFLP analysis using the FOA scoring method and Nei's genetic distance (Tang, *et al*, 1995). Nei's genetic distances among the three Chinese larch taxa were estimated to be zero, providing only a preliminary insight into the evolutionary relationship among them.

Therefore there is at least one advantage for studying sequence of some specific regions in cpDNA over the RFLP analysis in that it can give us a clear picture of the difference between the three taxa in terms of specific cpDNA regions. Furthermore, comparison of cpDNA sequences may directly avoid the drawbacks of RFLP analysis and may provide evidence in support of genetic relationship between the three larch taxa.

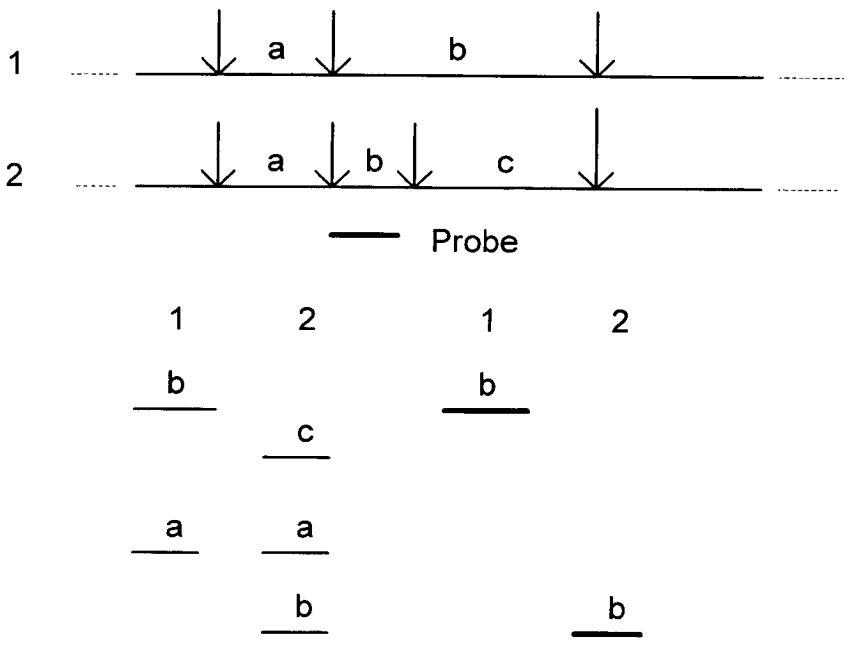


Fig.4.3 Schematic presentation of restriction, electrophoresis and Southern hybridisation. Top: Two individuals 1, 2. Arrows show cutting sites. Letters a, b and c indicate the fragment size. The homologous region to the probe is shown. Botton left: Fragments are separated on agarose gel by electrophoresis. Bottom right: Radio-labelled bands after Southern hybridisation.

Neutral theory predicts that the non-coding regions of DNA are likely to exhibit higher point mutation rates than the coding region (Kimura, 1968), because mutation in this region does not lead to a change in biological function, while the coding region is usually more conservative because of natural selection. Analysis of noncoding regions of cpDNA could be more informative in studies at lower taxonomic levels. Thus, markers for the non-coding region may reveal greater levels of variation than coding region markers, and therefore, may be more useful for elucidation of relationships between closely related species (Wolfe and Sharp, 1988; Gielly and Taberlet, 1994).

For example, one non-coding region in the *trn* L intron ($tRNA^{Leu}$ (UAA)-intron) of cpDNA has already been employed in analysis of population structure in *Silene alba*, a dioecious angiosperm (McCauley, 1994). Using PCR-RFLP analysis, McCauley (1994) found the variation in the *trn*L intron and estimated that Wright's F_{st} to be 0.67 over a 25 × 25 km portion of the species range. In a separate study, the *trn*L intron was used for population structure in *Quercus robur* and *Q. petraea* in Demark (JØhnk and Siegismund, 1997), showing significant population differentiation among the seventeen Danish population of *Q. robour* ($G_{st} = 0.6$). In reconstructing phylogenies of the three closely related genera *Hordeum*, *Triticum* and *Aegilops*, which cannot be resolved by *rbc*L gene but by *trn*L markers, Gielly and Taberlet (1994) argued that these three non-coding regions seem to be well suited for infering plant phylogenies between closely related taxa since (i) double-stranded cpDNA can be easily be amplified for a wide taxonomic range of plant species, and (ii) the size of these noncoding regions is small enough to allow us to get the whole sequence by using only the amplification primers.

Based on the above considerations, use of the non-coding regions of cpDNA is an appropriate experimental method for elucidating the genetic relationship among the three larch taxa. On the one hand, it provides a larger opportunity to find variation between them. On the other hand, it will probably provide an estimation of divergence over a long time scale because of a lower point mutation rate in cpDNA compared with the nuclear genome.

4.1.2. Aim of the present study

It is likely that the difference between the three *Larix* taxa may be evident in many traits such as in branch, seedling and cone traits (Appendix I) due to environmental modification.

The three taxa can also be distinguished using some allozymes (nuclear markers; Chapter 3), showing quite close relationship between them. However, results already obtained using RFLP analysis of cpDNA, show that there is no difference between the three taxa (Tang, *et al.*, 1995). These results motivate further study to elucidate the divergence of the three Chinese larch taxa by using cpDNA sequence data, particularly from the non-coding regions of the molecule.

The objective of this part of the study is to elucidate the genetic relationship among the three Chinese larch taxa using cpDNA markers. Three specific non-coding regions of cpDNA, from *trn* T (UGU) to *trn* F(GAA) (Taberlet, *et al.*, 1991), will be first investigated using PCR-RFLP analysis, and be further sequenced so as to find the genetic relationship between the three larch taxa. Since complete sequence of cpDNA of black pine, which is also from family Pinaceae, is available, black pine cpDNA is used as a reference for comparison with larch. The evolutionary relationship between larch and black pine are not studied here, but the extent of divergence between them at the cpDNA sequence level can be ascertained.

The practical significance of this analysis is twofold. First, the results may provide information concerning historical events related to the formation of the three taxa, and maybe be used for providing evidence on current population genetic structure. Second, the genetic relationship may provide useful information on how large the divergence is between taxa and hence on how to take advantage of these divergences. For example, if the genetic relation among them are quite close, there is little advantage in hybridisation of taxa. However, if there are large difference in genetic composition, heterosis is possible between the taxa.

4.2 Materials and Methods

4.2.1 Materials

Fresh buds were collected from *L. gmelinii* and *L. principis-rupprechtii* in the Royal Botanic Garden (RBG), Edinburgh, for DNA extraction. Needles were obtained from *L. olgensis* seedlings, germinated for about three months and then used for DNA extraction (Table 4.1). Two individuals were analysed from each of the three taxa.

Table 4.1 Sample size and locations of the three Chinese larch species

Species	Samples	Location
L. gmelinii	2	19795307, and 54.0089, China adj. E.U.S.S.R., RBG, Edinburgh.
L. olgensis	2	Changbei Shan Mountain, China (1994)
L. principis-rupprechtii	2	19731138 and 19793328, RBG, Edinburgh.

4.2.2 DNA extraction

The method used in this study for extracting DNA is given in Appendix III.1.

4.2.3 PCR-RFLP analysis

4.2.3.1 PCR principle

The polymerase chain reaction (PCR) is based upon cycling between three different steps a number of times. In the first step, genomic DNA is denatured by heating (Fig.4.5), resulting in production of single stranded DNA. A pair of primers for amplification of the specific region of interest is then annealed to the DNA, in the presence of buffer, enzyme and free nucleotides, at a reduced temperature. Then a new strand of DNA is synthesised between the primers by Taq DNA polymerase I, resulting in a double increase in concentration of the required DNA fragments. Many cycles of the same procedure (25-40) are repeated, resulting in an exponential increase in the concentration of the required fragment of DNA. In the initial cycle copies are made of the target sequence and, thereafter, copies are made from these copies. An agarose gel, 1.0 ~2.0%, was run to examine the PCR products.

4.2.3.2 Setting up the PCR reaction

See Appendix III.2.

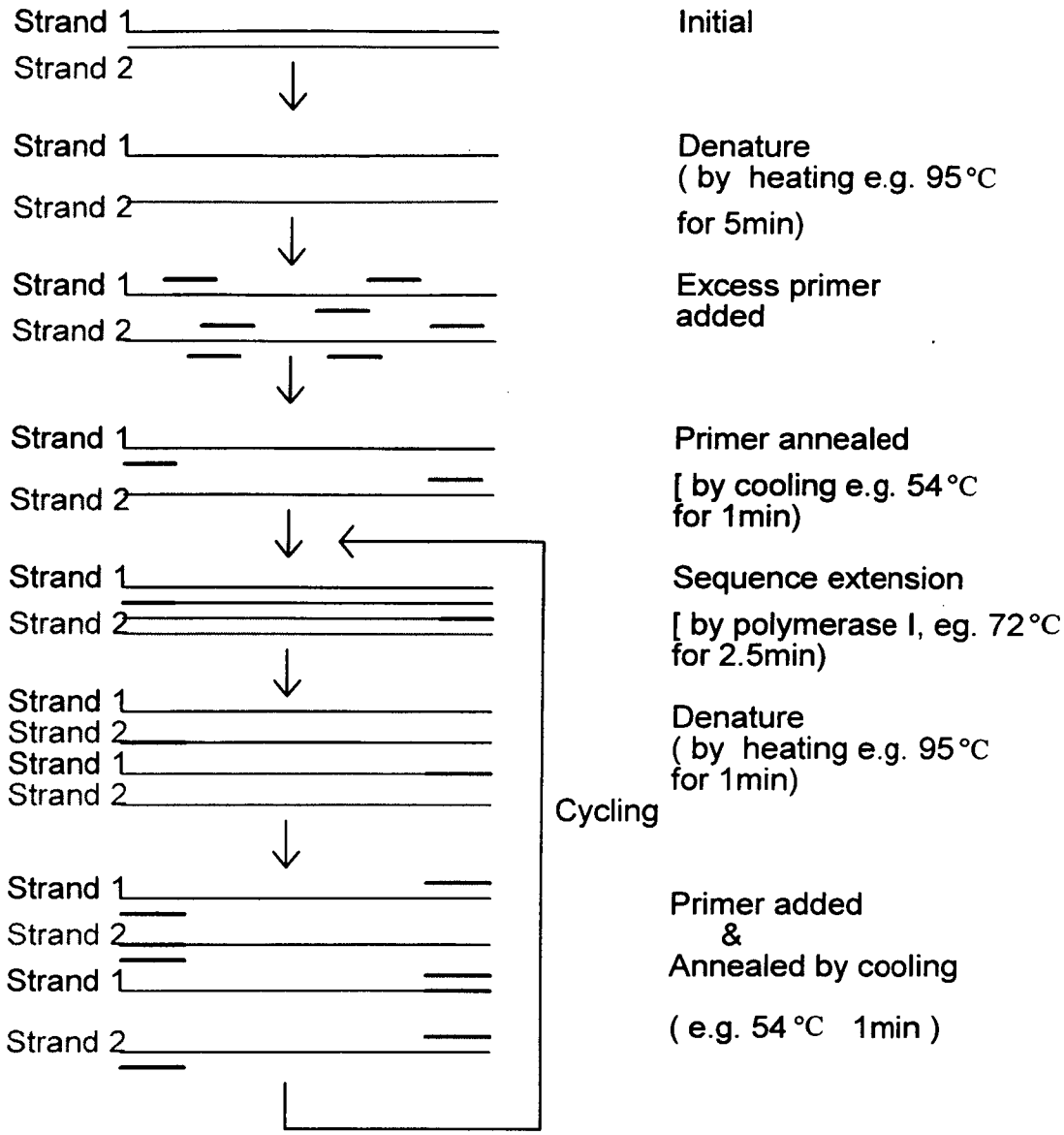


Fig 4.4 The basic Polymerase Chain Reaction (PCR) process for amplifying up specific regions of DNA flanked by primers.

4.2.3.3 Chloroplast DNA primers

Three pairs of universal chloroplast DNA primers for the amplification of three non-coding regions of *cp*DNA were used in this study (Fig.4.2; Taberlet, *et al*, 1991).

4.2.3.4 Digestion with restriction endonuclease

Restriction enzymes with 4bp recognition sites were employed to digest the amplified products, to detect differences between the three taxa. Eight four-base-cutter restriction enzymes were employed (see Appendix III.3). The recipes used for the digestion are listed in Appendix III.3.

4.2.3.5 Preparation of agarose gel and electrophoresis

See Appendix III.4.

4.3. DNA sequencing

The DNA sequencing method used in this study is the chain-termination method developed by Sanger *et al.* (1977). The DNA to be sequenced acts as a template for the enzymatic synthesis of new DNA, starting at a defined primer binding site. A mixture of both deoxy- and dideoxynucleotides is used in the reaction. There is a finite probability that a dideoxynucleotide (ddNTP) will be incorporated in place of the usual deoxynucleotide at each nucleotide position in the growing chain. Once a ddNTP is incorporated into the growing chain, elongation is terminated, resulting in lots of fragments of varying length. The nucleotide sequence can, therefore, be determined by running four separate reactions; each of which contains a single dideoxynucleotide (ddATP, ddCTP, ddGTP, ddTTP). The resulting fragments of varying length are then separated on a polyacrylamide gel, and the sequence is determined by correlating the order of the bands on the gel with the dideoxynucleotide used to generate each band. An alternate approach to that used in this study, is to attach a different nucleotide-specific label to each dideoxynucleotide. The reaction can then be run in one tube instead of four and analysed on a single gel lane. For detail, see QIAGEN sequencing guide (QIAGEN,1995).

The automated sequencing (fluorescent) method was employed in the present study. This method, unlike manual sequencing methods that generally use a radioactive label and visualise the banding pattern by autoradiography (QIAGEN,1995; Sambrook, *et al.*, 1989), uses a scanning laser to detect DNA fragments labeled with fluorescent dyes. The total sequencing procedure is composed of four steps: (i) purification of PCR products; (ii) cycle sequencing; (iii) purification of extension products, and (iv) electrophoresis, followed by data collection. See Apppendix III.5.

4.4 Results

4.4.1. PCR-RFLP

4.4.1.1 PCR amplification

Figure 4.5 shows the three fragments amplified using the three pairs of universal cpDNA primers **a** and **b**, **c** and **d**, and **e** and **f** (Fig. 4.2) with the three Chinese larch taxa. Results show that there is no detectable variation in band size for the three non-coding regions of cpDNA between the three Chinese larch taxa (Fig.4.5).

The sizes of the amplified bands were estimated according to non-linear regression of length on distance for the standard fragments (1kbp ladder marker), using a least-squares method (Weir, 1990). The approximate sizes of amplified fragments for the three larch species were: 475bp between *trn*T(UGU) and *trn*L(UAA), amplified by primer pairs **a** and **b**; 546 bp between *trn*L(UAA) 5' exon and *trn*L(UAA) 3' exon, amplified by primer pairs **c** and **d**; and 469 bp between *trn*L (UAA) 3' exon and *trn* F(GAA), amplified by primer pairs **e** and **f**. The total length between *trn*T and *trn*F was about 1490bp, which is similar to that of black pine cpDNA whose sequence was obtained from GenBank, and to other species such as *Pinus nigra* and *Robinia pseudacacia* (Taberlet, *et al.*,1991).

4.4.1.2 RFLP analysis

Eight 4bp cutter restriction endonucleases were used to digest the three larch taxa (Appendix III.3). Enzymes were chosen based on the pine cpDNA sequence information already obtained from the Genbank database. Digestion patterns indicated that there were no differences in the three non-coding regions between the three Chinese larch taxa. For

example, figures 4.6(a), (b) and (c) show some of the patterns produced by the different restriction enzymes. The bands sizes estimated from the gels are summarised in Table 4.2.

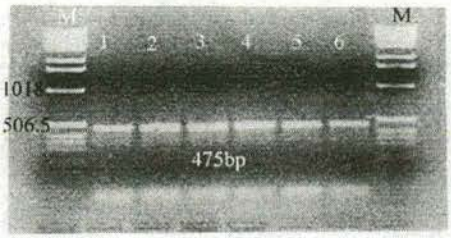
However, three restriction site differences were found in these regions that differentiated *Larix* cpDNA from that of black pine. One of these mutations was that found in the region between *trn*T and *trn*L(5'), which lacks a cutting site for *Tru* 9 I in pine, while a cutting site was evident in larch. A second mutation was detected in the region between *trn* L(5') and *trn*L(3'), which possessed a *Rsa* I cutting site in pine but not in larch. The third mutation was present in the fragment amplified between *trn*L (3') and *trn*F ; there was a *Tru* 9I cut site in pine but not in larch.

Table 4.2 Fragment length estimated according to its migration distance on the gel after digestion for each of the three Chinese larch taxa.

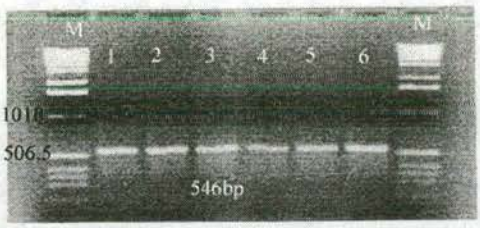
Fragments	*trnT--trnL*(5')(**ab** band)	*trnL*(5')-*trnL* (3')(**cd** band)	*trnL* (3')- *trnF*(**ef** band)
Size	475bp	546bp	469bp
Alu I	460bp	382,130bp	407bp
Cfo I	364bp	undigested	undigested
Hsp92 II	undigested	421bp,146bp	310bp
Mbo I	235,167bp	278bp	not resolved
Msp I	undigested	434,130bp	425bp
Rsa I	undigested	undigested*	366bp
Taq I	249,200bp	218, 98bp	210bp
Tru9 I	363,185bp*	357bp	undigested*

*: indicates that the fragment could / could not be digested by the enzyme in pine cpDNA, but could not/ could be digested by the same enzymes in larch cpDNA.

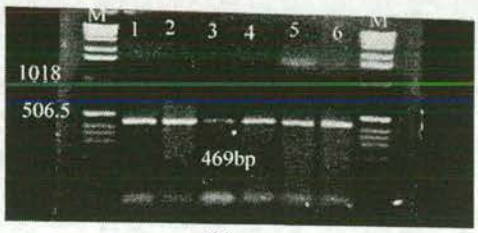
Fig. 4.5. PCR products obtained using three pairs of universal cpDNA primers for the three larch taxa in this study. Photographs (a), illustrates band amplification obtained by primer pair a and b, the region between *trn*T and *trn*L(5'); photograph (b) illustrates band amplification obtained by primer pair c and d, the region between *trn* L(5') and *trn*L(3'); photograph (c) illustrates band amplification obtained by primer pair e and f, the region between *trn*L (3') and *trn*F. In each figure, lanes 1 and 2 are *L. gmelinii*, 3 and 4 are *L. olgensis*, and 5 and 6 are *L. principis-rupprechtii*. The lane marked 'M ' is 1kb ladder marker.



(a)



(b)



(c)

Fig. 4.6a. PCR products amplified by universal primers a and b; the region between *trn*T and *trn*L(5'), restricted by endonuclease C*fo* I. Lanes 1 and 2 are *L. gmelinii* ; 3 and 4 are *L. olgensis*; 5 and 6 are *L. principis-rupprechtii*. Lanes 7-11 are other larch species which are not focused upon in this study. Lane CK, containing original PCR products and no restriction enzyme, is used for control. The 364bp band estimated according to its migration distance was resolved after digestion, but the expected presence of the small band (about 111bp) was not detected. The bottom band is thought not to be restriction products but primer artfacts which can be inferred from the presence of the same band in control lane. The lane marked 'M ' is 1kb ladder marker.

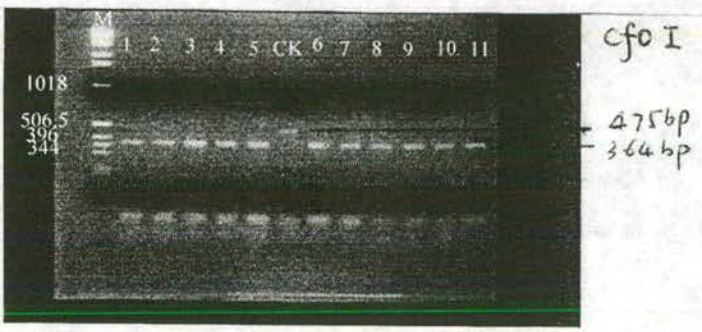


Fig. 4.6b. PCR products amplified by universal primers c and d: the region between *trn L*(5') and *trnL*(3'), digested by eight restriction endonuclease in *L. olgensis*. Lanes 1 to 7 represent the digestion by restriction endonuclease *Alu* I, *Cfo* I, *Msp* I, *Mbo* I, *Hsp* 92II, *Rsa* I, *Taq* I and *Tru* 9I, respectively. The lane marked 'CK' represents uncut PCR products for control. The lane marked 'M ' is 1kb ladder marker. Results of the remaining larch taxa in this study are the same as *L. olgensis*.

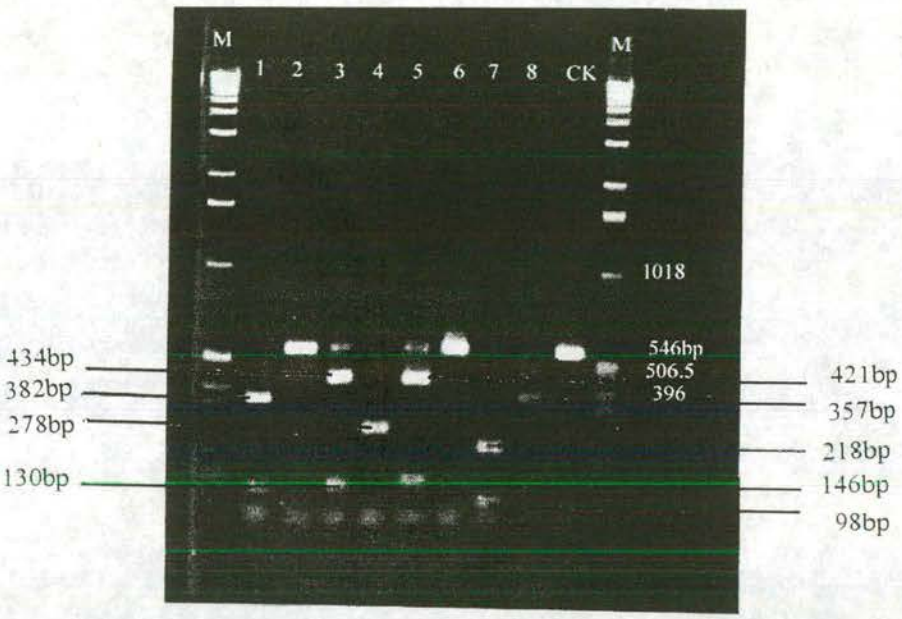
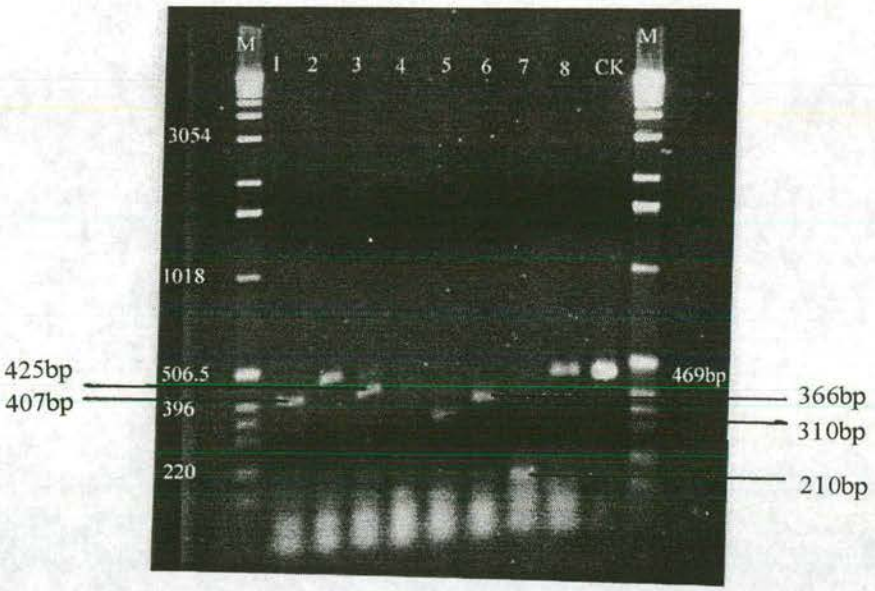


Fig. 4.6c. PCR products amplified by universal primers e and f; the region between *trn*L (3') and *trn*F, digested by eight restriction endonuclease in *L. olgensis*. Lanes 1 to 7 represent the digestion by restriction enzymes *Alu* I, *Cfo* I, *Msp* I, *Mbo* I, *Hsp* 92II, *Rsa* I, *Taq* I and *Tru* 9I, respectively. The lane marked 'CK' represents uncut PCR products for control. The lane marked 'M ' is 1kb ladder marker. Results of the remaining larch taxa are the same as *L. olgensis*.



The objective of the digestion analysis is roughly to try to find variation among the three larch taxa, caused by point mutation. No difference in length scored by restriction implies that the three larch taxa are quite related in terms of these three non-coding regions. However, only a small proportion of sites is investigated using this method. Thus, comparison of full sequences of these three non-coding regions is clearly required in order to find variation.

4.4.2 Sequencing

In order to further investigate any differences between the three Chinese larch taxa, a sequencing analysis was carried out on each amplified fragment. One individual from each species was used for sequencing. Three fragments were sequenced separately and the results combined, to form a single sequence for the region *trn*T(exon) to *trn* F (exon). The DNA sequence was determined according to the corresponding strongest signal at each base position, showing that the sequencing error is less than 10% and hence the results are reliable.

Table 4.3 Base compositions for the three regions of cpDNA not including primers, and the total length including primers in *Larix*, as compared with black pine cpDNA (*Pinus thunbergii*).

Region	Species	Length(bp)	Frequency of four bases A	C	G	T
Intergenic spacer (**ef**)						
	Larch	415	0.340	0.164	0.183	0.313
	Black pine	420	0.348	0.159	0.181	0.312
trnL intron (**cd**)						
	Larch	517	0.284	0.222	0.174	0.320
	Black pine	528	0.287	0.212	0.164	0.336
Intergenic spacer(**ab**)						
	Larch	440	0.275	0.218	0.157	0.350
	Black pine	448	0.270	0.239	0.156	0.335
Total region(**af**)						
	Larch	1452	0.297	0.202	0.167	0.324
	Black pine	1477	0.298	0.205	0.173	0.324

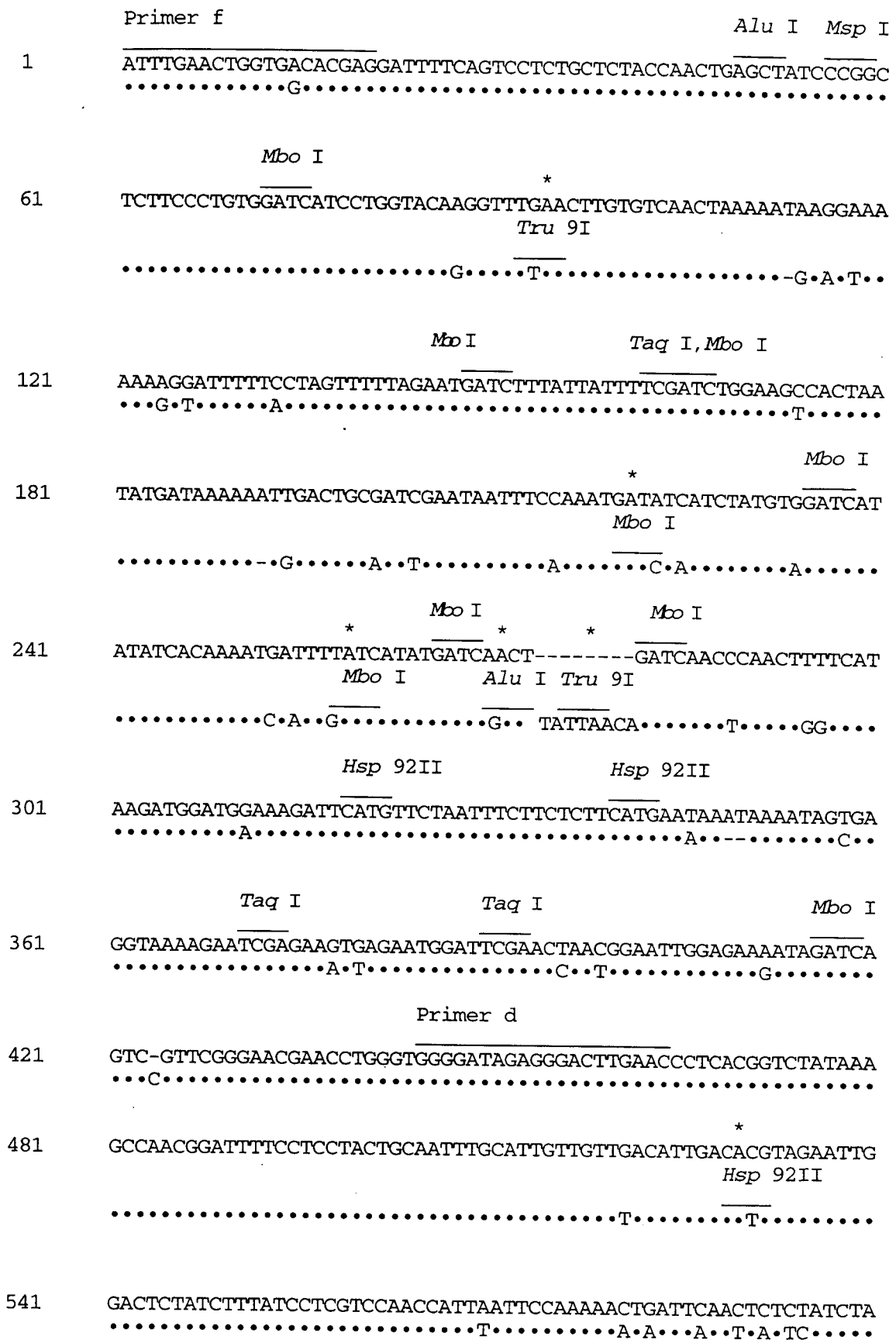
No differences were observed between the three larch taxa within the three regions of cpDNA that were sequenced (Fig. 4.7). The base composition of these three non-coding regions of cpDNA was also compared with the same regions of black pine (Table 4.3), showing a similar sequence in both taxa and with more than 60% of A+T content. The total length of the three fragments was 1452bp, which is smaller than that estimated from the gel 1490bp (Table 4.2).

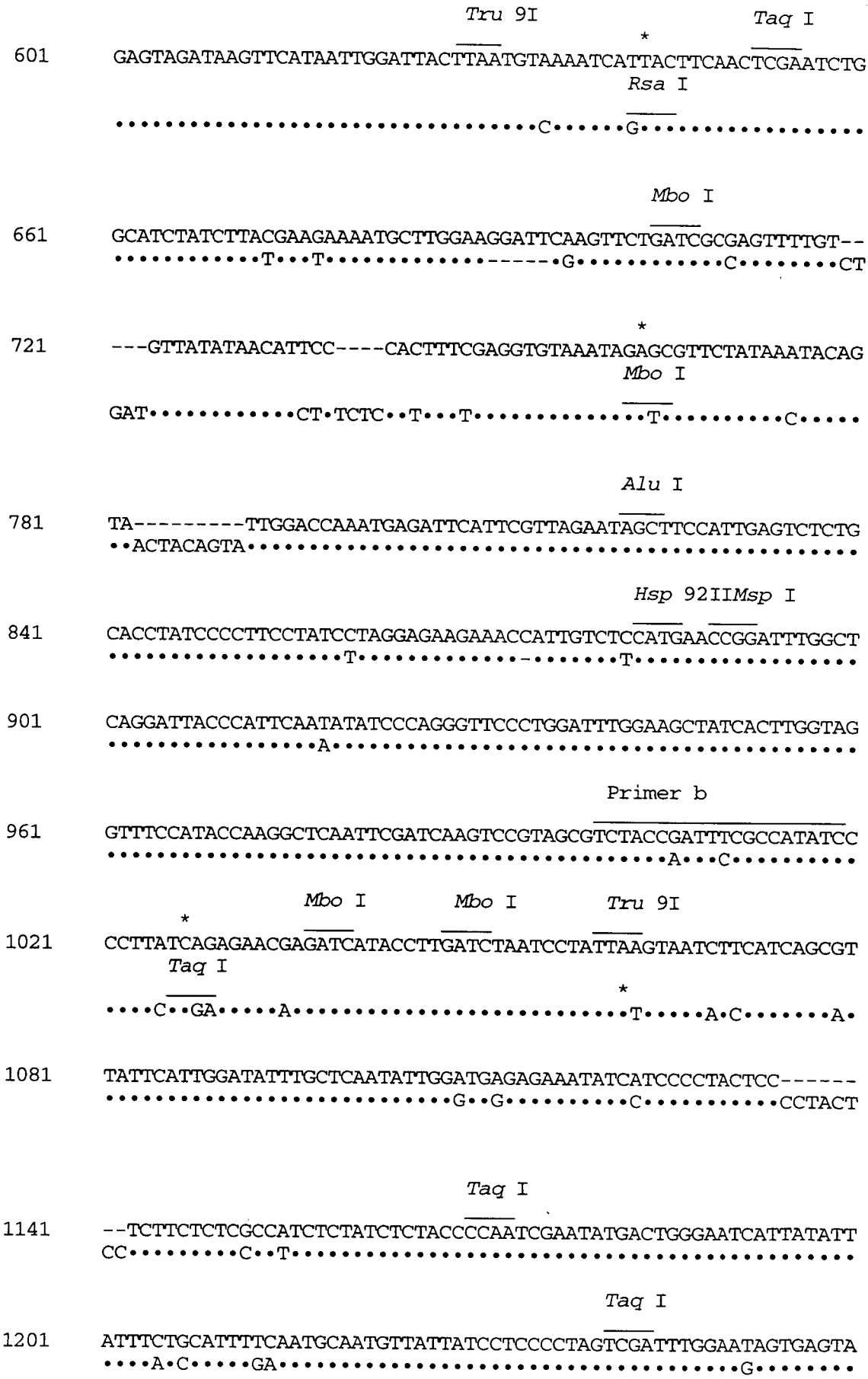
A sequence alignment analysis conducted between larch and black pine cpDNA using the Gene Jockey package (Fig. 4.7), indicated many differences between these two taxa, including both base substitution (92bp) and insertion/deletion events. In addition, the total length of the region (**af**) was shorter in *Larix* than that detected in black pine (Table 4.3).

In the **ab** band, there were two cutting sites for the *tru* 9I in larch cpDNA but not in that of pine (Fig. 4.7). One is located at the 1063th base position; where 'TTAA' in larch has become 'TTAT' in black pine by mutation (transversions). Another cutting site is at the 1343th base position; where 'TTAA' in larch has become 'GTAA' in black pine by mutation (transversions). Comparison of sequence also reveals the difference between larch and pine cpDNA for the digestion by *Taq* I in the 1027th and 1028th base, and the 1445th base (Fig.4.7).

In the region between *trn* L (5') and *trn*L (3'), amplified by primer pair **c** and **d,** the reason for the *Rsa* I not digesting larch cpDNA but digesting black pine is due to the substitution of base G by T, at the 633th base position (Fig. 4.7). Although both larch and black pine cpDNA can be digested by either *Mbo* I or *Hsp* 92II, there are differences between them. For the *Mbo* I, the 764th base position is different: the base is G for larch and T for pine. For the *Hsp* 92II, the 531th base position is different: the base is C for larch and T for pine.

Comparison of sequence of the region amplified by primer pair **e** and **f** indicated that the reason that the *tru*9I could not digest larch but could digest pine is due to mutation occurring at two sites. One is at the 93rd site where T is substituted by G in larch (transversions). The second is at the 273rd base where TTAA is missing in larch cpDNA. Furthermore, there are also differences in the digestion patterns obtained by *Mbo* I and *Alu*I due to base substitution (Fig.4.7). The *Mbo* I cannot digest at the 258th base position in larch but can in pine, due to


Primer f
Alu I
Msp I
1
ATTTGAACTGGTGACACGAGGATTTTCAGTCCTCTGCTCTACCAACTGAGCTATCCCGGC
Mbo I
61
TCTTCCCTGTGGATCATCCTGGTACAAGGTTTGAACTTGTGTCAACTAAAAATAAGGAAA
Tru 9I
Mbo I
Taq I, Mbo I
121
AAAAGGATTTTTCCTAGTTTTTAGAATGATCTTTATTATTTTCGATCTGGAAGCCACTAA
Mbo I
181
TATGATAAAAAATTGACTGCGATCGAATAATTTCCAAATGATATCATCTATGTGGATCAT
Mbo I
Mbo I
Mbo I
241
ATATCACAAAATGATTTTATCATATGATCAACT--------GATCAACCCAACTTTTCAT
Mbo I
Alu I
Tru 9I
Hsp 92II
Hsp 92II
301
AAGATGGATGGAAAGATTCATGTTCTAATTTCTTCTCTTCATGAATAAATAAAATAGTGA
Taq I
Taq I
Mbo I
361
GGTAAAAGAATCGAGAAGTGAGAATGGATTCGAACTAACGGAATTGGAGAAAATAGATCA
Primer d
421
GTC-GTTCGGGAACGAACCTGGGTGGGGATAGAGGGACTTGAACCCTCACGGTCTATAAA
481
GCCAACGGATTTTCCTCCTACTGCAATTTGCATTGTTGTTGACATTGACACGTAGAATTG
Hsp 92II
541
GACTCTATCTTTATCCTCGTCCAACCATTAATTCCAAAAACTGATTCAACTCTCTATCTA


Tru 9I
Taq I
601 GAGTAGATAAGTTCATAATTGGATTACTTAATGTAAAATCATTACTTCAACTCGAATCTG
Rsa I
Mbo I
661 GCATCTATCTTACGAAGAAAATGCTTGGAAGGATTCAAGTTCTGATCGCGAGTTTTGT--
721 ---GTTATATAACATTCC----CACTTTCGAGGTGTAAATAGAGCGTTCTATAAATACAG
Mbo I
Alu I
781 TA---------TTGGACCAAATGAGATTCATTCGTTAGAATAGCTTCCATTGAGTCTCTG
Hsp 92IIMsp I
841 CACCTATCCCCTTCCTATCCTAGGAGAAGAAACCATTGTCTCCATGAACCGGATTTGGCT
901 CAGGATTACCCATTCAATATATCCCAGGGTTCCCTGGATTTGGAAGCTATCACTTGGTAG
Primer b
961 GTTTCCATACCAAGGCTCAATTCGATCAAGTCCGTAGCGTCTACCGATTTCGCCATATCC
Mbo I
Mbo I
Tru 9I
1021 CCTTATCAGAGAACGAGATCATACCTTGATCTAATCCTATTAAGTAATCTTCATCAGCGT
Taq I
1081 TATTCATTGGATATTTGCTCAATATTGGATGAGAGAAATATCATCCCCTACTCC------
Taq I
1141 --TCTTCTCTCGCCATCTCTATCTCTACCCCAATCGAATATGACTGGGAATCATTATATT
Taq I
1201 ATTTCTGCATTTTCAATGCAATGTTATTATCCTCCCCTAGTCGATTTGGAATAGTGAGTA

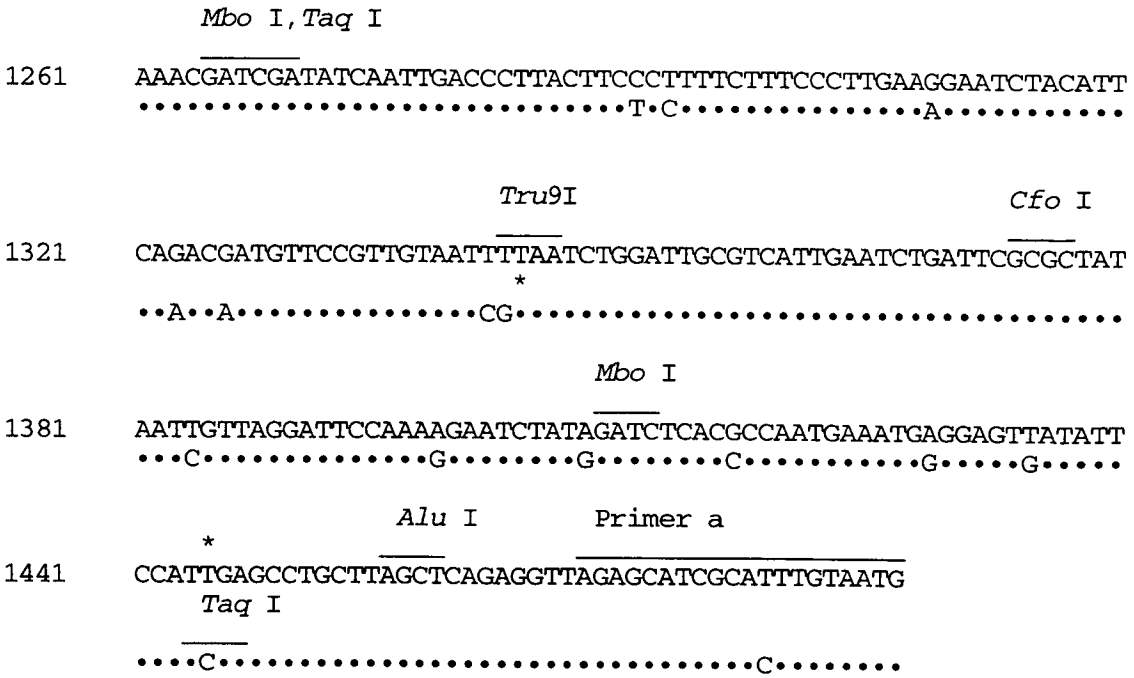


Fig.4.7 DNA sequences for the three noncoding regions of *Trn* L(UGU) to *Trn* F (GAA) for each of the three larch taxa (Top), and their alignment with their corresponding regions of black pine cpDNA (bottom). The three pairs of universal cpDNA primers, together with the cutting sites for the eight restriction enzymes used in the study are indicated. The * symbol means that the labelled enzyme can cut the larch at that site(s) but not that of pine, or *vice versa*. The -/+ symbols in any DNA sequence stand for deletion/insertion of a base. The • symbol represents the same base in pine cpDNA appearing at that position in the larch cpDNA.

the mutation from T in larch to G in pine. The *Alu*I cannot digest at the 271st base position in larch, but can in pine, due to the mutation from A in larch to G in pine.

Sequencing also revealed that the nucleotide sequences in primers **a**, **b** and **f**, the conservative region (Taberlet, *et al.*, 1991), exhibit base substitutions between the three larch species and pine. The difference between black pine and Chinese larch taxa in terms of the three non-coding regions of cpDNA sequence was summarised in Table 4.4.

If the mutation rates by transitions and by transversions are assumed to be equal to each other, Juke and Cantor's (1969) one-parameter model can be used to calculate the genetic distance, K, i.e.,

$$K = \frac{3}{4}\ln\left(\frac{3}{4q-1}\right) \qquad (4.1a)$$

$$\approx 2\mu t \qquad (4.1b)$$

where q is the proportion of the bases that are same between two sequences, μ is mutation rate, and t is the divergence time. Using the sequence not including bases due to insertion and deletion, there are 92 bases different between Chinese larch and black pine within the three non-coding regions. Thus, q is estimated to be 1-92/1442 = 0.9362, and the genetic distance is K= 0.06667 according to equation (4.1a). This distance is much larger than that found between larxh taxa using RFLP analysis (Tang, *et al.*, 1995), where the maximum genetic distance is 0.0096.

In summary, no variation was observed between the three Chinese larch taxa in terms of the sequence of the three non-coding regions *trn*T(exon) to *trn* F (exon), amplified by three universal cpDNA primers. These results are also consistent with those obtained by PCR-RFLP analysis. Many differences between larch and pine cpDNA were observed in terms of sequence, including substitution and insertion/deletion.

Table 4.4. Difference between black pine and Chinese larch taxa in terms of cpDNA sequence within the three non-coding regions. Black pine was used as a reference for comparison. The numbers of base difference for larch in this study are sumarised below.

Region	Insertion(bp)	Deletion(bp)	Substitution (bp)
Primer			
f			1
d			0
b			2
a			1
Fragment			
ef	4	9	31
cd	6	18	25
ab	0	8	32

4.5 Discussion

The aim of this study was to use cpDNA markers to elucidate the evolutionary relationships between the three Chinese *Larix* taxa, defined as the *L. gmelinii* complex (Chapter 1). Two methods were used to resolve differences between the three taxa of three non-coding regions of the cpDNA molecule: PCR-RFLP and straightforward sequencing analysis. No variation was observed between the three taxa, which provides evidence, in part, to support the idea that *L. olgensis* and *L. principis-rupprechtii* are considered to be two varieties of *L. gmelinii* (Ostenfeld and Larsen, 1930), but to refute the opposite view that *L. olgensis* and *L. principis-rupprechtii* are two different species (Zheng, *et al.*, 1983; Wang, *et al.*, 1995)

The experimental results are also consistent with those obtained by Tang *et al.* (1995), who showed that the Nei's genetic distances were zero between the three Chinese larch taxa according to the RFLP analysis. More specific information is obtained in this study, namely that there is no difference observed between them in terms of nucleotide sequence of the three non-coding region between *trn*T (UGU) and *trn*F(GAA) of cpDNA. The same sequence in the three non-coding regions indicates that divergence between the three larch

taxa might have occurred recently. Thus, the quite close genetic relationship among the taxa is further proved and can be judged strongly by this study.

The mutation rate of non-coding regions of cpDNA is greater than that in coding regions, thus non-coding regions are usually used to elucidate genetic variation at lower taxonomic level such as at the generic level (Gielly and Taberlet, 1994). Furthermore, the point mutation rate in cpDNA is lower than that in the nuclear genome (Wolfe, *et al.*, 1987). Thus, use of non-coding regions of cpDNA as a genetic marker might help to elucidate the length of time since the taxa diverged. Wolfe and Sharp (1988) argued that the average rates of synonymous substitution for cpDNA protein coding genes vary approximately from 0.2 to 1.0×10^{-9} substitutions per site per year. If the mutation rate in the non-coding regions is not smaller than, or at least larger than, that of synonymous substitution, the time for the presence of the same sequence in the three Chinese *Larix* taxa can be estimated as the following. If there is one base substitution between any pair of the three larch taxa, the q in equation (4.1a) is 1451/1452= 0.9993112, and the genetic distance is 0.000689 according to equation (4.1a). The diverged time is much less than half a million years, which is roughly inferred according to equation (4.1b) under the hypothesis of a molecular clock that may be violated among plant families and orders (Clegg, *et al.*, 1994), i.e.

$$t = K / 2\mu$$
$$= 0.000689 / 2\times 1.0\times 10^{-9}$$
$$= 0.3445\times 10^{6} \text{ years.}$$

The diverged time inferred from the above calculation seems appropriate. When analysing fossil records of *Larix*, LePage and Basinger (1995) argued that " The general absence of long-braced in the fossil record may reflect their adaptation to alpine habitats, where chance of entry into the fossil record is remote. The distribution of the living larches that the short-bracted species commonly occupy habitats at lower altitudes, where chance of preservation is greater." The three larch taxa in this study are short-bracted, while other larch species in Asia, *L.kaempferi*, *L. griffithian*, *L. potaninii* and *L. mastersiana*, are long-bracted. LePage and Basinger (1991) postulated that the long-bracted species diverged from the short-bracted species early in their evolution. Lepage and Basinger (1995) argued that past and present distribution of the short-bracted larches provide good evidence that larches used the

Beringian Route at least as the Oligocene (25 to 36 million years ago). The Beringian Route is believed to have been an effective floral and faunal conduit between North America and Asia from Albian time (about 100 million years ago, LePage and Basinger, 1995). From these analyses, they also inferred that the distribution pattern of the living long-bracted larches in Asia indicated that displacement occurred across the Beringian Corridor prior to the time when the European climate became cooler in the Miocene (about 5 to 25 million years ago) and Pliocene (2 to 5 million years ago), during which the North Atlantic land routes had been destroyed by sea-floor spreading and were no longer available. According to these analyses, the time for the displacement of long-bracted larches from high altitudes to low altitudes probably occurred 25 million years ago. Lepage and Basinger (1995) inferred that the current distribution pattern was probably established by the late Tertiary. Thus divergence of the short-bracted larches probably lagged far behind divergence between the long-bracted larch species. Similar inference may infer that the southward displacement of *L. gmelinii* might lag far behind the long-bracted species *L. potaninii*, *L. griffithiana* and *L. mastersiana*. This may indicate that establishment of the current populations and taxa in China probably occurred in the late Tertiary.

After analysing the results obtained by other researchers, LePage and Basinger (1995) stated that " ....., In fact, taxa such as *Larix*, *Picea*, and *Pseudolarix* do not appear in Europe until the Miocene and Pliocene.", which is between 3 and 25 million years ago. If a comparable time scale occurred for the southward extension of *L. gmlinii* into China, this time is much smaller than that required for occurrence of a mutation in cpDNA.

The quite close genetic relationship, elucidated by both sequence analysis of cpDNA and allozyme markers (Chapter 3), could reflect the short history for the formation of the three taxa. As we know that larches are well adapted to the regions with poor soils, cold climate, short growing season, etc.. These properties enable larches to become pioneer species. One likely hypothesis is that first *L. gmelinii* migrated to northeast China. Thus pioneer populations were gradually established. Compared with the original population (*L. gmelinii*), the genetic polymorphism in these newly established populations was reduced because of the occurrence of genetic drift. However, large migration occurred later from the southern population of *L. gmelinii* to the northeast population, the former populations of *L. olgensis*, may reduced the difference between source and recipient populations. When the Changbei

Mountains were gradually formed, these newly established populations gradually become the current *L. olgensis* due to evolution in isolation (Fig. 4.8).

Formation of the *L. principis-rupprechtii* could have occurred by two routes. One is from populations of *L. olgensis*, and the other is from straightforward populations of *L. gmelinii*. Because of the warm climate which blocked further southward extension, the newly established populations underwent speciation into *L. principis-rupprechtii*, and retreated north and to higher elevation in the mountains when climate warming occurred. The backward colonisation became difficult due to late clear logging for farm land (Fig.4.8). Formation of the current distribution of the three larch taxa probably occurred in late Tertiary according to the inference of LePage and Basinger (1995).

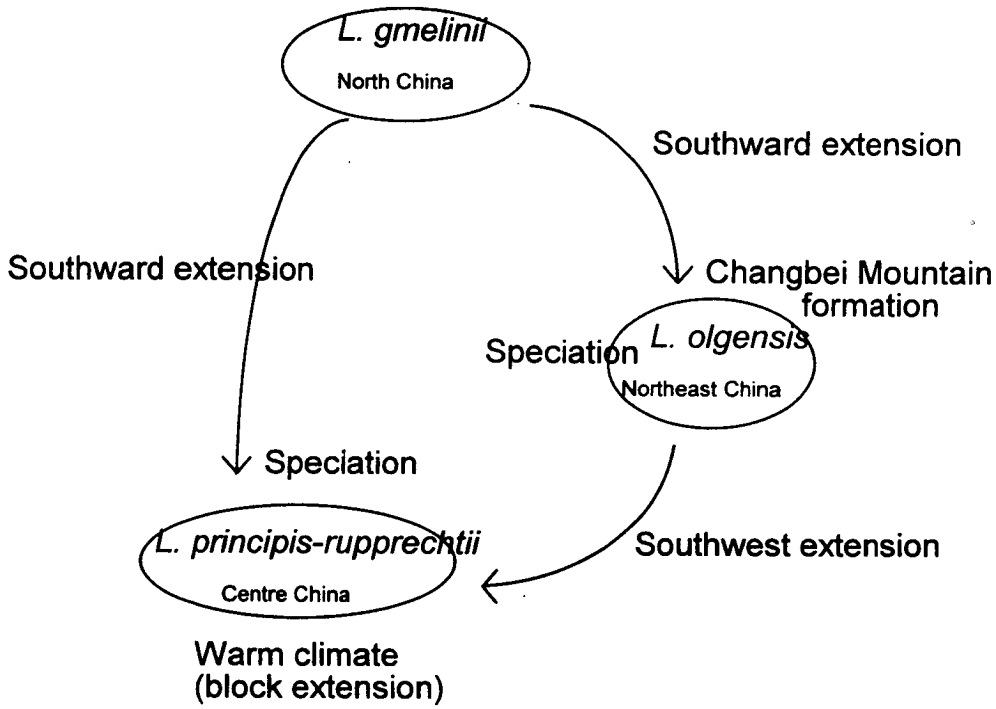


Fig.4.8 A hypothensis for the formation of the three Chinese larch taxa in this study.

Another event that may help to explain why there are no differences between the three larch taxa in terms of cpDNA sequence, is hybridisation. Hybridisation among these larches can

be carried out in nature. For example, *Larix lubarsikii* Suk is the natural hybrid between *L. olgensis* and *L. principis-rupprechtii* (Yang, 1995). Larch is a wind-pollinated conifer species. Transmission of the cpDNA can be mediated by pollen flow because of its paternal inheritance mode (Szmidt, 1990). Thus extensive hybridisation or pollen flow among the three larch taxa may homogenise the difference in cpDNA composition. In analysing phylogenetic consequences of cytoplasmic gene flow in plant, Rieseberg and Soltis (1991) also pointed out that "... reticulation is perhaps the most likely to lead to faulty phylogenetic conclusions in plants due to their high potential for interspecific gene flow.....". From the dominant outcrossing and very small population genetic structure obtained using allozymes (Chapter 2), it can be inferred that migration via pollen flow may be extensive in these three larch taxa. Thus, the reticulation is likely to be an important factor leading to the current homogeneity in cpDNA composition.

The very close genetic relationship among these three larch taxa could in part explain a very similar performance for the man-made hybrids among these three larch taxa. Few reports are given on the existence of obvious heterosis between any pairs of the three taxa except one report (Yang, *et al.*,1985). Yang *et al.*(1985) reported that hybrid of *L. principis-rupprechtii* × *L. olgensis* exhibited an increase of 7% over *L. principis-rupprechtii* in terms of seedling growth. However, heterosis was also displayed in hybrids of each of the three taxa with Japanese larch (*L. kaempferi*) with respect to growth performance (Wang, *et al.* 1989; Yang, *et al.* 1985). There is no great potential for utilising hybrids between any pair of the three larch taxa in practical forestry.

Formation of a species is a dynamic process in nature. It is important to distinguish the concepts between species and subspecies or variety. According to the biological definition of species (Riger, *et al.*, 1991, p458), a species refers to " groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups." Thus, a species is " the largest and most inclusive reproductive communities of sexual and cross-fertilisation individuals that share in a common gene pool." However, subspecies refers to "an aggregate of (local) breeding populations of a given species that occupy a geographic subdivision of the species range." (Riger, et al., 1991, p465). " A subspecies usually differs from other similar breeding groups of the same species both taxonomically and with respect to certain gene pool characteristics (such as the frequency or prevalence of certain genes)". Appling these concepts to the three larch taxa in this study, it would be

reasonable to define *L. olgensis* and *L. principis-rupprechtii* as two varieties of *L. gmelinii* rather than as two separate species. The reasons are as follow:

First, all these three larch taxa can interbreed. Reproductive isolation is not formed. Thus all these three larch taxa can be treated as one species. Second, these three larch taxa occupy different geographic regions. They can be distinguished in terms of morphological traits (Appendix I) that may reflect the adaptive differences due to varying effect of environmental factors. There were shown to be a small genetic distances between the three taxa (0.01; see Chapter 3) in terms of allozyme markers that are not affected by natural selection. Moreover, there are no differences between the three taxa in terms of sequences of the three non-coding regions of cpDNA. These facts provide good evidences in support of points of Ostenfeld and Larsen (1930) who proposed that *L. olgensis* and *L. principis-rupprechtii* were two varieties of *L. gmelinii*.

A situation similar to the present study was reported by Shiraishi *et al.* (1996), who recently classifed larch trees occurring at Mt. Manokami, Japan. They studied the genetic relationship between the larch trees occurring at Mt. Manokami, which was previously defined as Japanese larch (*L. kaempferi*), and three other larch taxa: *L. gmelinii* var. *japonica*, *L. kaempferi* and *L. gmelinii* var. *olgensis*. They found that there were no differences between Mt. Manokami larch and *L. kaempferi*, but there were differences between the Mt. Manokami larch and two other larch taxa (*L. gmelinii* var. *japonica*, *L. gmelinii* var. *olgensis*) in terms of sequence of the *rbc*L gene. However, they also found genetic differentiation between the Mt Manokami larch and *L. kaempferi* regarding nuclear genome composition. Based on these results and the morphological characters, Shiraishi *et al.* (1996) proposed that the Mt. Manokami larch should be classified as a variety of *L. kaempferi*, rather than a new species.

In order to further elucidate the evolutionary relationship between the three taxa, it may be necessary to get more information by investigating variation of the mtDNA genome, which is maternally inherited (DeVerno, *et al.*, 1993), since cpDNA (paternally inherited) and allozymes (biparentally inherited) have already been investigated. The migration for maternal genes is mediated by seed flow only. Use of maternal genetic markers to investigate the differentiation between the three Chinese *Larix* taxa may provide additional information to confirm current results. However, mtDNA variation is mainly by large

rearrangements. Such types of variation are not useful for building phylogenies (Palmer, 1992).

4.6 Summary

In summary, the genetic relationships between the three Chinese larch taxa (*L. gmelinii, L. olgensis* and *L. principis-rupprechtii*) were studied by analysing three non-coding regions of the chloroplast genome, from *trn* T(UGU) to *trn* F(GAA), using universal primers PCR-RFLP and DNA sequencing. Results show that there is no difference in the DNA sequences between all three taxa. Combining the results obtained in this chapter with those obtained using allozyme markers, which show very small genetic distance between the three taxa, it may be inferred that the three taxa have diverged relatively recently. It is more reasonable to consider *L. olgensis* and *L. principis* to be two varieties of *L. gmelinii*, as was proposed by Ostenfeld and Larsen (1930), rather than two different species. Further study is required to investigate variation of the mitochondrial genome so as to provide additional information to elucidate phylogenetic relationships between the three taxa.

CHAPTER 5

Understanding the Genetic Structure of Populations

5.1. Introduction

The living world comprises a variety of species that are themselves composed of many populations, aggregates of individuals coexisting in space and time. Individuals within population show varying levels of genetic variation that is exchanged and reassorted as a consequences of sexual reproduction, genetic interchange among individuals. The extent and pattern of variation in any situation depends on factors such as the breeding system of the species and gene flow between populations, and is modified by processes of genetic drift and natural selection. Thus, beneath the numerical dynamics of populations, there exists the dynamic behaviour of genetic variation. Understanding such behaviour of genetic variation is the objective of population genetics. With an understanding of the behaviour of genetic variation comes the possibility of managing this variation for particular objectives. An understanding of population genetics therefore underlies such applied subjects as selective breeding and is essential for the management of genetic resources in conservation.

Factors involved in modifying the dynamics of genetic variation are quite diverse, including mutation, breeding system, genetic drift, gene flow and natural selection. Mutation is the source of all variation. The breeding system assorts this variation and determines the within population genetic structure (Wright, 1951). Genetic drift leads to irregular (random) fluctuations of gene frequency, to loss of variability, and to differentiation of isolated populations. This may lead to uniformity within sub-populations and increased homozygosity of the whole population (Falconer, 1989). Gene flow, however, may homogenise gene frequencies among different populations. Natural selection modifies gene frequency in relation to environment both within and between populations.

Among these factors, mutation, gene flow and natural selection tend to change gene frequency in a manner predictable both in amount and direction. They can cause *systematic change* in gene frequency (Wright, 1969). The effect of genetic drift , however, is predictable in amount but not in direction. It causes *random change* of gene frequency. Usually, these factors may act together in a population. Thus population genetics is essentially an understanding of the interplay between each of these factors.

Population genetics must take place within a population framework, within a particular spatial array of populations or individuals, not in the abstract. In order to model the real

world, we need to understand how the processes governing genetic change interact within a variety of spatial population types representing as far as possible the range of situations in the real world. No one model will describe all situations. A variety of population model frameworks therefore have been devised to study the dynamics of genetic structure. These include the island, stepping stone and isolation by distance models that are described below.

In addition to a framework we need a common language with which to describe and compare the behaviour of genetic variation. At its simplest we can describe changes in gene frequency, single locus genetic structure, and multilocus genetic structure within populations. Several ways of describing variation in structure over space are available (Wright, 1969), quantifying differentiation between populations (F_{st}) measured in terms of inbreeding coefficients.

With the processes, the framework and the common language in place, we can begin to describe the behaviour of genetic variation. The second part of this thesis will be devoted to doing this, with particular emphasis on plant populations and the behaviour of organelle genomes. I will begin by describing the range of models of population structure used as frameworks for developing population genetic theory.

5.2. Theoretical considerations

5.2.1 Island model and its variants

5.2.1.1 The island model

The island model is an important classical model designed to simulate the real world. It is perhaps the simplest way of representing a real population that is subdivided. It plays an important role in generating ideas for models of population structure. The theoretical results from this model are still widely employed in practical work. The original idea of the island model was introduced by Wright in an historic paper "Evolution in Mendelian Populations " (Wright,1931) and then was developed further (Wright, 1943, 1951).

The basic ideal for the model is that a population is subdivided into infinite number of local populations each with equal size (N). These local populations are discretely distributed in space. Each local population receives a small proportion of migrants (m) from the whole

population. It should be pointed out here that migrants are diploid individuals (nuclear loci). Both migration rate and migrant gene frequency are constant at any generation (Fig. 5.1). Because of the finite size in each local population, genetic drift effects may lead to differentiation among local populations. However, constant migration ensures that this does not lead to fixation in each population. Thus eventually, migration and drift reach a balance. The whole population structure is then maintained in a steady state. The distribution of gene frequencies among populations reach a steady state.

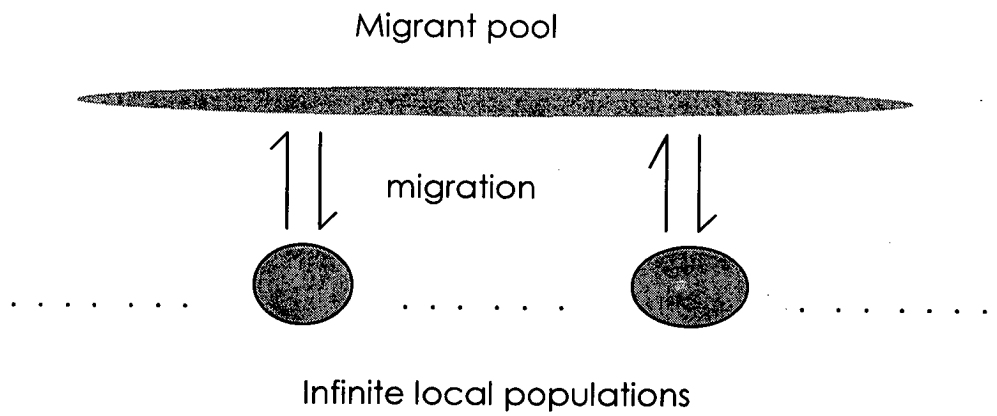


Fig. 5.1 Schematic representation of the island model. The dots below stand for an infinite number of local populations. The fine dots above stand for the migrant pool that comes from small contributions of migrants from all of the local population. The straight line arrow stands for long distance immigration or emigration between local population and migrant pool.

Using this model as a base, Wright (1951) introduced *F*- statistics, which offer a convenient way to summarise population structure, a common language for population genetics. Here, using the same notation as Wright's , let the F_{it} and F_{is} be the inbreeding coefficients in whole and local population, respectively. F_{st} is the inbreeding coefficient of any two gametes sampled from different local populations, and hence it can be used to measure local population differentiation at equilibrium. The first feature is that the inbreeding can be

partitioned into components, those within and between populations, and are related by $1-F_{it}=(1-F_{is})(1-F_{st})$. The second result is the relationship between the F_{st} and number of migrants (Nm). That is $F_{st}=(1+4Nm)^{-1}$. These two well known formulae lay down strong theoretical foundation in the analysis of population structure.

There are several assumptions involved in the ideal island model. These imply some distance between the model and the real world. These assumptions are: ① infinite number of sub-populations; ② constant population size (N), no extinction and recolonisation; ③ constant migration rate from the entire population; ④ constant migrant gene frequency; ⑤ two alleles of a locus, the simplest case; ⑥ random mating; ⑦ selective neutral genes; ⑧ diploid nuclear genes; ⑨ discrete generations.

5.2.1.2 Constraints and relaxation

Each of these assumptions has been relaxed in more recent models so as to approach more nearly the real world. For example, the assumption ① is unrealistic because the actual number of local populations should be finite. Condition ① has been released in finite island models (Nei 1975; Takahata, 1983; Takahata and Nei, 1984). In this case, the relationship between F_{st} and the number of migrants and the number of subpopulations is obtained. Other assumptions are also relaxed to different extents in a range of other more realistic models (Table 5.1).

Many genetic characteristics have been theoretically investigated for the island model structure, for example, the number or effective number of alleles maintained in the finite island model (Maruyama, 1970), and the effective population size (Nei and Takahata, 1993). One of the important applications in practice for the island model is that it provides a simple way to indirectly estimate gene flow (Nm) among local populations (Slatkin and Barton, 1989)

5.2.1.3 Limitation to plant population

One important shortcoming of the ideal model is that it does not describe adequately the behaviour of plant species. The ideal model assumes that migrants are all diploid

Table 5.1. More recent models relaxing some assumptions of Wright's island model

Assumption	Relaxed case	Reference
① Infinite number of sub-populations	Finite number	Nei,1975; Takahata, 1983 Takahata and Nei, 1984
② Constant population size	Extinction/recolonization	Maruyama and Kimura, 1980
③ Constant migration rate	Stochastic migration	Nagylaki, 1979;
④ Constant migrant gene frequency	Stochastic migration	Nagylaki, 1979;
⑤ Two allele a locus	Multilocus	Takahata, 1983;
⑥ Random mating	Partial selfing	———
⑦ Selectively neutral gene	Selected gene	Wright, 1978; Nagylaki, 1979;
⑧ Diploid nuclear gene	Haploid organelle gene	Birky, *et al*, 1988; Petit, *et al*, 1993; Ennos, 1994;
⑨ Discrete generation	Overlapping generation	———

individuals. However, in plants migration can be mediated either by seed flow (diploid) or by pollen flow (haploid). These two types of gene flow occur within the same generation, but are separate biological processes. Gene flow by the vector of pollen is successful only when migrant pollen grains fuse with ovules in the recipient population. This is quite different from the gene flow by the vector of seed flow.

Furthermore, the three plant genomes (chloroplast, mitochondrial and nuclear DNA) exhibit different inheritance. In some conifers, for example, chloroplast DNA exhibits paternal inheritance, mitochondrial DNA exhibits maternally inheritance (Mogensen, 1996), while the nuclear genome is bi-parentally inherited. For genes with biparental or paternal inheritance, migration occurs by both seed and pollen flow. For maternal genes, migration occurs only by the vector of seed flow. Thus, this results in asymmetric migration rates for the three different plant genomes (Petit, *et al.*, 1993b; Ennos, 1994). Therefore, the extension of the island model to plants is of particular interest in both theory and practice.

To date, different methods have been employed in the extension of the island model to plant populations. Petit and others (1993) addressed the finite island model using a method similar to that introduced by Birky and others (1989), which is quite different from those used by Wright (1951). Wright (1951, 1969) obtained the *F*-statistics using path analysis or variance of gene frequency. Petit and others (1993) obtained the G_{st}, a version of Wright's F_{st} in the case of one locus with multiple alleles, by analysing the components of gene diversity. Ennos (1994) addressed infinite island models by analysing the composition of the migration in plant populations. He found an important relationship between the ratio of pollen to seed flow and *F*-statistics with different inheritance modes. In plant population, for example, the population differentiation for uniparent genomes (maternal genes) is $F_{st} = (1 + 2Nm)^{-1}$ where $m = m_s$ (seed migration rate) that occurs by the vector of seeds only.

Thus, based on this analysis, there is a requirement to build models that describe the uniparentally inherited markers and the diploid nuclear markers.

5.2.1.4 Variants of the Island model

Suppose that we divide the entire local populations into two parts, one is a local population and the rest together are seen as another population. Then the island model is changed into as mainland-island or continental-island model (Hanski, 1994). Sometimes this structure consists of one mainland and several island populations surrounded the mainland population. As mentioned before, this kind of structure can be found in natural plant populations. Thus, the mainland-island model is actually a variation of Wright's island model (Wright, 1951). The obvious characteristic for the mainland-island model is the vast difference in size between populations. The mainland population exhibits stability and no extinction, but the island population exhibits extinction with high probability. Therefore, the relationship between them is like a source-sink if migration from island population to mainland is small enough to be ignored (Gaggiotti and Smouse, 1996). If there is no extinction for the island population, then stochastic migration may cause dynamic structure between mainland and island. Therefore, the mainland-island model may help us to find some interesting results related to genetic conservation.

Another variant of the island model is metapopulation structure that is introduced by Levin (1970). Most individuals in each local population are born and die. Systems of such local populations joined by dispersing individuals then make a metapopulation (Hanski and Gilpin, 1991; Hanski, 1994; Gilpin, 1991; Hansson, 1991). This kind of population structure is used for studies on extinction and recolonization established from some other local populations. Study of the influence of extinction/recolonization on genetic structure has begun only in recent years (Wade and McCauley, 1988; Whitelock, 1992; Rannala and Hartigan, 1995), and there will be great potential for these models in the future, especially for plant populations.

5.2.2. Stepping-stone model

The stepping stone model was introduced by Kimura (Kimura, 1953) and a theoretical foundation was then established (Kimura and Weiss, 1964; Weiss and Kimura, 1965). It presents a situation where a whole population is subdivided into infinite number of local populations. Random mating occurs in each local population. Exchange of individuals

between local populations is allowed to occur between adjacent ones and from the entire population (long range dispersal; Fig. 5.2).

When there is no migration from the adjacent populations, the stepping-stone model becomes the island model. Thus, the island model (Wright, 1951) is a specific case of the stepping stone model. The advantage of the stepping stone model over the island model is that it considers migration from neighbourhoods and thus presents a more realistic case.

It seems an obvious biological phenomenon that individuals living close to each other look more similar than individuals living apart in space. This idea can be described in terms of the change of gene correlation with distance. It is for this reason that the genetic correlation is employed to describe population structure in addition to the variance of gene frequency.

In the stepping stone model, it can be shown that the decrease of genetic correlation with distance can be approximated exponentially (Kimura and Weiss, 1964) and also depends very much upon the number of dimensions. This qualitative results is important in helping us to understand natural populations. Treating gene frequency as a continuous variable, Malécot (1948, 1969) obtained similar qualitative results. It should be noted that the three dimensional case is not appropriate for plant populations.

In practice, the change of the genetic correlation with distance is difficult to measure. This is because we cannot obtain the true expected value, $E(p)$, of gene frequencies in space and time. However, estimation of $E(p)$ can be approximated by an average of all gene frequencies in all populations investigated. The important thing here is that the value at this time refers to that at a particular time and position in space, and is not the theoretical expected value. Using this value, we can estimate how the gene frequency correlation changes with distance.

The stepping stone model is now widely used in population structure modelling, especially in theoretical studies. Some genetic properties have also been investigated, including extension of classical conditions. For example, the classical infinite number of subpopulations is modified to study a finite stepping stone model (Maruyama, 1970). Maruyama and Kimura (1980) showed that effective size of the whole population (species) is much reduced and that population differentiation is prevented if local extinction and recolonization occur frequently. It is shown that, if migration is frequent, then finite island

and stepping stone models exhibit a rather similar extent of genetic variability within and between subpopulations (Nagylaki, 1983; Crow and Aoki, 1984).

As in the island model, many genetic properties need to be studied in the future, such as effects of linkage and recombination and the effects of stochastic migration. An important extension of the model is to plant populations. This has not been available to date, because migration in animal population is different from plants where it can be mediated by either seed flow or pollen flow. Furthermore, there is asymmetric migration for three plant

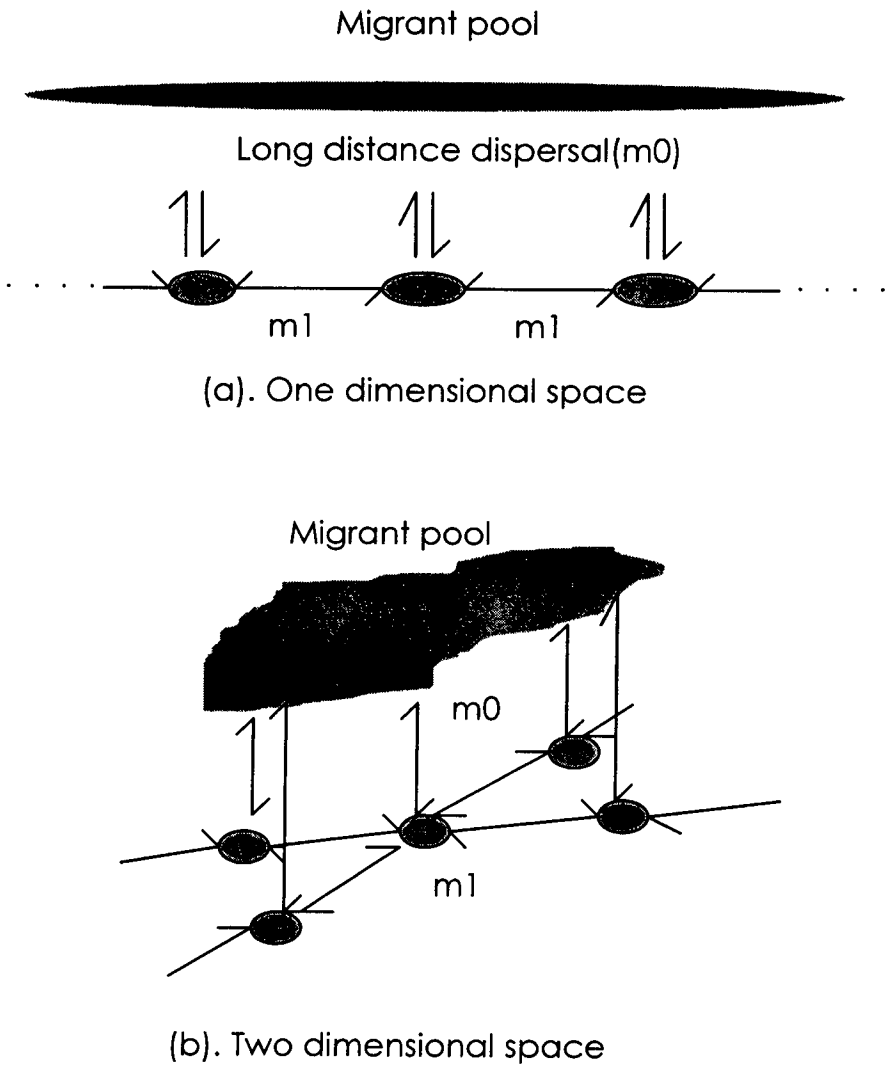


Fig.5.2. Stepping-stone model: (a). One dimension case. (b). Two dimension case. In both cases, the 'm1' stands for migration from neighbourhoods. The 'm0' stands for migration from long distance dispersal. The migrant pool comes from small contributions of migrants from all of the local population.

genomes with different inheritance modes (Petit, *et al*, 1993b; Ennos, 1994). Thus, a study of these characteristics of these genomes in a stepping stone model is clearly required.

5.2.3. Isolation by distance model and its related models

A contrast to the cases that the island and stepping stone models address is the situation where a single population is continuously distributed in space, but interbreeding is restricted to individuals within a restricted area due to short distance gene dispersal. Groups of individuals a large distance apart may then be differentiated merely due to limited dispersal of genes. This phenomenon is described in a model of *isolation by distance* (Wright, 1940, 1943). The model emphasises the importance of spatial distance in isolation and the development of population differentiation. Thus, it presents a framework for understanding the genetic structure of natural population that have a continuous distribution in space.

An important parameter in this model is the neighbourhood size (N_b), defined as the number of individuals within an area from which the parents of central individuals may be treated as if drawn at random (Wright, 1943). A key assumption in the model is that the neighbourhood size at generation t in the past is tN_b for area continuity and $\sqrt{t}N_b$ for linear continuity (1943, 1946; Fig. 5.3). Mathematical proof of this assumption is still not available to date. However, this assumption simplifies the model to describing continuously distributed populations and also provides a simple means for calculating inbreeding at any ancestral generation.

As with the island model, if there is long distance migration from the whole population to each neighbourhood at any ancestral generation t, there will be a balance between migration and drift. Thus differentiation among neighbourhoods can reach a steady state. The distribution of gene frequency also arrive at steady state. Measures of population differentiation in this case can be obtained by $F_{st} = (F_{it} - F_{is}) / (1 - F_{is})$ (Wright, 1943, 1969).

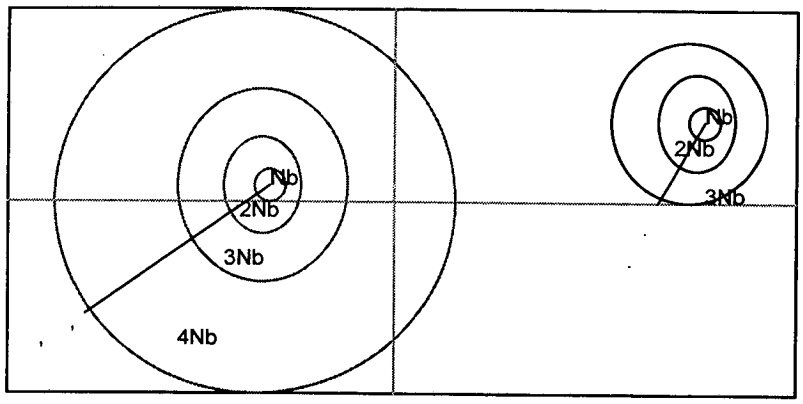


Fig. 5.3. A diagram showing the change of neighbourhood size (Nb) with generation t in the past in two dimension space. Neighbourhood size is linearly increased with generation in the past. The circles represent the size of neighbourhoods. The dots in the block stand for individuals uniformly distributing in space.

One important factor influencing differentiation of local neighbourhoods is the systems of mating that affects the inbreeding coefficient (Wright, 1921,1946) and hence affects population differentiation (Wright, 1943). Influences of several systems of mating were addressed by Wright (1946). One important assumption is that the system of mating is homogeneous in all neighbourhoods. However, the actual mating system maybe exhibits diversity in space because of influence of many factors, say the age structure.

Instead of addressing the case of uniform density throughout population, the isolation by distance model may be modified so that it is suitable for analysing patterns of randomly distributed clusters. Each cluster has a small amount of exchange with those that are closest (Wright, 1969, p320-23, 1978, p59-61). The previous isolation by distance model is thus shifted to patchily distributed populations (*Clusters Model*). This is similar to the stepping stone model but uses a completely different approach (Wright 1969, p320-323; 1978).

Several problems are involved in the isolation by distance model that need to be solved in the future. ① Assumptions for the random dispersal of offsprings and the generation of population at next generation will lead to a clumping in space due to there being a lack of local regulation of population density (Malécot, 1969; Felsenstein, 1975b). This violates assumption of uniform distribution in space. ② Neighbourhood size at any generation t needs to be proved in both mathematics and biology. When applied in plant populations,

calculation of the ancestral neighbourhood size becomes very complicated if we consider dispersal of both seed and pollen flow. Furthermore, if partial selfing together with seed and pollen flow is considered, calculation of the neighbourhood size at any generation in the past will be unmanageable. ③ The inbreeding coefficients are obtained under the hypothesis of *linear*, *complete* and *equal* effects of gametes on zygotes using path analysis (Wright, 1921, p118; 1968; Li, 1976, p290). This means that the path coefficients are obtained under assumption of equal effects from zygotic value to egg and to sperm, or from egg or sperm to zygote. If a dominance effect is considered between genes in egg and sperm, for example, we cannot get the key relationship in this model, i.e. $b_1 b_2 = (1 + F')/2$ where b_1 and b_2 are the path coefficients from zygote to the two gametes, and the F' is the correlation coefficient between uniting gametes.

However, the isolation by distance model is still widely used in populations of continuous distribution. The concept of neighbourhood size, presents us with a means of understanding natural populations with continuous distribution in space. For example, this idea has been employed to calculate the probability of coalescent time of a sample drawn from continuously distributed populations (Barton and Wilson, 1995).

It is easy to understand that there is a positive relationship between differentiation and geographical distance, or a negative relationship between differentiation and number of migrants (Slatkin, 1987). However, the analytic expression to describe these is difficult to obtain. Several other methods/models used to detect isolation by distance have been presented in recently years (Sokal, 1978a, b, 1979; Slatkin and Maddison, 1990). One of them correlates genetic data and geographical neighbourhoods — the spatial autocorrelation analysis. It can be used for detecting spatial pattern and thus the probability of isolation by distance and for estimating the patch size (Sokal, 1978a, b). However, it is not a genetic model but a purely statistical model. The analysis does not explain how the specific pattern is generated. However, Barbujani (1987) showed that the autocorrelation coefficient of gene frequencies at a given distance (Moran's I) is a direct function of the kinship at that distance (f), and an inverse function of the standardised gene frequency variance (F_{st}), i.e. $I = f / F_{st}$. This relationship, to some extent, presents us with a genetic underpining for the spatial autocorrelation coefficients.

Using computer simulation under diversity models of population structure (discretely distributed populations), Slatkin and Maddison (1990) found the linear relationship between effective migrants ($\hat{M}$=Nm) and geographical distance (d), i.e. $Log(\hat{M}) = a + bLog(d)$ where a and b are constant. This method can also be used to detect isolation by distance. If b is significantly less than 0, this means that the effect of isolation by distance is significant, otherwise the isolation is not obvious.

As in the island and stepping stone model, an important extension of the model to three plant genomes with different inheritance models has not been studied.

5.2.4. Cline: a specific population genetic structure

The three models described to date represent general models of population genetic structure built upon theoretical assumptions. Another way of studying population genetic structure is to measure patterns of genetic structure in the field and to build models to account for the observation. A commonly observed genetic structure in nature is that of the cline in which the gene frequency is a function of the geographical situation of populations studied. One typical characteristic is the gradient of change in gene frequency with geographical distance (Fig. 5.4). This is an old topic, but still is an important area of evolutionary interest particularly in relation to allopatric speciation. Clines can exist in the hybrid zone (Young, 1996) or in other situations (Millar, 1983).

The three classical models mentioned above are not able to explain the formation of clines. Understanding this natural phenomenon is also an important subject in theory. Study on this topic goes back to Fisher's pioneer work (1937). He studied the wave of advance of advantageous genes. Using dispersal behaviour together with the selection effect, Fisher (1937) found the relationship among the velocity of wave advance, the selective advantage, and the standard deviation of dispersal. That is $v = \sigma\sqrt{2m}$ where the v is the velocity, the σ is the standard deviation of scattering and the m is the selective advantage. The "length" of the wave is proportional to $\sqrt{k/m}$ where the k is diffusion coefficient. This value $\sqrt{k/m}$ has more recently been defined as the *characteristic length* within which there is no change in gene frequency (Slatkin, 1973).

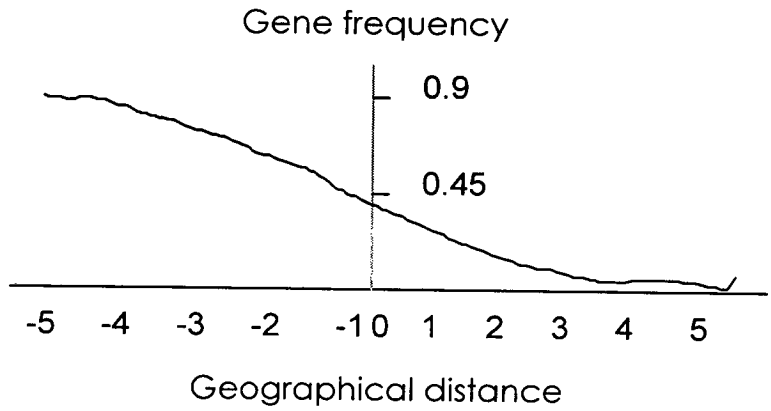


Fig. 5.4. A typical cline shows the change of gene frequency with geographical distance. The numbers labelled along geographical distance can also stand for positions at which different populations locate.

Since then, there have been extensive studies on this topic (Haldane, 1948; Fisher, 1950; Slatkin and Maruyama, 1975; Nagylaki, 1975, 1976, 1978a, b; Felsenstein, 1975a; Barton, 1979, 1983; Barton and Hewitt, 1989). Among most of these studies, the diffusion model is employed to describe the basic process though this is an approximation. Knowledge of cline formation is mainly based on the compound effect of dispersal and selection.

The genetic mechanism underlying a cline may be complicated. In nature, if a cline has existed for many generations, there must be some balance system maintaining it, otherwise the cline will eventually disappear. An important genetic mechanism is the balance of selection-migration-drift. If there is a pure drift process, or drift-migration process, a cline cannot exist. If there is selection but it is too weak, or if migration is too strong, a cline still cannot exist. Therefore, the intensity of natural selection that changes geographically is a very important factor in deciding the existence of a cline. Several conditions have been investigated in different clines (Nagylaki, 1975).

Two parameters are important in shaping a cline. If there is no drift or the drift effect is small enough to be ignored in a cline, the characteristic length decides the width of a cline under a given selection system. The larger the characteristic length, the wider the cline. If drift cannot be ignored in a cline, Nagylaki (1978a) obtained another important dimensionless parameter, β, governing the relative strength of selection and random drift. That is $\beta = 2\rho\sigma^2 / c$ where the ρ is population density in space, the σ^2 is dispersal variance and the c is characteristic length in a deterministic cline. β is the ratio of the natural distance for migration and selection ($2\rho\sigma^2$) to that for the deterministic cline (c). Selection is strong (weak) compared to random drift if $\beta >> 1$ ($\beta << 1$).

Like a barrier to gene flow, the cline prevents or delays complete exchange of genes between adjacent populations. A cline can move forward or backward and exhibits dynamics (Barton and Hewitt, 1989). The genetic structure within a cline may provide much information related to population history.

Many theoretical problems need to be considered in the future study of clines. An important one is the extension of current cline theories to plant species which may exhibit a complex history of colonisation influenced by seed and pollen flow. Furthermore, genes with different inheritance mode may exhibit different clines even in the same swarm. Analysis of these aspects of cline formation in plants may help us to understand the population history in some depth.

We can conclude that all models mentioned above present us a framework of how to think of different types of natural population structure and how to understand them. The obvious problems with them is that those results cannot be simply applied in plant populations. Furthermore, the behaviour of organelle genes with uniparental modes of inheritance have not been considered. The second part of this thesis addresses this area.

5.3. Testing our understanding

Population genetic theory can be used to develop models of population genetic structure that makes testable predictions. They also allow us to interpret observed population genetic structure in terms of underlying biological process and infer parameters such as gene flow among populations. However, in order to test population genetic theory, and to make biological inferences, we need to be able to measure population genetic structure in real population. For these purposes, we need to use naturally occurring genetic markers within taxa.

The markers are chosen including morphological traits, physiological (or biochemical) variants, karyotypic variants, allozymes and differences in DNA sequence. They have been applied in studies on population genetic structure to different extents (Avise, 1994; Mallet, 1996).

In recent years, with the introduction of PCR (Polymerase Chain Reaction; Mullis *et al*, 1986), many different techniques related to PCR have been developed, such as RAPD (Random Amplified Polymorphic DNA) and SSRs (Simple Sequence Repeats) (Rafalski and Tingey, 1993). With the development of a variety of primers (Strand, *et al*, 1997; Dumolin-Lapegue, *et al*, 1997; White, 1996; Morgante and Olivieri, 1993; Hadrys, *et al.*, 1992), many different markers will appear. This may help us to investigate the genetic structure of population at a fundamental level. Here we do not attempt to review these methods further, but note that the potential for practical measurement of population genetic structure is ever increasing.

5.4. Extension of the classical theories

5.4.1. Significance in population genetics

As is emphasised before, all models mentioned above present us with several ways to look at population genetic structure. However, few of them addresses genetic structure of the plant population(Wright, 1969; Petit, *et al*, 1993; Ennos, 1994). At first sight, it is surprising that few theories are involved in this important area of plant species. The probable reason is that plant population genetics always lags behind animal population genetics and the usual

situation is the application of animal population genetic theories in plants. However, some important problems in plant species cannot be solved using those theories suitable for animal populations. A number of obvious differences between plant and animal population genetic structure are:

①***Vectors of gene flow in plant species are different from those in animals.*** In most plant species, gene migration can be seen to occur in two stages, seed and pollen flow. These two forms of gene flow occur in the same generation. Pollen flow first, leading to seed formation and subsequently seed flow takes place. Thus, gene flow for plant species can be mediated by either seed migration or pollen flow. The extent of seed and pollen flow is quite variable in space and time, and also among species and populations.

For example, in natural populations of one conifer species *Abies amabilis* on Vancouver island, British Columbia, levels of inbreeding were very variable, ranging from zero to 27%. Allele frequency in the pollen pool is variable from population to population (Davidson and EI-Kassaby, 1997). In this situation pollen mediated gene flow is far from constant.

The classical models only consider migration of diploid individuals, and hence can be applied to model seed flow in plants, but not haploid migrants travelling via pollen.

② ***The migration rate contained in the formulae of traditional population structure models cannot be substituted linearly by seed and pollen flow if rates of seed and pollen flow are not too small.*** This is an important reason and can be easily understood. Pollen behaves differently from seed. The migration of pollen grains is carried out by mating with ovules. It is affected by the mating system. If migration rates are not too small, say from its neighbouring populations, interaction between pollen and seed flow may act on population structure. This interaction is rather like 'epstasis' effect among different loci. If the mating system departs dramatically from random mating, these effects may become marked. However, if both migration rates are small under random mating, an approximation by linear substitution is appropriate (Ennos, 1994). If there are multilocus interactions, the joint effect of seed flow and pollen flow on them may be larger than when only diploid individuals are migrating.

③ ***Population genetic structures for three plant genomes are different.*** Three plant genomes (nuclear DNA, cpDNA and mtDNA) exhibit different inheritance modes (see review by Mogensen, 1996). For example, for most conifers like in loblolly pine (Neale, *et al.*, 1989a,b) and *Larix* (Azmidt, *et al.*, 1987; DeVerno, *et al.*, 1993), nuclear DNA , cpDNA and mtDNA exhibit biparental, paternal and maternal inheritance, respectively. The different inheritance modes influence their migration mechanism. For the paternal and bi-parental genomes, migration occurs by vector of both seed and pollen flow. However, for the maternal genome, only seed flow contributes to the migration.

As is mentioned before, theoretical results show that the bi-parentally inherited genome exhibits the lowest population differentiation and that maternally inherited genome has the largest differentiation among the three genomes (Petit, *et al*, 1993; Ennos, 1994). Using information for nuclear-organelle genomes differentiation can provide estimation of seed and pollen flow (Ennos, 1994; McCauley, 1995). This cannot be obtained from models suitable for animal populations.

④ ***Differences in population structure among three plant genomes can provide important information on estimation of seed and pollen flow, colonisation history, etc..*** Cytonuclear relationship may present insight into population structure. Pollen and seed migration can influence cytonuclear structure (Asmussen and Schnabel, 1991). Asmussen and others (1989) investigated the effects of nonrandom mating and continued immigration of the parental species on the cytonuclear dynamics in a hybrid zone. They find that permanent cytonuclear disequilibria can be generated by continued migration in the hybrid zone. The joint nuclear-cytoplasmic frequency data can provide particularly sensitive estimates of gene flow into a hybrid zone.

In metapopulations with frequent turnover of local population, population genetic structure can be influenced by the founder effect together with seed and pollen effects. The consequence of genetic structure for the three genomes may be different (McCauley, 1995). Thus this difference is related to the colonisation history. If we separately consider three plant genomes and assume random mating between pollen and ovule, then the differences in population differentiation among them may provide some information of colonisation history. For example, the post-glacial history of the oak species, *Quercus petraea* (Matt.) Liebl. were extensively surveyed using biparental and maternal inheritence markers

throughout the natural range (Le Corre, *et al*, 1997a,b) and within Denmark (Jøhnk and Siegismund, 1997). Both results indicate that population differentiation is larger for maternal inheritance markers (cpDNA markers) than for biparental inheritence markers (allozyme and RAPD markers).

The cytonuclear population differentiation may probably provide the possibility to estimate pollen and seed flow. Furthermore, the differences in population differentiation among three genomes provide the possibility to use them to estimate the ratio of pollen to seed flow, which is an important index only in plant seed management.

Based on these considerations, it may be concluded that there is a pressing need to build theory suitable for describing plant population structure.

5.4.2. Purposes of this study

As can be seen, insights into the genetic structure of plant population are important. Understanding how seed and pollen flow affect population structure will help us to describe accurately the plant world, the different types of spatial structure, and the diverse behaviour of three plant genomes with different inheritance modes. The results are of direct relevance to the interpretation of genetic structure measured in nature and should influence activities such as genetic improvement and conservation.

The primary purposes of the second part of the thesis are therefore to develop such understanding, with a focus on the impacts of seed and pollen flow on different types of plant population structure for each of three genomes. Generally, these impacts include influences on differentiation of populations that are continuously or discretely distribute in space, on coalescent times, and on cline displacements between haploid organelle genes.

I will therefore study the areas listed below, which may help us to understand the relationships between the pollen and seed flow and the different types of genetic structure of plant populations. The major purposes are to formulate general solutions in theory to these questions. The choice of these topics allows us to study the influence of different types of plant population structures, the behaviour of different markers, and the behaviour of different types of genetic data (gene frequency or DNA sequence data).

Specific areas addressed

(a) *Using gene frequency data*

- Extension of three classical population structure models (island model, stepping stone model, and the isolation by distance model) to plant species. The objective is to find the analytical relationship between seed/pollen flow and population differentiation (Chapter 6 and 7).

- Estimation of the ratio of pollen to seed flow. The objective is to look for the possibility of estimating the ratio of pollen to seed flow using different genetic methods (Chapter 7).

- The structure of clines for haploid organelle plant genomes. The objective is to find how seed and pollen flow influence cline width and displacement for haploid organelle genes located on paternally and maternally inherited genomes (Chapter 9).

(b) *Using DNA sequence data*

- Relationship between gene genealogy and geography. The objective is to find how the seed and pollen migration rates influence mean coalescent times for a sample randomly drawn from a population that is subdivided into many local populations, and how to use the number of segregating nucleotide sites to estimate the ratio of pollen to seed flow (Chapter 8).

CHAPTER 6

Extension to Plant Populations of the Island and Stepping-stone Models

6.1 Introduction

A variety of models have been formulated to analyse the development of population genetic structure under a balance between drift and migration. To date they have concentrated on the problem of differentiation for nuclear genes and are appropriate for the situation in animals where diploid individuals migrate between populations. The *island model* comprises many discrete populations with a certain proportion of migrants interchanging between them irrespective of their spatial proximity (Wright, 1931,1969). At the other extreme *isolation by distance* (Wright,1943,1969) describes a large population which has a continuous distribution over a wide area, but in which mating is restricted to a "neighbourhood" of limited scale and migration occurs among neighbourhoods. The *stepping stone* model (Kimura and Weiss,1964), deals with an intermediate situation in which a certain proportion of migration occurs strictly between neighbouring populations, while the remainder takes place by long distance migration, with migrants being drawn randomly from a migrant pool. The applicability of these models to natural populations will of course depend on the population structure and reproductive ecology of the species concerned.

While these models are appropriate for investigating differentiation for nuclear markers in animal populations they are inadequate for fully describing genetic differentiation under drift/migration in plant populations. For these situations models are needed which explicitly incorporate seed and pollen flow as agents of migration. In addition the models must address the cases of differentiation for the uniparentally inherited (both maternal and paternal) chloroplast and mitochondrial markers that can now be detected in natural plant populations through the application of molecular techniques (Dong and Wanger, 1993,1994; Neale, *et al.*, 1986, 1989, 1991; Powell, *et al.*, 1995).

Recently the classical *island models* dealing with population differentiation for nuclear genes have been extended to consider differentiation for uniparentally inherited organelle genes in animal and plant populations (Takahata and Palumbi, 1985; Birky, *et al.*, 1989; Petit, *et al.*, 1993). Petit *et al.* (1993) showed that the effects of gene flow on G_{st} at equilibrium depends on the relative rates of pollen and seed migration, as well as the mode of inheritance of genes (McCauley, 1995). A further insight in this area was given by Ennos (1994) who used an island model to show that a comparison of F_{st} values for markers with different modes of inheritance could provide an estimate of the relative rate of pollen flow to seed pollen flow

among populations. These results can be applied in practical work (Furniers and Stine, 1995;McCauley, *et al.*, 1996; Strauss, et al., 1993;Wheer and Guries, 1982).

The purpose of this chapter is to provide further theory required for understanding and interpreting population genetic structure of nuclear, chloroplast and mitochondrial genes in plant populations under drift/migration equilibrium. I extend traditional island and stepping stone models firstly to incorporate seed and pollen flow as agents of migration, and secondly to contrast population differentiation for biparentally, maternally and paternally inherited markers. Differentiation for markers with contrasting patterns of inheritance is then investigated under the island and stepping stone models of population structure.

6.2 Island model

The rate of gene migration in the classical expression for F_{st}, derived by Wright (1969) for an island model, has a general meaning. It is relevant to situations where there is simple migration of diploid individuals between populations before mating takes place. When dealing with hermaphrodite plants it is necessary to model gene migration as a two step process, which occurs both by migration of haploid pollen before fertilisation, and also by migration of diploid seeds after fertilisation.

Drift/migration balance can be reached either by seed flow, or by pollen flow, or by both forms of gene migration combined. In the following we will re-derive a formula for F_{st} under drift/migration balance which applies to hermaphrodite plants following the method used by Wright (1969). We also investigate whether complex expressions for F_{st} derived for each genome can be reduced to Wright's general formula by substitution of appropriate values for effective population size and migration rate specific to the different genomes.

6.2.1 Assumptions

The model deals with an hermaphrodite population of plants showing random mating. Paternally and maternally inherited genes are assumed to be haploid, while biparentally inherited genes are considered to be diploid. Two neutral alleles per gene are present in each case. The mutation rate for each gene is assumed to be much smaller than migration rate and

is therefore not considered. There is no linkage among the genes differing in mode of inheritance. We consider initially that all the populations have already become established and contain the same effective number of adult plants, N. The effective number of paternal and maternal genes is considered to be N since they are effectively haploid. This assumption can be relaxed if they are not the same by letting $N = N_f$, the effective female donors for maternal genes, and $N = N_m$, the effective male donors for paternal genes.

Figure 6.1 illustrates the processes that occur from generation to generation which influence rates of gene migration and genetic drift among hermaphrodite plant populations. The gene frequencies in migrating pollen grains or seeds are equal to the mean gene frequency over all populations in that generation. The male gametes including those from migrated pollen grains are assumed to combine randomly with female gametes (ovules) during the formation of seeds. The gene frequency in ovules before mating with pollen grains is assumed to be the same as that in the preceding generation.

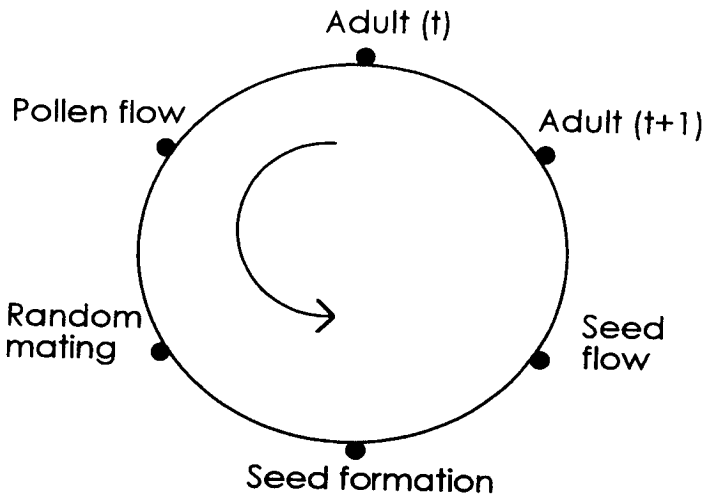


Fig.6.1 The basic processes of pollen flow and seed flow occurring among populations within the life cycle of an hermaphrodite plant population.

6.2.2 Biparentally inherited diploid genes

The following derivation is based on the method used by Wright (1969, p292). Suppose that there are infinite number of populations. At generation t $(t \geq 1)$, let $p_{i,t}$ be the gene frequency in adults in population i. Each population contains the same number of adults, N. After *pollen flow*, the gene frequency in male gametes (pollen) of population i at generation $t+1$, $p^{p}_{i,t+1}$, is

$$p^{p}_{i,t+1} = m_p \bar{p} + (1 - m_p) p_{i,t}, \tag{6.1}$$

where m_p is the rate of pollen flow and $\bar{p}$ is the mean gene frequency over all populations. It can be inferred that the gene frequency in seeds formed by random combination between pollen and ovules at generation $t+1$, $p^{s}_{i,t+1}$, is the arithmetic average of the gene frequencies in male and female gametes, i.e.

$$\begin{aligned} p^{s}_{i,t+1} &= \frac{1}{2}(p^{p}_{i,t+1} + p_{i,t}) \\ &= \frac{1}{2}[m_p \bar{p} + (2 - m_p) p_{i,t}] \end{aligned} \tag{6.2}$$

Similarly, after *seed flow* the gene frequency in seeds of population i, $p'_{i,t+1}$, is

$$p'_{i,t+1} = m_s \bar{p} + (1 - m_s) p^{s}_{i,t+1} \tag{6.3}$$

where m_s is the rate of seed flow. The variance of gene frequencies over infinite number of populations in seeds after *seed flow*, $\sigma^2_{p'_{t+1}}$, is

$$\sigma^2_{p'_{t+1}} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (p'_{i,t+1} - \bar{p})^2 \tag{6.4a}$$

$$\approx E_{\Phi}\left[(p'_{i,t+1} - \bar{p})^2\right] \tag{6.4b}$$

$$= \int_0^1 (p'_{t+1} - \bar{p})^2 \varphi(p'_{t+1}) dp'_{t+1} \qquad (6.4c)$$

where E_Φ stands for an operator for taking expectation with respect to gene frequency distribution among populations. $\varphi(p'_{t+1})$ is the probability density of gene frequency at generation t+1. n is the number of populations. Equation (6.4a) is the expression of the variance of gene frequencies in the case of an island model. The method for the use of equation (6.4b) to approximate (6.4a) can be found in Kimura and Weiss (1964, pp562-563), and also in Nagylaki (1979, p168) in derivation of his equation(22), although they did not indicate this approximation clearly.

The variance of gene frequencies among populations after *seed flow* can be obtained via (6.2) and (6.3), i.e.

$$\sigma^2_{p'_{t+1}} = (1 - m_s)^2 (1 - \frac{1}{2} m_p)^2 \sigma^2_{p_t} \qquad (6.5)$$

It can be seen from equation (6.5) that the variance of gene frequencies among populations is reduced due to different contributions of seed flow $(1 - m_s)$ and pollen flow $(1 - \frac{1}{2} m_p)$. However, after *randomly sampling N* seeds in each population, the variance of gene frequencies among populations will increase because of genetic drift. Thus the gene frequency in adults, p_{t+1}, which can be regard as the same as that in a corresponding sample of seeds for selectively neutral genes, is,

$$p_{i.t+1} = p'_{i.t+1} + \delta_{p'_{t+1}} \qquad (6.6)$$

where $\delta_{p'_{t+1}}$ is the change due to sampling. Similarly, we can obtain the variance of gene frequencies among populations in adults, i.e.

$$\sigma^2_{p_{t+1}} = \sigma^2_{p'_{t+1}} + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 2 \cdot (p'_{i.t+1} - \bar{p}) \delta_{p'_{i.t+1}} + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \delta^2_{p'_{i.t+1}} \qquad (6.7)$$

The second item on the left-hand side of equation (6.7) is equal to zero because of the independence between $(p'_{i.t+1} - \bar{p})$ and $\delta_{p_{i.t+1}}$. The third item on the left-hand side of equation (6.7) is

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\delta^2_{p_{i.t+1}} \approx E_\Phi[E_\delta(\delta^2_{p'_{i.t+1}}|p'_{i.t+1})] \tag{6.8a}$$

$$= E_\Phi\left[\frac{p'_{i.t+1}(1-p'_{i.t+1})}{2N}\right] \tag{6.8b}$$

$$= \int_0^1 \frac{p'_{t+1}(1-p'_{t+1})}{2N}\varphi(p'_{t+1})dp'_{t+1} \tag{6.8c}$$

$$= \frac{1}{2N}\cdot[\bar{p}(1-\bar{p}) - \sigma^2_{p'_{t+1}}] \tag{6.8d}$$

where E_δ stands for an operator for taking expectation with respect to δ distribution within population. Use of equation(6.8) to approach the variance of δ's over infinite number of populations can also be found in Kimura and Weiss (1964, p 563) and Nagylaki (1979, p168).

Therefore, the total variance of gene frequencies among populations after *random sampling,* can be obtained by putting equation (6.8d) into equation (6.7), i.e.

$$\sigma^2_{p_{t+1}} = \sigma^2_{p'_{t+1}} + \frac{1}{2N}[\bar{p}(1-\bar{p}) - \sigma^2_{p'_{t+1}}] \tag{6.9}$$

Substituting equation (6.5) into equation (6.9), we obtain

$$\sigma^2_{p_{t+1}} = \frac{1}{4}(1-m_s)^2(2-m_p)^2(1-\frac{1}{2N})\sigma^2_{p_t} + \frac{\bar{p}(1-\bar{p})}{2N} \tag{6.10}$$

According to equation (6.10), the variance of gene frequencies among populations at steady state can be obtained by letting $\sigma^2_{p_{t+1}} = \sigma^2_{p_t} = \sigma^2$, i.e.

$$\sigma^2 = \frac{\bar{p}(1-\bar{p})}{2N - \frac{1}{4}(1-m_s)^2(2-m_p)^2(2N-1)} \tag{6.11}$$

Using Wight's notation (Wright, 1969, p295 & p299), according to equation (6.11) we can obtain $F_{st(b)}$ (= $\sigma^2/\bar{p}(1-\bar{p})$), the population differentiation for biparentally inherited diploid genes, i.e.

$$F_{st(b)} = \frac{1}{2N - \frac{1}{4}(1-m_s)^2(2-m_p)^2(2N-1)} \tag{6.12a}$$

$$\approx \frac{1}{1+4N(m_s + \frac{1}{2}m_p)} \tag{6.12b}$$

if migration rates of seed and pollen flow are small. In order to obtain a comparable expression between genes differing in mode of inheritance, denote the effective number of genes by $\tilde{N} = 2N$, which is total effective number of genes in each population in adults. Denote the effective migration rate by $\tilde{m} = m_S + m_P / 2$, which is due to diploid seed flow and haploid pollen flow. Thus a simpler expression of equation (6.12b) can be obtained, i.e.

$$F_{st(b)} = \frac{1}{1+2\tilde{N}\tilde{m}} \tag{6.12c}$$

6.2.3 Paternally inherited haploid genes

By the same method as in the case of biparentally inherited diploid genes, the $F_{st(p)}$ at steady state for paternally inherited genes can be obtained,

$$F_{st(p)} = \frac{1}{N-(1-m_s)^2(1-m_p)^2(N-1)} \tag{6.13a}$$

$$\approx \frac{1}{1+2N(m_s+m_p)} \tag{6.13b}$$

if m_s and m_p are small. It is necessary to note that in deriving equation (6.13a) the gene frequency in seed after random combination between male and female gametes, $p^S_{i,t+1}$, is equal to that after *pollen flow*, $p^P_{i,t+1}$. This is different from the case of biparentally inherited genes (equation (6.2)). Similarly, let $\tilde{N} = N$ (haploid) be the effective number of genes. Let $\tilde{m} = m_S + m_P$ be the effective migration rate. This is because both seed and pollen flow contribute to migration of paternally inherited haploid genes. Using these equalities $F_{st(p)}$ can be written in the same form as equation (6.12c).

6.2.4 Maternally inherited haploid genes

Similarly for maternally inherited haploid genes, we can obtain

$$F_{st(m)} = \frac{1}{N-(1-m_s)^2(N-1)} \tag{6.14a}$$

$$\approx \frac{1}{1+2Nm_s} \tag{6.14b}$$

if m_s is small. It is necessary to note that in deriving equation (6.14a) pollen flow and random mating are not considered because only *seed flow* contributes to migration of maternally inherited haploid genes. Let $\tilde{N} = N$ be the effective number of haploid genes and $\tilde{m} = m_s$ be effective migration rate. Then $F_{st(m)}$ has the same form as equation (6.12c). The above results show that the complex expressions for F_{st} derived for each genome can be reduced to Wright's general formula by substitution of appropriate values for effective population size and migration rate specific to the different genomes

6.3 Stepping stone model

In this section we incorporate seed and pollen flow into the stepping stone model using similar considerations and mathematical methods to those of Kimura and Weiss (1964),Weiss and Kimura (1965).

6.3.1 Assumptions

The basic assumptions are similar to those in the classical stepping stone model (Kimura and Weiss, 1964; Weiss and Kimura, 1965). An infinite array of populations lie on a Cartesian grid. Only one- and two-dimensions are in turn considered. Both forms of migration have two components: that between populations one step apart (m_{p1} for pollen and m_{s1} for seeds), and long distance migration ($m_{p\infty}$ for pollen and $m_{s\infty}$ for seeds) that draws pollen and seed from all populations. For the one step migration, half of this comes from each side. The number of seeds produced in each population is assumed to be large enough for sampling effects of pollen and ovules before seed formation to be ignored.

6.3.2 One dimensional case

6.3.2.1 Biparentally inherited diploid genes

Using the same notation as Weiss and Kimura (1965), let $p(i)$ be the gene frequency in population i and $p(i+k)$ be the gene frequency in the population k steps away from population i. Initially we assume that all populations comprise adult plants. When reaching reproductive stage, they produce pollen. Let $p^P(i)$ be the gene frequency in pollen grains after *pollen flow*, which can be written by the following equation according to the stepping stone model (Kimura and Weiss ,1964),

$$p^P(i) = (1 - m_{p1} - m_{p\infty})p(i) + \frac{1}{2}m_{p1}[p(i-1) + p(i+1)] + m_{p\infty}\bar{p} \qquad (6.15)$$

where m_{p1} stands for the rate of pollen migration per generation one step away from the donor population, $m_{p\infty}$ stands for the rate of long distance pollen dispersal per generation. Here the long- and short-distance pollen migrations are assumed to occur at each generation. For simplicity in mathematics the same shift operator S used by Weiss and Kimura (1965) is also employed to express equation (6.15), i.e.

$$\tilde{p}^P(i) = L_p\tilde{p}(i) \qquad (6.16)$$

where $\tilde{p}^p(i) = p^p(i) - \bar{p}$, $\tilde{p}(i) = p(i) - \bar{p}$ and $L_p = (1 - m_{p1} - m_{p\infty}) + \frac{1}{2} m_{p1}(S^{-1} + S)$.

The shift operator S is defined by the properties: $Sp(i) = p(i+1)$, $S^{-1}p(i) = p(i-1)$ (Weiss and Kimura, 1965, p132).

As in the case of the island model, after random combination between pollen and ovules the gene frequency in seeds so formed, $p^s(i)$, is

$$p^s(i) = \frac{1}{2}[p^p(i) + p(i)] \tag{6.17}$$

which is the arithmetic average of the gene frequencies in male and female gametes. Substituting equation (6.15) into equation (6.17), then we obtain

$$\tilde{p}^s(i) = L_s \tilde{p}(i) \tag{6.18}$$

where $L_s = \frac{1}{2}(2 - m_{p1} - m_{p\infty}) + \frac{1}{4} m_{p1}(S^{-1} + S)$

Similarly, after *seed flow* and then *sampling* the gene frequency in adults at the next generation, $p'(i)$, which is assumed to be the same as in seeds after *seed flow*, can also be expressed by

$$p'(i) = (1 - m_{s1} - m_{s\infty}) p^s(i) + \frac{1}{2} m_{s1}[p^s(i-1) + p^s(i+1)] + m_{s\infty}\bar{p} + \xi_s(i) \tag{6.19}$$

where m_{s1} stands for the rate of seed migration per generation one step away from the donor population, $m_{s\infty}$ stands for the rate of long distance seed migration per generation, and the $\xi_s(i)$ for the change of gene frequency due to *sampling*, with mean $E[\xi_s(i)] = 0$ and variance $V[\xi_s(i)] \approx p^s(i)[1 - p^s(i)] / 2N$. Again, the long- and short-distance seed migrations are assumed to occur at each generation. Putting equation (6.17) into equation (6.19), we can obtain

$$\tilde{p}'(i) = L\tilde{p}(i) + \xi_s(i) \tag{6.20}$$

where $$L = \sum_{j=0}^{2} \beta_j (S^{-j} + S^j) \tag{6.21}$$

in which $$2\beta_0 = 1 - \alpha_0 - 2\beta_1 - 2\beta_2$$

$$\alpha_0 = \frac{1}{2} m_{p\infty}(1 - m_{s\infty}) + m_{s\infty}$$

$$\beta_1 = \frac{1}{4}[m_{s1}(2 - m_{p1} - m_{p\infty}) + m_{p1}(1 - m_{s1} - m_{s\infty})]$$

$$\beta_2 = \frac{1}{8} m_{s1} m_{p1}$$

It can be seen from equation (6.21) that the gene frequency in adults at the next generation is ultimately affected by populations up to two steps away due to the two processes of gene flow (pollen and seed flow), even though only one step migration is considered for each process. This is because these two processes of gene flow are connected via the stage of *random mating*. Obviously, the situation is different from animal populations where only the two neighbouring populations exchange genes with the studied population if only one step migration is considered (Weiss and Kimura, 1965).

Since the L in equation (21) satisfies the relationship,

$$E[L\tilde{p}(k) L\tilde{p}(0)] = L^2 \rho(k) \qquad (k \neq 0) \tag{6.22}$$

where $\rho(k)$ is the unnormalized correlation function (Weiss and Kimura, 1965, p133), we can directly obtain the solution of the correlation of gene frequencies between populations k steps apart at steady state, $r(k)$, by substituting

$$H(\cos\theta) = \sum_{j=0}^{2} 2\beta_j \cos j\theta \tag{6.23}$$

into the equation (3.10) of Weiss and Kimura (1965, p134). The equation (6.23) was developed by Weiss and Kimura (1965, p134) to obtain the exact solution of $r(k)$.

According to Weiss and Kimura (1965), the general solution to the $r(k)$ can be obtained, i.e.

$$r(k) = \frac{A_1(k) + A_2(k)}{A_1(0) + A_2(0)} \tag{6.24}$$

where

$$A_1(k) = \frac{1}{4\pi}\int_0^{2\pi} \frac{\cos k\theta}{1 - H(\cos\theta)} d\theta \tag{6.25a}$$

$$A_2(k) = \frac{1}{4\pi}\int_0^{2\pi} \frac{\cos k\theta}{1 + H(\cos\theta)} d\theta . \tag{6.25b}$$

The equation (6.24) provides an exact solution for $r(k)$. However, ignoring the very small part β_2, we can obtain

$$A_1(k) = \frac{1}{2\sqrt{\alpha_0(\alpha_0 + 4\beta_1)}}\left(1 + \frac{\alpha_0}{2\beta_1} - \sqrt{\frac{\alpha^2{}_0}{4\beta^2{}_1} + \frac{\alpha_0}{\beta_1}}\right)^k \tag{6.26a}$$

$$A_2(k) = \frac{1}{2\sqrt{(2-\alpha_0)(2-\alpha_0 - 4\beta_1)}}\left(\sqrt{\frac{(2-\alpha_0-2\beta_1)^2}{4\beta^2{}_1} - 1} - \frac{2-\alpha_0-2\beta_1}{2\beta_1}\right)^k \tag{6.26b}$$

Justification of equation (6.26) can be easily obtained by comparison of equation (6.26) with the equation (4.4) of Weiss and Kimura(1965).

If the rates of short-distance migration are much larger than those of long-distance migration for both seed and pollen flow, i.e. $m_{s1} >> m_{s\infty}, m_{p1} >> m_{p\infty}$, according to discussion of Weiss and Kimura (1965, p136) we can see that $A_1(k)$ is much greater than $A_2(k)$. Therefore, the $r(k)$ can be approximated by

$$r(k) \approx A_1(k)/A_1(0) \tag{6.27a}$$

$$= e^{-\sqrt{\frac{\alpha_0}{\beta_1}}k} \tag{6.27b}$$

The equation (6.27b) is equivalent to the equation (1.13) developed by Kimura and Weiss (1964) for animal populations.

Now, consider population differentiation. The same notations as Weiss and Kimura (1965) are used. Let $\rho(0)$ be the variance of gene frequencies among populations. Using the method similar to Weiss and Kimura (1965), according to equation (6.18) we can obtain the variance of gene frequencies in seeds among infinite number of populations after *pollen flow* and *random mating*, $\rho^s(0)$, i.e.

$$\rho^s(0) = L_s^2 r(0)\rho(0) \tag{6.28}$$

Similarly, according to equation (6.20) we can obtain the variance of gene frequencies among populations in adults at the next generation, $\rho'(0)$, i.e.

$$\rho'(0) = L^2 r(0)\rho(0) + \frac{1}{2N}\left[\bar{p}(1-\bar{p}) - \rho^s(0)\right] \tag{6.29}$$

Putting equation (6.28) into equation (6.29) and letting $\rho'(0) = \rho(0)$, the general solution of the $F_{st(b)}$ $(= \rho(0)/\bar{p}(1-\bar{p}))$ at steady state can be obtained, i.e.

$$F_{st(b)} = \frac{1}{1 + 2N[(1-L^2)r(0) - \frac{1}{2N}(1-L_s^2)r(0)]} \tag{6.30}$$

If $m_{s1} = m_{p1} = 0$, but $m_{s\infty} \neq 0$ and $m_{p\infty} \neq 0$, according to equation(6.25) we can obtain

$$[(1-L^2)r(0)]^{-1} = \frac{1}{2\tilde{m}_\infty} + \frac{1}{2(2-\tilde{m}_\infty)} \approx \frac{1}{2\tilde{m}_\infty}, \quad \frac{1}{2N}(1-L_s^2)r(0) = \frac{1}{2N}\tilde{m}_\infty(2-\tilde{m}_\infty) \approx 0.$$

Thus the equation (6.30) reduces to equation (6.12b) in the infinite *island model.*

If the rates of short-distance migration are much larger than those of long-distance migration, i.e. $m_{s1} >> m_{s\infty}, m_{p1} >> m_{p\infty}$, and the β_2 is very small, according to the equation (3.11) and the discussions from equation (4.3) to equation (4.6) in Weiss and

Kimura (1965, p134) we can obtain $\{(1-L^2)r(0)\}^{-1} = A_1(0) + A_2(0) \approx A_1(0) = \frac{1}{2\sqrt{\alpha_0(\alpha_0+4\beta_1)}}$. Ignore the small part of the $\frac{1}{2N}(1-L_s^2)r(0)$, which is introduced by pollen flow. Then the equation (6.30) becomes

$$F_{st(b)} \approx \frac{1}{1+4N\sqrt{\alpha_0(\alpha_0+4\beta_1)}} \tag{6.31a}$$

$$\approx \frac{1}{1+2\tilde{N}\sqrt{2\tilde{m}_1\tilde{m}_\infty}} \tag{6.31b}$$

where $\tilde{m}_1 = m_{s1} + m_{p1}/2, \tilde{m}_\infty = m_{s\infty} + \tilde{m}_{p\infty}/2$, and $\tilde{N} = 2N$. The equation(6.31b) is equivalent to the equation (1.12) of Kimura and Weiss (1964) for animal populations.

6.3.2.2 Paternally inherited haploid genes

In order to avoid repeating procedures similar to those used for biparentally inherited genes, the main results are listed below. After *pollen flow*, the gene frequency in pollen of population i, $p^p(i)$, is

$$\tilde{p}^p(i) = L_p\tilde{p}(i) \tag{6.32}$$

where $\tilde{p}^p(i) = p^p(i) - \bar{p}$, $\tilde{p}(i) = p(i) - \bar{p}$, and the L_p is the same as that in equation (6.16). The gene frequency in resident seeds formed by random combination between pollen and ovules is the same as in male parents (pollen), i.e. $p^s(i) = p^p(i)$. After *seed flow* the gene frequency in adults at the next generation is

$$\tilde{p}'(i) = L\tilde{p}(i) + \xi_s(i) \tag{6.33}$$

where $L = \sum_{j=0}^{2}\beta_j(S^{-j} + S^j)$

in which $2\beta_0 = 1 - \alpha_0 - 2\beta_1 - 2\beta_2$

$\alpha_0 = (1-m_{s\infty})m_{p\infty} + m_{s\infty}$

$$\beta_1 = \frac{1}{2} m_{s1}(1 - m_{p1} - m_{p\infty}) + \frac{1}{2} m_{p1}(1 - m_{s1} - m_{s\infty})$$

and $$\beta_2 = \frac{1}{4} m_{s1} m_{p1}$$

The $\xi_s(i)$ is the change of gene frequency by sampling, with mean $E[\xi_s(i)] = 0$ and variance $V[\xi_s(i)] \approx p^s(i)[1 - p^s(i)] / N$. The correlation of gene frequencies between populations k steps apart at steady-state can be obtained by substituting the α_0 and β_1 in equation (6.33) into equation (6.27). However, $F_{st(p)}$ at steady state is different from the case of biparentally inherited genes because of paternally inherited haploid genes, and is shown to be

$$F_{st(p)} = \frac{1}{1 + N[(1 - L^2)r(0) - \frac{1}{N}(1 - L_p^2)r(0)]} \qquad (6.34)$$

If $m_s = m_p = 0$, but $m_{s\infty} \neq 0$ and $m_{p\infty} \neq 0$, then $F_{st(p)}$ is the same as that in the *island model* (equation (6.13)). If $m_{s1} >> m_{s\infty}$ and $m_{p1} >> m_{p\infty}$, and let $\tilde{m}_1 = m_{s1} + m_{p1}$, $\tilde{m}_\infty = m_{s\infty} + \tilde{m}_{p\infty}$, $\tilde{N} = N$, then $F_{st(p)}$ has the same form as equation (6.31b).

6.3.2.3 Maternally inherited haploid genes

The case of maternally inherited genes is exactly the same as the case of paternally inherited haploid genes except that no pollen flow occurs. The correlation of gene frequencies between populations k steps apart, $r(k)$, and the population differentiation, $F_{st(m)}$ can be immediately obtained by $m_{p1} = m_{p\infty} = 0$ in corresponding equations of paternally inherited haploid genes. For simplicity, these results obtained above are summarised in Table 6.1.

Table 6.1 Comparison of genetic differentiation F_{st} and genetic correlation $r(k)$ for three genomes with different modes of inheritance in an island and a one dimensional stepping stone model. Parameters $\tilde{N}$ and $\tilde{m}$ represent effective number of genes and effective rate of migration in each case.

Model	Biparental	Paternal	Maternal
Infinite island model			
$F_{st} = \frac{1}{1+2\tilde{N}\tilde{m}}$	$\tilde{N} = 2N$ $\tilde{m} = m_{s\infty} + m_{p\infty}/2$	$\tilde{N} = N$ $\tilde{m} = m_{s\infty} + m_{p\infty}$	$\tilde{N} = N$ $\tilde{m} = m_{s\infty}$
One-dimensional stepping stone model			
① If $m_{s1} = m_{p1} = 0, m_{s\infty} \neq 0, m_{p\infty} \neq 0$, F_{st} 's are the same as in island model			
② If $m_{s1} >> m_{p\infty}$ and $m_{p1} >> m_{p\infty}$.			
$F_{st} = \frac{1}{1+2\tilde{N}\sqrt{\tilde{m}_1\tilde{m}_\infty}}$	$\tilde{N} = 2N$ $\tilde{m}_1 = m_{s1} + m_{p1}/2$ $\tilde{m}_\infty = m_{s\infty} + m_{p\infty}/2$	$\tilde{N} = N$ $\tilde{m}_1 = m_{s1} + m_{p1}$ $\tilde{m}_\infty = m_{s\infty} + m_{p\infty}$	$\tilde{N} = N$ $\tilde{m}_1 = m_{s1}$ $\tilde{m}_\infty = m_{s\infty}$
$r(k) = e^{-\sqrt{\frac{2\tilde{m}_\infty}{\tilde{m}_1}}k}$	$\tilde{N} = 2N$ $\tilde{m}_1 = m_{s1} + m_{p1}/2$ $\tilde{m}_\infty = m_{s\infty} + m_{p\infty}/2$	$\tilde{N} = N$ $\tilde{m}_1 = m_{s1} + m_{p1}$ $\tilde{m}_\infty = m_{s\infty} + m_{p\infty}$	$\tilde{N} = N$ $\tilde{m}_1 = m_{s1}$ $\tilde{m}_\infty = m_{s\infty}$

6.3.3 Two dimensional case

6.3.3.1 Biparentally inherited diploid genes

As in the case of the one-dimensional stepping-stone model, the gene frequency in pollen after *pollen flow* in the population located at position (i,i), $p^p(i,i)$, can be written by

$$\tilde{p}^p(i,i) = L_p \tilde{p}(i,i) \tag{6.35}$$

where $\tilde{p}^p(i) = p^p(i) - \bar{p}$,

$$L_p = (1 - m_{p1x} - m_{p1y} - m_{p\infty}) + \frac{1}{2} m_{p1x}(S_1^{-1} + S_1)S_2^0 + \frac{1}{2} m_{p1y} S_1^0 (S_2^{-1} + S_2),$$

in which m_{p1x} and m_{p1y} stand for the rate of pollen migration per generation, and the S_1 and S_2 stand for shift operators along the x and y axes ($S_1 p(i,j) = p(i+1,j)$, $S_2 p(i,j) = p(i,j+1)$), respectively.

After random combination between male (pollen) and female (ovules) gametes, the gene frequency in seeds so formed, $p^s(i,i)$, is the arithmetic average of the gene frequencies in male and female gametes. Then we can show

$$\tilde{p}^s(i,i) = L_s \tilde{p}(i,i) \tag{6.36a}$$

where

$$L_s = \frac{1}{2}(2 - m_{p1x} - m_{p1y} - m_{p\infty}) + \frac{1}{4} m_{p1x}(S_1^{-1} + S_1)S_2^0 + \frac{1}{4} m_{p1y} S_1^0 (S_2^{-1} + S_2) \tag{6.36b}$$

After *seed flow*, the gene frequency in adults at the next generation is

$$\tilde{p}'(i,i) = L\tilde{p}(i,i) \tag{6.37}$$

where

$$L = \sum_{i=0}^{1}\sum_{j=0}^{1} \beta_{ij}(S_1^{-i} + S_1^i)(S_2^{-j} + S_2^j) + \beta_{02} S_1^0 (S_2^{-2} + S_2^2) + \beta_{20}(S_1^{-2} + S_1^2)S_2^0 \tag{6.38}$$

in which $2\beta_{00} = 1 - \alpha_0 - 2\beta_{01} - 2\beta_{10} - 4\beta_{11} - 2\beta_{02} - 2\beta_{20}$

$$\alpha_0 = \frac{1}{2}[(1 - m_{s\infty})m_{p\infty} + 2m_{s\infty}]$$

$$\beta_{01} = \frac{1}{4}[m_{s1y}(2 - m_{p1x} - m_{p1y} - m_{p\infty}) + m_{p1y}(1 - m_{s1x} - m_{s1y} - m_{s\infty})]$$

$$\beta_{10} = \frac{1}{4}[m_{s1x}(2 - m_{p1x} - m_{p1y} - m_{p\infty}) + m_{p1x}(1 - m_{s1x} - m_{s1y} - m_{s\infty})]$$

$$\beta_{11} = \frac{1}{8}(m_{s1x}m_{p1y} + m_{s1y}m_{p1x})$$

$$\beta_{20} = \frac{1}{8}m_{s1x}m_{p1x}$$

and $$\beta_{02} = \frac{1}{8}m_{s1y}m_{p1y}$$

It can be shown that the equation (6.22) still holds for the L in equation (6.38) (see Appendix IV). The equation (3.11) of Weiss and Kimura (1965) still holds under this case. Therefore, the correlation of gene frequencies between populations which are k_1 steps apart in the X direction and k_2 steps apart in the Y direction, $r(k_1, k_2)$, can be obtained by substituting

$$H(\cos\theta_1 \cos\theta_2) = \sum_{i=0}^{1}\sum_{j=0}^{1} 2\beta_{ij} \cos i\theta_1 \cos j\theta_2 + 2\beta_{20}\cos 2\theta_1 + 2\beta_{02}\cos 2\theta_2 \qquad (6.39)$$

into the equation (3.11) of Weiss and Kimura (1965). The analytic expression for the solution to $r(k_1, k_2)$ is not attempted due to the difficulties of calculation.

The general solution for $F_{st(b)}$ at steady state has the same form as equation (6.30), and can be obtained by substituting the L in equation (6.38) and the L_s in equation (6.36b) into equation (6.30), and replacing $r(0)$ in equation (6.30) with $r(0,0)$. Ignoring the small part of the $\frac{1}{2N}(1 - L_s^2)r(0,0)$ and using the equation (3.11) of Weiss and Kimura(1965, p134), we can show that $F_{st(b)}$ can be approximated by

$$F_{st(b)} \approx \left\{1+4\widetilde{N}\pi\left(\int_0^{2\pi}\int_0^{2\pi}\frac{d\theta_1 d\theta_2}{1-H^2(\cos\theta_1\cos\theta_2)}\right)^{-1}\right\}^{-1} \tag{6.40}$$

where $\widetilde{N} = 2N$. An analytic expression for equation (6.40) is not calculated further.

6.3.3.2 Paternally inherited haploid genes

After *pollen flow* the $\tilde{p}^p(i,i)$ has the same form as equation (6.35). The L_p is also the same as that in equation (6.35). After *seed flow* and *sampling* the L in the gene frequency in adults at next generation, $p'(i,i)$, is the same as that in equation (6.38) except that

$$\alpha_0 = (1-m_{s\infty})m_{p\infty} + m_{s\infty} \tag{6.41a}$$

$$\beta_{01} = \frac{1}{2}[m_{s1y}(2-m_{p1x}-m_{p1y}-m_{p\infty}) + m_{p1y}(1-m_{s1x}-m_{s1y}-m_{s\infty})] \tag{6.41b}$$

$$\beta_{10} = \frac{1}{2}[m_{s1x}(2-m_{p1x}-m_{p1y}-m_{p\infty}) + m_{p1x}(1-m_{s1x}-m_{s1y}-m_{s\infty})] \tag{6.41c}$$

$$\beta_{11} = \frac{1}{4}(m_{s1x}m_{p1y} + m_{s1y}m_{p1x}) \tag{6.41d}$$

$$\beta_{20} = \frac{1}{4}m_{s1x}m_{p1x} \tag{6.41e}$$

$$\beta_{02} = \frac{1}{4}m_{s1y}m_{p1y} \tag{6.41f}$$

The change of gene frequency by sampling, $\xi_s(i)$, is the same as that in equation(6.33), with mean $E[\xi_s(i)] = 0$ and variance $V[\xi_s(i)] \approx p^s(i)[1-p^s(i)]/N$. The $r(k_1,k_2)$ can be calculated using equation (3.11) of Weiss and Kimura(1965). $F_{st(p)}$ at steady state has the same form as equation(6.34) and can be approximated by the same formula as equation (6.40) except substituting α_0 and β's in equation (6.41) and $\widetilde{N} = N$ (haploid) into them.

6.3.3.3 Maternally inherited haploid genes

Because only seed flow contributes to migration, following similar procedure as in the case of biparentally or paternally inherited genes, we can obtain $\alpha_0 = m_{s\infty}$, $\beta_{01} = \frac{1}{2} m_{s1y}$ and $\beta_{10} = \frac{1}{2} m_{s1x}$. Compared with paternally inherited genes, the $r(k_1, k_2)$ and $F_{st(m)}$ can be obtained by putting $m_{p1}x = m_{p1y} = m_{p\infty} = 0$ into the corresponding equations.

6.4 Some properties of $r(k)$

The objective of this section is to find how the correlation of gene frequencies between populations k steps apart, $r(k)$, varies with seed flow and pollen flow in a one-dimensional stepping-stone model. It can be seen from equation (6.27) that $r(k)$ decreases monotonically with the ratio of long- to short-distance migration (α_0 / β_{10}). Thus, the ratio of α_0 / β_{10} plays a more important role than either of them separately in determining the correlation of gene frequencies. Generally an increase in the rate of the long distance migration (α_0) may reduce the correlation ($r(k)$), while an increase in migration from neighbouring populations may strengthen the genetic correlation. Therefore completely different effects on $r(k)$ are carried out by long- and short- distance migration.

Denote the correlation of gene frequencies between populations k steps apart by $r_b(k), r_p(k)$ and $r_m(k)$ for biparentally, paternally and maternally inherited genes, respectively. We can roughly obtain

$$r(k) = e^{-\sqrt{\frac{2\bar{m}_\infty}{\bar{m}_1}}k} \qquad (6.44)$$

according to equation (6.27) for each of the three genomes. If $m_{p1}/m_{s1} > m_{p\infty}/m_{s\infty}$, the correlation of gene frequencies for maternally inherited haploid genes is smaller than that for biparentally inherited diploid genes, which in turn is smaller than that for paternally

inherited haploid genes, i.e. $r_m(k) < r_b(k) < r_p(k)$. However, if $m_{p1}/m_{s1} < m_{p\infty}/m_{s\infty}$, then $r_m(k) > r_b(k) > r_p(k)$. In order to confirm these inferences about the properties of $r(k)$ outlined above, I directly calculated the genetic correlation according to the equation (3.10) of Weiss and Kimura (1965). Let $m_{p1} = 10^{-2}$, $m_{s1} = 10^{-4}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-6}$, i.e. $m_{p1}/m_{s1} > m_{p\infty}/m_{s\infty}$. The results indicate that $r_p(k) > r_b(k) > r_m(k)$ (Fig.6.2a). Letting $m_{p1} = 10^{-2}$, $m_{s1} = 10^{-3}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-7}$, i.e. $m_{p1}/m_{s1} < m_{p\infty}/m_{s\infty}$, the results indicate that $r_m(k) > r_b(k) > r_p(k)$ (Fig.6.2b). These are consistent with the inferences above. The correlation for biparentally inherited diploid genes, however, is very close to that for paternally inherited haploid genes.

In order to find the separate effects of pollen flow and seed flow on $r(k)$, we hold pollen flow and change seed flow, or fix seed flow and change pollen flow. Figure 6.3a shows that an increase in one-step seed flow (m_{s1}) may increase the genetic correlation for each of the three inherited types of genes. Similarly, figure 6.3b shows that an increase in one-step pollen flow (m_{p1}) can increase the $r(k)$ for paternally and biparentally inherited genes, but has no effect on maternally inherited genes.

Under the condition $m_{p1}/m_{s1} > m_{p\infty}/m_{s\infty}$, we can show that the correlation of gene frequencies changes faster with both seed migration (m_{s1}) and distance for maternally inherited haploid genes than for biparentally inherited diploid genes, which in turn is faster than for paternally inherited haploid genes. However, the change of the correlation of gene frequencies with pollen migration is faster for paternally inherited haploid genes than for biparentally inherited diploid genes.

A further important result is that for biparentally inherited diploid genes, the correlation of gene frequencies between populations changes faster with seed flow than with pollen flow. Therefore seed flow can potentially have much more influence than pollen flow in determining population structure of biparentally inherited genes. However, the correlation of gene frequencies between populations is affected to the same extent by seed and pollen flow for paternally inherited genes.

6.5 Estimation of the ratio of pollen flow to seed flow

For answering many practical questions about gene flow in plant populations it may be important both to estimate the ratio of pollen to seed flow between neighbouring populations and to estimate the same ratio for long-distance gene dispersal. The ratio from long-distance dispersal can be estimated with the help of the island model (Wright,1969; Ennos,1994).

In the case of one dimensional stepping-stone model, let $A=[\ln r_b(k)]^2, B=[\ln r_p(k)]^2$ and $C=[\ln r_m(k)]^2$. If the correlations of gene frequencies between populations can be obtained for both biparentally inherited gene such as nuclear DNA markers and for paternally inherited genes such as cpDNA markers in some conifers (Dong and Wanger, 1994; Neale, *et al.*,1986,1991), then the ratio of short range pollen to seed flow can be approximated by

$$\frac{m_{p1}}{m_{s1}} \approx \frac{A(1+m_{p\infty}/m_{s\infty})}{2B-A+(B-A)\cdot m_{p\infty}/m_{s\infty}} - 1 \qquad (6.45)$$

Where the correlations of gene frequencies between populations can be obtained for both biparentally and maternally inherited genes, the ratio of short range pollen to seed flow can be approximated by

$$\frac{m_{p1}}{m_{s1}} \approx \frac{C}{A}(2+\frac{m_{p\infty}}{m_{s\infty}}) - 2 \qquad (6.46)$$

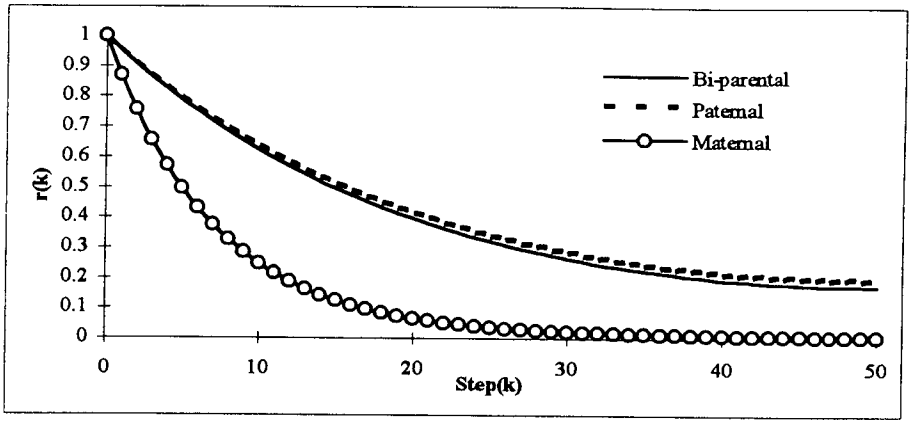
If the correlations for both paternally and maternally inherited genes are measured, the ratio of pollen to seed flow between neighbouring populations can be approximated by

$$\frac{m_{p1}}{m_{s1}} \approx \frac{C}{B}(1+\frac{m_{p\infty}}{m_{s\infty}}) - 1 \qquad (6.47)$$

If the long range migration can be ignored, the ratio of pollen to seed flow from short range migration can be estimated from equations (6.45), (6.46) and (6.47). However, it seems

Fig. 6.2 Comparison of the genetic correlation with distance between populations $r(k)$ for biparentally, paternally and maternally inherited genes at migration/drift equilibrium. Levels of short distance pollen migration, short distance seed migration, long distance pollen migration and long distance seed migration are: (a) $m_{p1} = 10^{-2}$, $m_{s1} = 10^{-4}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-6}$; (b) $m_{p1} = 10^{-2}$, $m_{s1} = 10^{-3}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-7}$.

(a)



(b)

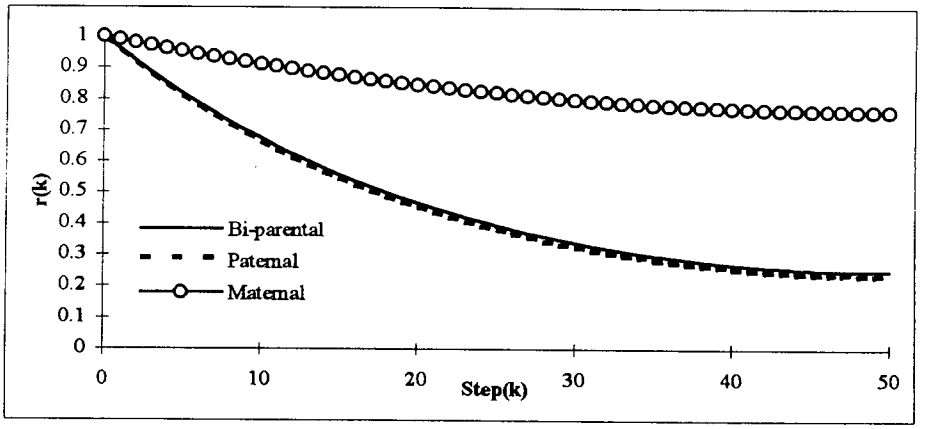
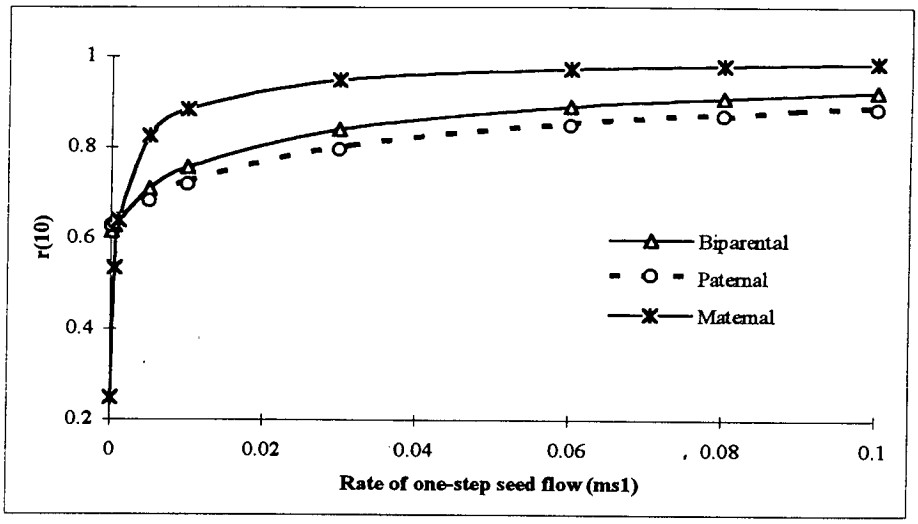
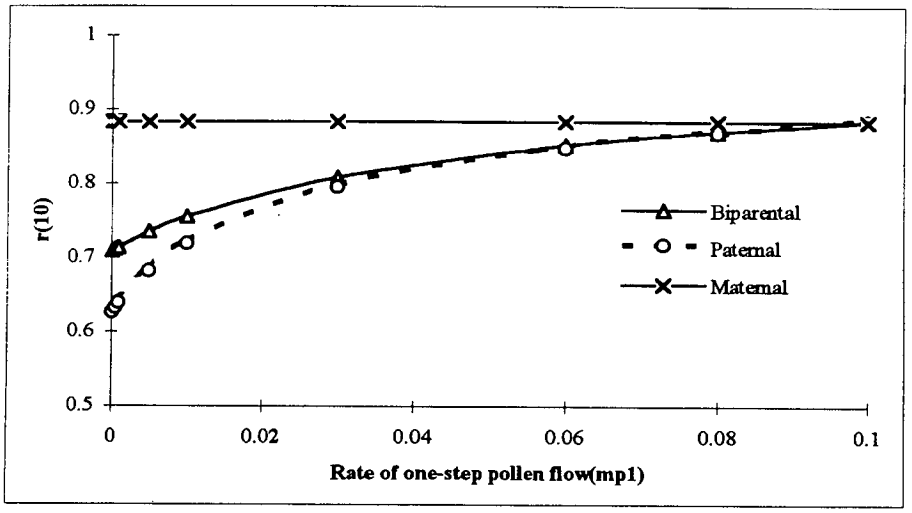


Fig. 6.3 Effects of short distance seed flow (a) and short distance pollen flow (b) on genetic correlation with distance $r(10)$ for biparentally, paternally and maternally inherited genes at migration/drift equilibrium. Fixed values for short and long distance pollen and seed migration are: (a) $m_{p1} = 10^{-2}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-6}$; (b) $m_{s1} = 10^{-2}$, $m_{p\infty} = 10^{-5}$, and $m_{s\infty} = 10^{-6}$

(a)



(b)



impossible to calculate correlation of gene frequencies between populations because the true expected gene frequency ($\bar{p} = E(p)$) is difficult to estimate in practice. Therefore, in this sense it is impossible to use the correlation of gene frequencies between populations to estimate the ratio of pollen flow.

6.6 Discussion

The aim of this chapter is to extend the existing island and stepping-stone models to plants in terms of gene frequency. Gene flow within and among plant populations is fundamentally different from gene flow in most animal populations in that the haploid generation (pollen in higher plants) is well adapted to, and may comprise the major means of gene dispersal. In order to understand fully the development of population structure in plants, dispersal via both seed and pollen needs to be incorporated explicitly into population genetic models. It is surprising that despite its importance, there have been relatively few attempts to take plant dispersal biology into account in developing population genetic models. (Wright, 1969; Crawford, 1984; Petit, *et al.*, 1993; Ennos, 1994; McCauley, 1995).

Studies of plant population genetic structure which take into account both pollen and seed flow have thus far concentrated on the island model (Petit, *et al.*, 1993; Ennos, 1994). Using F_{st} or G_{st} for measuring population differentiation, they show that the level of population differentiation at mutation/drift equilibrium decreases from maternally to paternally to biparentally inherited markers. Here we rederive this result for the island model from a consideration of variance in gene frequency among populations. We also show that the result is true under a stepping stone model of population structure.

In addition to quantifying levels of population differentiation at drift/migration equilibrium the stepping stone model can be used in theory to predict the spatial genetic structure of populations in terms of genetic correlations with distance. An important result to emerge is that the extent of genetic correlation for genomes with contrasting modes of inheritance is dependent upon the relative ratio of pollen to seed migration rates for short and long distance gene dispersal. Indeed the ordering of the size of genetic correlations for different genomes may be reversed with a change in the relative importance of pollen and seed in bringing about short and long distance gene flow.

Recent interest in this area has been stimulated by the application of molecular techniques in plant population biology, allowing the estimation of genetic structure for genes with different modes of inheritance (biparental, maternal and paternal). With an understanding of the influence of pollen and seed flow on genetic structure, estimates of the relative amounts of pollen and seed flow among populations can be inferred from comparison of F_{st} values for maternally and biparentally inherited markers (Ennos, 1994). However, consideration of the stepping stone model outlined here suggests that if data on genetic correlation with distance $r(k)$ can be collected, as well as data on F_{st} for genes with contrasting modes of inheritance, it is still difficult to estimate the relative importance of pollen and seed in both long and short distance gene dispersal among plant populations.

It should be remembered however that there are considerable theoretical and practical difficulties in applying models such as those outlined above to real populations for the purpose of estimating levels of pollen and seed flow. The first of these is that the assumptions of the models may not be met. In particular populations may not be at migration/drift equilibrium (McCauley, *et al.*, 1995) and founding events may contribute substantially to observed population structure.

The second major difficulty lies with obtaining estimates of F_{st} and $r(k)$ for the maternally and paternally inherited organelle genomes. Such genomes do not recombine and behave effectively as single genetic loci. As a consequence estimates of F_{st} and $r(k)$ are based on unreplicated sets of data. Errors in estimated values are unquantifiable and must be treated with considerable caution .

Despite these practical difficulties with testing and applying models of population structure under drift/migration equilibrium, they are extremely useful in providing the theoretical foundation for interpreting the growing number of studies in which genetic structure for nuclear and organelle genomes in plants are contrasted.

6.7 Summary

Gene flow occurs in two ways for hermaphrodite plants; seed flow and pollen flow. This produces asymmetrical migration for biparentally, paternally and maternally inherited genes, and may lead to different levels of population differentiation among them. In this chapter I incorporate seed flow and pollen flow into the classical island and stepping stone models of population structure. I evaluate their effects on population differentiation, and (in the stepping stone model) the correlation in gene frequency with distance for biparentally, paternally and maternally inherited genes. For both the island and stepping stone models differentiation for maternally inherited markers at migration/drift equilibrium is greater than for paternally inherited genes, which in turn is greater than that for biparentally inherited nuclear genes. In the stepping stone model the rate of decline of genetic correlation with distance is influenced by the relative values of long and short distance migration by seed and pollen. Differences in genetic correlation with distance among the differently inherited genes are conditional on the values of long and short distance migration for pollen and seeds.

CHAPTER 7

Estimation of the Ratio of Pollen to Seed Flow

7.1 Introduction

A variety of models can be used indirectly to estimate gene flow among populations of a species using data on genetic structure for selectively neutral markers (Barton, *et al*, 1986; Slatkin, *et al*, 1989; Slatkin 1989; Hudson, *et al*, 1992). When applied in plant species, especially hermaphrodite plants, gene flow should distinguish both pollen and seed flow. As mentioned in Chapter 6, seed flow and pollen flow may lead to asymmetrical migration for the biparentally inherited (nuclear), and maternally inherited (chloroplast and mitochondrial) genes that occur in angiosperm species and the paternally inherited (chloroplast) genes that occur in conifer species (Neale, *et al*, 1986, 1989, 1991). This produces different levels of population differentiation for the three variously inherited genomes. If the behaviour of genes with different modes of inheritance can be modelled, analysis of differences in genetic differentiation for these genes may allow estimation of the relative rates of pollen flow and seed flow (Ennos, 1994).

It is certain that the ratio of pollen to seed flow is an important parameter in plant population genetics, and is also useful in practice. Knowledge of the ratio of pollen to seed flow in natural populations can aid decision-making in establishing plantations of particular function, for example a seed orchard (Zobel and Talbert, 1984). If the ratio is very high in natural populations, this suggests that control of pollen flow may have to be practiced so as to avoid contamination of foreign genes.

Thus, the objective of this chapter focuses on methods for using a variety of population genetic statistics for estimating the ratio of pollen to seed flow in addition to those in Chapter 5. As a supplement to the extension of the classical model in Chapter 6, Wright's isolation by distance model is extended to plant populations. The first method for estimating the ratio employs data on F_{1s} measured in populations having a continuous distribution in space according to Wright's isolation by distance model (Wright 1943,1946). We then consider a simple model which describes the development of genetic distance between populations (Nei and Feldman, 1972), and relate Nei's distance to levels of seed and pollen flow. Finally, we briefly address the possible estimation of the ratio of pollen to seed flow from data on differences in DNA sequence between populations, and from gene phylogenies.

7.2 Wright's Isolation by Distance Model

In the isolation by distance model (Wright, 1943, 1946), an important parameter is the neighbourhood size which is defined as an area from which the parents of central individuals may be treated as if drawn at random. The calculation of neighbourhood area is relatively complicated when both pollen and seed dispersal are considered. Crawford (1984a,b) presented a modified formula for calculating neighbourhood size for a plant population, which will be used here. For both pollen and seed the distribution of dispersal distances between parents and offspring is assumed to be normal with mean zero. We assume that the nuclear biparentally inherited genes are diploid, and the paternally and maternally inherited genes are haploid, and only consider selectively neutral genes. We will use the same method of path analysis as Wright (1968) to analyse the population structures of the three differently inherited genes. Some of these results were, in fact, presented by Wright (1943,1946).

7.2.1 Biparentally inherited genes

Let σ_m^2 and σ_f^2 be the variance of the distances between male parents and offspring, and between female parents and offspring respectively. Also let σ_s^2 be the variance of seed dispersal, and σ_p^2 be the variance of the dispersal of pollen grains before seed formation. The number of individuals in the neighbourhood is $N_{(b)} = 4\pi(\frac{1}{2}\sigma_p^2 + \sigma_s^2)d$ in area continuity according to Crawford (1984b) and $2\sqrt{(\frac{1}{2}\sigma_p^2 + \sigma_s^2)\pi}\ d$ in linear continuity. Let N_p be the number of individuals in a neighbourhood after pollen dispersal and before seed formation which is equal to $2\pi\sigma_p^2 d$ (area). The number of individuals after seeds formation and dispersal in the neighbourhood is $N_f = 4\pi\sigma_s^2 d$ (area). Similarly the number of individuals after pollen flow and before seed is formed at ancestors of generation X is XN_p (area) or $\sqrt{X}N_p$ (linear), and for the individuals after seed dispersal is XN_f (area) or

$\sqrt{X}N_f$ (linear). The total number of individuals in the neighbourhood for both parents at ancestral generation X are $4\pi(\frac{1}{2}\sigma_p^2+\sigma_s^2)Xd$ (area) or $2\sqrt{(\frac{1}{2}\sigma_p^2+\sigma_s^2)\pi X}\ d$ (linear).

7.2.1.1 Drift case

Let F_{1s} be the correlation between ovules and pollen grains that contribute to zygotes after pollen and seed dispersal. According to the same considerations as Wright (1943,1946), the F_{1s} in area continuity can be approximately written by

$$F_{1s}=\frac{1}{N_{(b)}}b^2+(1-\frac{1}{N_{(b)}})F_{2s} \tag{7.1}$$

$$b^2=\frac{1+F_1'}{2} \tag{7.2}$$

Therefore the recurrence equations at ancestor of generation X in area continuity is

$$F_{XS}=\frac{1}{XN_{(b)}}b^2+(1-\frac{1}{XN_{(b)}})F_{(X+1)S} \tag{7.3}$$

For simplicity the calculation of F_{1s} after infinite generations can be expressed by

$$F_{1s}=\sum_1^\infty t/(2-\sum_1^\infty t) \tag{7.4}$$

where $t_1=\frac{1}{N_{(b)}}$, and $t_X=\frac{(X-1)N_{(b)}-1}{XN_{(b)}}t_{X-1}$.

For linear continuity the recurrence equations can be obtained by substituting the X in (7.3) by $\sqrt{X}$. We suppose that all populations are initially present as adults and produce pollen grains for dispersal, and the boundary condition is $F_{ks}=0$ after a large number of generations back (k).

7.2.1.2 Balance case

Where there is a balance between drift and long range dispersal of seeds and pollen grains, i.e. drift / migration equilibrium, let $m_{p\infty}$ be the proportion of male parents (pollen grains) replaced by pollen migration when random mating with ovules, and $m_{s\infty}$ be the proportion of both parents replaced by seed migration. If both long range dispersal and reversible mutation are considered, then $m_{p\infty}$ or $m_{s\infty}$ are just substituted by $m_{p\infty}+u$ or $m_{s\infty}+u$. Consider random sampling of size $N_{(b)}$, the proportion of male parents which makes a contribution to F_{1s} is $1-m_{p\infty}-m_{s\infty}$, while the proportion of female parents which contributes to F_{1s} is $1-m_{s\infty}$. Therefore , after seeds and pollen grains disperse,

$$F_{1s}=(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})[\frac{1}{N_{(b)}}b^2+(1-\frac{1}{N_{(b)}})F_{2s}]$$

$$F_{2s}=(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})[\frac{1}{2N_{(b)}}b^2+(1-\frac{1}{2N_{(b)}})F_{3s}] \text{ etc.} \qquad (7.5)$$

At steady state,

$$F_{1s}=\sum t/(2-\sum t) \qquad (7.6)$$

where $$t_1=(1-m_{s\infty})(1-m_{p\infty}-m_{s\infty})\frac{1}{N_{(b)}},$$

$$t_X=(1-m_{s\infty}-m_{p\infty})(1-m_{s\infty})\frac{(X-1)N_{(b)}-1}{XN_{(b)}}t_{X-1}$$

In the cases where only the pollen grains or seeds disperse, F_{1s} can be obtained by letting $N_f\to\infty, m_{s\infty}=0$ and $N_p\to\infty, m_{p\infty}=0$ respectively.

7.2.2 Paternally inherited genes

The number of individuals in the neighbourhood is $N_{(p)}=2\pi(\sigma_p^2+\sigma_s^2)d$ (area) or $\sqrt{(\sigma_s^2+\sigma_p^2)\pi}\ d$ (linear) due to individuals being haploid after the dispersal of both seeds

and pollen grains. The number of individuals in the neighbourhood after the dispersal of pollen grains but before seed formation is $N_p = 2\pi\sigma_p^2 d$ (area) or $\sqrt{\pi}\sigma_p d$ (linear), but the number of individuals in the neighbourhood after seeds dispersal is $N_f = 2\pi\sigma_s^2 d$ (area) or $\sqrt{\pi}\sigma_s d$ (linear). Similarly the number of individuals in the neighbourhood at ancestor of generation X is $2\pi X(\sigma_p^2 + \sigma_s^2)d$ (area) or $\sqrt{\pi X(\sigma_p^2 + \sigma_s^2)}d$ (linear).

7.2.2.1 Drift case

Here define F_{1s} as the correlation between adjacent individuals.

$$F_{1s} = \sum_{1}^{\infty} t_i \tag{7.7}$$

where $t_1 = \frac{1}{N_{(p)}}$ and $t_X = \frac{(X-1)N_{(p)} - 1}{XN_{(p)}} t_{X-1}$

7.2.2.2 Balance case

At steady state(drift / migration equilibrium),

$$F_{1s} = \sum_{1}^{k-1} t_i \tag{7.8}$$

where $t_1 = \frac{1}{N_{(p)}}(1 - m_{p\infty} - m_{s\infty})$ and $t_X = (1 - m_{p\infty} - m_{s\infty})\frac{(X-1)N_{(p)} - 1}{XN_{(p)}} t_{X-1}$.

For linear continuity the recurrence equations can be obtained by substituting $\sqrt{X}$ in place of X in (7.8). The boundary condition is $F_{ks} = 0$ a large number of generations (k) back.

In the case where only the pollen grains or seeds disperse, F_{1s} can be found by letting $N_f \to \infty, m_{s\infty} = 0$ and $N_p \to \infty, m_{p\infty} = 0$ respectively.

7.2.3 Maternally inherited genes

Since both paternally and maternally inherited genes are considered to be haploid or uniparental, the number of individuals in the neighbourhood is $N_{(m)} = 2\pi\sigma_s^2 d$ (area). Wright (1943) also addressed this case.

$$F_{1s} = \sum_{1}^{\infty} t_i \tag{7.9}$$

where $t_1 = \frac{1}{N_{(m)}}$ and $t_X = \frac{(X-1)N_{(m)} - 1}{XN_{(m)}} t_{X-1}$.

7.2.3.1 Balance case

At steady state (drift / migration equilibrium),

$$F_{1s} = \sum_{1}^{\infty} t_i \tag{7.10}$$

where $t_1 = (1 - m_{s\infty})\frac{1}{N_{(m)}}$ and $t_X = (1 - m_{s\infty})\frac{(X-1)N_{(m)} - 1}{XN_{(m)}} t_{X-1}$.

7.2.4 Comparison of population differentiation

In order to compare population differentiation among three genomes, we use the same notation as Wright (1943, p124). Consider a total population of size N_t, subdivided into H groups of intermediate size N_i and these are subdivided into K random groups of size N_u. Next we will compare the levels of population differentiation relative to N_i among the three genomes in the drift/migration balance case.

7.2.4.1 Biparental vs paternal genes

It can be seen that the neighbourhood size of biparentally inherited genes is greater than that of paternally inherited genes, i.e. $N_{(b)} > N_{(p)}$, and also t_i (i= *1, 2, ..., K*) in the case of

paternal genes is greater than that in the case of biparental genes according to (7.6) and (7.8). Therefore after going back to the ancestral generation K, $\sum_{1}^{K-1} t_i$ of paternal genes is greater than that of biparental genes. It can be shown that the correlation of paternally inherited genes, $F_{1s(p)}$, is greater than $F_{1s(b)}$ for biparentally inherited genes, i.e., $F_{1s(p)} > F_{1s(b)}$.

Similarly after going back to ancestral generation *KH*, it can be shown that the correlation of paternal genes, $F_{it(p)}$ is greater than $F_{it(b)}$ of biparentally inherited genes. We can also prove that

$$\frac{F_{it(p)} - F_{1s(p)}}{1 - F_{1s(p)}} > \frac{F_{it(b)} - F_{1s(b)}}{1 - F_{1s(b)}},$$

i.e.,

$$F_{st(p)} > F_{st(b)} \tag{7.11}$$

7.2.4.2 Paternal vs maternal genes

As above, we can prove the relationship

$$\frac{F_{it(m)} - F_{1s(m)}}{1 - F_{1s(m)}} > \frac{F_{it(p)} - F_{1s(p)}}{1 - F_{1s(p)}},$$

i.e.,

$$F_{st(m)} > F_{st(p)} \tag{7.12}$$

In summary the population differentiation of maternal genes is greater than that of paternal genes, which in turn is greater than that of biparental genes as long as the dispersal of seeds and pollen grains take place.

7.2.5 Ratio of pollen to seed flow

In this part we consider how to estimate the ratio of pollen to seed flow from long range dispersal. According to the Taylor expansion, $\sum_{1}^{\infty} t_i$ in (7.6) can be written using a simple formula, $\sum_{1}^{\infty} t = 1-[1-(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})]^{1/N_{(b)}}$. Similarly expressions can also be obtained for (7.8) and (7.10).

Let $A = 1-\left(\frac{1-F_{1s}}{1+F_{1s}}\right)^{N_{(b)}}$, $B = 1-\left(1-F_{1s}\right)^{N_{(p)}}$ and $C = 1-\left(1-F_{1s}\right)^{N_{(m)}}$ for biparentally, paternally and maternally inherited genes respectively. Then the ratio of pollen to seed flow from long range distance can be approximated by

$$\frac{m_{p\infty}}{m_{s\infty}} = \frac{A-B^2}{B-A} \text{ ,or } \frac{C^2-A}{C(1-C)} \text{ , or } \frac{C-B}{1-C} \qquad (7.13)$$

7.3 Nei's Genetic Distance

In this part we will incorporate seed flow and pollen flow into Nei's genetic distance (Nei, 1972) for three differently inherited genomes based on the assumptions of mutation / migration / drift equilibrium, as addressed by Nei and Feldman (1972) and Chakraborty and Nei (1974). Here we will use Nei and Feldman's model because of its simplicity and practicality.

Suppose that a population splits into two incompletely isolated populations and thereafter gene migration occurs in every generation between the two populations with a constant rate of both pollen and seed flow. Let N_1 and N_2 be the sizes of populations 1 and 2 respectively and assume that effective size is the same as the actual size.

Let m_{s1} and m_{p1} be rates of seed and pollen migration in population 1 respectively, and m_{s2} and m_{p2} be the rates of seed and pollen flow in population 2. Using the same notation as

Chakraborty and Nei (1974), let $J_{11}^{(t)}$ and $J_{22}^{(t)}$ be the probabilities of identity of two randomly chosen genes from population 1 and 2 respectively at generation t. Let $J_{12}^{(t)}$ be the probability of identity of two randomly chosen genes, one from each of the two populations. Each new mutation is different from the alleles pre-existing in any of the two populations. Only selectively neutral alleles are considered. Therefore the only way in which two genes can be the same "allele" is if they are identical by descent.

7.3.1 Biparentally inherited genes

Male parents for the biparental genes come from two sources: one comes from migration with frequency $m_{s1}+m_{p1}$, denoted by B; the other is from within populations with frequency $(1-m_{s1}-m_{p1})$, denoted by A. Similarly, female parents come from two sources: m_{s1} from migration, denoted by D, and $(1-m_{s1})$ from the population itself , denoted by C. The probabilities of two randomly chosen genes from population 1 come from A and A, B and B, A and B, etc. are $(1-m_{s1}-m_{p1})^2$, $(m_{s1}+m_{p1})^2$, $(1-m_{s1}-m_{p1})(m_{s1}+m_{p1})$, etc. respectively. Following Malecot (1969), we can derive the recurrence equation for $J_{11}^{(t)}$, which is:

$$\begin{aligned} J_{11}^{(t+1)} = (1-u_b)^2 \{ & \frac{1}{4}(AA+CC+2AC)[\frac{1}{2N_1}+(1-\frac{1}{2N_1})J_{11}^{(t)}] \\ & +\frac{1}{4}(2AD+2AB+2CD+2CB)J_{12}^{(t)} \\ & +\frac{1}{4}(BB+DD+2BD)[\frac{1}{2N_2}+(1-\frac{1}{2N_2})J_{22}^{(t)}]\} \end{aligned} \qquad (7.14a')$$

where u_b is the mutation rate for biparental genes. Substituting for A, B, C and D in (7.14a'), we can obtain (7.14a)

$$\begin{aligned} J_{11}^{(t+1)} = (1-u_b)^2 \{ & (1-m_{s1}-\frac{1}{2}m_{p1})^2[\frac{1}{2N_1}+(1-\frac{1}{2N_1})J_{11}^{(t)}] \\ & +2(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s1}+\frac{1}{2}m_{p1})J_{12}^{(t)} \end{aligned}$$

$$+(m_{s1}+\frac{1}{2}m_{p1})^2[\frac{1}{2N_2}+(1-\frac{1}{2N_2})J_{22}^{(t)}]\} \tag{7.15a}$$

Similarly, we can derive the recurrence equations for $J_{12}^{(t)}$ and $J_{22}^{(t)}$.

$$\begin{aligned}J_{12}^{(t+1)} &= (1-u_b)^2\{(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2})[\frac{1}{2N_1}+(1-\frac{1}{2N_1})J_{11}^{(t)}] \\ &+ (1-m_{s1}-\frac{1}{2}m_{p1})(1-m_{s2}-\frac{1}{2}m_{p2})+(m_{s1}+\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2})]J_{12}^{(t)} \\ &+(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s1}+\frac{1}{2}m_{p1})[\frac{1}{2N_2}+(1-\frac{1}{2N_2})J_{22}^{(t)}]\}\end{aligned} \tag{7.15b}$$

$$\begin{aligned}J_{22}^{(t+1)} &= (1-u_b)^2\{(m_{s2}+\frac{1}{2}m_{p2})^2[\frac{1}{2N_1}+(1-\frac{1}{2N_1})J_{11}^{(t)}] \\ &+2(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s2}+\frac{1}{2}m_{p2})J_{12}^{(t)} \\ &+(1-m_{s2}-\frac{1}{2}m_{p2})^2[\frac{1}{2N_2}+(1-\frac{1}{2N_2})J_{22}^{(t)}]\end{aligned} \tag{7.15c}$$

When $m_{p1}=m_{p2}=0$, the above equations reduces to those of Chakraborty and Nei (1974).

Using matrix notations, formulas (7.15a), (7.15b) and (7.15c) may be written as

$$\mathbf{J}^{(t+1)}=(1-u_b)^2\mathbf{T}+(1-u_b)^2\mathbf{M}\mathbf{J}^{(t)} \tag{7.16}$$

where

$$\mathbf{J}^{(t)'}=\left(J_{11}^{(t)},J_{12}^{(t)},J_{22}^{(t)}\right),$$

$$\mathbf{T}=\begin{pmatrix} \dfrac{(1-m_{s1}-\frac{1}{2}m_{p1})^2}{2N_1}+\dfrac{(m_{s1}+\frac{1}{2}m_{p1})^2}{2N_2} & \cdot \\ \dfrac{(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2})}{2N_1}+\dfrac{(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s1}+\frac{1}{2}m_{p1})}{2N_2} & \\ \dfrac{(1-m_{s2}-\frac{1}{2}m_{p2})^2}{2N_2}+\dfrac{(m_{s2}+\frac{1}{2}m_{p2})^2}{2N_1} & \end{pmatrix}$$

$$\mathbf{M}=\begin{pmatrix} (1-m_{s1}-\frac{1}{2}m_{p1})^2(1-\frac{1}{2N_1}) & 2(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s1}+\frac{1}{2}m_{p1}) & (m_{s1}+\frac{1}{2}m_{p1})^2(1-\frac{1}{2N_2}) \\ (1-m_{s1}-\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2}) & (1-m_{s1}-\frac{1}{2}m_{p1})(1-m_{s2}-\frac{1}{2}m_{p2}) & (1-m_{s2}-\frac{1}{2}m_{p2})(m_{s1}+\frac{1}{2}m_{p1}) \\ \cdot(1-\frac{1}{2N_1}) & +(m_{s1}+\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2}) & \cdot(1-\frac{1}{2N_2}) \\ (m_{s2}+\frac{1}{2}m_{p2})^2(1-\frac{1}{2N_1}) & 2(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s2}+\frac{1}{2}m_{p2}) & (1-m_{s2}-\frac{1}{2}m_{p2})^2(1-\frac{1}{2N_2}) \end{pmatrix}$$

Under steady state, the vector of equilibrium identity probabilities is given by letting $\mathbf{J}^{(t+1)} = \mathbf{J}^{(t)}$ in (7.16), i.e.

$$\mathbf{J} = (1-u_b)^2\{\mathbf{I}-(1-u_b)^2\mathbf{M}\}^{-1}\mathbf{T} \qquad (7.17)$$

Since Chakraborty and Nei (1974) have already discussed this equation in detail, we can use their results in later sections.

7.3.2 Paternally inherited genes

Here again suppose that the paternal gene is haploid. Its migration can also be mediated by both pollen flow and seed flow. Following similar consideration to those for biparental genes, the vector of equilibrium identity probabilities is

$$\mathbf{J} = (1-u_p)^2 \{\mathbf{I} - (1-u_p)^2 \mathbf{M}\}^{-1} \mathbf{T} \tag{7.18}$$

where u_p is the mutation rate of paternal genes, and

$$\mathbf{T} = \begin{pmatrix} \dfrac{(1-m_{s1}-m_{p1})^2}{N_1} + \dfrac{(m_{s1}+m_{p1})^2}{N_2} \\ \dfrac{(1-m_{s1}-m_{p1})(m_{s2}+m_{p2})}{N_1} + \dfrac{(1-m_{s2}-m_{p2})(m_{s1}+m_{p1})}{N_2} \\ \dfrac{(1-m_{s2}-m_{p2})^2}{N_2} + \dfrac{(m_{s2}+m_{p2})^2}{N_1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} (1-m_{s1}-m_{p1})^2(1-\dfrac{1}{N_1}) & 2(1-m_{s1}-m_{p1})(m_{s1}+m_{p1}) & (m_{s1}+m_{p1})^2(1-\dfrac{1}{N_2}) \\ (1-m_{s1}-m_{p1})(m_{s2}+m_{p2}) \cdot (1-\dfrac{1}{N_1}) & (1-m_{s1}-m_{p1})(1-m_{s2}-m_{p2}) + (m_{s1}+m_{p1})(m_{s2}+m_{p2}) & (1-m_{s2}-m_{p2})(m_{s1}+m_{p1}) \cdot (1-\dfrac{1}{N_2}) \\ (m_{s2}+m_{p2})^2(1-\dfrac{1}{N_1}) & 2(1-m_{s2}-m_{p2})(m_{s2}+m_{p2}) & (1-m_{s2}-m_{p2})^2(1-\dfrac{1}{N_2}) \end{pmatrix}$$

7.3.3 Maternally inherited genes

Consider that the maternally inherited genes are haploid. Only seed flow contributes to their migration. Under this case, the vector of equilibrium identity probabilities is

$$\mathbf{J} = (1-u_m)^2 \{\mathbf{I} - (1-u_m)^2 \mathbf{M}\}^{-1} \mathbf{T} \tag{7.19}$$

where the u_m is mutation rate of maternal genes, and

$$\mathbf{T} = \begin{pmatrix} \frac{(1-m_{s1})^2}{N_1} + \frac{m_{s1}^{\ 2}}{N_2} \\ \frac{(1-m_{s1})m_{s2}}{N_1} + \frac{(1-m_{s2})m_{s1}}{N_2} \\ \frac{(1-m_{s2})^2}{N_2} + \frac{m_{s2}^{\ 2}}{N_1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} (1-m_{s1})^2(1-\frac{1}{N_1}) & 2(1-m_{s1})m_{s1} & m_{s1}^{\ 2}(1-\frac{1}{N_2}) \\ m_{s2}(1-m_{s1})(1-\frac{1}{N_1}) & (1-m_{s1})(1-m_{s2})+m_{s1}m_{s2} & (1-m_{s2})m_{s1}(1-\frac{1}{N_2}) \\ m_{s2}^{\ 2}(1-\frac{1}{N_1}) & 2(1-m_{s2})m_{s2} & (1-m_{s2})^2(1-\frac{1}{N_2}) \end{pmatrix}$$

7.3.4 Ratio of pollen to seed flow

Here consider a special case where $u << m_{s1}, m_{p1}, m_{s2}, m_{p2} << 1$, which was addressed by Chakraborty and Nei (1974). Nei's distance for the three genomes are

$$D_b \approx \frac{2u_b}{m_{s1} + m_{s2} + \frac{1}{2}m_{p1} + \frac{1}{2}m_{p2}} \tag{7.20a}$$

$$D_p \approx \frac{2u_p}{m_{s1} + m_{s2} + m_{p1} + m_{p2}} \tag{7.20b}$$

$$D_m \approx \frac{2u_m}{m_{s1} + m_{s2}} \tag{7.20c}$$

where D_b , D_p and D_m stand for Nei's distance of biparental, paternal and maternal genes respectively.

Let $\tilde{m}_s = m_{s1} + m_{s2}$ and $\tilde{m}_p = m_{p1} + m_{p2}$. The ratio of pollen to seed flow is given by

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{2(a-1)}{2-a}, \text{ where } a = \frac{D_b u_p}{D_p u_b} \tag{7.21a}$$

or

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{2(1-a)}{a}, \text{ where } a = \frac{\dot{D}_b u_m}{D_m u_b} \tag{7.21b}$$

or

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{1-a}{a}, \text{ where } a = \frac{D_p u_m}{D_m u_p} \tag{7.21c}$$

7.4 Number of Nucleotide Differences

The variation in DNA sequence within and between populations contains much information on population evolution. Every sequence may be unique, and all the information is contained in the genealogical relationship between sequences (Barton and Wilson, 1995). Differences at the DNA level can be measured by the number of segregating sites among DNA sequences sampled (Watterson, 1975) or by the average number of (pairwise) nucleotide differences between DNA sampled (Tajima, 1983). For simplicity, only the average number of (pairwise) nucleotide differences between DNA is considered. If only two DNA sequences are sampled from a population, the expectation of the average number of nucleotide differences is equal to the expected number of segregating sites (Tajima, 1989).

Under a balance of migration / mutation / drift, the average number of pairwise nucleotide differences sampled within a population is independent of migration, but is related to migrations for pairwise DNA sampled between populations (Strobeck, 1987). This provide the foundation for estimating the ratio of pollen to seed flow.

From above the migration rate for biparental genes can be obtained directly , i.e. $m_s + \frac{1}{2}m_p$, while the migration rates for paternal and maternal genes are $m_s + m_p$ and m_s , respectively. We use similar notation to Strobeck (1987). In the island model with finite number of subpopulations, n, let $A = \frac{(n-1)u_b}{\hat{\xi}_{ij.b} - \hat{\xi}_{ii.b}}$, $B = \frac{(n-1)u_p}{\hat{\xi}_{ij.p} - \hat{\xi}_{ii.p}}$ and $C = \frac{(n-1)u_m}{\hat{\xi}_{ij.m} - \hat{\xi}_{ii.m}}$, where u represents mutation rate, $\hat{\xi}_{ij}$ and $\hat{\xi}_{ii}$ stand for the expected number of nucleotide differences between two randomly chosen DNA sequences from the same subpopulation and

from two different subpopulations respectively. Subscripts b, p and m stand for biparentally, paternally and maternally inherited genes respectively. The ratio of pollen to seed flow can be obtained by

$$\frac{m_p}{m_s} \approx \frac{B-C}{C}, \text{or } \frac{2(A-C)}{C}, \text{or } \frac{2(B-A)}{2A-B} \tag{7.22}$$

In the circular stepping-stone model, let $A = \frac{i(n-i)u_b}{\hat{\xi}_{i.b} - \hat{\xi}_{0.b}}$, $B = \frac{i(n-i)u_p}{\hat{\xi}_{i.p} - \hat{\xi}_{0.p}}$ and $C = \frac{i(n-i)u_m}{\hat{\xi}_{i.m} - \hat{\xi}_{0.m}}$, where $\hat{\xi}_i (i = 1,2,\ldots)$ stands for the expected number of nucleotide differences between two randomly chosen DNA sequences from two subpopulations which are i steps apart, and $\hat{\xi}_0$ from the same subpopulation. Under the balance of mutation / migration / drift, the ratio of pollen to seed flow can be obtained according to Strobeck (1987), which has the same formula as (4.1) except for different A, B and C.

7.5. Phylogenies

Another method that also uses DNA sequence information for estimating of the ratio of pollen to seed flow is based on the phylogenies of gene. Slatkin *et al* (1989, 1990) and Hudson *et al.* (1992) introduced a method for analyzing phylogenies of genes sampled from a geographically structured population. Using simulation, they showed that the minimum number of migration events (s) is a simple function of Nm based on phylogenies of alleles and genes under a variety of population structure models. This method depends on knowing the phylogeny of the nonrecombining segments of DNA that are sampled , but does not require complete sequences though it does assume that an accurate phylogeny can be inferred from the segments of DNA sampled (Slatkin, *et al*, 1989). Although the analytical expression, $s = f(Nm)$ has not been obtained to date, this nevertheless provides an additional potential method for estimating the ratio of pollen to seed flow among plant populations.

Following similar considerations to those above, for the biparentally inherited genome (nuclear DNA), both seed and pollen contribute to the migration events. Thus the

relationship between s_b, the minimum number of migration events between pairs of populations sampled, and number of migrants may be written:

$$s_b = f[N(m_s + \frac{1}{2}m_p)] \qquad (7.23)$$

Similarly, the minimum number of migration events between pairs of populations sampled should be related to both seed and pollen flow for paternally inherited genes, and to seed flow only in maternal genes. Therefore, there may be the following relationships,

$$s_p = f[N(m_s + m_p)] \qquad (7.24)$$

and

$$s_m = f(Nm_s) \qquad (7.25)$$

where s_p and s_m stand for the minimum number of migration events consistent with phylogeny for paternal and maternal genomes, respectively. By combining (7.23), (7.24) and (7.25) , it will be possible to estimate the ratio of pollen to seed flow once any two of these three relationships are available.

7.6 Ratio of movement in space

Basing on field observation, Bateman (1947) presented a simple formula to describe the distribution of pollen density with distance for either insect- or wind-pollination plants. Similar relationship between pollen density and distance was observed by J.W. Wright (1952). However, this formula is difficult to be used to estimate ratio of pollen to seed flow at population level. Slatkin (1993) pointed out an approximate log-log linear relationship between number of migrants and geographical distance. This relationship was tested by simulation under a variety of models and can be detected using different genetic markers (see Chapter 2). If isolation by distance exists for each of the three genomes, then it may provide the chance to explore the relationship between the ratio of pollen to seed flow (Nm) and the geographical distance(d).

For biparental genes, according to equation $Log_{10}(Nm) = a + bLog_{10}(d)$, we can obtain

$$N(m_s + m_P / 2) = 10^{a_1} \cdot d^{b_1} \tag{7.26}$$

where a_1 and b_1 are constants.

Similarly, we can obtain

$$N(m_S + m_P) = 10^{a_2} \cdot d^{b_2} \tag{7.27}$$

for paternal genes, and

$$Nm_S = 10^{a_3} \cdot d^{b_3} \tag{7.28}$$

for maternal genes.

Since the constant b is negatively related to the number of migrants *Nm*, larger *Nm* may lead to lower b value. Thus, it can be inferred that the following relationship exists

$$|b_3| < |b_1| < |b_2| \tag{7.29}$$

This is due to $Nm_S < N(m_S + m_P / 2) < N(m_S + m_P)$. Therefore, combining (7.26) with (7.27), we can obtain the relationship between the ratio of pollen to seed flow and geographical distance. That is

$$\frac{m_P}{m_S} = \left(2A \cdot d^B - 1\right)^{-1} - 1 \tag{7.30}$$

where $A = 10^{a_1 - a_2}$ and $B = b_1 - b_2$. The B is greater than zero, i.e. $B>0$.

Similarly, combining (7.26) and (7.28) we can obtain

$$\frac{m_P}{m_S} = 2\left(A \cdot d^B - 1\right) \tag{7.31}$$

where $A = 10^{a_1 - a_3}$ and $B = b_1 - b_3$ ($B<0$). Combining (7.27) with (7.28), we can obtain

$$\frac{m_P}{m_S} = A \cdot d^B - 1 \qquad (7.32)$$

where $A = 10^{a_2 - a_3}$ and $B = b_2 - b_3$ ($B<0$).

From (7.30), (7.31) and (7.32), it can be seen that the ratio of pollen to seed flow is reduced with geographical distance. One recent paper proves this qualitative relationship (McCauley, 1997). McCauley (1997) found that the ratio of pollen to seed movement was estimated as 6.4 at the largest spatial scale and 124.0 at the finest scale in *Silene alba.*

According to McCauley's (1997) hypothesis (Fig.7.1a), three critical values of geographical distance are important in describing the ratio of pollen to seed flow in space. These are the minimum (d_1) and maximum distance (d_2) where the relative contribution of pollen and seed flow is equal, i.e. $m_p / m_s = 1$, and the distance at which the ratio is maximum (d_m) (Fig. 7.1b). The above equations (7.30), (7.31) and (7.32), may likely reflect the relationship between d_m and d_2 (Fig.7.1b). Thus one critical value d_2 can be obtained according to equations (7.30), (7.31) and (7.32), i.e.

$$d_2 = \sqrt[B]{3/4A} \qquad \text{(from (7.30))}, \qquad (7.33a)$$

or

$$d_2 = \sqrt[B]{3/2A} \qquad \text{(from (7.31))}, \qquad (7.33b)$$

or

$$d_2 = \sqrt[B]{2 / A} \qquad \text{(from (7.32))} \qquad (7.33c)$$

It should be remembered here that this method is based on the existence of isolation by distance for any pair of the three genomes. Thus the first necessity is to detect the existence of such isolation by distance before using this method. Estimates of the distances d_1 and d_m can not be obtained using Slatkin's model. However, these two distances may occur within population rather than between populations. Thus, independent measurements are required.

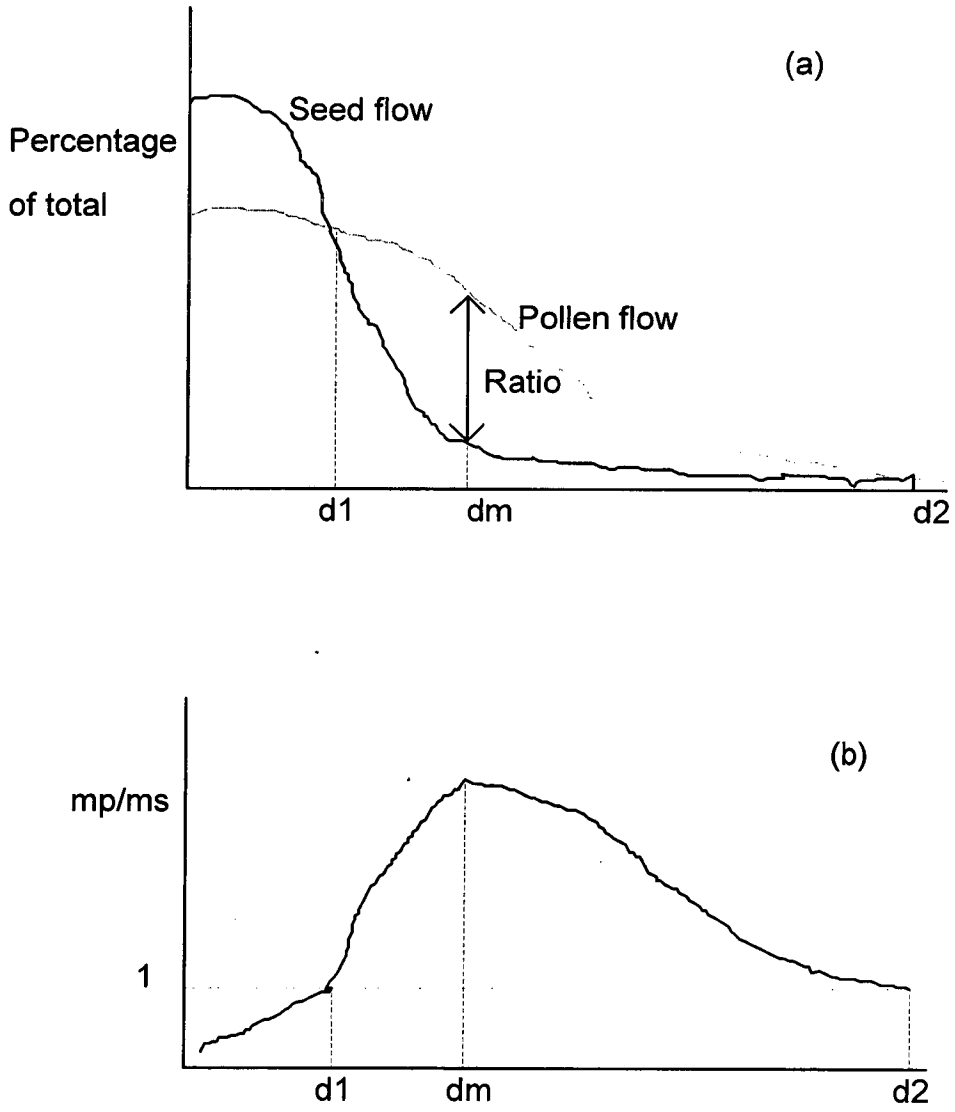


Fig.7.1 Hypothetical distance of dispersal distance of seeds and pollen illustrating how their relative contribution to total gene flow can vary with spatial scale (cited from McCauley, 1997; Fig.7.1a). Three critical values are indicated.

7.7. Discussion

One of the aims of this chapter has been to develop theory for population structure of plant genes with different modes of inheritance under isolation by distance. In the island model and the stepping stone models where populations are discretely distributed, differentiation for maternally inherited genes $F_{st(m)}$ is greater than for paternally inherited genes $F_{st(p)}$, which in turn is greater than for biparentally inherited genes $F_{st(b)}$ (Chapter 6). In this chapter we show that this relationship still holds in populations with a continuous distribution and limited dispersal of seeds and pollen.

In the isolation by distance case it is possible to obtain analytical expressions for estimating this ratio under the hypothesis of a balance between migration and drift (formula (7.13)). In practice this formula will be very difficult to apply. In the first place it requires estimates of neighbourhood size for the three different genomes. These are difficult to measure in the field (Levin & Kerster, 1968, 1971, 1974; Schaal, 1975; Crawford, 1984; Gliddon & Saleem, 1985). The model also assumes a random mating population, reaching an infinite number of generations back to its ancestors. If there is any self fertilisation, then F_{1s} will increase and the model assumptions will not be met.

Within the isolation by distance model it is possible to take into account deviations from random mating caused by self fertilisation. Let r be the proportion of the pollination randomly coming from the neighbourhood and 1-r be the proportion of self fertilisation. If there is no seed dispersal but pollen dispersal, the neighbourhood size at ancestors of generation X for the biparental genes is $4\pi((1+(X-1)r)\sigma_p^2/2+\sigma_s^2)d$ (area) or $2\sqrt{((1+(X-1)r)\sigma_p^2/2+\sigma_s^2)\pi d}$ (linear) according to Wright (1946).

Similarly the size of neighbourhood at ancestors of generation X for paternal genes is $2\pi((1+(X-1)r)\sigma_p^2+\sigma_s^2)d$ (area) or $\sqrt{((1+(X-1)r)\sigma_p^2+\sigma_s^2)\pi d}$ (linear). However, if both seed flow and pollen are considered, the calculation of neighbourhood size becomes very complicated.

Finally formula (7.13) will be difficult to apply in practice because the total number of individuals sampled in experimental work is always less than infinite. For this reason therefore $\hat{F}_{1s}$ may be underestimated. Taking all these points into consideration it is much more difficult to estimate the ratio of pollen to seed flow in the isolation by distance case than in either the island or stepping stone models of population structure (Chapter 6).

The second method explored in this paper for estimating the ratio of pollen to seed flow involved analysis of Nei's genetic distance. In order to apply the formulae (7.21a-c) derived here we must assume neutrality of mutations (Tajima, 1989b) and must possess estimates of the mutation rates in the three different genomes. There is evidence from analysis of rates of sequence divergence over evolutionary time that mutation rates differ significantly among the three plant genomes, with mutation rates being higher for nuclear genes than for chloroplast genes which in turn are higher than for mitochondrial genes (Birky, 1988). If mutation rates of three genomes were equal, genetic distances among the different genomes would vary according to the relationship $D_m > D_b > D_p$. Deviations from this predicted ordering of genetic distances could provide further evidence for large differences in the mutation rates of the three genomes.

The use of DNA sequence data to estimate the ratio of pollen to seed flow suffers from the same limitation as Nei's distance measure; we need to estimate mutation rate of the genes in the three genomes before the ratio of pollen to seed flow can be measured.

Furthermore it may be also be necessary to test the neutral mutation hypothesis before the formulae derived above can be applied. For these reasons it may be more practical to utilise statistics which rely only on the detection of differences between alleles i.e. F statistics rather than those which require measurement of the extent of genetic differences between alleles when indirectly estimating the ratio of pollen to seed flow. Great care should be taken even with these methods since their usefulness may only be judged once their variances, $Var(\frac{m_p}{m_s})$, are available. Finally we must remember that the assumption of strict maternal and paternal inheritance of organelle genomes underlies the models developed above. Further experimental data are required to confirm the general validity of these assumptions.

7.8 Summary

Gene flow occurs in two ways for hermaphrodite plants; seed flow and pollen flow. Dispersal of biparentally inherited (nuclear) and paternally inherited (conifer chloroplast) genes can be mediated by both seed and pollen, while for maternally inherited (angiosperm chloroplast and most mitochondrial) genes only seed flow contributes to dispersal. This produces asymmetrical migration for biparentally, paternally and maternally inherited genes and may lead to different levels of population differentiation among them. This chapter explores the effects of contrasting patterns of gene flow for different plant genes on their population structure under isolation by distance, on Nei's genetic distance measure, on divergence in nucleotide sequence between populations and on gene phylogenies. We discusses the possibilities of using data on population structure, genetic distance, sequence divergence and gene phylogenies as a basis for estimating the ratio of pollen to seed flow among subpopulations. One important general result from the isolation by distance model is that population differentiation for maternally inherited genes is greater than that for paternally inherited genes, which in turn is greater than that for biparentally inherited genes as long as the dispersal of seeds and pollen grains take place. This is consistent with results obtained previously for the island and stepping stone models in which populations are discretely distributed. If there is isolation by distance for any pair of the three genomes, it is possible to obtain the relationship between the ratio of pollen to seed movement with geographical distance.

CHAPTER 8

Genealogies and Geography

8.1 Introduction

With the development of modern molecular biology, use of DNA sequence data in population genetics will become popular in the future. In order to exploit this, theory that uses DNA sequence data to investigate plant population structure is clearly required. A new area of theoretical population genetics, the coalescent model, has been developed in recent years to facilitate this. One of the coalescent models applied in microevolution is gene genealogy, or the gene tree approach.

The whole family tree structure, the genealogy or coalescent process, is an important way of describing the evolutionary process of a population. Unlike traditional population theory, which was developed in terms of inbreeding coefficients (Wright, 1921) or probabilities of identities by descent (Malécot, 1969), analysis of the genealogy focuses on the times at which two or more genes have a common ancestor in the past. The results of traditional and genealogical theories are equivalent since they describe the same phenomenon of biological evolution (Slatkin, 1991), i.e. the consequences of inheritance, mutation and genetic drift. However, different types of genetic data are required for these two methods. Traditional theory uses allele frequencies while the genealogy analysis uses DNA sequence data.

The basic hypotheses that are usually employed in developing gene genealogy or phylogeny are: (i) Ideal Wright-Fisher model, which assumes that the number of offspring produced by each parent individual follows a Poisson distribution. Each individual of the previous generation has an equal probability of being the parent of any individual of the current generation. The population size is constant and generations are discrete (non-overlapping). For the details, see Ewens (1979). (ii) Constant neutral mutation process (Kimura, 1983), i.e. molecular clock hypothesis. (iii) Infinite-site model for the gene (Kimura, 1969), which means that any new mutation is assumed to be different from any pre-existing mutation.

As an example, under the above hypotheses and without recombination and selection, consider a sample of 5 individual diploid nuclear genes from a population with effective population size, N. If the probability for coalescence of more than two genes at the same time in the past is ignored, the coalescent process may look like Fig. 8.1. First, coalescence of one pair of genes among the five individual genes occurred at generation t in the past. The probability that any pair of genes comes from the same ancestor at previous generation is

$\binom{5}{2}/2N$. Thus, the distribution of the coalescent probability follows the geometric distribution, i.e. there is no coalescence for *t-1* generations in the past until it occurs at generation *t*. The probability for the coalescence at generation *t* in the past is $\binom{5}{2}\cdot\frac{1}{2N}\cdot\left(1-\binom{5}{2}\cdot\frac{1}{2N}\right)^{t-1}$, which can be approximated by an exponential distribution with mean $2N/\binom{5}{2}$ (=E(T5)). After the first coalescent of two genes with a mean number of generations E(T5) in the past, there are four distinct ancestors left. Similar consideration continues until coalescence of last two genes occurs. Theoretical results show that if there is a sample of n, the expected time of the whole genealogy is $4N(1-1/n)$ (Tajima, 1983).

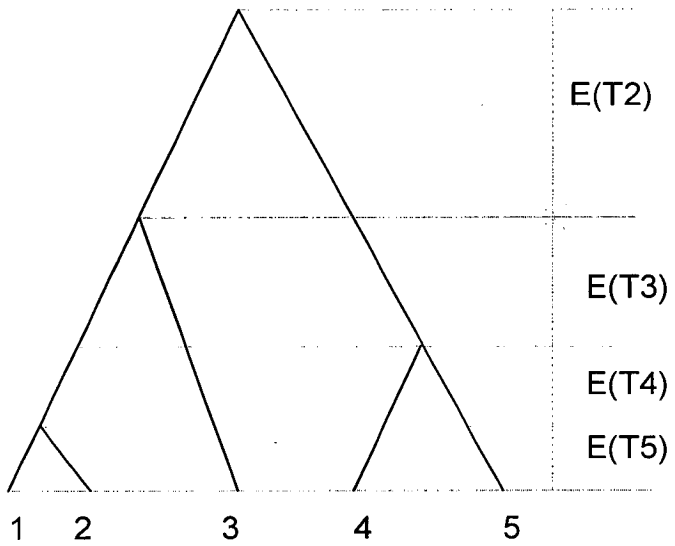


Fig. 8.1 A hypothetical coalescent process of a sample of 5 individual genes. E(T_i) (i=2,3,..5) stands for mean coalscent time of i distinct acestors.

A key important parameter in any of the coalescent analyses is the number of segregating sites (S) which can be *observed* from DNA sequence data (Tajima, 1993). Thus any complex evolutionary process will eventually be simplified into a formula for the number of segregating nucleotide sites. Since the molecular clock is assumed, the distribution of S is completely decided by the coalescent time (Hudson, 1992).

Thus, several advantages are involved in the gene genealogy over other models, such as the diffusion model in population genetics (Harding, 1995). Harding (1995) pointed out that coalescent models were appropriate for studying a wide rang of demographic histories including subdivision, constancy, expansion and fluctuation. For example, subdivision may slow down the rate of coalescence and will stretch the tree further back into the past.

With respect to the subject of population genetic structure, there have been extensive studies on this process (Watterson, 1984; Tavaré, 1984; Takahata and Nei, 1985) since the introduction of coalescence theory (Kingman, 1982a,b). These studies focused at first on completely isolated populations. The coalescent process for samples randomly chosen from partially isolated populations was then addressed, such as two partially isolated populations (Takahata, 1988; Takahata and Slatkin, 1990) or several partially isolated populations (Takahata, 1991; Nei and Takahata, 1993), using either island or stepping stone model (Notohara, 1990; Slatkin, 1991). Slatkin (1991) obtained the relationship between probabilities of identity by descent and the distribution of coalescence times, indicating that the coalescent model and traditional population genetic model are equivalent because both describe biological phenomena involving inheritance over time. However, the populations addressed by these authors were discretely distributed in space (island model and stepping stone models). Barton and Wilson (1995) recently developed a method which can be used in populations with a continuous distribution.

If the coalescent process is considered for a sample taken from partially isolated or continuously distributed populations of hermaphrodite plants, the situation becomes more complicated. This is because gene flow among plant populations can be mediated by seed flow, pollen flow, or both seed and pollen. Furthermore, there is asymmetric migration for the three differently inherited plant genomes (Petit, *et al*, 1993b; Ennos, 1994; and previous Chapters). Therefore a study of the influence of seed and pollen flow on the genealogy of the three plant genomes maybe provide an important insight into evolutionary process of

geographically structured or unstructured populations. Models of the coalescent process which incorporate both seed and pollen flow are clearly required.

This chapter will thus extend these existing results on genealogy theory to plant populations. We will first consider a simple case, the genealogy of two partially isolated populations (Takahata, 1988; Takahata and Slatkin, 1990), and demonstrate how to incoporate seed/pollen flow into the coalescent process when the Markov chain method is used. Then a general case of L ($L \geq 2$) partially isolated populations is considered using Nei and Takahata's method (Nei and Takahata, 1993). After that the results of the coalescent process for continuously distributed populations (Barton and Wilson, 1995) is extended to plant species. Practical implications of these theoretical results are then discussed.

8.2 General assumptions

For the three genomes, paternally and maternally inherited organelle genomes (cpDNA and mtDNA respectively in most conifers) are assumed to be haploid. Bi-parentally inherited nuclear genomes (nDNA) are assumed to be diploid. Only selective neutral genes without recombination are considered. There are no linkage disequilibrium among the three genomes. These assumptions are the same as in Ennos (1994).

The basic biological framework for investigating genealogy in discretely distributed population of plants linked by seed and pollen flow is outlined below. Our considerations begin with adults in each subpopulation at generation t. These adults produce pollen grains and ovules. Pollen dispersal occurs among subpopulations. In each subpopulation, pollen grains including the migrant fraction, randomly fertilise ovules (randomly mating assumption). Seeds so formed disperse among subpopulations, the process of seed flow. Each subpopulation contains a small proportion of migrant seeds. After seed flow, a fixed number of seeds is sampled and these grow up to form adults at the next generation $t+1$. The same process continues from generation to generation.

8.3 Populations with discrete distributions

In this section, we first demonstrate how to incoporate seed and pollen flow into the coalescent process and into the results presented by Takahata (1988) for two two partially isolated populations. Then, a general analytic expression for coalescent time suitable for any L (≥ 2) partially isolated populations is addressed.

8.3.1 Two partially isolated populations

The objective of the following is to extend the results obtained by Takahata (1988) to plant genomes. Therefore, all the assumptions in that paper are valid in here. The results presented in Takahata (1988) are applicable to maternal genes in the present paper where the migration of maternal genes is mediated by seed flow only. Therefore, substituting the migration rate in Takahata (1988) by effective migration rates of seeds, mean and variance of coalescent times are available immediately. For example, if the sample size is $n = 2$, the probability density of coalescence time can be given by

$$f(t) = \frac{2M}{A}\{\exp[-(1+2M-A)t] - \exp[-(1+2M+A)t]\}$$

where $A = \sqrt{1+4M^2}$, $M = 4Nm$, and t is meaured in units of $2N$ generations (Eq. (25)) of Takahata, 1988; or p332 of Takahata and Slatkin , 1990). In the following, we present the case of paternally and bi-parentally inherited genes.

8.3.1.1 Paternally inherited haploid organelle genomes

The migration can be mediated by both seed and pollen flow. Let m_s and m_p be effective migration rates of seed and pollen per generation, respectively, between the two populations (X and Y). Suppose that the generations are discrete and counted backward from the time at which n_0 individuals (adults) are randomly sampled without replacement from these two subpopulations (T=0). The remainder of the assumptions are the same as in Takahata (1988). For a given generation T, there may be n subsets ($1 \leq n \leq n_0$) in which any two individuals share a common ancestor. The configuration of n ancestral lineages T generations ago can be

described by an integer j in $S_n = \{0,1,2,\ldots,n\}$. Here j is the number of individuals drawn from population X at ancestral generation T. Because there are two episodes of migration (pollen and seed flow) within a generation, there are two transitions from state i ($i \in S_n$) before migration to state j ($j \in S_n$) after migration via state l ($l \in S_n$).

Following a similar derivation to that of Takahata (1988, p214), the transition probability matrix , $\mathbf{M_P}$, can be obtained after *pollen flow*.

$$\left.\begin{aligned} M_{p.ll} &= 1 - nm_p + O(m_p / N) \\ M_{p.l,l+1} &= (n-l)m_p + O(m_p / N) \\ M_{p.l,l-1} &= lm_p + O(m_p / N) \\ M_{p.li} &= O(m_p / N) \end{aligned}\right\} \qquad (8.1)$$

where $O(x)$ stands for the order of magnitude of x, and $M_{p.li}$ stands for the transition probability from state i before pollen migration to state l after pollen migration. It should be mentioned here that, in deriving equation (8.1), the configuration after migration is determined by two probabilities which follow hypergeometric distributions (Takahata, 1988, p214).

Similarly, after *seed flow*, the transition probability matrix from state l to state j ($l, j \in S_n$), $\mathbf{M_S}$, can be obtained following similar considerations. Elements of the $\mathbf{M_S}$ can be obtained by replacing m_p with m_s in equation (8.1).

It should also be noted here that a key assumption is made. This is that state l still belongs to one element of configuration of sets S_n. An alternative to this may be to assume that there is a temporary configuration, S_n', after pollen flow, which returns to S_n after seed flow.

Combining the two migrations, the transition probability matrix , **M**, is the product of $\mathbf{M_P}$ and $\mathbf{M_S}$, i.e.,

$$\mathbf{M} = \mathbf{M_P}\,\mathbf{M_S} \qquad (8.2)$$

Considering m's <<1 and ignoring all items including products of $m_s m_p$ or less than, the elements of matrix **M** can be obtained by replacing m_p with $m_s + m_p$ in equation (8.1). Therefore, Takahata's (1988) results still hold in plant populations for paternally inherited haploid organelle genes by using the above minor modification.

8.3.1.2 Biparentally inherited diploid nuclear genomes

When considering biparental genes (diploid) and using a Markov chain to model the coalescent process of a sample from a structured population, the calculation is slightly different from uniparentally inherited haploid organelle genomes. It is incorrect to consider half the number of haploid genes to be the number of the diploid individuals sampled in order to keep the number of genes analysed constant. Likewise it is incorrect to consider the total number of diploid genes to be double the number of haploid genes. The probabilities are different for sampling the same number of genes from adults and from gametes. For example, consider the probability for sampling n_0 genes that comprise a number n_A of allele A, and a number n_a of allele a from a population under Hardy-Weinberg equilibrium. The probability for a random sample from adults is $P(n_{AA}, n_{Aa}, n_{aa}) = 2^{n_{Aa}} p^{n_A} q^{n_a} n!/n_{AA}!n_{Aa}!n_{aa}!$ where the p and q are frequencies of allele A and a, respectively, and $2n_{AA} + n_{Aa} = n_A$ and $2n_{aa} + n_{Aa} = n_a$. However, the probability for a random sample from gametes is $P(n_A, n_a) = (2n)! p^{n_A} q^{n_a} / n_A! n_a!$. They are different patterns of distribution. The suggestion of Takahata and Slatkin (1990, p332) and Hudson (1992, p8) who consider the number of diploid individuals to be half the number of haploid genes, is a very approximate treatment for diploid genes.

Under this case, we straightforwardly consider a sample of $2n_0$ individual genes from a pool of gametes, the mixed ovules and pollen pool. Thus, the coalescent analysis introduced by Takahata (1988) still holds except that double the sample size is used. Therefore, as for the analysis in the case of paternally inherited haploid genes, we can obtain the transition probability matrix, **M**,

$$\left.\begin{aligned} M_{jj} &= 1 - n(\frac{1}{2}m_p + m_s) \\ M_{j,j+1} &= (n-j)(\frac{1}{2}m_p + m_s) \\ M_{j,j-1} &= j(\frac{1}{2}m_p + m_s) \\ M_{ji} &= 0 \end{aligned}\right\} \qquad (8.3)$$

Therefore, by substituting m and N by $m_s + \frac{1}{2}m_p$ and $2N$, respectively, Takahata's (1988) results holds for biparentally inherited diploid genes in plant populations. However, the above sampling method is difficult to follow in practice because the extraction of DNA from pollen grains is difficult.

Although the incorporation of seed and pollen flow into the coalescent process using the Markov chain method has been demonstrated above, it is very difficult to obtain a simple analytical expression suitable for practical use that makes use of data on DNA polymorphism of the three differently inherited plant genomes. In the following, a simple analytic expression for the coalescent process is obtained following the method used by Nei and Takahata (1993).

8.3.2 L (L ≥ 2) partially isolated populations

For a population with constant effective size (N_e individuals) per generation (Wright-Fisher's model), the mean coalescent time back for a sample of n individual genes, $E(T)$, is

$$E(T) = 4N_e(1 - 1/n) \qquad (8.4)$$

and its variance, $V(T)$,

$$V(T) = (4N_e)^2 \sum_{i=2}^{n} [1/i(i-1)]^2 \qquad (8.5)$$

which was shown by Tajima (1983). If constant mutation rate, μ, is assumed, the total number of segregating sites, $E(S)$, and its variance, $V(S)$, can be obtained

$$E(S) = 4N_e\mu a \tag{8.6}$$

$$V(S) = E(S)[1 + E(S)b / a^2] \tag{8.7}$$

where $a = \sum_{i=1}^{n-1} 1/i$ and $b = \sum_{i=1}^{n-1} 1/i^2$ (Watterson, 1975; Hudson, 1992). The above formulae are standard results of coalescent theory.

The effective population size of haploid genes is assumed to be half that of diploid genes. However, this assumption can be eliminated by letting N_m and N_f be effective population size of maternally and paternally inherited genes. For simplicity, this assumption is used in the following analysis. Therefore, estimates of the above parameters for paternally and maternally inherited haploid organelle genes can be obtained by replacing N_e with $N_e / 2$ in equations from (8.4) to (8.7), respectively.

Now we consider the case where the population is subdivided into L subpopulations. It can be seen from equations (8.4) to (8.7) that the coalescent analysis of a sample of n individuals randomly drawn from the population can be obtained by replacing the different effective population size in them. This is the method used by Nei and Takahata (1993). The key problem is to calculate the effective population size of the population that is divided into L subpopulations. Nei and Takahata (1993) pointed out that the effective population size of the whole population in this case was obtained by Wright (1943), i.e.

$$N_e = \frac{LN}{1 - G_{st}} \tag{8.8}$$

where N is the effective subpopulation size. Equation (8.8) still holds for haploid genes (see Appendix V.1).

Thus, once the expression of the G_{st} for plant populations is obtained, the effective population size for each of the three plant geneomes can be calculated according to equation (8.8). Derivation of the G_{st} for finite island model is given in Appendix V.2. Following considerations similar to those of Nei and Takaha (1993), the effective population size is

$$N_e = LN\left[1+\frac{(L-1)^2}{2\tilde{N}\tilde{m}L^2}\right] \tag{8.9}$$

where

$$\tilde{m} = \begin{cases} \frac{1}{2}m_p + m_s, & \text{biparental genes} \\ m_p + m_s, & \text{paternal genes} \\ m_s, & \text{maternal genes} \end{cases}$$

$$\tilde{N} = \begin{cases} 2N & \text{biparental genes} \\ N & \text{paternal / maternal genes} \end{cases}$$

Therefore, the mean coalescent time and its variance, and the expected number of segregating sites and its variance for each of the three plant genomes can be immediately obtained by substituting the equation (8.9) into equations (8.4),(8.5), (8.6) and (8.7), respectively. For example, the mean coalescent time and the mean number of segregating sites are

$$E(T) = 2\tilde{N}L\left[1+\frac{1}{2\tilde{N}\tilde{m}}\left(1-\frac{1}{L}\right)^2\right]\cdot\left(1-\frac{1}{n}\right) \tag{8.10}$$

and

$$E(S) = 2\tilde{N}\tilde{\mu}aL\left[1+\frac{1}{2\tilde{N}\tilde{m}}\left(1-\frac{1}{L}\right)^2\right] \tag{8.11}$$

where $\tilde{\mu}$ is μ_b for biparentally inherited nuclear genes, μ_p for paternally inherited genes and μ_m for maternally inherited genes.

In particular, if only two genes are randomly sampled from the L subpopulations, then the a in equation (8.11) is equal to 1. According to equation (8.11), the expected number of segregating sites is

$$E(S) = 2\tilde{N}\tilde{\mu}L + \frac{L\tilde{\mu}}{\tilde{m}}\left(1-\frac{1}{L}\right)^2 \qquad (8.12)$$

If the L is large enough that $\left(1-\frac{1}{L}\right)^2 \approx 1$, then equation (8.12) is approximately the same as Strobeck (1987) for biparentally inherited nuclear genes in the finite island model. Thus the equation (8.11) provides a general case for sampling n (n >1) individual genes.

8.4 Population with a continuous distribution

The coalescent times described above are based on populations of discrete distribution. Barton and Wilson (1995) presented a method for calculating coalescent time in a continuously distributed population. Because of the difficulties in modelling populations that are continuously distributed, an ideal mathematical model to describe the biological situation is not available (Wright, 1943; Malécot, 1948, 1969; Felsenstein, 1975b). Both Wright's *isolation by distance* model (1943) and Malécot's model (Malécot 1969) cannot avoid clumping of population because there is lack of regulation of population density (Felsenstein, 1975b). However, the clumping can be avoided by considering the dispersal behaviour of offspring (Kawata, 1995). In this section, we first consider incorporation of seed and pollen dispersal into Barton and Wilson's model (1995). Then the coalescence process is re-analysed purely based on Wright's isolation by distance model (1943).

A key parameter in calculating the coalescent times (Barton and Wilson 1995, E.q. (11a), p54) is the neighbourhood size (N_b). When applied in plant population, the N_b's for the three plant genomes are not the same , i.e.,

$$N_b = \begin{cases} 4\pi(\frac{1}{2}\sigma_p^2 + \sigma_s^2)d, & \text{bi - parental genes} \\ 2\pi(\sigma_p^2 + \sigma_s^2)d, & \text{paternal genes} \\ 2\pi\sigma_s^2 d, & \text{maternal genes} \end{cases} \qquad (8.13)$$

which can be obtained by following Crawford's calculation (Crawford,1984; Hu and Ennos, 1997). Here the σ_p^2 and σ_s^2 in (8.13) stand for the variance of the distances between parents and offspring in pollen and seeds in two dimensional space, respectively. Dispersals of both pollen and seed are assumed to follow the normal distribution with mean zero and variance of σ_p^2 and σ_s^2, respectively. The d is the effective population density. Suppose that there is random mating between pollen and ovules in any neighbourhood at each generation. Following Wright's idea, the neighbourhood size at ancestral generation t is the product of t and N_b in two dimensional space, i.e. tN_b. These are the same assumptions as Barton and Wilson (1995) used in deriving their equation (11a). Thus, putting these parameters into the formula obtained by Barton and Wilson (1995, p54), the probability of coalescent times of any pair of genes at any generation is immediately available.

However, if we base the analysis on Wright's isolation by distance model (Wright, 1943), an alternative simple way to calculate the coalescent times can be obtained immediately. In the following we consider bi-parentally inherited nuclear genes. For the case of maternally and paternally inherited haploid genes, the following analyses require modification by replacing the half neighbourhood size of diploid nuclear genes ($2N_b$) with that for haploid organelle genes, the N_b in equation (8.13). Let $f(t)$ be the probability of coalescence at generation t in the past. For a sample of n individual genes, according to Wright's isolation by distance model, the probability of n distinct ancestor at generation k in the past, $g(k)$, is

$$g(k) = \prod_{i=1}^{n-1}\left(1 - \frac{i}{2N_b k}\right) \approx 1 - \binom{n}{2} \cdot \frac{1}{2N_b k} \qquad (8.14)$$

If there is no occurrence of coalescence of any pair of the two genes in the past *t-1* generations but one common ancestor occurs at *t* generation in the past, the probability of coalescent time, $f(t)$, can be obtained, i.e.

$$f(t) = [1 - g(t)] \cdot \prod_{i=1}^{t-1} g(i) \tag{8.15a}$$

$$= \binom{n}{2} \cdot \frac{1}{2N_b t} \prod_{i=1}^{t-1} \left[1 - \binom{n}{2} \cdot \frac{1}{2N_b i} \right] \tag{8.15b}$$

$$\approx \binom{n}{2} \cdot \frac{1}{2N_b t} \left[1 - \binom{n}{2} \cdot \frac{1}{2N_b} H_{t-1} \right] \tag{8.15c}$$

where $H_{t-1} = \sum_{i=1}^{t-1} 1/i$. If the population size is fixed per generation (Wright-Fisher's model), the equation (15a) reduces to $f(t) = \binom{n}{2} \cdot \frac{1}{2N_b} \exp\left[-\binom{n}{2} \cdot \frac{1}{2N_b} t \right]$, which is the standard results for coalescent theory in a completely isolated population. Compared with Barton and Wilson's (1995) model, equation (8.15) cannot provide additional information regarding geographical positions for the sampled genes, but it is the extension of the original coalescent theory to a plant population that is continuously distributed in space.

8.5 Implication and discussion

The aims of this chapter are to extend to plant (hermaphrodite) populations the existing coalescent theories describing geographically structured or unstructured (continously distributed) populations, and to compare the difference among three plant genomes differing in modes of inheritance. Two obvious implications from the above results can be obtained.

First, the above results provide the possibility of addressing a question of theoretical interest, that may not hold in some species (Kenneth, *et al.*, 1987), i.e. how the mode of inheritance and seed/pollen flow influence the coalescent process if mutation rates are assumed to the same between different genomes. Since there are different extents of migration rate and population size (discrete distribution model), or neighbourhood sizes (continuous distribution model) among the three genomes, these genomes should differ in

mean coalescent time. Denote the mean coalescent times of biparentally, paternally and maternally inherited genes, by $E(T_b)$, $E(T_p)$ and $E(T_m)$ respectively.

In the case of populations that are discretely distributed in space, according to the equation (8.10), it can be shown that $E(T_m) > E(T_p)$ and $E(T_b) > E(T_p)$. If the condition, i.e. $2NLm_s(\frac{2m_s}{m_p}+1) < \left(1-\frac{1}{L}\right)^2$, or roughly $2NLm_s < 1$ (if $m_p >> m_s$), is satisfied, then we can obtain $E(T_m) > E(T_b)$. For example a specific case is modelled in which we let n=10 , L=16 and N=30, the migration rate of pollen is fixed at $m_p = 0.0001$ and the rates of seed flow are changed from 10^{-6} to 0.01. The $E(T_m)$'s are always larger than the $E(T_b)$'s until $2NLm_s > 1$ (Fig.8.2). Therefore, the value of $2NLm_s$ is important in affecting the relative evolutionary processes of bi-parental and maternal genes. Therefore, mean coalescent times is shortest for paternal genes among the three genomes, and, under particular conditions, mean coalescent time is longest for maternally inherited genes.

In the case of a population that is continuously distributed in space, according to equation (8.15c), the mean coalescent time for biparentally inherited diploid nuclear genes is

$$E(T_b) = \sum_t t \cdot \binom{n}{2} \cdot \frac{1}{2N_b t}\left[1-\binom{n}{2}\cdot\frac{1}{2N_b}H_{t-1}\right] \qquad (8.16)$$

Expressions similar to equation (8.16) for the mean coalescent time of paternally and maternally inherited genes can be obtained. However, althought the sum of the left-hand side of equation (8.16) is convergent, it is difficult to make a judgement on the relationship among the three plant genomes in terms of the mean coalescent times. This is also the case in Barton and Wilson's (1995) model.

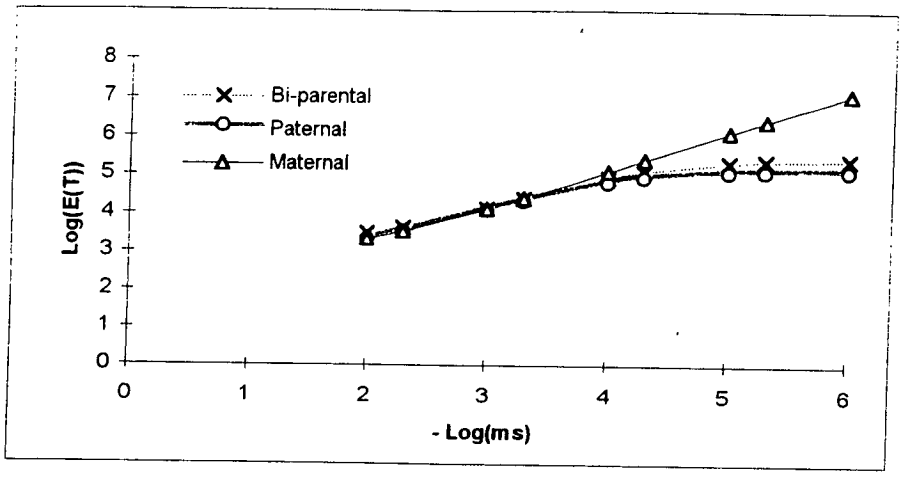


Fig.8.2 Comparison of mean coalescent times of three differently inherited genes with seed migration rates from 10^{-6} to 0.01. Other parameters are $n = 10$, $L = 16$, $N = 30$ and $m_p = 0.0001$.

Second, the above results also provide the possibility of estimating the ratio of pollen to seed flow, which is an important indicator of the relative contribution to migration between seed and pollen flow. Consider the case where populations are discretely distributed in space. If the same number of selectively neutral genes or markers randomly drawn from L subpopulations are sequenced among the three plant genomes, it is possible to estimate the number of segregating sites (Tajima, 1993; Watterson, 1975). Denote the expected total number of segregating sites within subpopulations investigated by $E(S_b)$ $(= 4N\mu_b La)$, $E(S_p)$ $(= 2N\mu_p La)$ and $E(S_m)$ $(= 2N\mu_m La)$ for biparentally, paternally and maternally inherited genes, respectively. These parameters can be estimated using DNA sequence data (Tajima, 1993) and are denoted by $\hat{S}_b'$, $\hat{S}_p'$ and $\hat{S}_m'$ respectively. Therefore, the ratio of mutation rates between two different genomes can be estimated under certain assumption. For example, the ratio of mutation rates between biparentally and paternally inherited genes is

$$\frac{\mu_b}{\mu_p} = \frac{1}{2} \cdot \frac{4N\mu_b La}{2N\mu_p La} = \frac{E(S_b)}{2E(S_p)} \approx \frac{\hat{S}_b'}{2\hat{S}_p'} \tag{8.17}$$

and its variance

$$V\left(\frac{\mu_b}{\mu_p}\right) = \frac{1}{4}\cdot\left[\frac{1}{(\hat{S}'_p)^2}V(\hat{S}'_b) + \frac{(\hat{S}'_b)^2}{(\hat{S}'_p)^4}V(\hat{S}'_p)\right] \tag{8.18}$$

Equation (8.18) is obtained according to Kendall and Stuart's (1969, p232) formula and the independence hypothesis between different genomes.

Similarly, let the estimates of the expected number of segregating sites among subpopulations be $\hat{S}_b$, $\hat{S}_p$ and $\hat{S}_m$ for biparentally, paternally and maternally inherited genes. Using equation (8.11), we can obtain

$$\frac{m_p}{m_s} = \left(1 - \frac{\hat{S}'_p}{\hat{S}'_b}\cdot\frac{\hat{S}_b - \hat{S}'_b}{\hat{S}_p - \hat{S}'_p}\right)^{-1} - 2, \tag{8.19a}$$

or

$$= \frac{\hat{S}'_p}{\hat{S}'_m}\cdot\frac{\hat{S}_m - \hat{S}'_m}{\hat{S}_p - \hat{S}'_p} - 1, \tag{8.19b}$$

or

$$= \frac{\hat{S}'_b}{\hat{S}'_m}\cdot\frac{\hat{S}_m - \hat{S}'_m}{\hat{S}_b - \hat{S}'_b} - 2 \tag{8.19c}$$

If the numbers of sampled individual genes are different among the three plant genomes, it is still possible to estimate of the ratio of pollen to seed flow by modificating the above equation. This result extends those obtained by Hu and Ennos (1997) to n ($n \geq 2$) genes investigated.

Using DNA sequence data to estimate gene flow has been reported before (Slatkin, *et al*, 1989,1990; Hudson, *et al*, 1992). They showed that the minimum number of migration events (s) is a simple function of Nm based on phylogenies of alleles and gene trees by computer simulation. This relationship existing in a phylogeny of alleles investigated can be reflected in terms of number of segregating sites among populations, which is shown in equation (8.11).

Among the variety of methods presented for estimating mean coalescent time in populations with discrete distributions, use of effective population size is the simplest for population

geneticists (Nei and Takahata, 1993; Takahata, 1991). However, it should be indicated that the N_e depends on the assumption that the whole population keeps the same structure for a long evolutionary time. This can be seen from the derivation of Wright (1943, p132-133). The detailed discussion can be found in Nei and Takahata (1993, p243).

In the method which uses the discrete- or continuous-time Markov chain for plants, an important assumption is that seed and pollen flow occurs together, or there is no new configuration of (S_n). This may not hold in natural populations. For instance, there is the possiblity that for a given sample, coalescence will occur after pollen flow at generation t in the past but before seed flow. However, since both pollen and seed flow occur within the same generation, and if effective rates of pollen and seed dispersal are much smaller than 1, this treatment is reasonable and simple.

When relating genealogies to geography, patterns of migration should also be considered besides number of migrants. In Wright's isolation by distance model, this can be reflected in the neighbourhood size in terms of variances of distance between parents and offspring in space. It is difficult to use the neighbourhood size to resolve two cases where both have the same variance of seed and pollen flow but different dispersal patterns, for example gaussian and exponential distributions. In island or stepping stone models, the results above can be used to represent constant migration rate per generation. Stochastic migration may increase the variance of estimated coalescent times. For example, stochastic migration may lead to increase in population differentiation, which in turn may cause increase in effective population size (N_e) of the whole population, and thus lead to longer mean coalescent time. However, qualitative relationship between coalescent times and migration should still hold.

Finally, we must remember that an important assumption for testing these results is that there exist selectively neutral genes or markers for each of the three genomes. Effects of migration, mutation and genetic drift are considered in the above analyses. Other effects, such as selection and recombination, are not considered. Thus, it is important to carry out a test for selective neutrality prior to using equation (8.19) to estimate the ratio of pollen to seed flow. A set of universal primers for amplification polymorphic non-coding regions of mtDNA and cpDNA in plants was reported recently (Taberlet, *et al.*, 1991; Demesure, *et al*, 1995). These primers provide a convenient way to amplify non-coding regions of plant

organelle genomes and to obtain their DNA sequences by PCR (polymerase chain reaction) based methods. Use of selectively neutral markers of cpDNA to address population structure has been reported in plant species. For example, reports were given for the use of the non-coding regions of cpDNA to investigate plant population structure (McCauley, 1994; JØhnk and Siegismund, 1997) and to infer postglacial migration (Ferris, *et al.*, 1995). These amplified non-coding regions may be selectively neutral. Thus, application of the theoretical results obtained in this chapter in practical work is possible in the foreseeable future. However, it is still remembered that the selectively neutral region are possibly linked to potentially selected loci. Thus, effects of hitchhiking and selection-sweeping deserve considerable attention when the above theoretical results are applied in practice.

8.6 Summary

This chapter extends to plants the existing theories on coalescence times for genotypes randomly chosen from geographically discrete or continuously distributed populations. Three plant genomes (nuclear DNA, chloroplast DNA and mitochondrial DNA), with different modes of inheritance are considered separately due to the differences in migration rate that they show. Results indicate that in the discrete model of populations, mean coalescent time is shortest for the paternally inherited genome (cpDNA in conifers) and, given certain conditions, is longest for the maternally inherited genome (cpDNA in angiosperms and mtDNA in conifers and angiosperms). Estimation of the ratio of pollen to seed flow from a sample of n ($n \geq 2$) individual genes is presented in terms of the number of segregating sites between and within populations. These results are difficult to obtain in a model of a population that is continuously distributed in space.

CHAPTER 9

Cline Theory for Haploid Organelle Plant Genomes

9.1 Introduction

The genes studied in previous chapters (6-8) are assumed to be selectively neutral. The roles of seed and pollen flow under the influence of natural selection have not been considered so far. As is mentioned in Chapter 5, clines are one of the most important characteristics of population genetic structure, which are associated with the effects of natural selection. In this chapter, genes under selection are considered, and the specific population structure, a cline, will be investigated so as to find the role that seed and pollen flow play in cline formation for genes located on plant genomes.

9.1.1 Definition

A cline has been defined as a gradient (decrease or increase) within a continuous population in the frequencies of different genotypes ("genocline") or phenotypes ("phenocline") in different localities (Rieger, *et al.*, 1991). However, the cline investigated in this chapter refers to the gradient change (increase or decrease) of *gene frequency*, not phenotypes or genotypes, with geographical distance.

9.1.2 Origin of clines

The origin of clines are very complex. Clinal situations are often associated with speciation (summarized in Endler, 1977). According to the description given by Endler (1977, p13), a cline is a temporal phase in the process of speciation. Three types of speciation can generate this temporal phase: sympatric, parapatric and allopatric speciation (Fig.9.1). An ancestral species may either spread or not spread over a spatially heterogeneous area. If it does, spatial divergence will occur and may result in two situations. One is that populations remain in contact (continuous range), and the genetic differentiation proceeds in adjacent contacting areas (parapatry) and further leads to formation of shallow clines (gradation) and steep clines (conjunction). The other is that populations become separated (disjunct), and the genetic differentiation proceeds in isolation (allopatry). Populations differentiated to some extent may meet again (secondary contact) and hence produce a cline. If the ancestral species does not spread over a spatially heterogeneous area, no spatial divergence occurs, but sympatric genetic divergence may occur due to ecological (e.g. habitat selection) and temporal segregation. Thus, clines may form in the adjacent contacting areas between these

diverged populations. The three paths for the formation of a cline are summarised in Fig. 9.1.

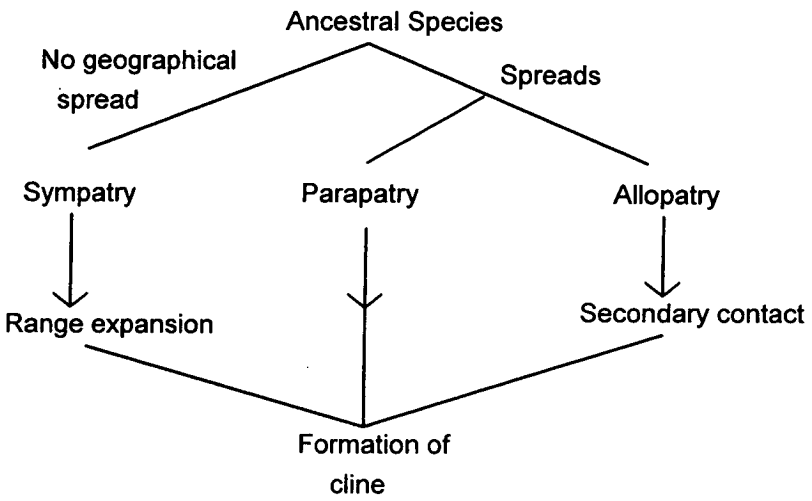


Fig.9.1 Possible paths to cline, redrawn after Endler (1977).

9.1.3 Modelling of clines

Theoretical studies of clinal situation go back to Fisher's pioneering work (1937). He studied the wave of advance of advantageous genes. The results were then extended to continuous changes in selection intensity (Fisher,1950). Haldane (1948) indicated that where one phenotype is favoured in one area and another phenotype in a neighbouring area (discontinuous selective intensity), the character in question may be expected to show a cline in the neighbourhood of the boundary. Moreover, on certain assumptions, he demonstrated that the selection intensities in the cline could be calculated (Haldane,1948).

Since then, there have been extensive studies on many aspects of clines. These include the effects of geographical barriers (Slatkin, 1973), genetic drift (Felsenstein,1975a; Slatkin and Maruyama, 1975; Nagylaki, 1978), conditions for existence of a cline (Nagylaki, 1975), variable migration (Nagylaki, 1976), clines for selective neutral genes affected by closely linked and weakly selected loci, i.e. the hitchhiking effects (Barton, 1979), and multilocus clines (Barton, 1983). In most of these studies, however, the basic process employed is approximated by a diffusion model, which is integrated with different factors, selection, genetic drift, linkage/recombination, etc..

9.1.4 Previous practical work

Early studies of natural clines concentrated on changes in morphological and physiological traits. For example, *Eucalyptus urnigera* (Thomas and Barber, 1974) has glaucous, waxy leaves for resisting freezing above 1000m and bright green leaves below, displaying a narrow cline.

With the application of molecular techniques to natural populations, many different markers such as allozyme and DNA markers have become available (Avise, 1994) and clines in such markers have been recorded (Millar 1983; Tsumura, *et al.*, 1994). For example, Millar (1983) used one allozyme marker (GOT) to investigate a cline less than 3 km width, existing between two northern Califonia bishop pine populations differing in stomatal form, monoterpene composition and flowering times. She found that allele frequencies changed from 0.97 in north of the cline to 0.23 south of the cline. Miller further indicated that difference in allelic frequencies between mature trees and embryos are attributed to long distance pollen flow across the cline.

In two hybrid zones of the pacific coast *irises* (Iridaceae), Young (1996) found, using cpDNA (chloroplast DNA) markers and morphological traits, that the cpDNA marker cline (maternally inherited markers) is displaced 1-2 km relative to the morphological cline in all three transects across the *I. douglasianal/ I. innominata* hybrid zone. One possible explanation for this cline displacement is due to asymmetric migration existing between cpDNA marker mediated by seed flow, and morphological traits mediated by either seed or pollen flow (Young, 1996). In a separate study, Brubaker *et al.* (1993) also found that nuclear introgression is more geographically wide spread and more frequently detected than cytoplasmic introgression in *Gossypium barbadense* and *G. hirsutum* species. Interpretation of these findings requires a exploration of cline theory which incorporates the three plant genomes.

It is now time to re-examine the cline at the molecular level and to extend our understanding of its spatial population genetic structure. In plant species, three differently inherited types of markers are available; nuclear, chloroplast and mitochondrial (Ennos, 1994). A small proportion of these markers may be selectively important. These markers may be under different selection in different environments, and cline formation is likely. In order to model

such clines in plant populations it is necessary to recognise that there is asymmetric migration among the three plant genomes possessing different modes of inheritance (Ennos, 1994; Petit, *et al*, 1993; Mogensen, 1996, and references therein). Thus differences in cline characteristics are expected among markers on these plant genomes, and it is important to understand the effects in theory, and test the theory in practice.

9.1.5 Application of previous models to plant clines

Attempts to apply existing cline models reveal two shortcomings. They do not take into account seed and pollen flow as separate modes of gene flow. Neither do they consider the uniparentally inherited organelle genomes. Some of these deficiencies have been addressed in a recent paper by Nagylaki (1997). Nagylaki (1997) showed that reparametrization may render the previous cline theory suitable for diploid plant nuclear genes. However, uniparentally inherited organelle markers are not considered. The theoretical results obtained in previous chapters cannot be used to explain cline formation even if there is some relationship between them. Therefore, it is of practical significance to build the theory suitable for explaining haploid cline formation in plants.

9.1.6 Aim of this chapter

The objective of this study is to fill the gap between haploid and diploid cline theories of plant genomes, i.e. to explore the cline theory for haploid organelle genes and to study the impacts of seed flow and pollen flow on cline difference between paternally and maternally inherited haploid genes.

We first consider a general case where population size is not large and the effect of genetic drift must be considered. Then a simple case where the effect of genetic drift can be ignored is considered. Some numerical cases are then given to illustrate the influence of seed and pollen flow on cline width for genes located on paternally and maternally inherited genomes.

9.2. Model analysis with genetic drift

9.2.1 Assumptions

General assumptions are:

① A single locus with two alleles (A_1, A_2) is considered in turn for paternally and maternally inherited haploid organelle genes.

② Interaction between any pair of the two genes on genomes differing in mode of inheritance is ignored.

③A hermaphrodite plant population is distributed in an infinite chain of equally spaced colonies each with the same population size in adults in one dimensional space.

④ Migration is symmetrical between colonies, such that the migration between colony i and j $(i \neq j)$ is $m_{ij} = m_{ji}$ (i or $j = 0, \pm 1, \pm 2, \ldots$).

⑤ The life cycling scheme for colony i follows figure 9.2 and occurs within a short time interval Δt.

⑥The population distribution is assumed to be uniform after selection, thus the m_{ij} also represents the probability of migration from colony j to i in the time Δt (Nagylaki, 1978a, p424).

⑦Random mating between pollen and ovules is assumed in each colony.

⑧Density-independent selection of the offspring takes place at each location independently after seed flow.

If colony size is not very large, then genetic drift effects must be considered. Usually, it is complicated to incorporate drift into a cline. There are some studies that are not suitable for plant species (Nagylaki, 1978a; Slatkin and Maruyama, 1975; Felsenstein, 1975a). However, the approximation by diffusion model is still useful for plant species. Thus, in the following the method used by Nagylaki (1978a) is employed to address the plant case. The method used by Nagylaki (1978a) is first to formulate a discrete model in time and space, and then to transform this into a diffusion model that is continuous in time and space. Since detailed derivations can be found in Nagylaki (1978a), we here merely outline the main steps at which seed and pollen flow are incorporated.

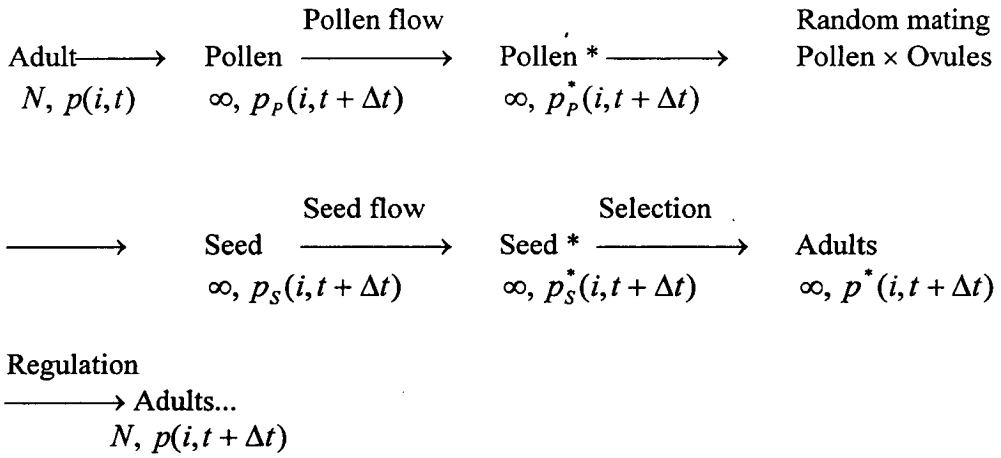


Fig. 9.2 Basic cycle scheme for modelling cline existing in hermaphrodite plant species that is discretely distributed in one dimensional space. The effect of genetic drift is considered. The corresponding gene frequencies to each stage within time interval Δt in colony i are marked below. The subscripts S and P stand for seed and pollen.

9.2.2 Paternally inherited haploid organelle genes

Let $p(i,t)$ be the gene frequency of allele A_1 in colony i at time t. Suppose that the number of pollen grains produced by adults within time interval Δt is large enough that the frequency in pollen, $p_P(i,t+\Delta t)$, is the same as in adults, i.e.

$$p_P(i,t+\Delta t)) = p(i,t) \tag{9.1}$$

After *pollen flow*, the gene frequency in pollen in colony i is

$$p_P^*(i,t+\Delta t) = \sum_j m_{p,ij} p_P(j,t+\Delta t) \tag{9.2}$$

where $m_{P,ij}$ is the migration rate of pollen from colony j to i, and $\sum_j m_{P,ij} = 1$.

After *random mating* between pollen and ovules and formation of seeds, the gene frequency in seeds, $p_S(i,t+\Delta t)$, is the same as that after pollen flow due to haploid assumption, i.e.

$$p_S(i,t+\Delta t) = p_P^*(i,t+\Delta t) \tag{9.3}$$

Similarly, after *seed flow*, the gene frequency in seeds, $p_S^*(i,t+\Delta t)$, is

$$p_S^*(i,t+\Delta t) = \sum_j m_{S,ij} p_S(j,t+\Delta t) \tag{9.4}$$

where the $m_{S,ij}$ is the migration rate of seed flow between colony i and j, and $\sum_j m_{S,ij} = 1$.

Thus, putting equations (9.1) to (9.3) into (9.4), and ignoring the items involving the product of migration rates of seed and pollen, we can obtain

$$p_S^*(i,t+\Delta t)=\left(1-\sum_{j\neq i}m_{S.ij}-\sum_{j\neq i}m_{P.ij}\right)p(i,t)+\sum_{j\neq i}(m_{S.ij}+m_{p.ij})p(j,t)+O(m_sm_P)$$

(9.5')

Let $\tilde{m}_{ij}=m_{s.ij}+m_{p.ij}$ when $j\neq i$; $\tilde{m}_{ii}=m_{S.ii}+m_{P.ii}-1$ when $j=i$. In order to make $\tilde{m}_{ii}$ be at reasonable level, we must assume that $\sum_{j\neq i}m_{S.ij}$ or $\sum_{j\neq i}m_{P.ij}<<1$. The distribution of new migration rate between colony j and i is also symmetric and inclusive, i.e. $\tilde{m}_{ij}=\tilde{m}_{ji}$ and $\sum_j\tilde{m}_{ij}=1$. The equation (9.5') can be rewritten by

$$p_S^*(i,t+\Delta t)=\sum_j\tilde{m}_{ij}p(j,t) \qquad (9.5)$$

Now consider the effects of natural selection. Let $1+\omega_{pat}g_{pat}(i)\Delta t$ and $1-\omega_{pat}g_{pat}(i)\Delta t$ be the fitnesses of genotypes A_1 and A_2 respectively. The subscript *pat* refers to paternally inherited markers. ω_{pat} is the selection coefficient, and $g_{pat}(i)$ is a function that describes the spatial variation in selection coefficient. Then the change of gene frequency due to selection can be obtained , i.e.

$$p^*(i,t+\Delta t)=p_S^*(i,t+\Delta t)+2\omega_{pat}g_{pat}(i)p_S^*(i,t+\Delta t)\left[1-p_S^*(i,t+\Delta t)\right]\Delta t$$

(9.6)

After *regulation* (sampling), let $p(i,t+\Delta t)$ be the gene frequency in adults, which can be expressed by

$$p(i,t+\Delta t)=p^*(i,t+\Delta t)+\varsigma(i) \qquad (9.7)$$

where the $\varsigma(i)$ is the change of gene frequency due to sampling, with expectations $E[\varsigma(i)|p^*(i,t+\Delta t)]=0$ and $E[\varsigma(i)\varsigma(j)|p^*(i,t+\Delta t)]=$ $p^*(i,t+\Delta t)[1-p^*(i,t+\Delta t)]\delta(i,j)/N$ in which $\delta(i,j)$ is the Kronecker delta.

Let $P(i,t)$ be the expected gene frequency, that is $P(i,t) = E[p(i,t)]$. Let $V(i,j;t)$ be the covariance of gene frequencies between colonies i and j. Setting

$$p(i,t) = P(i,t) + \pi(i,t) \tag{9.8a}$$

so that

$$E[\pi(i,t)] = 0 \tag{9.8b}$$

$$V(i,j;t) = E[\pi(i,t)\pi(j,t)] \tag{9.8c}$$

In the next section, following Nagylaki (1978a), the diffusion approximation is employed to model the above process. According to assumption ⑥, the $\tilde{m}_{ij}$ is equivalent to the transition probability from state (colony here) j to i in a Markov process (Feller, 1971, p322). Assumption of the uniform distribution (⑥) indicates that only homogeneous migration is approximated by the diffusion model. Using assumptions ③ and ④, let $x = i\varepsilon$ and $y = j\varepsilon$ be positions of any two colonies, where the ε is assumed to be the spacing between colonies. Conditions similar to equation (8) of Nagylaki (1978a) can be obtained. Let $\Delta t \to 0$ and $\varepsilon \to 0$, positing that, for any $\theta > 0$,

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \sum_{j} {}_{|j-i| \geq \theta/\varepsilon} \ \tilde{m}_{ij} = 0 \tag{9.9a}$$

$$\lim_{\Delta t \to 0} \frac{\varepsilon}{\Delta t} \sum_{j} {}_{|j-i| < \theta/\varepsilon} \ (j-i)\tilde{m}_{ij} = 0 \tag{9.9b}$$

$$\lim_{\Delta t \to 0} \frac{\varepsilon^2}{\Delta t} \sum_{j} {}_{|j-i| < \theta/\varepsilon} \ (j-i)^2 \tilde{m}_{ij} = \tilde{\sigma}^2_{pat} \tag{9.9c}$$

Equation (9.9a) states that any large displacement is impossible, the necessary and sufficient condition for continuity of the sampling function. Equation (9.9b) is the infinitesimal mean and equals zero according to assumption ②(symmetry), and equation (9.9c) is the infinitesimal variance. Equations (9.9a), (9.9b) and (9.9c) are equivalent to equations(4.2), (4.3) and (4.4) of Feller (1971, p333).

Equation (9.9) can be decomposed into two components due to seed and pollen flow. Since the seed and pollen flow between colony j and i is assumed to be symmetric, the equation (9.9b) becomes

$$\lim_{\Delta t\to 0}\frac{\varepsilon}{\Delta t}\sum_{j\ |j-i|<\theta/\varepsilon}(j-i)\tilde{m}_{ij} = \lim_{\Delta t\to 0}\left[\frac{\varepsilon}{\Delta t}\sum_{j\ |j-i|<\theta/\varepsilon}(j-i)(m_{s.ij}+m_{P.ij})\right]$$

$$+\lim_{\Delta t\to 0}\frac{\varepsilon}{\Delta t}\sum_{j=i\ |j-i|<\theta/\varepsilon}(j-i)(-1)$$

$$=0 \qquad (9.10a)$$

The equation (9.9c) becomes

$$\lim_{\Delta t\to 0}\frac{\varepsilon^2}{\Delta t}\sum_{j\ |j-i|<\theta/\varepsilon}(j-i)^2\tilde{m}_{ij} = \lim_{\Delta t\to 0}\left[\frac{\varepsilon^2}{\Delta t}\sum_{j\ |j-i|<\theta/\varepsilon}(j-i)^2(m_{s.ij}+m_{P.ij})\right]$$

$$+\lim_{\Delta t\to 0}\frac{\varepsilon^2}{\Delta t}\sum_{j=i\ |j-i|<\theta/\varepsilon}(j-i)^2(-1)$$

$$=\sigma_S^2+\sigma_P^2 \qquad (9.10b)$$

Using assumption of equation (9.9) and the results obtained by Feller (1971, p334-335), for any function $\psi(j\varepsilon)$, the transformation of $\sum_j \tilde{m}_{ij}\psi(j\varepsilon)$ satisfies the Kolmogorov backward equation, i.e.

$$\sum_j \tilde{m}_{ij}\psi(j\varepsilon) = \psi(i\varepsilon)+\frac{\tilde{\sigma}^2_{pat}}{2}\frac{d^2}{dx^2}\psi(i\varepsilon)\Delta t + O(\Delta t) \qquad (9.11a)$$

Thus, as $\Delta t \to 0$, then

$$\sum_j \tilde{m}_{ij}\psi(j\varepsilon) = \psi(i\varepsilon)+O(\Delta t) \qquad (9.11b)$$

which is the same as equation (10) of Nagylaki (1978a). Thus, following arguments similar to those of Nagylaki (1978a) and using equations (9.5), (9.6),(9.7), (9.8) and (9.11b), we can obtain the expected allele frequency after the time interval Δt, i.e.

$$P(i,t+\Delta t) = E\{E[p(i,t+\Delta t)|p^*(i,t+\Delta t)]\} \tag{9.12a}$$

$$= E[p^*(i,t+\Delta t)] \tag{9.12b}$$

$$= E\left\{\sum_j \tilde{m}_{ij} p(j,t) + \omega' g_{pat}(i) \sum_j \tilde{m}_{ij} p(j,t)[1 - \sum_j \tilde{m}_{ij} p(j,t)]\Delta t\right\} \tag{9.12c}$$

$$= \sum_j \tilde{m}_{ij} E[p(j,t)] + \omega' g_{pat}(i) E\{p(i,t)[1 - p(i,t)]\}\Delta t \tag{9.12d}$$

$$= \sum_j \tilde{m}_{ij} P(j,t) + \omega' g_{pat}(i)\{P(i,t)[1 - P(i,t)] - V(i,i;t)\}\Delta t \tag{9.12e}$$

where $\omega' = 2\omega_{pat}$. Equation (9.12) is equivalent to equation (12) of Nagylaki (1978a). Derivation of equation (9.12d) requires the substitution of equation (9.11b) into equation (9.12c).

Similar to the derivation of (13a) of Nagylaki (1978a), we can show

$$\pi(i,t+\Delta t) = p(i,t+\Delta t) - P(i,t+\Delta t) \tag{9.13a}$$

$$= \sum_j \tilde{m}_{ij} p(j,t) + \omega' g_{pat}(i) \sum_j \tilde{m}_{ij} p(j,t)[1 - \sum_j \tilde{m}_{ij} p(j,t)]\Delta t + \varsigma(i) - P(i,t+\Delta t) \tag{9.13b}$$

$$= \sum_j \tilde{m}_{ij}[P(j,t) + \pi(j,t)] + \omega' g_{pat}(i) p(i,t)[1 - p(i,t)]\Delta t + \varsigma(i) - P(i,t+\Delta t) \tag{9.13c}$$

$$= \sum_j \tilde{m}_{ij} \pi(j,t) + \varsigma(i) + b_i \tag{9.13d}$$

where the $b_i = \omega' g_{pat}(i)\{\pi(i,t)[1 - 2P(i,t) - \pi(i,t)] + V(i,i;t)\}$. The equation (9.13d) is obtained by substituting (9.8a) and (9.11b) into (9.13c). The equation (9.13d) is the same as equation (13) of Nagylaki (1978a).

Similarly, during the transformation from a discrete to a continous model, random drift is assumed to accumulate in increment (Nagylaki, 1978a, equation (11)), thus

$$E[\varsigma(i)|p^*(i,t+\Delta t)] = 0 \tag{9.14a}$$

$$E[\varsigma(i)\varsigma(j)|p^*(i,t+\Delta t)] = p^*(i,t+\Delta t)[1-p^*(i,t+\Delta t)]\delta(i,j)\Delta t / N \tag{9.14b}$$

Therefore,

$$E[\varsigma(i)\varsigma(j)] = E\{E[\varsigma(i)\varsigma(j)|p^*(i,t+\Delta t)]\} \tag{9.15a}$$

$$= E\{p^*(i,t+\Delta t)[1-p^*(i,t+\Delta t)]\delta(i,j)\Delta t / N\} \tag{9.15b}$$

$$= \{P(i,t)[1-P(i,t)] - V(i,i,t)\}\delta(i,j)\Delta t / N + O(\Delta t^2) \tag{9.15c}$$

Derivation of (9.15c) from (9.15b) requires in turn application of equations (9.7), (9.6), (9.5), (9.11b) and (9.8). The equation (9.15c) is the same as the result of the last expectation of the equation (14) of Nagylaki (1978a).

Therefore, from here all the following analysis can be connected to Nagylaki (1978a). Using considerations similar to those of Nagylaki (1978a), two important equations can be immediately obtained

$$\frac{\partial P(x,t)}{\partial t} = \frac{1}{2}\tilde{\sigma}^2 \cdot \frac{\partial^2 P(x,t)}{\partial x^2} + \omega' g(x)[h(P) - V(x,x;t)] \tag{9.16a}$$

$$\frac{\partial V(x,y;t)}{\partial t} = \frac{1}{2}\tilde{\sigma}^2\left[\frac{\partial^2 V(x,y;t)}{\partial x^2} + \frac{\partial^2 V(x,y;t)}{\partial y^2}\right] + \omega' F(x,y,t)V(x,y;t)$$

$$+[h(P) - V(x,x;t)]\rho^{-1}\delta(x-y) \tag{9.16b}$$

where the definition of the $h(P)$ and the $F(x,y,t)$ are the same as Nagylaki (1978a), i.e., $h(P) = P(x,t)[1-P(x,t)]$ and $F(x,y,t) = g_{pat}(x)[1-2P(x,t)] + g_{pat}(y)[1-2P(y,t)]$. The ρ is population density ($\rho = N/\varepsilon$) and the $\delta(x-y)$ is the Dirac delta function, and $\tilde{\sigma}^2 = \tilde{\sigma}_{pat}{}^2 = \sigma_S^2 + \sigma_P^2$, and $\omega' = \omega'_{pat} = 2\omega_{pat}$

9.2.3 Maternally inherited haploid organelle genes

Similarly, let $\tilde{m}_{ij} = m_{S.ij}$ for maternally inherited genes. We can also obtain differential equations by substituting $\tilde{\sigma}^2 = \tilde{\sigma}_{mat}{}^2 = \sigma_S^2$, $\omega' = \omega'_{mat} = 2\omega_{mat}$ $g_{pat}(x) = g_{mat}(x)$ into equation (9.16).

9.2.4 Comparison

If the covariance, $V(i,j;t)$, is investigated in terms of the average position $((x+y)/2)$ and separation $((x-y)/2)$ of two points, an important parameter β was obtained by Nagylaki (1978a). This parameter governs the relative strength of selection and random drift, which can be obtained immediately in plants after similar transformations to those of Nagylaki (1978a). That is $\beta = \rho\tilde{\sigma}^2 / c$ for paternally or maternally inherited organelle genes, where c is the characteristic length.

The meaning of the parameter β is obvious. It is the ratio of two distances (Nagylaki, 1978a). One is the natural distance for migration and random drift $\rho\tilde{\sigma}^2$ for paternally or maternally inherited organelle genes. The other c is the characteristic length. According to Nagylaki's (1978a, p425) argument, selection is strong (weak) compared to random drift if $\beta >> 1$ ($\beta << 1$).

If the $g(\xi)$, where ξ is the transform of x ($\xi = x / c$), is set to be $-\alpha^2$ when $\xi < 0$ and to be 1 when $\xi > 0$, which is the same as equation (25) of Nagylaki (1978a), using similar transformations to equation (19) of Nagylaki (1978a), the parameters become

$$\beta_{pat} = 2\rho\tilde{\sigma}_{pat}\sqrt{\omega_{pat}} \qquad (9.17a)$$

$$\beta_{mat} = 2\rho\tilde{\sigma}_{mat}\sqrt{\omega_{mat}} \qquad (9.17b)$$

From equation (9.17), it can shown that if $\frac{\omega_{mat}}{\omega'_{pat}} \leq 1+\frac{\sigma_P^2}{\sigma_S^2}$, then $\beta_{pat} \geq \beta_{mat}$, otherwise $\beta_{pat} < \beta_{mat}$. Thus, the relative values between that ratio of pollen to seed dispersal and the ratio of slection coefficients between two types of genes are important in determining the relative strength of selection to random drift for paternally and maternally inherited genes.

9.3 Model analysis without genetic drift

9.3.1 Stationary cline

If the colony size is large enough that the effect of genetic drift can be ignored, then we let $P(i,t) = p(i,t)$, $\pi(i,t) = 0$ and $V(i,j;t) = 0$ according to equation (9.8). Thus, the equation (9.16b) will vanish, and equation (9.16a) becomes

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\tilde{\sigma}^2 \cdot \frac{\partial^2 p(x,t)}{\partial x^2} + \omega' g(x)p(x,t)[1-p(x,t)] \qquad (9.18)$$

Under balance of migration and selection, letting $p(x,t) = p(x)$, we can obtain

$$\frac{\partial^2 p(x)}{\partial x^2} = -\frac{2\omega'}{\tilde{\sigma}^2} \cdot g(x)p(x)[1-p(x)] \qquad (9.19)$$

where $\tilde{\sigma}^2 = \tilde{\sigma}_{pat}^{\ 2} = \sigma_S^2 + \sigma_P^2$, $\omega' = \omega'_{pat} = 2\omega_{pat}$, $g(x) = g_{pat}(x)$ for paternally inherited organelle genes, and $\tilde{\sigma}^2 = \tilde{\sigma}_{mat}^{\ 2} = \sigma_S^2$, $\omega' = \omega'_{mat} = 2\omega_{mat}$ $g(x) = g_{mat}(x)$ for maternally inherited organelle genes.

9.3.2 Characteristic length

An important parameter in a cline is the characteristic length within which the gene frequency does not change (Slatkin, 1973). From analysis described above the characteristic lengths of the two genomes can be obtained immediately according to Slatkin (1973). Let l_{pat} and l_{mat} be the characteristic lengths of paternally and maternally inherited genes

respectively within a cline. The characteristic length for haploid organelle genes can be obtained from equation (9.19).

$$l_{pat} = \sqrt{\frac{\tilde{\sigma}^2_{pat}}{\omega'_{pat}}} = \sqrt{\frac{\sigma_S^2 + \sigma_P^2}{2\omega_{pat}}}, \tag{9.20a}$$

$$l_{mat} = \sqrt{\frac{\tilde{\sigma}^2_{mat}}{\omega'_{mat}}} = \sqrt{\frac{\sigma_S^2}{2\omega_{mat}}}, \tag{9.20b}$$

If the intensities of natural selection are equal between the two types of genes, i.e. $\omega_{pat} = \omega_{mat} = \omega$, it is easy to see that the characteristic length for paternally inherited organelle genes is equal to that of maternally inherited organelle genes if $\sigma_P^2 = 0$, but larger than that if $\sigma_P^2 \neq 0$, i.e.

$$l_{pat} \geq l_{mat} \tag{9.21}$$

If the intensities of selection are not equal to one another, the above relationship (9.21) will not hold. However, if $\frac{\omega_{mat}}{\omega_{pat}} > 1 + \frac{\sigma_P^2}{\sigma_S^2}$, then $l_{mat} > l_{pat}$. Here again, we can see that the relative values of the ratio of selection intensities between genes differing in inheritance mode and the ratio of pollen to seed flow play an important role in determining the characteristic cline lengths.

9.3.3 Infinite cline

Suppose that the function $g(x)$ has a similar pattern for each of the two types of genes. Following Nagylaki (1976), let

$$g(x) = \begin{cases} 1, & x < 0, \\ -\alpha^2, & x > 0. \end{cases} \tag{9.22}$$

where α^2 is the ratio of selection intensities in the two parts of the habitat. Actually, this is a variation of Haldane's one step selection (1948).

Let $K = 4\omega_{pat} / \tilde{\sigma}_{pat}^{\ 2}$, $k = \alpha_{pat}^{\ 2} K$, and $\alpha = \alpha_{pat}$ for paternally inherited organelle genes, $K = 4\omega_{mat} / \tilde{\sigma}_{mat}^{\ 2}$, $k = \alpha_{mat}^{\ 2} K$ and $\alpha = \alpha_{mat}$ for maternally inherited genes. The cline equation can be decomposed into three parts according to equation (9.19).

$$\frac{d^2 p(x)}{dx^2} = -Kp(x)[1 - p(x)], \qquad x < 0 \tag{9.23a}$$

$$\frac{d^2 p(x)}{dx^2} = kp(x)[1 - p(x)], \qquad x > 0 \tag{9.23b}$$

$$p(0-) = p(0+), \qquad \frac{dp(x)}{dx}(0-) = \frac{dp(x)}{dx}(0+) \tag{9.23c}$$

Since there have been extensive studies on equations similar to (9.23) (see Nagylaki, 1975), the solution to equation (9.23) can be easily obtained. Let b be the boundary value at $x = 0$. Following the method used by Haldane (1948), we can obtain the iterative equation (9.24) for calculating the b value, and the cline equation (9.25).

$$b = \left\{3(1+\alpha^2)\left[1 - \frac{2}{3}b\right]\right\}^{-\frac{1}{2}} \tag{9.24}$$

$$\begin{cases} x = K^{-\frac{1}{2}} \int_{p(x)}^{b} \left\{\frac{1}{3} - u^2\left[1 - \frac{2}{3}u\right]\right\}^{-\frac{1}{2}} du, & x < 0, \\ x = k^{-\frac{1}{2}} \int_{b}^{p(x)} u^{-1}\left[1 - \frac{2}{3}u\right]^{-\frac{1}{2}} du, & x > 0 \end{cases} \tag{9.25}$$

It is difficult to obtain simple analytic expression from (9.25), but equation (9.25) provides a convenient way to carry out numerical calculation.

9.3.4 Impacts of seed and pollen dispersal

In the following section three numerical examples are used to explain the impacts of seed and pollen flow on cline width and the differences between two types of genes.

According to equations (9.24) and (9.25), gene frequencies in a cline can be calculated. Let $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\alpha^2 = 1.0$ and $\sigma_P^2 = \sigma_S^2 = 1.0$. This means that the two types of genes are the same in selection intensity, and that the dispersal variances of pollen and seed are also the same. Under this case, the results are shown in Fig.9.3a. It can be seen that the cline is wider for paternally inherited genes than for maternal genes within a given gene frequency interval [0.1, 0.9].

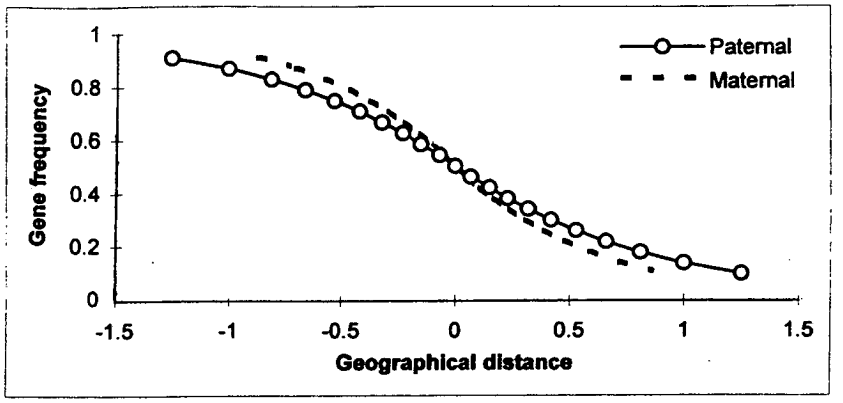
If pollen dispersal is much larger than seed dispersal, this will influence the relative cline width between the two types of genes. For example, let $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\alpha^2 = 1.0$, $\sigma_P^2 = 5.0$ and $\sigma_S^2 = 0.5$. The result (Fig. 9.3b) shows that the difference in cline width is larger than that in Fig. 9.3a.

Alternatively, if seed dispersal is much larger than pollen dispersal, the cline widths of both types of genes will become wider. Let $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\alpha^2 = 1.0$, $\sigma_p^2 = 0.5$ and $\sigma_S^2 = 5.0$. It can be seen that the difference between cline widths of these two types of genes is reduced, although the absolute values of cline widths of both types of genes increase (Fig.9.3c), compared with those in Fig. 9.3a.

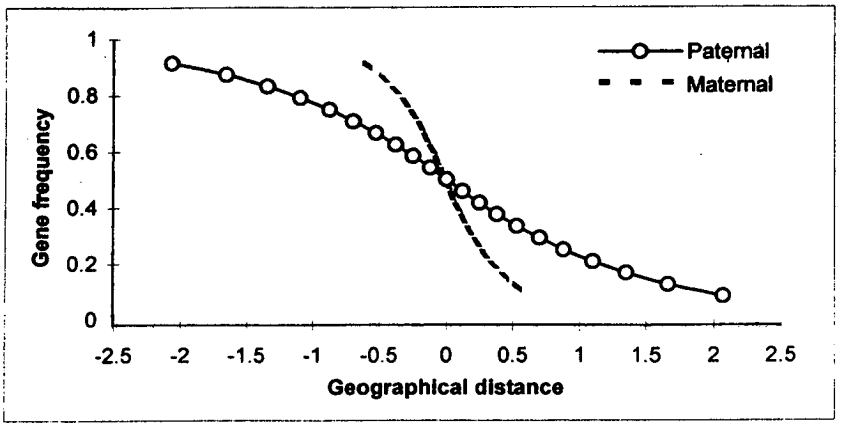
9.4 Discussion

Cline theory for a single locus two alleles model has been developed for haploid organelle genes possessing different inheritance modes. The results show that reparametrization may render previous cline theories applicable to plant haploid organelle genes. Both the ratio of pollen to seed dispersal, and the ratio of selection coefficients between paternally and maternally inherited genes, play a critical role in determining cline width and its difference between these two types of genes.

(a)



(b)



(c)

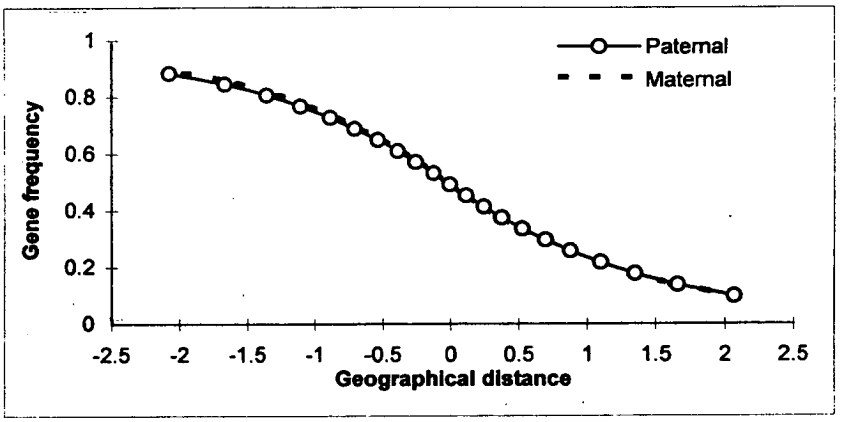


Fig. 9.3 Comparison of cline width between the three plant genomes within a given frequency interval [0.1, 0.9] in the infinite cline.

(a) Parameters settings are $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\alpha^2 = 1.0$ and $\sigma_P^2 = \sigma_S^2 = 1.0$;

(b) Parameters settings are $\alpha^2 = 1.0$, $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\sigma_P^2 = 5.0$ and $\sigma_S^2 = 0.5$;

(c) Parameters settings are $\alpha^2 = 1.0$, $\omega = \omega_{pat} = \omega_{mat} = 1.0$, $\sigma_P^2 = 0.5$ and $\sigma_S^2 = 5.0$.

Since different types of fitnesses function for genotypes are used by Nagylaki (1997) and by the present author, comparison between biparentally and uniparentally inherited genes are not carried out. However, the method for incorporating the effect of selection in a cline used in this chapter cannot be simply extended to diploid nuclear genes. This is because Hardy-Weinberg equilibrium is required for this method when diploid genes are considered after pollen and seed flow, which is difficult to be satisfied. Knowledge of genotype frequencies after pollen and seed flow rather than merely gene frequencies is required. After the processes of seed and pollen flow, it will not be legitimate to assume that genotype frequencies can be obtained from gene frequencies by assuming Hardy-Weinberg equilibrium. This problem was avoided by Nagylaki (1997), using different type of fitness function. However, this problem does not affect the analysis of haploid organelle genes. Thus, the method used in this chapter presents a simpler way to address the effects of genetic drift and seed and pollen flow in a cline for haploid organelle genes.

However, some results obtained by Nagylaki (1997) can be compared with those obtained in the present study. For example, if there are no effects of genetic drift and selfing, according to Nagylaki (1997, p425) the characteristic length for biparentally inherited nuclear genes, l_{nuc}, can be given by

$$l_{nuc} = \sqrt{\frac{2\sigma_S^2 + \sigma_P^2}{2A}} \qquad (9.26)$$

where $A = (U_{11} - U_{22})/2$, in which U_{11} and U_{22} are the scaled fitnesses of genotypes A_1A_1 and A_2A_2. Thus, comparing the equation (9.26) with equation (9.20), we can also conclude that both the ratio of pollen to seed dispersal, and the ratio of selection coefficients between biparentally and uniparentally inherited genes play a critical role in determining cline difference between them.

There are several limitations inherent in the present study when it is used to understand cline theory of plant genomes. First, the three plant genomes are separately considered. The effect of linkage disequilibrium between nuclear genes and chloroplast or mitochondrial genes has not been considered either by Nagylaki (1997) or by the present author. In a cline formed after secondary contact, linkage disequlibrium is likely to be substantial. Theoretical studies have shown using a different model, that linkage disequilibrium among loci on different

genomes may exist in a hybrid zone (Assumssen and Schnabel, 1991; Asmussen and Arnold, 1991). Thus, further work is required to extend the present analyses to situation where linkage disequilibrium exists between genes for biparentally, paternally, and maternally inherited markers.

Secondly, the influence of a barrier to gene flow is of specific interest in clines of plant genes. The barrier can be due to biological and non-biological factors. Biological factors generally include premating barriers, such as pollinator behavior and flowering times, post pollination barrier, such as self-incompatibility and incongruity, and post-fertilization, such as viability and survivorship (review by Arnold, 1997). Non-biological barriers are also variable, such as the different physical obstacles that block seed and pollen flow. These factors clearly influence migration and may lead to asymmetric migration of haploid genes, which may violate the assumption of symmetric migration between colonies.

Third, if the assumption of homogenous haploid chloroplast and mitochondrial genes holds, one particular problem that the model suffers from is that the cline for one locus under strong selection will seriously influence cline for loci under weak selection on the same haploid genomes. The effect of linkage between loci on haploid genomes must be considered in the future work. Thus, current single locus model needs to extend to multilocus models as well.

Finally, the cline modelled in this chapter is maintained by the compound action of dispersal, migration and genetic drift. The theoretical results obtained in this chapter may help us to look at the role that seed and pollen flow play in haploid gene cline. It is also important to understand that some clines can be maintained by selection alone. Dispersal-independent models include those in which the hybrid is more fit than either parent phenotype within a restricted geographic area (Endler, 1977; see review by Harrison, 1992). Thus the influence of pollen and seed flow on clines cannot be inferred in these models. It is critically important to distingush different types of cline maintained in practice before elucidating the impact of seed and pollen flow on cline formation.

9.5 Summary

Cline theory for haploid plant organelle genes is developed in this chapter, using a diffusion model. Results show that reparametrization may render previous cline theory suitable for plant organelle genes. This is the same conclusion drawn by Nagylaki (1997) for diploid plant nuclear genes. One additional important result is that both the ratio of pollen to seed flow and the ratio of selection coefficients between paternally and maternally inherited genes play a critical role in determining cline displacement of these two types of genes.

CHAPTER 10

General Conclusion and Discussion

10.1. Introduction

The thesis consists of two parts that are related via the theme of genetic markers. The first part is the application of genetic markers to study Chinese *Larix*. This is also part of the co-operative project between UK and China," Early Establishment and Tree Improvement of *Larix* in China". The objective is to survey genetic variation in natural populations of native Chinese larch species so as to provide background information for further genetic improvement. Based on current research progress achieved in population genetic improvement of larch species in China, the genetic structure of natural populations and the genetic relationship between three native larch taxa, *L. gmelinii*, *L. olgensis* and *L. principis-rupprechtii*, are studied using molecular markers.

The second part is the development of the theory for using genetic markers to elucidate the impacts of seed and pollen flow on population genetic structure of hermaphrodite plant species, under a variety of models. Migration occurs in two ways for hermaphrodite plant, seed flow and pollen flow. This produces asymmetrical migration for biparentally, paternally and maternally inherited genes. Perhaps due to historical reasons, the particular role that seed and pollen flow play in influencing population genetic structure and in genetic improvement of plant species has been seldomly recognised in theory. However, theories developed to explain the genetic structure of animal populations cannot be simply applied to plant populations because the mechanism of migration is different from animal population. Moreover, application of molecular techniques provides us with many useful markers suitable for surveying population genetic structure where the impacts of pollen and seed flow are marked. Thus, the requirement of theory to address these impacts is critical.

10.2. Application of molecular genetic marker to study genetic variation of the *Larix gmelinii* complex.

Native larch species, especially three larch taxa *L.gmelinii*, *L. olgensis* and *L. principis-rupprechtii*, are important forest tree species in China. At least 10 years of provenance trials have been carried out for each of the three taxa. Many important quantitative traits have been studying during the last "Seventh Five-Year Project", the " Eighth Five-Year Project" and current the "Ninth Five-Year Project". Seed zones have been delineated according to field growth performance.

However, genetic structure and mating system of natural populations have not been surveyed using molecular markers. This gap is filled up by the present thesis. The genetic variation within seventeen populations, eight in *L. gmelinii*, seven in *L. olgensis* and two in *L. principis-rupprechtii*, representing three Chinese larch taxa was quantified and studied using eight polymorphic allozyme loci: PGI, 6-PGD, MDH-I, AAT-I, AAT-II, AAT-III, PGM and SDH. These allozyme loci were shown to be in linkage equilibrium in each taxa. Most populations were found to be in Hardy-Weinberg equilibrium for these allozymes, with the exception of a few populations of *L. olgensis* and *L. gmelinii* due to heterozygote deficiency.

Mating systems of the three Chinese *Larix* taxa were scored using these allozyme markers. Population Jiagedaqi of *L. gmelinii* exhibited nearly total outcrossing (t_m= 0.986 ± 0.081). Two populations of *L. principis-rupprechtii*, Fengning and Hunyuan, possessed significant outcrossing (t_m=0.847± 0.427 ~ 0.792 ± 0.169). However, mating system was variable between populations of *L. olgensis*. Two populations of *L. olgensis*, Xiaobeihu and Changbei, exhibited significant levels of selfing ($t_m \approx 0.705$). One population, Dahailin, exhibited biparental inbreeding in addition to selfing, with t_m being 0.684 ±0.107. However, the other three populations of *L. olgensis*, Beihe, Beidaoshan and Dongfanghong, exhibited predominantly outcrossing, t_m = 0.847 ± 0.427 ~ 1.203 ± 0.371. These results are comparable to those for other reported conifers including *L. laricina* and *L. decidua*. It may be concluded that the three larch taxa possess predominantly outcrossing mating system, but in certain populations significant self-fertilisation can occur.

Less then 2% of total genetic variation occurred between populations investigated in each of the three taxa. Analyses of spatial patterns indicated that the distribution of genetic variation did not correlate with geographic pattern in *L. gmelinii*, but a weak correlation caused by isolation by distance was found in *L. olgensis*. As a result there is a higher level of population differentiation present in *L. olgensis* than in *L. gmelinii*.

Nei's genetic distances within each larch taxa were very small, about 0.002, while distances between taxa were larger than within, about 0.01, five times the distance within taxa. A dendrogram was reconstructed to elucidate evolutionary relationship between the three larch taxa, using these eight polymorphic enzyme loci. The results indicate that *L. gmelinii* is

more closely related to *L. olgensis* than to *L. principis-rupprechtii.*

Relationships among the three larch species was further studied using PCR-RFLP analysis of three noncoding regions of cpDNA from *Trn* L (UGU) to *Trn* F (GAA), and showed that there were no detectable difference in restriction sites within their regions. This was further supported by sequence analysis of these three non-coding regions, indicating that there were no differences at all within these regions among the three larch taxa.

Different levels of variation were surveyed to resolve the three taxa, using morphological traits, allozyme markers and the sequence of three noncoding regions of cpDNA. It is easy to use morphological traits such as cone size (Appendix I) to distinguish the three taxa. However, low genetic distances were detected among them using allozyme markers, and no difference was observed in terms of three noncoding sequence of cpDNA. Thus, it is reasonable to conclude that *L. olgensis* and *L. principis-rupprechtii* should be classified as two varieties of *L. gmelinii* rather than two separate species. This conclusion also suggests that it is also unnecessary to further define new varieties of *L. olgensis* and *L. principis-rupprechtii* because variable morphological characters mainly reflect adaptive evolution within these taxa.

One possible process involved in the formation of *L. olgensis* and *L. principis-rupprechtii* is that both of them were consequences of the southward colonisation of *L. gmelinii.* Formation of *L. principis-rupprechtii* is due to adaption to the warmer climate that ultimately blocked the southward clonisation of *L. gmelinii.* However, one question emerged form this process is: Does *L. principis-rupprechtii* come from *L. olgensis* or from *L. gmelinii*, or both? According to the genetic relationship elucidated by allozyme markers (Chapter 3), *L. principis-rupprechtii* is more distant from *L. gmelinii* than is *L. olgensis,* implying that *L. principis-rupprechtii* likely comes from *L. gmelinii* via *L. olgensis* according to their geographical distributions (Chapter2). However, this issue cannot be judged strongly at the moment until more research has been carried out.

The colonisation process is mediated by seed movement first and then by both seed and pollen movement in conifers. A low level of genetic structure among populations and the type of predominantly outcrossing mating system of the *L. gmelinii* complex may indicate more extensive pollen flow than seed flow. The bottleneck effect caused by the pioneer seed

colonisation could be swept away by later high levels of pollen flow, thus leading to a small difference in level of polymorphism from *L. gmelinii* to *L. olgensis* to *L. principis-rupprechti* for nuclear markers. Therefore, use of maternally inherited markers may help to elucidate this history more clearly because migration of this marker is mediated by seed movement only. Any difference established during the early migration phase of *L. gmelinii* are not affected by subsequent pollen flow, and may be detectable in present day populations.

The close genetic relationship among the three Chinese larch taxa may explain in part the limitation of using hybrids between taxa in practice, which has already been reflected by more recent experiments (Yang, *et al.*, 1991). In a word, the findings obtained in this thesis contribute to a better understanding of the genetic variation of natural Chinese larch populations and their relationship with one another.

10.3. Development of the theory for using genetic marker to infer plant population genetic structure

In plants, migration by seed and pollen flow represents different biological process although both types of gene flow have the same consequence of homogenising genetic differences between populations. When allied with genetic markers possessing different modes of inheritance, asymmetrical migration is generated for biparentally, paternally and maternally inherited genes, with associated consequences for population genetic structure.

If a population is distributed as an array of subpopulations (local population or colony) in space, its genetic structure can be modelled by the island or the stepping stone model. Under certain assumptions, it is shown that population differentiation is different among the three plant genomes. Differentiation is greater for maternally inherited genes than for paternally inherited genes, which in turn is greater than for biparentally inherited genes at equilibrium between migration and drift. If migration rates of seed and pollen flow are very small, a general formula for population differentiation in the island model, using Wright's F-statistics, can be expressed by

$$F_{st} = \frac{1}{1+2\tilde{N}\tilde{m}} \qquad (10.1)$$

for the three plant genomes. The $\widetilde{N}$ and $\widetilde{m}$ in (10.1) and in the following formulae stand for the same meaning for the three genomes: $\widetilde{N} = 2N_e$, $\widetilde{m} = m_S + m_P / 2$ for biparentally inherited nuclear genes; $\widetilde{N} = N_m$, (effective population size of male), $\widetilde{m} = m_S + m_P$ for paternally inherited organelle genes; $\widetilde{N} = N_f$, $\widetilde{m} = m_S$ for maternally inherited organelle genes. For one locus with many alleles in the finite island model, population differentiation can be approximated by a general formula,

$$G_{st} = \left(1 + 2\widetilde{N}\widetilde{m}\left(1 - \frac{1}{L}\right)^2\right)^{-1} \tag{10.2}$$

where the L is number of subpopulations investigated.

In the case of a one dimensional stepping-stone model, if rates of one-step migration for both seed and pollen flow are much larger than those of long distance migration, a general formula for population differentiation can be expressed by

$$F_{st} = \frac{1}{1 + 2\widetilde{N}\sqrt{\widetilde{m}_1 \widetilde{m}_\infty}} \tag{10.3}$$

In addition, an important result in the stepping stone model is that the rate of decline of genetic correlation with distance is influenced by the relative values of long and short distance migration by seed and pollen, which can be expressed in a general formula:

$$r(k) = \exp\left(-\sqrt{\frac{2\widetilde{m}_\infty}{\widetilde{m}_1}}k\right) \tag{10.4}$$

Differences among these three differently inherited genes in genetic correlation with distance are conditional on the values of long and short distance migration for pollen and seeds.

If a population is continuously distributed in space, its genetic structure can be modelled by

Wright's isolation by distance model. In this case, results similar to those obtained in the island and stepping-stone models are shown for the relative levels of population differentiation among the three genomes possessing different inheritance modes.

Different levels of population differentiation among the three genomes provide us with a theoretical foundation for estimating the ratio of pollen to seed flow, an important parameter in measuring relative contribution to the gene flow between seed and pollen. This possibility is explored in the three classical models using Wright's F-statistics, and in the Nei-Feldman (Nei and Feldman, 1972) two populations model using Nei's genetic distance.

DNA sequence data will be widely used to investigate population genetic structure in the future. Once the DNA sequence data for the three plant genomes (nuclear, chloroplast and mitochondria DNA) are available, estimation of the ratio of pollen to seed flow from a sample of n ($n \geq 2$) individual genes can be inferred in terms of the number of segregating nucleotide sites between and within populations. This provides us with a very useful tool to infer the relative contribution to migration of seed and pollen flow in the future. Another result of theoretical interest is that if the mutation rates are the same among different genomes, in the discrete model of populations the mean coalescent time is the shortest for the paternally inherited genome (cpDNA in conifers) and, given certain conditions, is the longest for the maternally inherited genome (cpDNA in angiosperms and mtDNA in conifers and angiosperms.

The mean coalescent time and number of segregation sites can be expressed by a general formula for a discretely distributed population for the three plant genomes, i.e.

$$E(T) = 2\tilde{N}L\left[1+\frac{1}{2\tilde{N}\tilde{m}}\left(1-\frac{1}{L}\right)^2\right]\cdot\left(1-\frac{1}{n}\right) \qquad (10.5)$$

and

$$E(S) = 2\tilde{N}\tilde{\mu}aL\left[1+\frac{1}{2\tilde{N}\tilde{m}}\left(1-\frac{1}{L}\right)^2\right] \qquad (10.6)$$

However, these results are difficult to obtain in a model of a population that is continuously distributed in space.

The characteristic of asymmetric migration also provides a basis for exploring cline theory of the three plant genomes. In this case the effects of natural selection are considered. The cline theory for haploid plant organelle genes is developed in this thesis, using a diffusion model. Results show that reparametrization may render previous cline theory suitable for plant organelle genes. This is the same conclusion drawn by Nagylaki (1997) for diploid plant nuclear genes. One additional important results is that both the ratio of pollen to seed flow, and the ratio of selection coefficients between paternally and maternally inherited genes play a critical role in determining cline displacement of these two types of genetic markers.

The above theoretical results may provide us with a framework to understand the genetic structure of plant populations, but they are obtained generally under conditions of ① no linkage disequlibrium between any pair of the three types of genes, ② one locus with two alleles, and ③ half of the effective population size of diploid genes for haploid genes. These are limitations of these findings.

Effects of the mode of inheritance of genetic markers on population genetic structure have been emphasised either in animal or in plant populations in recent years (Ennos, 1994; Birky, *et al*, 1989; Petit, *et al*, 1993; Chesser and Baker, 1996). In plant species, it can be seen from the theoretical results in this thesis that these impacts are marked. The assumption of no linkage disequilibrium holds between biparentally inherited nuclear and uniparentally inherited organelle genes if the three plant genomes are separately investigated. However, Asmussen *et al.* (1991) showed that pollen flow may affect linkage equilibrium between cytoplasmic and nuclear genes in a hybrid zone. Thus, this effect must be considered in the future.

The second condition needs to be relaxed so as to be valid for many loci if there are interactions among them. However, the effect of seed and pollen flow together with the impact of recombination rate on plant population genetic structure presents a new challenge in the future.

The third condition is important in the models. Organelle genomes (paternally and maternally inherited) are assumed to be haploid. Thus the rate of pure drift is equal to half that of nuclear genes. This assumption may be violated in some realistic cases. However, it

can be solved simply by substituting the population size N by the effective male N_m and female N_f population sizes.

For most of results obtained in the second part of the thesis, only selectively neutral genes are considered. Thus these results can be applied to practical work which uses neutral markers, such as some allozymes and non-coding regions of DNA sequence. However, an important genetic phenomena is the hitchhiking effect, the effect of linkage between selective genes and selectively neutral genes, which may lead to modification of the above theoretical results to different extents.

10.4 Future study

Future study associated with the thesis is clear in both theory and practice. In theory it is necessary to release these constraints mentioned above, including the effects of linkage disequilibrium between biparentally inherited nuclear genes and uniparentally inherited organelle genes, the effects of recombination rate, and the effects of hitchhiking. This work may help us to investigate the role of pollen and seed flow in greater deepth.

In practice, application of the results in the second part of the thesis is clearly required. Thus additional analysis, using maternally inherited markers (mtDNA markers), may provide evidence to clarify the genetic relationship between these three Chinese larch taxa. This work may also allow us to estimate the ratio of pollen to seed flow existing among natural populations of these three Chinese *Larix* taxa.

References

Adams, W. T. 1983 Application of isozymes in tree breeding. pp 381-400. In: *Isozymes in Plant Genetics and Breeding, Part A.* S.D. Tanksley and T.J. Orton (Eds.), Elsevier Science Publishers B.V., Amsterdam.

Adams,W. T. 1992 Gene dispersal within forest tree populations. *New Forests* **6**:217-240

Adams, W. T., V.D. Hipkins, J. Burczyk and W.K. Randall 1997 Pollen contamination trends in a maturing Douglas-fir seed orchard. *Canadian Journal of Forest Research* **27**: 131-134

Arnold, M. L. 1997 *Natural Hybridization and Evolution.* Oxford University Press, New York.

Asmuswsen, M. A. and J. Arnold 1991 The effects of admixture and population subdivision on cytonuclear disequilibrium. *Theoretical Population Biology* **39**: 273-300

Asmussen, M. A., J. Arnold and J. C. Avise 1989 The effects of assortative mating and migration on cytonuclear associations in hybrid zones. *Genetics* **122**: 923-934

Asmussen, M. A. and A. Schnabel 1991 Comparative effects of pollen and seed migration on the cytonuclear structure of plant populations. I. Maternal cytoplasmic inheritance. *Genetics* **128**: 639-654

Avise, J. C. 1994. *Molecular markers, natural history and evolution.* Chapman and Hall, London.

Barbujani, G. 1987 Autocorrelation of gene frequencies under isolation by distance. *Genetics* **117**: 777-782

Barton, N.H. 1979 Gene flow past a cline. *Heredity* **43**: 333-339

Barton, N.H. 1983 Multilocus clines. *Evolution* **37**: 454-471.

Barton, N.H. and G.M. Hewitt 1989 Adaptation, speciation and hybrid zones. *Nature* **341**: 497-503.

Barton, N.H. and M. Slatkin 1986 A quasiequilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity* **56**: 409-415

Barton, N.H. and M. Turelli 1989 Evolutionary quantitative genetics: how little do we know? *Annual Review of Genetics* **23**: 337-380

Barton, N. H. and I. Wilson 1995 Genealogies and geography. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences* **349**: 49-59

Bateman, A. J., 1947 Contamination in seed crops. III. Relation with isolation distance. *Heredity* **1**: 303-336

Birky, C.W., P. Fuerst and T. Maruyama 1989 Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* **121**: 613-627.

Birky, C. W. 1988 Evolution and variation in plant chloroplast and mitochondrial genomes. pp23-53. In: *Plant Evolutionary Biology*. Gottlieb, L. D. and S.K. Jain (Eds.), Chapman and Hall, New York.

Bremer, B. 1991 Restriction data from chloroplast DNA for phylogenetic reconstruction: is there only one accurate way of scoring ? *Plant Systematics and Evolution* **175**: 39-54

Brown, A. H. D. 1990 Genetic characterization of plant mating systems, pp 145-162. In: *Plant Population Genetics, Breeding, and Genetic Resources*. Brown, A.H.D., Clegg, M.T., Kahler, A.L. and Weir B.S. (Eds) Sinauer Associates, Inc., Sunderland, Ma.

Brown, A.H.D. and R.W. Allard 1970 Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics* **66**: 133-145

Brown, A. H.D. and B.S. Weir, 1983 Measuring genetic variability in plant populations. pp 219-239. In. *Isozymes in Plant Genetics and Breeding, Part A*. S.D. Tanksley and T.J. Orton (Eds.), Elsevier Science Publishers B.V., Amsterdam.

Brown, A.H.D., D. Zohary and E. Nevo, 1978 Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. *Heredity* **41**: 49-62

Brubaker, C. L., J. A. Koontz and J. F. Wendel, 1993 Bidirectional cytoplasmic and nuclear introgression in the new world cottons, *Gossypium barbadense* and *G. hirsutum* (*Malvacea*). *American Journal of Botany* **80**: 1203-1208.

Chakraborty, R. and M. Nei 1974 Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theoretical Population Biology* **5**: 460-469

Cheliak, W. M., B. P. Dancik, K. Morgan, F. C. H. Yeh and C. Strobeck 1985 Temporal variation of the mating system in a natural population of jack pine. *Genetics* **109**: 569-584

Cheliak, W. M. and S.A. Pitel 1984a *Techniques for starch electrophoresis of enzymes from forest tree species*. Information report PI-X-42, Petawawa National Forest Institute.

Cheliak, W. M. and J. A. Pitel 1984b Inheritance and linkage of allozymes in *Larix laricina*. *Silvae Genetica* **34**: 142-148

Cheliak, W. M., J. Wang and J. A. Pitel 1988 Population structure and genetic diversity in tamarack, *Larix laricina* (Du Roi) K.Koch. *Canadian Journal of Forest Research* **18**: 1318-1324.

Clegg, M.T. 1993 Chloroplast gene sequences and the study of plant evolution. *Proceedings of the National Academy of Sciences USA* **90**: 363-367

Clegg, M.T. and G. Zurawski, 1992 Chloroplast DNA and the study of plant phylogeny: present status and future prospects. pp1-13. In: *Moleuclar Systematics of Plants*. P.S. Soltis and D.E. Soltis (Eds.), Chapman and Hall, New York.

Clegg, M.T., B.S. Gaut, G.H. Jr Learn and B.R. Morton 1994 Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences USA* **91**: 6795-6801.

Clegg, M.T., A.L. Kahler and R.W. Allard 1978 Estimation of life cycle components of selection in an experimental plant population. *Genetics* **89**: 765-792

Chesser, R.K. and R.J. Baker 1996 Effective size and dynamics of uniparentally and diparentally inherited genes. *Genetics* **144**: 1225-1235

Crawford, D.J. 1983 Phylogenetic and systematic inferences from electrophoretic studies. pp 267-287. In: *Isozymes in Plant Genetics and Breeding, Part A*. S.D. Tanksley and T.J. Orton (Eds.), Elsevier Science Publishers B.V., Amsterdam.

Crawford,T.J. 1984a What is a population ? pp135-173 In: *Evolutionary Ecology*, edited by B. Shorrocks. Blackwell Scientific Publications, Oxford.

Crawford, T. J. 1984b The estimation of neighbourhood parameters for plant populations. *Heredity* **52**: 273-283

Crawford, D.J. and Landolt 1995 Allozyme divergence among species of *Wolffia* (Lemnaceae). *Plant Systematics and Evolution* **197**: 59-69

Crow, J.F. and K. Aoki, 1984 Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences USA* **81**: 6073-6077

Davidson, R. and Y.A. El-Kassaby, 1997 Genetic diversity and gene conservation of pacific silver fir (*Abies amabilis*) on vancouver island, British Columbia. *Forest Genetics* **4**: 85-98.

Demesure, B., N. Sodzi and R.J. Petit, 1995 A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Molecular Ecology* **4**: 129-131

DeVerno, L. L., P. J. Charest and L. Bonen 1993 Inheritance of mitochondrial DNA in the conifer *Larix*. *Theoretical and Applied Genetics* **86**: 383-388.

Dong, J. and D. B. Wanger 1993 Taxonomic and population differentiation of mitochondrial diversity in *Pinus banksiana* and *Pinus contorta. Theoretical and Applied Genetics* **86**: 573-578

Dong, J. and D. B. Wanger 1994 Paternally inherited chloroplast polymorphism in *Pinus:* Estimation of. diversity and population subdivision, and tests of disequilibrium with a maternally inherited mitochondrial polymorphism. *Genetics* **136**: 1187-1194.

Dow, B. D. and M.V. Ashley 1996 Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa. Molecular Ecology* **5**: 615-627

Downie, S.R. and J.D. Palmer 1990 Use of chloroplast DNA rearrangements in recostructing plant phylogeny. pp14-35 In: *Molecular Systematics of Plants*. P.S. Soltis and D.E. Soltis (Eds.). Chapman and Hall, New York.

Dumolin-Lapegue, S. , M.H. Pemonge and R.J. Petit 1997 An enlarged set of consensus primers for the study of organelle DNA in plants. *Molecular Ecology* **6**: 393-397

Endler, J.A. 1977 *Geographic variation, Speciation, and Clines*. In: Monographs in population biology, R.M. May (Ed.). Princeton University Press, Princeton, N.J.

Ennos, R.A. 1994 Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72**: 250-259

Ennos, R.A. 1996 Utilising genetic information in plant conservation programmes. pp 278-291. In: *Aspects of the Genesis and Maintenance of Biological Diversity*. Hochberg, M.E., Clobert J., Barbault R.(Eds.), Oxford University Press.

Ennos, R.A. and M. T. Clegg 1982 Effect of population substructuring on estimates of outcrossing rate in plant populations. *Heredity* **48**: 283-292

Ennos, R.A. and Q. Tang 1994 Monitoring the output of a hybrid larch seed orchard using isozyme markers. *Forestry* **67**: 63-74

Epperson, B.K. and R.W. Allard 1984 Allozymes analysis of the mating system in lodgepole pine populations. *Journal of Heredity* **75**: 212-214

Erikson, V.J. and W.T. Adams 1989 Mating success in a coastal Douglas-fir seed orchard as affected by distance and floral phenology. *Canadian Journal of Forest Research* **19**: 1248-1255

Erikson, V.J. and W.T. Adams 1990 Mating system variation among individual ramets in a Douglas-fir seed orchard. *Canadian Journal Forest Research.* **20**: 1672-1675

Ewens, W.J. 1979 *Mathematical Population Genetics*. Springer, Berlin

Falconer, D.S. 1989 *Introduction to Quantitative Genetics* (third edition) Longman, London.

Farris, M.A. and J.B. Mitton 1984 Population density, outcrossing rate, and heterozygote superiority in ponderosa pine. *Evolution* **38**: 1151-1154

Feller, W. 1971 *An Introduction to Probability Theory and Its Applications*. Wiley, New York, Vol.2 2nd Ed.

Felsenstein, J. 1975a Genetic drift in clines which are maintained by migration and natural selection. *Genetics* **81**: 191-207.

Felsenstein, J. 1975b A pain in the torus: some difficulties with the model of isolation by distance. *American Naturalist* **109**: 359-368

Ferris, C., R. P. Oliver, A. J. Davy and G.M. Hewitt 1995 Using chloroplast DNA to trace postglacial migration routes of oaks into Britain. *Molecular Ecology* **4**: 731-738

Fins, L. and L.W. Seeb 1986 Genetic variation in allozymes of western larch. *Canadian Journal of Forest Research* **16**: 1013-1018

Fisher, R.A. 1937. The wave of advance of advantageous genes. *Annals of Eugenics* **7**: 355-369

Fisher, R.A. 1950 Gene frequencies in a cline determined by selection and diffusion. *Biometrics* **6**: 353-361

Friedman S.T. and W.T. Adams 1985 Estimation of gene flow into seed orchards of loblolly pine (*Pinus taeda* L.). *Theoretical and Applied Genetics* **69**: 609-615

Furnier, G. R. and W.T. Adams 1986 Mating system in natural populations of Jeffrey pine. *American Journal of Botany* **73**: 1009-1015

Furnier, G. R. and L. Paule 1992 Inferences on mating system and genetic composition of a seed orchard crop in European larch (*Larix decidua* Mill.). *Journal of Genetics and Breeding* **46**: 309-314

Furnier, G. R. and M. Stine 1995 Interpopulation differentiation of nuclear and chloroplast loci in white spruce. *Canadian Journal of Forest Research* **25**: 736-742

Fyfe, J. T. and N. T. J. Bailey 1951 Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. *Journal of Agricultural Science* **41**: 371-378

Gaggiotti, Q. E. and P. E. Smouse 1996 Stochastic migration and maintenance of genetic variation in sink populations. *American Naturalist* **147**: 919-945.

Gillies, A. C. M. 1994 Molecular Analysis of Genetic Diversity and Evolutionary Relationships in the Tropical Legume Genus *Stylosannus* (Aubl.) SW. University of St. Andrews (Ph.D. Thesis).

Gielly, L. and P. Taberlet, 1994 The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus rbcL sequence. *Molecular Biology Evolution* **11**: 769-777.

Gilpin, M. E. 1991 The genetic effective size of a metapopulation. *Biological Journal of the Linnean Society* **42**: 165-175

Gilpin, M.E. 1993. Spatial structure and population vulnerability. pp.125-139 In: *Viable Populations for Conservation*. M.E. Soule (Ed.), Cambridge University Press, Cambridge.

Gitzedanner, M. A., E. E. White, B. M. Foord, G. E. Dupper, P. D. Hodgskiss and B. B. Kinloch, Jr. 1996 Genetics of Cronartium ribicola. III. Mating system. *Canadian Journal of Botany* **74**: 1852-1859

Gliddon C. and M. Saleem 1985 Gene-flow in *Trifolium repens*- an expanding genetic neighbourhood. pp293-309. In. *Genetic Differentiation and Dispersal in Plants*. Jacquard P., G. Heim and J. Antonovics (Eds.) NATO ASI Series. vol.**G5**. Springer Verlag, Berlin.

Gömöry, D. and L. Paule 1992 Inferences on mating system and genetic composition of a seed orchard crop in European larch (*Larix decidua* Mill.). *Journal of Genetics and Breeding* **46**: 309-314

Goudet, J. 1995 Fstat v-1.2: a computer program to calculate F-statistics. *Journal of Heredity* **86**: 485-486

Govindaraju, D.R. 1989 Estimates of gene flow in forest trees. *Biological Journal of the Linnean Society* **37**: 345-357.

Grierson, D. and S.N. Covey 1988 *Plant Molecular Biology* (2nd edition), Chapman and Hall, New York.

Hadrys, H., M. Balick and B. Schierwater 1992 Application of random amplified polymorphic DNA (RAPD) in molecular ecology. *Molecular Ecology* **1**: 55-63

Haldane, J.B.S. 1948 The theory of a cline. *Journal of Genetics* **48**: 277-284

Haldane, J.B.S., 1954 An exact test for randomness of mating. *Journal of Genetics* **52**: 631-635

Hamrick, J.L. 1989 Isozymes and analysis of genetic structure of plant populations. pp 87-105. In: *Isozymes in Plant Biology*, ed. D. Soltis and P. Soltis Dioscorides Press, Washington, D.C.

Hamrick, J.L. 1994 Genetic diversity and conservation in tropical forests. pp1-9. In: *Proceedings International Symposium on Genetic Conservation and Production of Tropical Forest Tree Seed.* Drysdale, R.M., John, S.E.T. and Yapa, A.C. (Eds.). Asen-Canada Forest Tree Seed Centre.

Hamrick, J. L., and M. J. Godt 1989 Allozymes diversity in plant species. pp43-63. In: *Plant Population Genetics, Breeding, and Genetic Resources* A.H.D. Brown, M.T. Clegg, A.L.Kahler, and B.S. Weir (Eds.). Sinauer, Sunderland, Mass.

Hamrick, J.L., M.J.W. Godt, D.A. Murawski, and M.D. Loveless, 1991 Correlations between species traits and allozyme diversity: implications for conservation biology. pp75-86. In: *Genetics and Conservation of Rare Plants*. (Eds) D.A. Falk and K.E. Holsinger. Oxford University Press, New York.

Hamrick, J. L. and M. J. Godt, 1996 Conservation genetics of endemic plant species. pp 281-304 In: *Conservation genetics:Case Histories from Nature*. Ed. J.C. Avise and J.L. Hamrick. Chapman and Hall, New York.

Hanski, I., and M. Gilpin 1991 Meatpopulation dynamics: a brief history and conceptual domain. *Biological Journal of Linnean Society* **42**: 3-16

Hanski, I. 1994 Patch-occupancy dynamics in fragmented landscapes. *Trends in Ecology and Evolution* **9**: 131-135

Hansson, L. 1991 Dispersal and connectivity in metapopulations. *Biological Journal of Linnean Society* **42**: 99-103

Harding, R.M. 1995 New phylogenies: an introductory look at the coalescent. pp15-22. In: *New Uses for New Phylogenies*. P.H. Harvey, A.J.L. Brown, J.M. Smith and S. Nee (Eds.) Oxford University Press.

Hare, P. M. and J. C. Avise, 1996 Molecular genetic analysis of a stepped multilocus cline in the American oyster (*Crassostera virginica*). *Evolution* **50**: 2305-2315

Harris, S. A. and R. Ingram 1991 Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plastid transmission. *Taxon* **40**: 393-412

Harrison, S. 1991 Local extinction in a metapopulation context: an empirical evaluation. *Biological Journal of Linnean Society* **42**: 73-88

Harrison, R.G. 1992 Hybrid zones: windows on evolutionary process. *Oxford Surveys in Evolutionary Biology* 7: 69-128

Heywood, V. H. 1967 *Plant taxonomy*. Edward Arnold Publishers Ltd, London

Hu, X.-S. and R.A. Ennos 1997 On estimation of the ratio of pollen to seed flow among plant populations. *Heredity* **79**: 541-552

Hubby, J. L. and R.C. Lewontin, 1966 A molecular approach to the study of genetic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54**: 577-594

Hudson, R. R. 1992 Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* 7: 1-44

Hudson, R. R., M. Slatkin and W. P. Maddison 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583-589.

John, B. and K.B. Lewis 1968 The Chromosome Complement. Springer-Verlag, Wien.

Jøhnk, N. and H. R. Siegismund 1997 Population structure and post-glacial migration routes of *Quercus robur* and *Quercus petraea* in Denmark, based on chloroplast DNA analysis. *Scandinavian Journal of Forest Research* **12**:130-137

Jones, S. B. and A. E. Luchsinger 1979 *Plant Systematics*. McGraw-Hill Book Company. New York.

Jukes, T. H. and C. R. Cantor 1969 Evolution of protein molecules. pp 21-132. In: *Mammalian Protein Metabolism*, H.N. Munro (ed.), Academic Press, New York.

Kawata, M. 1995 Effective population size in a continuously distributed population. *Evolution* **49**: 1046-1054.

Kendall, M. G. and A. Stuart 1969 The Advanced Theory of Statistics. Vol. 1 *Distribution Theory.* Charles Griffin and Company Limited, London.

Kenneth, H. W., W. S. Li and P. M. Sharp 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA* **84**: 9054-9058

Kimura, M. 1953 " Stepping-stone" model of population. *Annual Report of National institute of Genetic* **3**: 62-63

Kimura, M. 1964 Diffusion models in population genetics. *Journal of Applied Probability* **1**: 177-232.

Kimura, M. 1968 Evolutionary rate at the molecular level. *Nature* **217**: 624-626

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

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**: 111-120

Kimura, M. 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

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

King, J. N., B. P. Dancik and N. K. Dhir 1984 Genetic structure and mating system of white spruce (*Picea glauca*) in a seed production area. *Canadian Journal of Forest Research* **14**: 639-643

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

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

Knowles, P., G. R. Furnier, M. A. Aleksiuk and D. J. Perry 1987 Significant levels of self-fertilisation in natural populations of tamarack. *Canadian Journal of Botany* **65**: 1087-1091

Le Corre, N. Machon, R.J. Petit and A. Kremer, 1997 Colonization with long-distance seed dispersal and genetic structure of maternally inherited genes in forest trees: a simulation study. *Genetic Research* **69**: 117-125.

Le Corre, V. S. Dumolin-Lapegue and A. Kremer 1997 Genetic variation at allozyme and RAPD loci in sessile oak *Quercus petraea* (Matt.) Liebl.: the role of history and geography. *Molecular Ecology* **6**: 519-529

LePage, B.A. and J.F. Basinger 1991 A new species of *Larix* (Pinaceae) from the early Tertiary of Axel Heiberg Island, Arctic Canada. *Review of Palaeobotany and Palynology* **70**: 89-111

LePage, B.A. and J.F. Basinger 1995 The evolutionary history of the genus *Larix* (Pinaceae). pp. 19-29 In: *Ecology and Management of Larix Forests: A look Ahead. Proceeding of an international symposium*, UT, USA.

Levin, D.A. and H.W. Kerster 1968 Local gene dispersal in *Phlox*. *Evolution* **22**: 130-139.

Levin, D.A. and H.W. Kerster 1971 Neighbourhood structure in plants under diverse reproductive methods. *American Naturalist* **105**: 345-354.

Levin, D.A. and H.W. Kerster 1974 Gene flow in seed plants. *Evolutionary Biology* 7: 139-220.

Levins, R. 1970 Extinction. *American Mathematics Society* **2**: 75-108.

Lewandowski, A., J. Burczyk and L. Mejnartowicz 1991 Genetic structure and mating system in an old stand of Polish larch. *Silvae Genetica* **40**: 75-79.

Li, C. C. 1976 *First Course in Population Genetics*. The boxwood Press, Pacific grove, California.

Li, L., C.P. Yang, Y.X. Liu, and L.Z. Qi 1991 A study on the geographic variation of the peroxide isoenzyme of *Larix gmelinii. J. Northeast Forest University* **19** (Supp.): 84-89 (In Chinese)

Li, W.S. 1976 Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. *Theoretical Population Biology*. **10**: 303-308

Liu, Z.W. and P. Knowles 1991 Pattern of allozymes variation in tamarack (*Larix laricina*) from northern Ontario. *Canadian Journal of Botany* **69**: 2468 - 2474.

Liu,Z. and G.R. Furnier 1993 Comparison of allozyme, RFLP, and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. *Theoretical and Applied Genetics* **87**: 97-105

Ma, C.G. 1992 Brief Introduction of research on *Larix* spp. pp1-8. In: *Selection of Optimum Species and Seed Sources for Plantations of Larch.* C.G. Ma (Ed). Beijing Agricultural Uni. Press.(In Chinese)

Ma, C.G, and J.H.,Wang 1992 The optimum planting region delimitation for *Larix principis-rupprechtii.* pp30-37. In: *Selection of Optimum Species and Seed Sources for Plantations of Larch.*, C.G. Ma (Ed). Beijing Agricultural Uni. Press. (In Chinese)

Ma, C.G. and H. Tao 1992 Selection of species and provenances and delimitation of silvicultural regions of larch in China. pp9-21. In: *Selection of Optimum Species and Seed Sources for Plantations of Larch.* C.G. Ma (Ed.). Beijing Agric. Univ. Press.(In Chinese)

Malécot, G. 1948 Les mathematiques de l'heredite. Masson, Paris

Malécot G. 1969 " *The Mathematics of Heredity* " translated by D.M. Yermanos, Freeman, San Francisco.

Mallet, J. 1996 The genetics of biological diversity: from varieties to species. pp13-53. In. *Biodiversity: A biology of numbers and difference.* Gaston, K.J. (Ed.). Blackwell Science Ltd.

Mantel, N. 1967 The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**: 209-220

Martinsson, O. (Ed.) 1995a *Proceedings Larch Genetics and Breeding*, UMEÅ, Sweden.

Martinsson, O. 1995b Systematics and differentaition in the genus *Larix* in Eurasia-Proposal for an international research project. p93-98 In: *Proceedings Larch Genetics and Breeding* , O. Martinsson (Ed.), UMEÅ, Sweden.

Maruyama, T. 1970 Effective number of alleles in a subdivided population. *Theoretical Population Biology* **1**: 273-306

Maruyama, T. and M. Kimura 1980 Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *Proceedings of the National Academy of Sciences USA* 77: 6710-6714

Mason-Gamer, R.J., K.E. Holsinger and R.K. Jansen 1995 Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae). *Molecular Biology and Evolution* **12**: 371-381

McCauley, D.E. 1994 Contrasting the distribution of chloroplast DNA and allozyme polymorphism among local populations of *Silene alba*: implications for studies of gene flow in plants. *Proceedings of the National Academy of Sciences USA* **91**: 8127-8131

McCauley, D.E. 1995 The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends in Ecology and Evolution* **10**: 198-202

McCauley, D.E. 1997 The relative contributions of seed and pollen movement to the local genetic structure of *Silene alba*. *Journal of Heredity* **88**: 257-263

McCauley, D.E., J. Raveill and J. Antonovics 1995 Local founding events as determinants of genetic structure in a plant metapopulation. *Heredity* **75**: 630-636.

McCauley, D.E., J.E. Stevens, P.A. Peroni and J.A. Raveill 1996 The spatial distribution of chloroplast DNA and allozyme polymorphisms within a population of *Silene alba* (*Caryophllaceae*). *American Journal of Botany* **83**: 727-731

Milgroom, M.G., S.E. Lipari, R.A. Ennos and Y.-C. Liu, 1993 Estimation of the outcrossing rate in the chestnut blight fungus, *Cryphonectria parasitica*. *Heredity* **70**: 385-392

Millar, C.I. 1983 A step cline in *Pinus muricata*. *Evolution* **37**: 311-319

Mitton, J.B. 1983 Conifers. pp 443-472. In: *Isoazymes in Plant Genetics and Breeding*, Part B. S.D. Tanksley and T.J. Orton (Eds.), Elsevier Science Publishers B.V., Amsterdam.

Mitton, J.B. 1992 The dynamic mating system of conifers. *New Forests* **6**: 197-216

Mitton, J.B., Y.B. Linhart, M.L. Davis, and K.B. Sturgeon 1981 Estimation of outcrossing in ponderosa pine, *Pinus ponderosa* Laws., from patterns of segregation of protein polymorphisms and from frequencies of albino seedlings. *Silvae Genetica* **30**: 117-121.

Mogensen, H.L. 1996 The hows and whys of cytoplasmic inheritance in seed plants. *American Journal of Botany* **83**: 383-404

Morgante, M. and A.M. Olivieri 1993 PCR-amplifed microsatellites as markers in plant genetics. *Plant Journal* **3**: 175-182

Mullis, K., F. Faloona, S. Scharf, R. Saiki, G.Horn and H. Erlich, 1986 Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbour Symposia on Quantitative Biology* **LI**: 263-274

Mummenhoff, K, E. Kuhnt , M. Koch and K. Zunk 1995 Systematic implications of chloroplast DNA variation on *Lepidium* sections *Cardamon*, *Lepiocardamon* and *Lepia* (Brassicaceae). *Plant Systematics and Evolution* **196**: 75-88

Nagylaki, T. 1975 Conditions for the extence of clines. *Genetics* **80**: 595-615.

Nagylaki, T. 1976 Clines with variable migration. *Genetics* **83**: 867-886.

Nagylaki, T. 1978a Random genetic drift in a cline. *Proceedings of the National Academy of Sciences USA* **75**: 423-426.

Nagylaki, T. 1978b Clines with asymmetric migration. *Genetics* **88**: 813-827

Nagylaki, T. 1979 The island model with stochastic migration. *Genetics* **91**: 163-176

Nagylaki, T. 1983 The robustness of neutral models of geographical variation. *Theoretical Population Biology* **24**: 268-294

Nagylaki, T. 1997 The diffusion model for migration and selection in a plant population. *Journal of Mathematical Biology* **35**: 409-431

Neale, D.B. and W.T. Adams 1985a Allozyme and mating system variation in balsam fir (*Abies balsamea*) across a continuous elevational transect. *Canadian Journal of Botany* **63**: 2448-2453

Neale, D.B. and W.T Adams 1985b The mating system in natural and shelterwood stands of Douglas-fir. *Theoretical and Applied Genetics* **71**: 201-207

Neale, D.B. and R.R. Sederoff 1989 Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. *Theoretical and Applied Genetics* 77: 212-216

Neale, D.B, N.C.Wheeler and R.W. Allard 1986 Paternal inheritance of chloroplast DNA in Douglas-fir. *Canadian Journal of Forest Research* **16**: 1152- 1154

Neale, D.B, K.A. Marshall and R.R. Sederoff 1989 Chloroplast and mitochondrial DNA are paternally inherited in *Sequoia sempervirens* D. Don Endl. *Proceedings of the National Academy of Sciences USA* **86**: 9347-9349

Neale, D.B., K.A. Marshall AND D.E. Harry 1991 Inheritance of chloroplast and mitochondrial DNA in incense ceder (*Calocedrus decurrens*). *Canadian Journal of Forest Research* **21**: 717-720

Nei M. 1972 Genetic distance between populations. *The American Naturalist* **106** , 283-292.

Nei, M. 1973 Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA* **70**: 3321-3323

Nei, M. 1975 *Molecular population genetics and evolution.* North Holland and American Elsevier, Amsterdam, New York.

Nei M. and M.W. Feldman 1972 Identity of genes by descent within and between populations under mutation and migration pressures. *Theoretical Population Biology* **3**: 460-465

Nei, M. and K. Syakudo 1958 The estimation of outcrossing in natural populations *Japanese Journal of Genetics* **33**: 46-51

Nei, M. and N. Takahata 1993 Effective population size, genetic diversity , and coalescence time in subdivided populations. *Journal of Molecular Evolution* **37**: 240-244

Nei, M. 1987 *Molecular evolutionary genetics*. Colombia University Press, New York.

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

Ohyama,K., T. Kohchi, T. Sano and Y. Yamada 1988 Newly identified groups of genes in chloroplasts *TIBS* **13**: 19-22

Ostenfeld, C.H. and C.S. Larsen 1930 The species of the genus *Larix* and their geographical distribution. *Biologiske meddelelser Kgl. danske videnskabemes* **9**: 1-107. *Kobenhavn, Andr. Fred. Host & son, Bianco Linco bogtryykkeri.*

Palmer, J.D. 1990 Mitochondrial DNA in plant systematics: applications and limitation. pp 36-49. In: *Molecular Systematics of Plants*, P.S. Soltis and D.E. Soltis (Eds.), Chapman and Hall, New York.

Palmer, J.D., R.K. Jansen, H.J. Michaels, M.W. Chase, and J.W. Manhart 1988 Chloroplast DNA variation and plant phylogeny. *Annals of Missouri Botanical Garden* **75**:1180-1206

Peakall, R. P.E. Smouse and D.R. Huff 1995 Evolutionary implication of allozyme and RAPD variation in diploid populations of dioecious buffalo grass *Buchloe dactyloides Molecular Ecology* **4**: 135-147

Petit, R.J., A. Kremer and D.B. Wagner, 1993a Geographic structure of chloroplast DNA polymorphisms in European oaks. *Theoretical and Applied Genetics* **87**: 122-128

Petit R.J., A. Kremer and D.B. Wagner 1993b Finite island model for organelle and nuclear genes in plants. *Heredity* **71**: 630-640

Pigliucci, M. and G. Barbujani, 1991 Geographical patterns of gene frequencies in Italian populations of *Ornithogalum montanum* (Liliaceae). *Genetic Research* **58**: 95-104

Porani, M. and A. Parida 1997 Allozyme and RAPD polymorphism in *Tylophora indica* (Burmf) Merr. *Journal of Plant Biochemistry and Biotechnology* **6**: 29-33

Potenko, V.V. and P.N. Razumov 1996 Genetic variation and population structure of *Larix gmelinii* in the Khabarovsk territory. *Lesovedenie* **5**: 11-18

Powell, W., M. Morgante, R. McDevitt, G. G. Vendramin and J.A. Rafalski 1995 Polymorphic simple sequence repeat regions in chloroplast genomes: application to

population genetics of pines. *Proceedings of the National Academy of Sciences USA* **92**: 7759-7763

QIAGEN Inc. 1995 *The QIAGEN Guide to Template Purification and DNA Sequencing.* pp 1-46.

QIAGEN Inc., 1997 *QIAquick Spin Handbook* p18.

Quicke, D.L.J. 1996 *Principles and techniques of contemporary taxonomy.* Blackie Academic & Professional, London.

Rafaalski, J.A. and S.V. Tingey 1993 Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. *Trends in Genetics* **9**: 275-280

Rannala, B. and J.A. Hartigan 1995 Identity by descent in island-mainland populations *Genetics* **139**: 429-437

Raubeson, L.A. and R.K.Jansen, 1992 A rare chloroplast-DNA structural mutation is shared by all conifers. *Biochemical Systematics and Evolution* **20**: 17-24.

Raymond, M. and F. Rousset 1995 Genepop (version 1.2) a population genetics software for exact tests and ecumeicism. *Journal of Heredity* **86**: 248-249

Rieger, R., A. Michaelis and M.M. Green, 1991 *Glossary of Genetics: Classical and Molecular* (Fifth Edition), Springer-Verlag, Berlin.

Rieseberg, L.H. and D.E. Soltis, 1991 Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* **5**: 65-84.

Ris, H. and Plaut, W. 1962 The ultrastructure of DNA-containing areas in the chloroplast of *Chlamydomans*. *Journal of Cell Biology* **13**: 19-22

Ritland, K. 1983 Estimation of mating systems pp. 289-302. In: *Isozymes in Plant Genetics and Breeding. Part A.*, Tanksley, S.D. and Orton, T.J. (Eds), Elsevier Science Publishers B.V., Amsterdam.

Ritland, K. 1990 A series of FORTRAN computer programs for estimating plant mating systems. *Journal of Heredity* **81**: 235-237

Ritland, K. and S. Jain 1981 A model for the estimation of outcrossing rate and gene frequencies using *n* independent loci. *Heredity* **47**: 35-52

Rousset, F. and M. Raymond 1995 Testing heterozygote excess and deficiency. *Genetics* **140**: 1413-1419

Sambrook, J., E.F. Fritsch and T. Maniatis 1989 *Molecular Cloning: A laboratory manual* 1, 2 , 3 (2nd edition). Cold Spring Harbor Laboratory Press, .

Samuel, R., W. Pinsker and F. Ehrendorfer, 1995 Electrophoretic analysis of genetic variation within and between populations of *Quercus cerris*, *Q. pubescens*, *Q. petraea* and *Q. robur* (Fagceae) from eastern Austria. *Botanica Acta* **108**: 290-299

Sanger, F., S. Nicklen and A.R. Coulson 1977 DNA sequencing with Chain-terminating inhibitors. *Proceedings of the National Academy of Sciences USA* **74**: 5463-5467

Schaal, B.A., W.J. Leverich and S.H. Rogstad 1991 A comparison of methods for assessing genetic variation in plant conservation biology. pp123-134. In: *Genetics and Conservation of Rare Plants*. D.A. Falk and K.E. Holsinger (Eds), Oxford University Press, New York.

Schmidt, W.C. , and McDonald, K.J. (Eds.) **1995** *Ecology and Management of Larix Forests*: *A Look Ahead. Proceedings of an International Symposium*. USDA, UT.

Sharma, S K, I.K. Dawson and R. Waugh 1995 Relationships among cultivated and wild lentils revealed by RAPD analysis. *Theoretical and Applied Genetics* **91**: 647-654

Shaw, D.V. and R.W. Allard 1982 Estimation of outcrossing rates in Douglas-fir using isozyme markers. *Theoretical and Applied Genetics* **62**: 113-120

Shaw, D.V., A.L. Kahler and R.W. Allard 1981 A multilocus estimator of mating system parameters in plant populations. *Proceedings of the National Academy of Sciences USA* **78**: 1298-1302

Shiraish, S. , K. Isoda and H. Kawasaki 1995 Phylogenetic studies on east asian *Larix* species using random amplified polymorphic DNA (RAPD) and *rbc*L sequence (Abstract). p183. In: *Proceedings Larch Genetics and Breeding*, O. Martinsson (Ed.), UMEÅ, Sweden.

Shiaishi, S., K. Isoda, A. Watanabe and H. Kawasaki 1996 DNA systematical study on the *Larix* relic forest at MT. Manokami, the Zao Mountains. *Japanese Journal of Forestry Society* **78**: 175-182.

Slatkin, M. 1973 Gene flow and selection in a cline. *Genetics* **75** : 733-756

Slatkin, M. 1977 Gene flow and genetic drift in species subject to frequent local extinctions. *Theoretical Population Biology* **12**:253-262

Slatkin, M. 1987 Gene flow and geographic structure of natural population. *Science* **236**: 787-792

Slatkin, M. 1989 Detecting small amounts of gene flow from phylogenies of alleles. *Genetics* **121**: 609-612

Slatkin, M. 1991 Inbreeding coefficients and coalescence times. *Genetic Research* **58**: 167-175

Slatkin, M. 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **39**: 53-65

Slatkin, M. and N.H. Barton, 1989 A comparison of three indirect methods for estimating average levels of gene flow. *Evolution* **43**: 1349-1368

Slatkin, M. and W.P. Maddison 1990 Detecting isolation by distance using phylogenies of genes. *Genetics* **126**: 249-260

Slatkin, M. and T. Maruyama 1975 Genetic drift in a cline. *Genetics* **81**: 209-222

Sokal, R.R. 1979 Testing statistical significance of geographical variation patterns. *Systematic Zoology* **28**: 227-232

Sokal, R.R. and N.L. Oden 1978a Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society* **10**: 199-228

Sokal, R.R. and N.L. Oden 1978b Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society* **10**: 229-248

Sorensen, F.C. 1982 The role of polyembryonal vitality in the genetic system of conifers. *Evolution* **36**: 725-120

Sorensen, F. C. and R.S. Miles 1974 Self-pollination effects on Douglas-fir and ponderosa pine seeds and seedlings. *Silvae Genetica* **23**: 135-138

Sorensen, F. C. and R.S. Miles 1982 Inbreeding depression in height, height growth, and survival of Douglas-fir, ponderosa pine , and noble fir to 10 years of age. *Forest Science* **28**: 283-292.

Spooner, D.M., J. Tivang, J. Nienhuis, J.T. Miller, D.S. Douches and M.A. Contreras 1996 Comparison of four molecular markers in measuring relationships among the wild potato relatives *Solanum* section *Etuberosum* (subgenus Potatoe). *Theoretical and Applied Genetics* **92**: 532-540

Stebbins, G.L. 1958 Longevity, habitat, and release of genetic variability in the higher plants. *Cold Spring Harbor Symposia on Quantitative Biology* **23**: 365-378

Strand, A.E., J. Leebens-Mack and B.G. Millgan, 1997 Nuclear DNA-based markers for plant evolutionary biology. *Molecular Ecology* **6**: 113-118

Strauss, S.H. and A.H. Doerksen, 1990 Restriction fragment analysis of pine phylogeny. *Evolution* **44**: 1081-1096.

Strauss, S.H., J.D. Palmer, G. Howe and A.H. Doerksen 1988 Chloroplast genomes of two conifers lack an inverted repeat and are extensively rearranged. *Proceedings of the National Academy of Sciences USA* **85**: 3898-3902

Strauss, S.H., Y.P. Hong and V.D. Hipkins 1993 High levels of population differentiation for mitochondrail DNA haplotypes in *Pinus radiata, muricata,* and *attenuata. Theoretical and Applied Genetics* **86**: 605-611

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

Swoford, D.L. and R.B. Selander 1981 Biosys-1: a FORTRAN program for the comprehensive analysis for electrophoretic data in population genetic and systematics. *Journal of. Heredity* **72**: 281-283

Szmidt, A.E., T. Alden and J.E. Hallgren 1987 Paternal inheritance of chloroplast DNA in *Larix. Plant Molecular Biology* **9**: 59-64

Szmidt, A.E. 1991 Phylogenetic and applied studies on the chloroplast genome in forest conifers. pp 185-196. In: *Biochemical markers in the population genetics of forest trees.* S.Fineschi, M.E. Malvolti, F. Cannata and H.H. Hattemer (Eds.), Academic Publishing, Netherlands.

Szmidt, A.E., X.R. Wang and M.Z. Lu 1996 Empirical assessment of allozyme and RAPD variation in *Pinus sylvestris* (L.) using haploid tissure analysis. *Heredity* **76**: 412-420

Taberlet, P., L. Gielly, G. Patou and J. Bouvet 1991 Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* **17**: 1105-1109

Tajima, F. 1983 Evolutionary relationship of DNA sequences in finite populations *Genetics* **105**: 437-460

Tajima, F. 1989a The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597-601

Tajima, F. 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595

Tajima, F. 1990 Relationship between migration and DNA polymorphism in a local population. *Genetics* **126**: 231-234

Tajima, F. 1993 Measurement of DNA polymorphism. pp37-59. In. *Mechanism of molecular Evolution.* Japan Scientific Societies Press, Tokyo.

Takahata, N. 1983 Gene identity and genetic differentiation of populations in the finite island model. *Genetics* **104**: 497-512.

Takahata, N. 1988 The coalescent in two partially isolated diffusion populations. *Genetic Research* **52**: 213-222

Takahata, N. 1991 Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129**: 585-595

Takahata, N. and M. Nei 1984 Fst and Gst statistics in the finite island model. *Genetics* **107**: 501-504

Takahata, N. and M. Nei 1985 Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**: 325-344.

Takahata, N. and S.R. Palumbi 1985 Extranuclear differentiation and gene flow in the finite island model *Genetics* **109**: 441-457.

Takahata, N. and M. Slatkin 1990 Genealogy of neutral genes in two partially isolated populations. *Theoretical Population Biology* **38**: 331-350

Tanksley, S.D. and T.J. Orton, T.J. 1983 (Eds) *Isozymes in Plant Genetics and Breeding*. Part A, B. Elsevier Science Publishers B.V., Amsterdam.

Tang, Q., R.A. Ennos and T.Helgason 1995 Genetic relationship among larch species based on analysis of restriction fragment variation for chloroplast DNA. *Canadian Journal of Forest Research* **25**: 1197-1202

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

Thomas, D.A and H.N. Barber 1974 Studies on leaf characteristics of a cline of *Eucalyptus urnigera* from Mount Wellington, Tasmania. I. Water repellancy and the freezing of leaves. *Australia Journal of Botany* **22**: 501-512

Tsumura, Y., H. Taguchi, Y. Suyama and K. Ohba 1994 Geographical cline of chloroplast DNA variation in *Abies mariesii. Theoretical and Applied Genetics* **89**: 922-926

Vickery, R.K. Jr 1990 Close correspondence of allozyme groups to geographic races in the *Mimulus glabratus* complex (Scrophulariaceae). *Systematic Botany* **15**: 481-496

Wang, X.R. and A.E. Szmidt 1993 Chloroplast DNA-based phylogy of Asian *Pinus* species (Pinaceae). *Plant Systematics and Evolution* **188**: 197-211

Wade, M.J. and D.E. McCauley 1988 Extinction and recolonization: their effects on the genetic differentiation of local populations. *Evolution* **42**: 995-1005

Wang, J.Z. and Z.F. Ding 1989 Study on the heterosis of *Larix* and its application. *Hereditas* **4**: 1-4

Wang, Y.C 1995 Physical ecology and regulation measurement for establishment of fast-growing and high-yield larch forests in northeastern China. pp. 79-80. In: *Ecology and Management of Larix Forests: A look Ahead. Proceeding of an international symposium.* USDA, UT.

Watterson, G.A. 1975 On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **10**: 256-276

Watterson, G.A. 1984 Lines of descent and the coalescent. *Theoretical Population Biology* **26**: 77-92.

Weir, B.S. and C.C. Cockerham 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **36**: 1358-1370

Weir, B.S. 1990 *Genetic Data Analysis*. Sinauer Associates, Inc., Publishers, Sunderland.

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

Wheeler, N.C. and R.P. Guries 1982 Population structure, genic diversity, and morphological variation in *Pinus contorta* Dougl. *Canadian Journal of Forest Research* **12**: 595-606

White, T.J. 1996 The future of PCR technology: diversification of technologies and applications. *Trends in Biotechnology* **14**: 478-483

Whitlock, M.C. 1992 Temporal fluctuation in demographic parameters and the genetic variance among populations. *Evolution* **46**: 608-615

Williams, J.G. K., A.R. Kubelik, J.A. Rafalski, S.V. Tingey 1990 DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nuclear Acid Research* **18**: 6531-6535.

Wolfe, K.H., M.Gouy, Y.W. Yang, P.M. Sharp and W.H. Li, 1989 Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proceedings of the National Academy of Sciences USA* **86**: 6201-6205

Wolfe, K.H., W.H. Li and P.M. Sharp 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA* **84**: 9054-9058

Wright, S. 1921 Systems of mating. *Genetics* **6**: 111-178

Wright, S. 1931 Evolution in mendelian populations. *Genetics* **16**: 97-159.

Wright, S. 1940 Breeding structure of populations in realtion to speciation. *Ammican Naturalist* **74**: 232-248

Wright, S. 1943 Isolation by distance. *Genetics* **28**: 114-138

Wright, S. 1946 Isolation by distance under diverse systems of mating. *Genetics* **31**: 39-59

Wright, S., 1948 On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution* **2**: 279-294

Wright, S. 1951 The genetical structure of populations. *Annals of Eugenics* **15**: 323-354

Wright, J. W. 1952 Pollen dispersion of some forest trees. *U.S. Forest Service, Northeast Forest Experimental Station Paper* No. **46**

Wright, S. 1968 Evolution and the Genetics of Populations. Vol.1. *Genetic and Biometric Foundations*. The University of Chicago Press, Chicago.

Wright, S. 1969 Evolution and the Genetics of Populations. Vol. 2. *The Theory of Gene Frequencies*. The University of Chicago Press, Chicago.

Wright, S. 1978 Evolution and the Genetics of Populations. Vol. 4. *Variability within and among Natural Populations.* The University of Chicago Press, Chicago.

Yang, C.P., S.H. Qin, W. Zhang, B.J. Yu, P. Zhang and P.G. Zhang 1990a A study on provenance test of dahurian larch in China (II) — a division of provenance. *Journal of Northeast University* **18** (Supp.): 25-33 (In Chinese)

Yang, C.P., W. Zhang, B.J. Yu, S.H.Qin, Z.X. Sang, L.X. Zhang, X.D. Sheng, and Q. Dong 1990b A study on provenance test of dahurian larch in China. *Journal of Northeast University* **18** (Supp.):18-24 (In Chinese)

Yang, C.P., S.W.Yang, D.A. Xia, G.F. Liu, Q.Y. Lu, H.S. Zheng and P.G. Zhang, 1991 Study on the geographic variation rule and pattern of growth characters of *Larix olgensis*. *Journal Northeast University* **19** (Supp.): 10-18 (In Chinese)

Yang, S. W., 1995 *Genetic Improvement of Larix* Northeast Forestry University Press, Harbin, Hei Rong Jiang Province, China. (In Chinese)

Yang, S.W., Y.G. Jin, S.Y. Zhang , Q.F. Lin, C.W. Han and F.R. Meng 1985 Research on larch hybrid vigor. *Journal of Northeastern Forestry College* **13**: 30-36 (In Chinese)

Ying, L. and E.K. Morgenstern 1991 The population structure of *Larix laricina* in New Brunswick, Canada. *Silvae Genetica* **40**: 180-184

Young, N.D. 1996 Concordance and discordance: A tale of two hybrid zones in the *Pacific Coast irises* (*Iridaceae*). *American Journal of Botany* **83**: 1623-1629

Zanetto, A., G. Roussel and A. Kremer 1994 Geographic variation of inter-specific differentiation between *Quercus robur* L. and *Quercus petraea* (Matt) Liebl. *Forest Genetics* **1:** 111-123

Zhang, S.Y and Z. Wang 1992 *Chinese Larix.* Chinese Forest Press. (In Chinese).

Zhang, X.F., L.B. Zhou and M.X. Li 1985 A study of Karyotype of 5 species in *Larix*. *Hereditas* 7: 9-11 (In Chinese)

Zheng, W.J. (Ed.) **1983** *Sylva Sinica* Vol.**1**: 237-253. Chinese Forestry Press, Beijing. (In Chinese)

Zhou, Jin, 1962 *Scienta Silvae Sinicae* **2**: 97-116 (In Chinese)

Zobel, B. and J. Talbert 1984 *Applied Forest Tree Improvement.* Waveland Press, Inc., New York.

Appendix I. Comprehensive Check List for Larix Species and Their Varieties (Translated from *CHINESE LARIX* ed. by Zhang, S Y.*et al.*, 1992)

1. Cones reniform, or longly reniform , or cup form, or ellipse; bracts-scales shorter than cone-scales, not exposed or slightly exposed for the basal bract-scales of cone; cone-scales smooth, shining, or pilose

.....Sect.1 *Larix*

2. Cone-scales smooth, shining;

3. Cone from cup form to ellipse, length 1.5--2.0 (2.5) cm; average number of cone-scales is 20, seldom 30

.....*L. gmelini*

3. Cone from reniform to widly reniform, length 2.0--2.7 cm, cone-scales average more than 30, seldom less than 30 ;

4. Cone-scales number 20--30; one-year-old shoots light-yellow, stout

.....*L. principis-ruppechtii* var. *wulingshaneneis*

4. Cone--scales number more than 30; one-year-old shoots colour from black-red-brown to brown-yellow

.....*L. principis-rupprechtii*

2. Cone-scales pilose;

5. Edges of cone-scales not recurved,or slightly recurved, or emarginate;

6. Cone-scales thinner, edges not recurved, or slightly recurved; ripe cones colour becomes thick,from red-brown to brown; cone length somewhat longer than , or near the same to its width;

7. Cone-scales upsidedown reniform,or widly reniform, or near round; gland tumor and slightly pilosity found on under-side of leaves;

8. Cone length 1.4--3.0 cm

.....*L. olgensis* var. *koreana*

8. Cone length above 3.0 cm

.....*L.olgensis* var. *changpaiensis*

7. Cone-scales pentagonal reniform; gland tumor and densely pilosity found on under side of leaves

.....*L.olgensis* var. *heilingensis*

6. Cone-scales thicker, edges emarginate, or straight; ripe cones colour becomes thinner, lightly brow, or lightly yellow-brown; cone length longer than its width;

9.Cone-scales roundly emarginate, triangularly reniform, densely light-purple pilosity found on under side; seeds wing not longer than cone-scale; one-year-old shoots light-yellow,stout

.....*L. sibrica*

9. Cone-scales on the mid of cones , their edges emarginate and slightly recurved, reniform, or widly reniform; densely grey-brown pilosity found on under side of leaves; one-year-old shoots light-yellow, thin

.....*L. decudua*

5. Edges of cone-scales obviously recurved

.....*L. kaempferi*

1. Cones cylindrical, or reniformly cylindrical; bracts-scales longer than cone-scales, exposed ..... Sect 2. *Multiseriales*

10. Bract-scales recurved;

11. Cones big, length 5.0--11.0 cm ..... *L. griffithiana*

11. Cones small, length 2.5--4.0 cm ..... *L. mastersiana*

10. Bract-scales straight, or recurved,or slightly recurved;

12. Cones small,length 4.0 cm or so;

13. Cone-scales flatly round, near 90 degree against to fruit axle; one-year-old shoots yellow, or light-brown ..... *L. chinensis*

13. Cone-scales round, small angle against to axle; one-year-old shoots red-brown ..... *L. potaninii*

12. Cones big, length 5.0--11.0 cm ..... *L. potaninii* var. *macrocarpa*

14. Cone-scales thinner,purple, or red-brown; one-year-old shoots red-brown ..... *L.speciosa*

14. Cone-scales thicker, grey-brown; one-year-old shoots yellow, or lightly yellow-brown ..... *L.himalaica*

Appendix II. Recipes for the enzyme systems employed in this study

Enzyme	Recipes†		Electrophoresis	Incubation conditions
Aspartate aminotransferase (AAT, E.C.2.6.1.1)	2mg	pyridoxal-5′-phosphate	System I	In the dark at 37°C for 30 min.
	50mg	fast blue BB salt		
	25ml	substrate solution (5.30g L-aspartic acid + 0.70g *a*-ketoglutaric acid; dissolved in 1.0 l of 0.2M Tris-HCl, pH 8.0)		
Malate dehydrogenase (MDH; EC.1.1.1.37)	12.5ml	0.2M Tris-HCl, pH 8.0	System II	In the dark at 37°C for 45 min.
	12.5ml	0.5M DL-malic acid, pH 7.0		
	0.5ml	NAD		
	0.5ml	NBT		
	0.5ml	PMS		
6-phosphogluconate dehydrogenase (6PGD; E.C. 1.1.1.44)	5ml	0.2M Tris-HCl, pH 8.0	System II	In the dark at 37°C for 45min.
	10mg	6-phospogluconic acid		
	1ml	1% $MgCl_2$ (w/v)		
	1ml	NADP		
	1ml	MTT		
	0.5ml	PMS		

Phosphoglucose isomerase (PGI; E.C.5.3.1.9)	5ml	0.2M Tris-HCl, pH8.0	System I	In the dark at 37°C for 60min.
	12.5mg	fructose-6-phosphate,		
	5units	glucose-6-phosphate dehydrogenase		
	0.5ml	1%$MgCl_2$ (w/v)		
	0.5ml	NADP		
	0.5ml	MTT		
	0.5ml	PMS		
Phosphoglucomutase (PGM; E.C.2.7.5.1)	25ml	0.2M Tris-HCl, pH 8.0	System II	In the dark at 37°C for 60min.
	150mg	glucose-1-phosphate,		
	25units	glucose-6-phosphate dehydrogenase		
	0.5ml	1%$MgCl_2$ (W/V)		
	0.5ml	NADP		
	0.5ml	MTT		
	0.5ml	PMS		
Shikimic acid dehydrogenase (SDH; E.C.1.1.1.25)	50ml	0.2M Tris-HCl, pH8.0	System II	In the dark at 37°C for 60min
	25mg	shikimic acid		
	1.0ml	1%$MgCl_2$ (w/v)		
	0.5ml	NADP		
	1.0ml	NBT		
	0.5ml	PMS		

†: stock solution: NAD (β-nicotinamide adenine dinucleotide), 10 mg/ml; NADP (β-nicotinamide adenine dinucleotide phosphate), 10 mg/ml; NBT (nitro blue tetrazolium), 10 mg/ml; MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide), 10 mg/ml; PMS (phenazine methosulfate), 10 mg/ml.

Appendix III. DNA techniques

III.1 DNA extraction

Total DNA was extracted from each species sample of buds or needles using a CTAB (cetyltrimethylammonium bromide) extraction method (Mosselar, *et.al.*, 1992). One to two grammes of buds or needles were ground to a very fine powder in liquid nitrogen so that the cell walls were broken to release the cellular constituents. 5ml of 2 × CTAB buffer (1.4M NaCl, 100mM Tris base pH8.0, 20mM EDTA, 2% (v/v) Hexadecyltrimerthylammonium bromide) at 65°C was then added and mixed thoroughly so that the cell membranes were disrupted and the DNA was released into the extraction buffer. These tubes were transferred to a water bath at 65°C, and were incubated for 45~60 minutes, then allowed to cool to room temperature. An equal volume of chloroform: isoamyl alcohol (24:1) was then added and mixed thoroughly to a single phase so as to denature and separate the proteins from DNA. Tubes were spun at 3,200rpm in bench centrifuge for 10 minutes. The upper aqueous layer was transferred to a fresh tube and treated with chloroform:isoamyl alcohol a second time. The upper aqueous layer was transferred to a fresh tube and an equal volume of isopropanol (-20°C) was added. Tubes were left at -20°C overnight to precipitate, then spun at 3,200rpm for 10 minutes. The resulting pellet of DNA was resuspended in 2ml TE buffer (10mM Tris-HCl pH7.6 and 1mM EDTA). Then 2ml of phenol:chloroform (1:1) was added and mixed for further separating proteins from DNA. After spinning at 3,200 rpm for 10 minutes, the upper aqueous layer was transferred to a fresh tube. One-tenth volume of 3M sodium acetate and two volumes absolute ethanol were added and mixed well for removing CTAB. Tubes were placed at -20°C for one hour, and spun at 3,200rpm for 10 minutes. The pellet of DNA was air-dried. The pellet was resuspended in 400ul of TE buffer plus RNAse (200 ug/ml) to digest RNA. DNA was then stored at -20°C for future use.

III.2 Setting up the PCR reaction

The template for PCR amplification in this study consisted of 40ng of total genomic DNA. The PCR reaction mixture contained: 10mM Tris-HCl (pH 8.8), 1.5mM $MgCl_2$, 50mM KCl, 0.1% Triton X-100, 100 μM each of dTTP, dGTP, dCTP, dATP, and 1 unit of DNA polymerase DyNAZyme ™ II or Promeag Taq. Amplification was carried out in a final volume of 50μl using 1 cycle of 5 minutes at 94°C (denature), followed by 40 cycles of one minute at 94°C (denature), one minute 54°C (annealing), two and half minutes at 72 °C (extension). After that, a final ten minutes extension was performed at 72 °C. Then samples were held at 4°C. PCR products were visualised by UV transillumination after electrophoresis in agrose gel (1%-2%), stained wih ethidum bromide (0.5 ug/ml).

III.3 Restriction enzymes used in the experiments, their recognition sequence and recipes.

Restriction endonuclease	Recognition sequence*	Reaction mixture	Incubation	Supplier
Alu I	5'...AG↓CT...3'	6μl DNA, 0.3μl enzyme (10 units /μl), 8.0μl H_2O 2μl buffer (10mM Tris-HCl pH7.5, 50mM NaCl, 6mM $MgCl_2$, 1mM DTT, 0.1m EDTA, 0.5mg/ml BSA, 50% v/v Glycerol)	37°C, overnight	Promega
Cfo I	5'...GCG↓C...3'	6μl DNA, 0.3μl enzyme (10units/μl), 8.0μl H_2O 2μl buffer (20mM Tris-HCl pH7.8, 10mM MgCl2, 1.0mM dithiothreitol, 0.1mM EDTA, 0.5mg/ml BSA, 50% v/v Glycerol)	37°C, overnight	GeneScience
Hsp 92 II	5'...CATG↓...3'	6μlDNA, 0.3μl enzyme(10units/μl),8.0μl H_2O 2μl buffer (10mMTris-HCl pH7.4 , 50mM NaCl, 10mM $MgCl_2$, 0.1mM EDTA, 0.5mg/ml BSA, 50% v/v Glycerol)	37°C, overnight	Promega
Mbo I	5'...↓GATC...3'	6μlDNA, 0.3μl enzyme (10units/μl), 8.0μl H_2O 2μl buffer (10mM Tris-HCl pH7.5, 50mM NaCl, 10mM $MgCl_2$,1mMDTT, 0.1mM EDTA, 0.5mg/ml BSA, 50% v/v Glycerol)	37°C, overnight	Promega

*: The symbol ' ↓ ' marks the cut site for each 4bp sequence recognised by different enzymes.

III. 3 (*Continued*)

Restriction endonuclease	Recognition sequence	Reaction mixture	Incubation	Supplier
Msp I	5'...C↓CGG...3'	6μl DNA, 0.3μl enzyme (14 units/μl), 8.0μl H_2O 2μl buffer (10mM Tris-HCl pH7.4, 50mM KCl, 10m $MgCl_2$, 1mM dithiothreitol, 0.1mM EDTA, 200μg/ml BSA, 50%v/v Glycerol)	37°C, overnight	Gene Science
Rsa I	5'...GT↓AC...3'	6μl DNA,0.3μl enzyme(12units/μl),8.0μl H_2O 2μl buffer (10mM Tris-HCl pH7.4, 50mM NaCl, 10mM $MgCl_2$, 0.1mM EDTA, 1mM DTT, 0.5mg/ml BSA, 50%v/v Glycerol)	37°C, overnight	Promega
Taq I	5'...T↓CGA...3'	6μl DNA, 0.3μl enzyme (12units/μl),8.0μl H_2O 2μl buffer (20mM Tris-HCl pH7.4, 50mM KCl, 10m $MgCl_2$, 0.1mM EDTA, 1mM DTT, 0.5mg/ml BSA, 50%v/v Glycerol)	65°C, overnight	Gene Science
Tru 9I	5'...T↓TAA...3'	6μl DNA,0.3μl enzyme(16units/μl),8.0μl H_2O 2μl buffer (10mM Tris-HCl pH7.4, 50mM NaCl, 10mM $MgCl_2$, 0.1mM EDTA, 1mM DTT, 0.5mg/ml BSA, 50%v/v Glycerol)	37°C, overnight	Promega

III.4 Preparation of agarose gel and electrophoresis

When setting up a gel, the required amount of agarose was added to the correct amount of TAE buffer (0.04M Tris-acetate, 0.001M EDTA pH8.0), to produce 1.0 to 2% gel. The mixture was heated for 1~2 min at full power in microwave to dissolve the agarose. The gel mould was sealed with tape. Then agarose solution, when handed cool a little bit, was poured into the mould. A comb(s) was then put into the gel, and it was left to set for 30min. Two sizes of agarose gels were used in this study: size 8 and 14 gels. For example, the required amounts of TAE buffer and agarose for 1% agarose gel are shown in the following table.

Amount of the TAE buffer and agarose required for 1% gel

Gel Size	TAE Buffer	Agarose
14	100ml	1.0g
8	25ml	0.25g

When electrophoresis was run, the required amounts of DNA sample, DNA ladder marker (0.2μg/μl), and gel loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF and 40% (w/v) sucrose in water) were different depending upon gel size. The amount of these indigents used in this study is listed in the following table. The DNA samples were loaded into a gel and electrophoresis took place under the appropriate voltages and times shown in the following table (also see Sambrook, *et al.*, 1989). Staining of the gel by ethidium bromide (0.5μg/ml) was carried out after electrophoresis.

Electrophoresis set up in this study

Gel size	DNA Sample	Ladder	Loading Buffer	Water	Voltage/Time
14	16.0µl		4.0µl	0.0µl	90V/3hrs.
		3.0µl	4.0µl	13.0µl	
8	10.0µl		2.0µl	0.0µl	65V/45min
		2.0µl	3.0µl	7.0µl	

III.5 DNA sequencing

The total sequencing procedure is composed of four steps: (i) purification of PCR products; (ii) cycle sequencing; (iii) purification of extension products, and (iv) electrophoresis, followed by data collection.

(i) *Purification of PCR products*

The QIAquick PCR purification kit protocol (QIAGEN, 1997) was employed to purify the amplified products. The method is simple and pure DNA products can be obtained easily for sequencing. Methodology is as follows: 200 ml of Buffer PB , provided by QIAquick for efficient recovery of DNA and removing contaminants, was added to the PCR reaction (about 40µl) and mixed. It was not necessary to remove the mineral oil layer. A QIAquick spin column was then placed in the provided 2-ml collection tube. Then sample was transferred to the QIAquick column and was spun at 13000rpm for 30 ~ 60 seconds to bind DNA to the column. The flow-through aqueous solution was discarded. The QIAquich column was retained in the same tube. 0.75ml of Buffer PE, provided by QIAquick, was then added to the column and it was spun at 13000rpm for 30~60 seconds, to wash the DNA. The flow-through liquid was discarded a second time. Then the QIAquick column was placed back in the same tube, spun for an additional 1min at 13,000rpm. Then the column was placed into a clean 1.5ml microfuge tube. 30µl of elution buffer (10mM Tris-HCl, pH7.5) was then added to the centre of the QIAquick column, left to stand for 1 min,

then spun for 1 min at 13000rpm. PCR products purified was then stored at -20 °C for sequencing analysis.

(ii) *PCR cycle sequencing*

A fluorescent dye terminator method was employed to sequence the PCR products. For each reaction, the following reagents were mixed in a labelled 0.6ml thin-wall tube:

Setting up of sequencing PCR

Reagent	Quantity
Terminator ready reaction mix	8 µl
Template DNA	*x* ng
Primer	5 pmol
H_2O	*y* µl
Final reaction volume	20µl

The amount of template DNA per reaction could be estimated according to the empirical formula: $x(\text{ng}) = \text{DNA Length (bp)} \times 0.08$. For example, if the length of PCR product is 500bp, then the required number of template DNA for cycle sequencing is 40ng. The *y* is the volume of water required to bring the reaction to a final total 20 µl.

After setting up the reaction, each sample was overlaid with 10 µl light mineral oil and placed in the thermal cycler. 20-30 cycles were then carried out using the following temperature sequence: 96°C (denature), 30s, 45°C (annealing), 15s, and 60°C (extension), 4min. Samples were then held at 4°C.

(iii) *Purifying extension products*

After the cycling sequencing reaction, the amount of the dye terminators in the PCR products were significantly reduced. Excess terminators were removed by the ethanol precipition method. The methodology is as follows: Each sample (20 µl) was transferred to a 1.5ml microcentrifuge tube. 2.0 µl of 3M sodium acetate pH4.6 and 50µl 95% ethanol

(stored at -20°C) were added to each reaction and mixed well. They were then placed on ice for 15 minutes to precipitate DNA. Each sample was centrifuged at maximum speed (13,000 rpm) for 15~30 minutes. The ethanol solution was carefully aspirated with a micropipetter as completely as possible. 250µl of 70% ethanol was added to each sample to rinse the DNA pellet, and brief centrifugation was required. The ethanol solution was then carefully aspirated again with a micropipetter and the pellet was dried at room temperature for 15 minutes, then stored at -20 or 4°C before use.

(iv) *Electrophoresis and data collection*

Each DNA pellet was resuspended in 4µl of loading dye, and samples were heated to 70°C for 2-5 minutes to denature the DNA. They were then placed on ice to prevent DNA renaturation, and immediately loaded into a gel, on a 377 DNA sequencer machine (ABI PRISM™). Base sequences were recorded according to the fluorescent signal and were analysed using the Gene Jockey II sequence processor.

Appendix IV. Proof of the validity of eqn(6.22) in two-dimensional stepping-stone model of plant population genetic structure

Supposing that L_1 and L_2 separately satisfy the eqn(6.22), then we will show that the $L_1 + L_2$ also satify the eqn(6.22).

$$
\begin{aligned}
&E[(L_1 + L_2)\tilde{p}(k)\cdot(L_1 + L_2)\tilde{p}(0)] \\
&\quad = E\{[L_1\tilde{p}(k) + L_2\tilde{p}(k)]\cdot[L_1\tilde{p}(0) + L_2\tilde{p}(0)]\} \\
&\quad = E[L_1\tilde{p}(k)\cdot L_1\tilde{p}(0) + L_1\tilde{p}(k)\cdot L_2\tilde{p}(0) + L_2\tilde{p}(k)\cdot L_1\tilde{p}(0) + L_2\tilde{p}(k)\cdot L_2\tilde{p}(0)] \\
&\quad = E[L_1\tilde{p}(k)L_1\tilde{p}(0)] + E[L_1\tilde{p}(k)L_2\tilde{p}(0)] + E[L_2\tilde{p}(k)L_1\tilde{p}(0)] + E[L_2\tilde{p}(k)L_2\tilde{p}(0)] \\
&\quad = L_1^2\rho(k) + L_1L_2\rho(k) + L_2L_1\rho(k) + L_2^2\rho(k) \\
&\quad = (L_1 + L_2)^2\rho(k)
\end{aligned}
$$

Thus, if $L_1 = \sum_{i=0}^{1}\sum_{j=0}^{1}\beta_{ij}(S_1^{-i} + S_1^{i})(S_2^{-j} + S_2^{j})$

and $L_2 = \beta_{02}S_1^0(S_2^{-2} + S_2^2) + \beta_{20}(S_1^{-2} + S_1^2)S_2^0$, then we can show that they separately satisfy eqn(6.22). Thus, the eqn(6.22) also holds for $L_1 + L_2$ (eqn(6.38)).

Appendix V. Effective population size and Gst calculation

V.1 Effective population size of subdivided population for haploid genes.

Consider a locus with allele of haploid gene A and a, with frequencies q and 1-q in a population. Define the average inbreeding coefficient of individuals, F, as the correlation between two haploid individuals randomly drawn from the population. The genetic compositions of any pair of genes randomly drawn from the population are the similar to those for diploid genes as shown by Wright (1943), i.e.

Geneotype pair	Frequency	
AA	$x_t = q^2(1-F)+qF$	
Aa	$y_t = 2q(1-q)(1-F)$	(A1)
aa	$z_t = (1-q)^2(1-F)+(1-q)F$	

where the inbreeding, F, is caused by population subdivision, and was defined as F_{st} by Wright (1951).

Suppose that the population is subdivided into L subpopulations each with effective population size N. Allele frequencies are q' for A and 1-q' for a, including migrants (seed and pollen). Thus, the frequency of a hetero-genotype pair, Aa, in the whole population is

$$y_t = 2\sum_{1}^{L} q'(1-q')/L \qquad \text{(A2)}$$

Now, consider sampling variance. In each subpopulation, the sampling variance is $q'(1-q')/N$. The average sampling variance within subpopulations is

$$\sigma^2_{\delta q'} = \sum_{1}^{L} q'(1-q')/LN \qquad \text{(A3)}$$

The sampling variance for mean gene frequency of the whole population, i.e. the sampling variance of the $q = \sum_{1}^{L} q' / L$, can be obtained as following.

$$q - \bar{q} = \sum_{1}^{L} (q' - \bar{q}) / L \qquad \text{(A4)}$$

According to (A4), we can obtain

$$\sigma_{\delta q}^2 = \sigma_{\delta q'}^2 / L$$

$$= \sum_{1}^{L} q'(1 - q') / L^2 N \qquad \text{(A5)}$$

Using (A2) and (A1)

$$\sigma_{\delta q}^2 = y_t / 2LN$$

$$= q(1 - q)(1 - F) / LN \qquad \text{(A6)}$$

Therefore, according to (A6), variance effective population size for a haploid gene can be obtained, i.e.

$$N_e = LN / (1 - F) \qquad \text{(A7)}$$

which has the same form as for diploid genes.

V.2 Derivation of the G_{ST} for three differently inherited genomes in the finite island model

The G_{ST} derived here refers to differentiation among subpopulations in adults. Consider a finite island model in which the entire population consists of L subpopulations, each with effective size N. Each subpopulation exchanges seeds and pollen grains at rates of m_s and m_p with equal likelihood of exchange with the remaining subpopulations, respectively. Using similar notation to Takahata (1983), let K be a fixed number of potential alleles at a locus and $v/(K-1)$ be the mutation rate from one to any of the other $K-1$ alleles. The total mutation rate is v. Let $x_k(i,t)$ be the frequency of the kth allele in adults in subpopulation i at generation t.

Biparentally inherited diploid nuclear genes

As mentioned before, the biological basis on which G_{ST} is derived is that adults in each subpopulation produce pollen and pollen dispersal occurs. We assume that pollen grains randomly mate with ovules and produces seeds. After seed formation, there is seed flow among subpopulations. Then a sample of N seeds contributes to adults at the next generation.

Suppose that the whole population comprises adults at generation t and the frequency of the kth allele in the ith subpopulation is $x_k(i,t)$. Denote by $x_{p.k}(i,t+1)$ the gene frequency in pollen grains in the ith subpopulation. The subscript p stands for gene frequency in pollen grains. After *pollen flow*,

$$x_{p.k}(i,t+1) = (1-m_p)x_k(i,t) + \frac{m_p}{L-1}\sum_{j\neq i}^{L} x_k(j,t), \qquad \text{(A8)}$$

Denote by $x_{s.k}(i,t+1)$ the gene frequency in seeds. After random mating between pollen and ovules, the gene frequency in seeds so formed is half of the sum of gene frequencies of male and female parents, i.e.,

$$x_{s.k}(i,t+1)=\frac{1}{2}[x_k(i,t)+x_{p.k}(i,t+1)] \tag{A9}$$

After *seed flow*, the gene frequency in seeds, $x_{s.k}(i,t+1)$ becomes $x'_{s.k}(i,t+1)$, which is

$$x'_{s.k}(i,t+1)=(1-m_s)x_{s.k}(i,t+1)+\frac{m_s}{L-1}\sum_{j\neq i}^{L}x_{s.k}(j,t+1) \tag{A10}$$

Then assume that there are N individuals in adults which are sampled from these seeds in each subpopulation. Therefore, the gene frequency in adults at the next generation $t+1$, is

$$x_k(i,t+1)=x'_{s.k}(i,t+1)+\delta \tag{A11}$$

where δ is the change due to sampling (genetic drift), with mean $E[\delta]=0$ and variance $V[\delta]=\frac{x'_{s.k}(i,t+1)[1-x'_{s.k}(i,t+1)]}{2N}$. Putting equations (A8), (A9), (A10) into (A11), and ignoring items involving in $m_s m_p$, we can obtain

$$x_k(i,t+1)=(1-m_s-\frac{1}{2}m_p)x_k(i,t)+\frac{m_s+\frac{1}{2}m_p}{L-1}\sum_{j\neq i}^{L}x_k(j,t)+\delta \tag{A12}$$

Let $\delta x_k(i)=x_k(i,t+1)-x_k(i,t)$, $\tilde{m}=m_s+\frac{1}{2}m_p$, $m^*=\frac{\tilde{m}}{L-1}$ and cancel variable t in the formulae, then the mean $M[\delta x_k(i)]$ is

$$M[\delta x_k(i)]=-Lm^*x_k(i)+m^*\sum_{j=1}^{L}x_k(j) \tag{A13}$$

Considering mutation and letting $v^*=\frac{v}{K-1}$, then

$$M[\delta x_k(i)]=v^*-(Lm^*+Kv^*)x_k(i)+m^*\sum_{j=1}^{L}x_k(j) \tag{A14}$$

which is the same as equation (1) of Takahata (1983). When both $m_s << 1$ and $m_p << 1$ and assuming that random sampling of seeds takes place independently in each subpopulation, we can approximately obtain equation (2) of Takahata(1983), i.e.,

$$V[\delta x_k(i)\delta x_{k'}(j)] = \frac{1}{2N} x_k(i)[\delta_{kk'} - x_{k'}(j)]\delta_{ij} \qquad \text{(A15)}$$

where δ_{ij} stands for the Kronecker delta function. Therefore, the results derived by use of the diffusion model (Takahata ,1983) can be directly applied in plant populations by minor modification. When K= ∞ , i.e. infinite alleles model (Kimura and Crow, 1964) and under equilibrium among migration/drift/mutation,

$$G_{ST} = \frac{1}{1 + 2\tilde{N} \cdot \frac{L}{L-1}(\frac{L}{L-1}\tilde{m} + v)} \qquad \text{(A16)}$$

where $\tilde{N} = 2N$ and $\tilde{m} = m_s + \frac{1}{2}m_p$

Paternally and maternally inherited haploid organelle genes

Following similar consideration to those for *bi-parental genes* , we can obtain similar formulae to (A16) except that $\tilde{m} = m_s + m_p$ for paternal genes , $\tilde{m} = m_s$ for maternal genes and $\tilde{N} = N$ for both. If $L=\infty$, equation (A16) is used for the infinite island model for the three genomes.

Heredity **79** (1997) 541–552 *Received 6 December 1996*

On estimation of the ratio of pollen to seed flow among plant populations

XIN-SHENG HU†‡ & R. A. ENNOS*‡

†*The Research Institute of Forestry, Chinese Academy of Forestry, Wan Shou Shan, Beijing 100091, China and*
‡*Institute of Ecology and Resource Management, University of Edinburgh, Darwin Building, King's Buildings, Mayfield Rd, Edinburgh EH9 3JU, UK*

Gene flow occurs in two ways for hermaphrodite plants; seed flow and pollen flow. Dispersal of biparentally inherited (nuclear) and paternally inherited (conifer chloroplast) genes can be mediated by both seed and pollen, whereas for maternally inherited (angiosperm chloroplast and most mitochondrial) genes only seed flow contributes to dispersal. This produces asymmetrical migration for biparentally, paternally and maternally inherited genes and may lead to different levels of population differentiation among them. This paper explores the effects of contrasting patterns of gene flow for different plant genes on their population structure under isolation by distance, on Nei's genetic distance measure, on divergence in nucleotide sequence between populations and on gene phylogenies. The possibilities are discussed of using data on population structure, genetic distance, sequence divergence and gene phylogenies as a basis for estimating the ratio of pollen to seed flow among subpopulations. One important general result from the isolation-by-distance model is that population differentiation for maternally inherited genes is greater than that for paternally inherited genes, which, in turn, is greater than that for biparentally inherited genes as long as the dispersal of seeds and pollen grains takes place. This is consistent with results obtained previously for the island and stepping-stone models in which populations are discretely distributed.

Keywords: biparental gene, maternal gene, paternal gene, pollen flow, seed flow.

Introduction

A variety of models can be used indirectly to estimate gene flow among populations of a species using data on genetic structure for selectively neutral markers (Barton & Slatkin, 1986; Slatkin, 1989; Slatkin & Barton, 1989; Hudson *et al.*, 1992). When applied in plant species, especially hermaphrodite plants, gene flow should distinguish both pollen and seed flow. Seed flow and pollen flow may lead to asymmetrical migration for the biparentally inherited (nuclear), and maternally inherited (chloroplast and mitochondrial) genes, which occur in angiosperm species, and the paternally inherited (chloroplast) genes, which occur in conifer species (Neale & Sederoff, 1989; Neale *et al.*, 1986, 1991). This produces different levels of population differentiation for the three variously inherited genomes. If the behaviour of genes with different modes of inheritance can be modelled, analysis of differences in genetic differentiation for these genes may allow estimation of the relative rates of pollen flow and seed flow (Ennos, 1994).

Theory for differentiation of biparentally, paternally and maternally inherited markers has already been developed for the island and stepping-stone models of population structure (Petit *et al.*, 1993; Ennos, 1994; Hu unpubl. data). In this paper we are again concerned with the population genetic consequences of having plant genomes with three different modes of inheritance, and focus on methods for using a variety of population genetic statistics for estimating the ratio of pollen to seed flow. The first employs data on F_{1s} measured in populations having a continuous distribution in space according to Wright's isolation-by-distance model (Wright, 1943, 1946). We then consider a simple model which describes the development of genetic distance between populations (Nei & Feldman, 1972), and relate Nei's distance to levels of seed and pollen flow. Finally, we briefly address the possible estimation of the ratio of pollen to seed flow from data on differences in DNA sequence between populations and from gene phylogenies.

*Correspondence. E-mail: rennos@ed.ac.uk

Wright's isolation-by-distance model

In the isolation-by-distance model (Wright, 1943, 1946), an important parameter is the neighbourhood size which is defined as an area from which the parents of central individuals may be treated as if drawn at random. The calculation of neighbourhood area is relatively complicated when both pollen and seed dispersal are considered. Crawford (1984a,b) presented a modified formula for calculating neighbourhood size for a plant population which will be used here. For both pollen and seed the distribution of dispersal distances between parents and offspring is assumed to be normal with mean zero. We assume that the nuclear biparentally inherited genes are diploid, and the paternally and maternally inherited genes are haploid, and only consider selectively neutral genes. We will use the same method of path analysis as Wright (1968) to analyse the population structures of the three differently inherited genes. Some of these results were, in fact, presented by Wright (1943, 1946, 1969).

Biparentally inherited genes

Let σ_m^2 and σ_f^2 be the variance of the distances between male parents and offspring, and between female parents and offspring, respectively. Also let σ_s^2 be the variance of seed dispersal, and σ_p^2 be the variance of the dispersal of pollen grains before seed formation. The number of individuals in the neighbourhood is $N_{(b)} = 4\pi(\frac{1}{2}\sigma_p^2 + \sigma_s^2)d$ in area continuity according to Crawford (1984a,b) and $2\sqrt{(\frac{1}{2}\sigma_p^2 + \sigma_s^2)}\,\pi\, d$ in linear continuity, where d is the population density of breeding individuals. Let N_p be the number of individuals in a neighbourhood after pollen dispersal and before seed formation which is equal to $2\pi\sigma_p^2 d$ (area). The number of individuals after seed formation and dispersal in the neighbourhood is $N_f = 4\pi\sigma_s^2 d$ (area). Similarly, the number of individuals after pollen flow and before seeds are formed at ancestors of generation X is XN_p (area) or $\sqrt{X}\,N_p$ (linear), and for the individuals after seed dispersal is XN_f (area) or $\sqrt{X}\,N_f$ (linear). The total numbers of individuals in the neighbourhood for both parents at ancestral generation X are $4\pi(\frac{1}{2}\sigma_p^2 + \sigma_s^2)\,Xd$ (area) or $2\sqrt{(\frac{1}{2}\sigma_p^2 + \sigma_s^2)}\pi X d$ (linear).

Drift case Let F_{1s} be the correlation between ovules and pollen grains that contribute to zygotes after pollen and seed dispersal. According to the same considerations as Wright (1943, 1946), the F_{1s} in area continuity approximates to:

$$F_{1s}' = \frac{1}{N_{(b)}} b^2 + \left(1 - \frac{1}{N_{(b)}}\right) F_{2s} \tag{1}$$

$$b^2 = \frac{1 + F_1'}{2}. \tag{2}$$

Therefore, the recurrence equations at ancestor of generation X in area continuity is:

$$F_{XS} = \frac{1}{XN_{(b)}} b^2 + \left(1 - \frac{1}{XN_{(b)}}\right) F_{(X+1)S}. \tag{3}$$

For simplicity the calculation of F_{1s} after infinite generations can be expressed by:

$$F_{1s} = \sum_1^{\infty} t_i \Big/ \left(2 - \sum_1^{\infty} t_i\right), \tag{4}$$

where $t_1 = \dfrac{1}{N_{(b)}}$ and $t_X = \dfrac{(X-1)N_{(b)} - 1}{XN_{(b)}}\, t_{X-1}$.

For linear continuity the recurrence equations can be obtained by substituting the X in eqn (3) by $\sqrt{X}$. We suppose that all populations are initially present as adults and produce pollen grains for dispersal, and the boundary condition is $F_{ks} = 0$ after a large number of generations (k) back.

Balance case Where there is a balance between drift and long-range dispersal of seeds and pollen grains, i.e. drift/migration equilibrium, let $m_{p\infty}$ be the proportion of male parents (pollen grains) replaced by pollen migration when random mating with ovules, and $m_{s\infty}$ be the proportion of both parents replaced by seed migration. If both long-range dispersal and reversible mutation are considered, then $m_{p\infty}$ or $m_{s\infty}$ are just substituted by $m_{p\infty}+u$ or $m_{s\infty}+u$. Considering random sampling of size $N_{(b)}$, the proportion of male parents which makes a contribution to F_{1s} is $1-m_{p\infty}-m_{s\infty}$, whereas the proportion of female parents which contributes to F_{1s} is $1-m_{s\infty}$. Therefore, after seeds and pollen grains disperse,

$$F_{1s}=(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})\left[\frac{1}{N_{(b)}}b^2+\left(1-\frac{1}{N_{(b)}}\right)F_{2s}\right],$$

$$F_{2s}=(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})\left[\frac{1}{2N_{(b)}}b^2+\left(1-\frac{1}{2N_{(b)}}\right)F_{3s}\right], \text{ etc.} \tag{5}$$

At steady state,

$$F_{1s}=\sum_{1}^{\infty}t_i/(2-\sum_{1}^{\infty}t_i), \tag{6}$$

where $t_1=(1-m_{s\infty})(1-m_{p\infty}-m_{s\infty})\dfrac{1}{N_{(b)}}$ and $t_X=(1-m_{s\infty}-m_{p\infty})(1-m_{s\infty})\dfrac{(X-1)N_{(b)}-1}{XN_{(b)}}t_{X-1}$.

In the cases where only the pollen grains or seeds disperse, F_{1s} can be obtained by letting $N_f\to\infty$, $m_{s\infty}=0$ and $N_p\to\infty$, $m_{p\infty}=0$, respectively.

Paternally inherited genes

The number of individuals in the neighbourhood is $N_{(p)}=2\pi(\sigma_p^2+\sigma_s^2)d$ (area) or $\sqrt{(\sigma_s^2+\sigma_p^2)\pi}\ d$ (linear) because of individuals being haploid after the dispersal of both seeds and pollen grains. The number of individuals in the neighbourhood after the dispersal of pollen grains but before seed formation is $N_p=2\pi\sigma_p^2 d$ (area) or $\sqrt{\pi}\sigma_p d$ (linear), but the number of individuals in the neighbourhood after seed dispersal is $N_f=2\pi\sigma_s^2 d$ (area) or $\sqrt{\pi}\sigma_s d$ (linear). Similarly, the number of individuals in the neighbourhood at ancestor of generation X is $2\pi X(\sigma_p^2+\sigma_s^2)d$ (area) or $\sqrt{\pi X(\sigma_p^2+\sigma_s^2)}d$ (linear).

Drift case Here define F_{1s} as the correlation between adjacent individuals.

$$F_{1s}=\sum_{1}^{\infty}t_i, \tag{7}$$

where $t_1=\dfrac{1}{N_{(p)}}$ and $t_X=\dfrac{(X-1)N_{(p)}-1}{XN_{(p)}}t_{X-1}$.

Balance case At steady state (drift/migration equilibrium),

$$F_{1s}=\sum_{1}^{k-1}t_i, \tag{8}$$

where $t_1=\dfrac{1}{N_{(p)}}(1-m_{p\infty}-m_{s\infty})$ and $t_X=(1-m_{p\infty}-m_{s\infty})\dfrac{(X-1)N_{(p)}-1}{XN_{(p)}}t_{X-1}$.

For linear continuity the recurrence equations can be obtained by substituting $\sqrt{X}$ in place of X in eqn (8). The boundary condition is $F_{ks}=0$ a large number of generations (k) back.

In the case where only the pollen grains or seeds disperse, F_{1s} can be found by letting $N_f\to\infty$, $m_{s\infty}=0$ and $N_p\to\infty$, $m_{p\infty}=0$, respectively.

Maternally inherited genes

Because both paternally and maternally inherited genes are considered to be haploid or uniparental, the number of individuals in the neighbourhood is $N_{(m)} = 2\pi\sigma_s^2 d$ (area). Wright (1943) also addressed this case:

$$F_{1s} = \sum_{1}^{\infty} t_i, \tag{9}$$

where $t_1 = \dfrac{1}{N_{(m)}}$ and $t_X = \dfrac{(X-1)N_{(m)}-1}{XN_{(m)}} t_{X-1}$.

Balance case At steady state (drift/migration equilibrium),

$$F_{1s} = \sum_{1}^{\infty} t_i, \tag{10}$$

where $t_1 = (1-m_{s\infty})\dfrac{1}{N_{(m)}}$ and $t_X = (1-m_{s\infty})\dfrac{(X-1)N_{(m)}-1}{XN_{(m)}} t_{X-1}$.

Comparison of population differentiation

In order to compare population differentiation among three genomes, we use the same notation as Wright (1943, p. 124). Consider a total population of size N_t, subdivided into H groups of intermediate size N_i and these are subdivided into K random groups of size N_u. Next we will compare the levels of population differentiation relative to N_i among the three genomes in the drift/migration balance case.

Biparental vs. paternal genes It can be seen that the neighbourhood size of biparentally inherited genes is greater than that of paternally inherited genes, i.e. $N_{(b)} > N_{(p)}$, and also t_i ($i = 1, 2, \ldots, K$) in the case of paternal genes is greater than that in the case of biparental genes according to eqns (6) and (8). Therefore after going back to the ancestral generation K, $\Sigma_1^{K-1} t$ of paternal genes is greater than that of biparental genes. It can be shown that the correlation of paternally inherited genes, $F_{1s(p)}$, is greater than $F_{1s(b)}$ for biparentally inherited genes, i.e. $F_{1s(p)} > F_{1s(b)}$.

Similarly, after going back to ancestral generation KH, it can be shown that the correlation of paternal genes, $F_{it(p)}$ is greater than $F_{it(b)}$ of biparentally inherited genes. We can also prove that

$$\frac{F_{it(p)}-F_{1s(p)}}{1-F_{1s(p)}} > \frac{F_{it(b)}-F_{1s(b)}}{1-F_{1s(b)}}, \text{ i.e., } F_{st(p)} > F_{st(b)}. \tag{11}$$

Paternal vs. maternal genes As above, we can prove the relationship

$$\frac{F_{it(m)}-F_{1s(m)}}{1-F_{1s(m)}} > \frac{F_{it(p)}-F_{1s(p)}}{1-F_{1s(p)}}, \text{ i.e., } F_{st(m)} > F_{st(p)}. \tag{12}$$

In summary, the population differentiation of maternal genes is greater than that of paternal genes, which, in turn, is greater than that of biparental genes as long as the dispersal of seeds and pollen grains takes place.

Ratio of pollen to seed flow

In this section we consider how to estimate the ratio of pollen to seed flow from long-range dispersal. According to the Taylor expansion, $\Sigma_1^{\infty} t_i$ in eqn (6) can be written using a simple formula,

$$\sum_{1}^{\infty} t_i = 1-[1-(1-m_{p\infty}-m_{s\infty})(1-m_{s\infty})]^{1/N_{(b)}}.$$

Similarly, expressions can also be obtained for eqns (8) and (10).

Let

$$A = 1-\left(\frac{1-F_{1s}}{1+F_{1s}}\right)^{N_{(b)}},\ B = 1-(1-F_{1s})^{N_{(p)}} \text{ and } C = 1-(1-F_{1s})^{N_{(m)}}$$

for biparentally, paternally and maternally inherited genes, respectively. Then the ratio of pollen to seed flow from long-range distance can be approximated by

$$\frac{m_{p\infty}}{m_{s\infty}} = \frac{A-B^2}{B-A}, \text{ or } \frac{C^2-A}{C(1-C)}, \text{ or } \frac{C-B}{1-C}. \tag{13}$$

Nei's genetic distance

In this section we will incorporate seed flow and pollen flow into Nei's genetic distance measure (Nei, 1972) for three differently inherited genomes based on the assumptions of mutation/migration/drift equilibrium, as addressed by Nei & Feldman (1972) and Chakraborty & Nei (1974). Here we will use Nei and Feldman's model because of its simplicity and practicality.

Suppose that a population splits into two incompletely isolated populations and thereafter gene migration occurs in every generation between the two populations with a constant rate of both pollen and seed flow. Let N_1 and N_2 be the sizes of populations 1 and 2, respectively, and assume that effective size is the same as the actual size. Let m_{s1} and m_{p1} be the rates of seed and pollen migration in population 1, respectively, and m_{s2} and m_{p2} be the rates of seed and pollen flow in population 2. Using the same notation as Chakraborty & Nei (1974), let $J_{11}^{(t)}$ and $J_{22}^{(t)}$ be the probabilities of identity of two randomly chosen genes from populations 1 and 2, respectively, at generation t. Let $J_{12}^{(t)}$ be the probability of identity of two randomly chosen genes, one from each of the two populations. Each new mutation is different from the alleles pre-existing in any of the two populations. Only selectively neutral alleles are considered. Therefore, the only way in which two genes can be the same 'allele' is if they are identical by descent.

Biparentally inherited genes

Male parents for the biparental genes come from two sources: one comes from migration with frequency $m_{s1}+m_{p1}$, denoted by B; the other is from within populations with frequency $1-m_{s1}-m_{p1}$, denoted by A. Similarly, female parents come from two sources: m_{s1} from migration, denoted by D, and $1-m_{s1}$ from the population itself, denoted by C. The probabilities of two randomly chosen genes from population 1 coming from A and A, B and B, A and B, etc. are $(1-m_{s1}-m_{p1})^2$, $(m_{s1}+m_{p1})^2$, $(1-m_{s1}-m_{p1})(m_{s1}+m_{p1})$, etc., respectively. Following Malécot (1969), we can derive the recurrence equation for $J_{11}^{(t)}$, which is:

$$J_{11}^{(t+1)} = (1-u_b)^2\left\{\frac{1}{4}(AA+CC+2AC)\left[\frac{1}{2N_1}+\left(1-\frac{1}{2N_1}\right)J_{11}^{(t)}\right]+\frac{1}{4}(2AD+2AB+2CD+2CB)J_{12}^{(t)}\right.$$
$$\left.+\frac{1}{4}(BB+DD+2BD)\left[\frac{1}{2N_2}+\left(1-\frac{1}{2N_2}\right)J_{22}^{(t)}\right]\right\}, \tag{14a'}$$

where u_b is the mutation rate for biparental genes. Substituting for A, B, C and D in eqn (14a'), we can obtain eqn (14a):

$$J_{11}^{(t+1)} = (1-u_b)^2\left\{\left(1-m_{s1}-\tfrac{1}{2}m_{p1}\right)^2\left[\frac{1}{2N_1}+\left(1-\frac{1}{2N_1}\right)J_{11}^{(t)}\right]+2\left(1-m_{s1}-\tfrac{1}{2}m_{p1}\right)\left(m_{s1}+\tfrac{1}{2}m_{p1}\right)J_{12}^{(t)}\right.$$
$$\left.+\left(m_{s1}+\tfrac{1}{2}m_{p1}\right)^2\left[\frac{1}{2N_2}+\left(1-\frac{1}{2N_2}\right)J_{22}^{(t)}\right]\right\}. \tag{14a}$$

Similarly, we can derive the recurrence equations for $J_{12}^{(t)}$ and $J_{22}^{(t)}$.

$$J_{12}^{(t+1)} = (1-u_b)^2\left\{\left(1-m_{s1}-\tfrac{1}{2}m_{p1}\right)\left(m_{s2}+\tfrac{1}{2}m_{p2}\right)\left[\frac{1}{2N_1}+\left(1-\frac{1}{2N_1}\right)J_{11}^{(t)}\right]\right.$$
$$+\left[\left(1-m_{s1}-\tfrac{1}{2}m_{p1}\right)\left(1-m_{s2}-\tfrac{1}{2}m_{p2}\right)+\left(m_{s1}+\tfrac{1}{2}m_{p1}\right)\left(m_{s2}+\tfrac{1}{2}m_{p2}\right)\right]J_{12}^{(t)}$$
$$\left.+\left(1-m_{s2}-\tfrac{1}{2}m_{p2}\right)\left(m_{s1}+\tfrac{1}{2}m_{p1}\right)\left[\frac{1}{2N_2}+\left(1-\frac{1}{2N_2}\right)J_{22}^{(t)}\right]\right\}. \tag{14b}$$

$$J_{22}^{(t+1)} = (1-u_b)^2\left\{\left(m_{s2}+\tfrac{1}{2}m_{p2}\right)^2\left[\frac{1}{2N_1}+\left(1-\frac{1}{2N_1}\right)J_{11}^{(t)}\right]+2\left(1-m_{s2}-\tfrac{1}{2}m_{p2}\right)\left(m_{s2}+\tfrac{1}{2}m_{p2}\right)J_{12}^{(t)}\right.$$
$$\left.+\left(1-m_{s2}-\tfrac{1}{2}m_{p2}\right)^2\left[\frac{1}{2N_2}+\left(1-\frac{1}{2N_2}\right)J_{22}^{(t)}\right]\right\}. \tag{14c}$$

When $m_{p1} = m_{p2} = 0$, the above equations reduce to those of Chakraborty & Nei (1974).

Using matrix notations, formulae (14a), (14b) and (14c) may be written as

$$\mathbf{J}^{(t+1)} = (1-u_b)^2\mathbf{T} + (1-u_b)^2\mathbf{M}\mathbf{J}^{(t)} \tag{15}$$

where

$$\mathbf{J}^{(t)'} = (J_{11}^{(t)}, J_{12}^{(t)}, J_{22}^{(t)}),$$

$$\mathbf{T} = \begin{pmatrix} \dfrac{(1-m_{s1}-\frac{1}{2}m_{p1})^2}{2N_1}+\dfrac{(m_{s1}+\frac{1}{2}m_{p1})^2}{2N_2} \\ \dfrac{(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2})}{2N_1}+\dfrac{(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s1}+\frac{1}{2}m_{p1})}{2N_2} \\ \dfrac{(1-m_{s2}-\frac{1}{2}m_{p2})^2}{2N_2}+\dfrac{(m_{s2}+\frac{1}{2}m_{p2})^2}{2N_1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} (1-m_{s1}-\frac{1}{2}m_{p1})^2\left(1-\dfrac{1}{2N_1}\right) & 2(1-m_{s1}-\frac{1}{2}m_{p1})(m_{s1}+\frac{1}{2}m_{p1}) & (m_{s1}+\frac{1}{2}m_{p1})^2\left(1-\dfrac{1}{2N_2}\right) \\ (1-m_{s1}-\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2})\cdot\left(1-\dfrac{1}{2N_1}\right) & (1-m_{s1}-\frac{1}{2}m_{p1})(1-m_{s2}-\frac{1}{2}m_{p2})+(m_{s1}+\frac{1}{2}m_{p1})(m_{s2}+\frac{1}{2}m_{p2}) & (1-m_{s2}-\frac{1}{2}m_{p2})(m_{s1}+\frac{1}{2}m_{p1})\cdot\left(1-\dfrac{1}{2N_2}\right) \\ (m_{s2}+\frac{1}{2}m_{p2})^2\left(1-\dfrac{1}{2N_1}\right) & 2(1-m_{s2}-\frac{1}{2}m_{p2})(m_{s2}+\frac{1}{2}m_{p2}) & (1-m_{s2}-\frac{1}{2}m_{p2})^2\left(1-\dfrac{1}{2N_2}\right) \end{pmatrix}.$$

Under steady state, the vector of equilibrium identity probabilities is given by letting $\mathbf{J}^{(t+1)} = \mathbf{J}^{(t)}$ in eqn (15), i.e.

$$\mathbf{J} = (1-u_b)^2\{\mathbf{I}-(1-u_b)^2\mathbf{M}\}^{-1}\mathbf{T}. \tag{16}$$

As Chakraborty & Nei (1974) have already discussed this equation in detail, we can use their results in later sections.

Paternally inherited genes

Here again suppose that the paternal gene is haploid. Its migration can also be mediated by both pollen flow and seed flow. Following similar considerations to those for biparental genes, the vector of equilibrium identity probabilities is

$$\mathbf{J} = (1-u_p)^2\{\mathbf{I}-(1-u_p)^2\mathbf{M}\}^{-1}\mathbf{T}, \tag{17}$$

where u_p is the mutation rate of paternal genes, and

$$\mathbf{T} = \begin{pmatrix} \dfrac{(1-m_{s1}-m_{p1})^2}{N_1}+\dfrac{(m_{s1}+m_{p1})}{N_2} \\ \dfrac{(1-m_{s1}-m_{p1})(m_{s2}+m_{p2})}{N_1}+\dfrac{(1-m_{s2}-m_{p2})(m_{s1}+m_{p1})}{N_2} \\ \dfrac{(1-m_{s2}-m_{p2})^2}{N_2}+\dfrac{(m_{s2}+m_{p2})}{N_1} \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} (1-m_{s1}-m_{p1})^2\left(1-\dfrac{1}{N_1}\right) & 2(1-m_{s1}-m_{p1})(m_{s1}+m_{p1}) & (m_{s1}+m_{p1})^2\left(1-\dfrac{1}{N_2}\right) \\ (1-m_{s1}-m_{p1})(m_{s2}+m_{p2}) & (1-m_{s1}-m_{p1})(1-m_{s2}-m_{p2}) & (1-m_{s2}-m_{p2})(m_{s1}+m_{p1}) \\ \cdot\left(1-\dfrac{1}{N_1}\right) & +(m_{s1}+m_{p1})(m_{s2}+m_{p2}) & \cdot\left(1-\dfrac{1}{N_2}\right) \\ (m_{s2}+m_{p2})^2\left(1-\dfrac{1}{N_1}\right) & 2(1-m_{s2}-m_{p2})(m_{s2}+m_{p2}) & (1-m_{s2}-m_{p2})^2\left(1-\dfrac{1}{N_2}\right) \end{pmatrix}.$$

Maternally inherited genes

Consider that the maternally inherited genes are haploid. Only seed flow contributes to their migration. Under this case, the vector of equilibrium identity probabilities is

$$\mathbf{J} = (1-u_m)^2\{\mathbf{I}-(1-u_m)^2\mathbf{M}\}^{-1}\mathbf{T}, \tag{18}$$

where u_m is the mutation rate of maternal genes, and

$$\mathbf{T} = \begin{pmatrix} \dfrac{(1-m_{s1})^2}{N_1}+\dfrac{m_{s1}^2}{N_2} \\ \dfrac{(1-m_{s1})m_{s2}}{N_1}+\dfrac{(1-m_{s2})m_{s1}}{N_2} \\ \dfrac{(1-m_{s2})^2}{N_2}+\dfrac{m_{s2}^2}{N_1} \end{pmatrix}$$

$$\mathbf{M}=\begin{pmatrix} (1-m_{s1})^2\left(1-\frac{1}{N_1}\right) & 2(1-m_{s1})m_{s1} & m_{s1}^2\left(1-\frac{1}{N_2}\right) \\ m_{s2}(1-m_{s1})\left(1-\frac{1}{N_1}\right) & (1-m_{s1})(1-m_{s2})+m_{s1}m_{s2} & (1-m_{s2})m_{s1}\left(1-\frac{1}{N_2}\right) \\ m_{s2}^2\left(1-\frac{1}{N_1}\right) & 2(1-m_{s2})m_{s2} & (1-m_{s2})^2\left(1-\frac{1}{N_2}\right) \end{pmatrix}.$$

Ratio of pollen to seed flow

Here consider a special case where $u \ll m_{s1}$, m_{p1}, m_{s2}, $m_{p2} \ll 1$, which was addressed by Chakraborty & Nei (1974). Nei's distances for the three genomes are:

$$D_b \approx \frac{2u_b}{m_{s1}+m_{s2}+\frac{1}{2}m_{p1}+\frac{1}{2}m_{p2}}, \tag{19a}$$

$$D_p \approx \frac{2u_p}{m_{s1}+m_{s2}+m_{p1}+m_{p2}} \text{ and} \tag{19b}$$

$$D_m \approx \frac{2u_m}{m_{s1}+m_{s2}}, \tag{19c}$$

where D_b, D_p and D_m are Nei's distances for biparental, paternal and maternal genes, respectively.

Let $\tilde{m}_s = m_{s1}+m_{s2}$ and $\tilde{m}_p = m_{p1}+m_{p2}$. The ratio of pollen to seed flow is given by:

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{2(a-1)}{2-a}, \text{ where } a = \frac{D_b u_p}{D_p u_b} \tag{20a}$$

or

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{2(1-a)}{a}, \text{ where } a = \frac{D_b u_m}{D_m u_b} \tag{20b}$$

or

$$\frac{\tilde{m}_p}{\tilde{m}_s} = \frac{1-a}{a}, \text{ where } a = \frac{D_p u_m}{D_m u_p}. \tag{20c}$$

Number of nucleotide differences

The variation in DNA sequence within and between populations contains much information on population evolution. Every sequence may be unique, and all the information is contained in the genealogical relationship between sequences (Barton & Wilson, 1995). Differences at the DNA level can be measured by the number of segregating sites among DNA sequences sampled (Watterson, 1975) or by the average number of (pairwise) nucleotide differences between sampled DNA (Tajima, 1983). For simplicity, only the average number of

(pairwise) nucleotide differences between DNA is considered. If only two DNA sequences are sampled from a population, the expectation of the average number of nucleotide differences is equal to the expected number of segregating sites (Tajima, 1989a). Under a balance of migration/mutation/drift, the average number of pairwise nucleotide differences sampled within a population is independent of migration, but is related to migrations for pairwise DNA sampled between populations (Strobeck, 1987). This provides the foundation for estimating the ratio of pollen to seed flow.

From above the migration rate for biparental genes can be obtained directly, i.e. $m_s+\frac{1}{2}m_p$, whereas the migration rates for paternal and maternal genes are m_s+m_p and m_s, respectively. We use similar notation to Strobeck (1987). In the island model with a finite number of subpopulations, n, let

$$A=\frac{(n-1)u_b}{\hat{\xi}_{ij.b}-\hat{\xi}_{ii.b}},\ B=\frac{(n-1)u_p}{\hat{\xi}_{ij.p}-\hat{\xi}_{ii.p}}\quad\text{and}\quad C=\frac{(n-1)u_m}{\hat{\xi}_{ij.m}-\hat{\xi}_{ii.m}},$$

where u represents mutation rate, $\hat{\xi}_{ij}$ and $\hat{\xi}_{ii}$ stand for the expected number of nucleotide differences between two randomly chosen DNA sequences from the same subpopulation and from two different subpopulations, respectively. Subscripts b, p and m on $\hat{\xi}_{ij}$, $\hat{\xi}_{ii}$ and u stand for biparentally, paternally and maternally inherited genes, respectively. The ratio of pollen to seed flow can be obtained by:

$$\frac{m_p}{m_s}\approx\frac{B-C}{C},\ \text{or}\ \frac{2(A-C)}{C},\ \text{or}\ \frac{2(B-A)}{2A-B}. \tag{21}$$

In the circular stepping-stone model, let

$$A=\frac{i(n-i)u_b}{\hat{\xi}_{i.b}-\hat{\xi}_{0.b}},\ B=\frac{i(n-i)u_p}{\hat{\xi}_{i.p}-\hat{\xi}_{0.p}}\quad\text{and}\quad C=\frac{i(n-i)u_m}{\hat{\xi}_{i.m}-\hat{\xi}_{0.m}},$$

where $\hat{\xi}_i(i=1, 2,\ldots)$ stands for the expected number of nucleotide differences between two randomly chosen DNA sequences from two subpopulations which are i steps apart, and $\hat{\xi}_0$ from the same subpopulation. Under the balance of mutation/migration/drift, the ratio of pollen to seed flow can be obtained according to Strobeck (1987), which has the same formula as (21) except for different values of A, B and C.

Phylogenies

Another method that also uses DNA sequence information for estimating the ratio of pollen to seed flow is based on the phylogenies of genes. Slatkin and coworkers (Slatkin & Barton, 1989; Slatkin & Maddison, 1990) and Hudson *et al.* (1992) introduced a method for analysing phylogenies of genes sampled from a geographically structured population. Using simulation, they showed that the minimum number of migration events (s) is a simple function of Nm based on phylogenies of alleles and genes under a variety of population structure models. This method depends on knowing the phylogeny of the nonrecombining segments of DNA that are sampled, but does not require complete sequences, although it does assume that an accurate phylogeny can be inferred from the segments of DNA sampled (Slatkin & Barton, 1989). Although the analytical expression, $s=f(Nm)$ has not been obtained to date, this nevertheless provides an additional potential method for estimating the ratio of pollen to seed flow among plant populations.

Following similar considerations to those above, for the biparentally inherited genome (nuclear DNA), both seed and pollen contribute to the migration events. Thus the relationship between s_b, the minimum number of migration events between pairs of populations sampled, and number of migrants may be written:

$$s_b=f[N(m_s+\tfrac{1}{2}m_p)]. \tag{22}$$

Similarly, the minimum number of migration events between pairs of populations sampled should be related to both seed and pollen flow for paternally inherited genes, and to seed flow only for maternal genes. Therefore, there may be the following relationships,

$$s_p = f[N(m_s + m_p)] \quad (23)$$

and

$$s_m = f(Nm_s), \quad (24)$$

where s_p and s_m stand for the minimum number of migration events consistent with phylogeny for paternal and maternal genomes, respectively. By combining eqns (22), (23) and (24), it will be possible to estimate the ratio of pollen to seed flow once any two of these three relationships are available.

Discussion

One of the aims of this paper has been to develop theory for population structure of plant genes with different modes of inheritance under isolation-by-distance. In the island model and the stepping-stone models where populations are discretely distributed, differentiation for maternally inherited genes $F_{ST(m)}$ is greater than for paternally inherited genes $F_{ST(p)}$, which, in turn, is greater than for biparentally inherited genes $F_{ST(b)}$ (Ennos, 1994; Hu, unpubl. data). In this paper we show that this relationship still holds in populations with a continuous distribution and limited dispersal of seeds and pollen.

Another aim of this paper has been to develop theory for indirectly estimating the ratio of pollen to seed flow among plant populations by a variety of methods. In the isolation-by-distance case it is possible to obtain analytical expressions for estimating this ratio under the hypothesis of a balance between migration and drift (formula (13)). In practice this formula will be very difficult to apply. In the first place it requires estimates of neighbourhood size for the three different genomes. These are difficult to measure in the field (Levin & Kerster, 1968, 1971, 1974; Schaal, 1975; Crawford, 1984a,b; Gliddon & Saleem, 1985). The model also assumes a random mating population, reaching an infinite number of generations back to its ancestors. If there is any self-fertilization, then F_{1s} will increase and the model assumptions will not be met.

Within the isolation-by-distance model it is possible to take into account deviations from random mating caused by self-fertilization. Let r be the proportion of the pollinations randomly coming from the neighbourhood and $1-r$ be the proportion of self-fertilization. If there is no seed dispersal but pollen dispersal, the neighbourhood size at ancestors of generation X for the biparental genes is $4\pi((1+(X-1)r)\sigma_p^2/2+\sigma_s^2)d$ (area) or $2\sqrt{((1+(X-1)r)\sigma_p^2/2+\sigma_s^2)\pi}d$ (linear) according to Wright (1946). Similarly, the size of neighbourhood at ancestors of generation X for paternal genes is $2\pi((1+(X-1)r)\sigma_p^2+\sigma_s^2)d$ (area) or $\sqrt{((1+(X-1)r)\sigma_p^2+\sigma_s^2)\pi}d$ (linear). However, if both seed flow and pollen are considered, the calculation of neighbourhood size becomes very complicated.

Finally, formula (13) will be difficult to apply in practice because the total number of individuals sampled in experimental work is always less than infinite. For this reason therefore $\hat{F}_{1s}$ may be underestimated. Taking all these points into consideration it is much more difficult to estimate the ratio of pollen to seed flow in the isolation-by-distance case than in either the island or stepping-stone models of population structure (Ennos, 1994; Hu, unpubl. data).

The second method explored in this paper for estimating the ratio of pollen to seed flow involved analysis of Nei's genetic distance. In order to apply the formulae (20a–c) derived here we must assume neutrality of mutations (Tajima, 1989b) and must possess estimates of the mutation rates in the three different genomes. There is evidence from analysis of rates of sequence divergence over evolutionary time that mutation rates differ significantly among the three plant genomes, with mutation rates being higher for nuclear genes than for chloroplast genes, which, in turn, are higher than for mitochondrial genes (Birky, 1988). If mutation rates of the three genomes were equal, genetic distances among the different genomes would vary according to the relationship $D_m > D_b > D_p$. Deviations from this predicted ordering of genetic distances could provide further evidence for large differences in the mutation rates of the three genomes.

The use of DNA sequence data to estimate the ratio of pollen to seed flow suffers from the same limitation as Nei's distance measure; we need to estimate mutation rate of the genes in the three genomes before the ratio of pollen to seed flow can be measured. Furthermore, it may be also be necessary to test the neutral mutation hypothesis before the formulae derived above can be applied. For these reasons it may be more practical to utilize statistics which rely only on the detection of differences between alleles, i.e. F-statistics rather than those which require measurement of the extent of genetic differences between alleles when

indirectly estimating the ratio of pollen to seed flow. Great care should be taken even with these methods because their usefulness may only be judged once their variances, $\mathrm{Var}(m_p/m_s)$, are available. Finally, we must remember that the assumption of strict maternal and paternal inheritance of organelle genomes underlies the models developed above. Further experimental data are required to confirm the general validity of these assumptions.

Acknowledgements

Thanks are given to the Overseas Development Administration (ODA), UK for supporting the study of the first author in the University of Edinburgh, and to Professor N. H. Barton for helpful comments on parts of this paper.

References

BARTON, N. H. AND SLATKIN, M. 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity*, **56**, 409–415.

BARTON, N. H. AND WILSON, I. 1995. Genealogies and geography. *Phil. Trans. R. Soc. B*, **349**, 49–59.

BIRKY, C. W. 1988. Evolution and variation in plant chloroplast and mitochondrial genomes. In: Gottlieb, L. D. and Jain, S. K. (eds) *Plant Evolutionary Biology*, pp. 23–53. Chapman and Hall, New York.

CHAKRABORTY, R. AND NEI, M. 1974. Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theor. Pop. Biol.*, **5**, 460–469.

CRAWFORD, T. J. 1984a. What is a population? In: Shorrocks, B. (ed.) *Evolutionary Ecology*, pp. 137–173. Blackwell Scientific Publications, Oxford.

CRAWFORD, T. J. 1984b. The estimation of neighbourhood parameters for plant populations. *Heredity*, **52**, 273–283.

ENNOS, R. A. 1994. Estimating the relative rates of pollen and seed migration among plant populations. *Heredity*, **72**, 250–259.

GLIDDON, C. AND SALEEM, M. 1985. Gene-flow in *Trifolium repens* — an expanding genetic neighbourhood. In: Jacquard, P., Heim, G. and Antonovics, J. (eds) *Genetic Differentiation and Dispersal in Plants*. NATO ASI Series, Vol. G5, pp. 293–309. Springer Verlag, Berlin.

HUDSON, R. R., SLATKIN, M. AND MADDISON, W. P. 1992. Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.

LEVIN, D. A. AND KERSTER, H. W. 1968. Local gene dispersal in *Phlox*. *Evolution*, **22**, 130–139.

LEVIN, D. A. AND KERSTER, H. W. 1971. Neighborhood structure in plants under diverse reproductive methods. *Am. Nat.*, **105**, 345–354.

LEVIN, D. A. AND KERSTER, H. W. 1974. Gene flow in seed plants. *Evol. Biol.*, **7**, 139–220.

MALÉCOT, G. 1969. *The Mathematics of Heredity*. Translated by D. M. Yermanos. W. H. Freeman, San Francisco.

NEALE, D. B. AND SEDEROFF, R. R. 1989. Paternal inheritance of chloroplast DNA and maternal inheritance of mitochondrial DNA in loblolly pine. *Theor. Appl. Genet.*, **77**, 212–216.

NEALE, D. B., WHEELER, N. C. AND ALLARD, R. W. 1986. Paternal inheritance of chloroplast DNA in Douglas fir. *Can. J. Forest Res.*, **16**, 1152–1154.

NEALE, D. B., MARSHALL, K. A. AND HARRY, D. E. 1991. Inheritance of chloroplast and mitochondrial DNA in incense cedar (*Calocedrus decurrens*). *Can. J. Forest Res.*, **21**, 717–720.

NEI, M. 1972. Genetic distance between populations. *Am. Nat.*, **106**, 283–292.

NEI, M. AND FELDMAN, M. W. 1972. Identity of genes by descent within and between populations under mutation and migration pressures. *Theor. Pop. Biol.*, **3**, 460–465.

PETIT, R. J., KREMER, A. AND WAGNER, D. B. 1993. Finite island model for organelle and nuclear genes in plants. *Heredity*, **71**, 630–640.

SCHAAL, B. 1975. Population structure and local differentiation in *Liatris cylindracea*. *Am. Nat.*, **110**, 511–528.

SLATKIN, M. 1989. Detecting small amounts of gene flow from phylogenies of alleles. *Genetics*, **121**, 609–612.

SLATKIN, M. AND BARTON, N. H. 1989. A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349–1368.

SLATKIN, M. AND MADDISON, W. P. 1990. Detecting isolation by distance using phylogenies of genes. *Genetics*, **126**, 249–260.

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

TAJIMA, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437–460.

TAJIMA, F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.

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

WATTERSON, G. A. 1975. On the number of segregating sites in genetic models without recombination. *Theor. Pop. Biol.*, **7**, 256–276.

WRIGHT, S. 1943. Isolation by distance. *Genetics*, **28**, 114–138.

WRIGHT, S. 1946. Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39–59.

WRIGHT, S. 1968. *Evolution and the Genetics of Populations*, vol. 1, *Genetic and Biometric Foundations.* University of Chicago Press, Chicago.

WRIGHT, S. 1969. *Evolution and the Genetics of Populations*, vol. 2, *The Theory of Gene Frequencies.* University of Chicago Press, Chicago.