# Population Genetics of Speciation in Nonmodel Organisms: I. Ancestral Polymorphism in Mangroves

*Renchao Zhou,*[1] *Kai Zeng,*[1] *Wei Wu,* *Xiaoshu Chen,* *Ziheng Yang,‡ Suhua Shi,* *and Chung-I Wu†*

*State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-Sen University, Guangzhou, China; †Department of Ecology and Evolution, University of Chicago; and ‡Department of Biology, University College London, Gower Street, London, United Kingdom

The level of DNA polymorphism in the ancestral species at the time of speciation can be estimated using DNA sequences from many loci sampled from 2 or more extant species. The comparison between ancestral and extant polymorphism can be informative about the population genetics of speciation. In this study, we collected and analyzed DNA sequences of ~60 genes from 4 species of *Sonneratia*, a common genus of mangroves on the Indo-Pacific coasts. We found that the 3 ancestral species were comparable to each other in terms of level of polymorphism. However, the ancestral species at the time of speciation were substantially more polymorphic than the extant geographical populations. This ancestral polymorphism is in fact larger than, or at least equal to, the level of polymorphism of the entire species across extant geographical populations. The observations are not fully compatible with speciation by strict allopatry. We suggest that, at the time of speciation, the ancestral species consisted of interconnected but strongly divided geographical populations. This population structure would give rise to high level of polymorphism across species range. This approach of studying the speciation history by genomic means should be applicable to nonmodel organisms.

## Introduction

The pattern of genetic polymorphism at the time of speciation can be useful for studying the population genetic theory of speciation, sometimes referred to as "paleopopulation biology" (Takahata 1993). For example, if the extant species are much less polymorphic than the ancestral species, then the extant species may have experienced bottlenecks or the ancestral species may be highly structured at the time of speciation.

Methods for estimating the level of ancestral polymorphism can be classified into 2 categories, referred to as the "2-species" and "3-species" methods. The 2-species methods, proposed by Takahata and colleagues (Takahata 1986; Takahata et al. 1995; Takahata and Satta 1997), utilize multilocus divergence data from 2 species. The distribution of levels of neutral divergence across loci are used to estimate the time of speciation as well as the effective population size when speciation occurred. These methods assume neutrality and use either moment estimators or maximum likelihood estimators to make inference about the parameters.

The 3-species methods (Nei 1987; Pamilo and Nei 1988; Wu 1991; Yang 2002; Rannala and Yang 2003) take advantage of a very different principle, namely the conflict between gene tree and species tree. The conflict happens when, for example, 3 species are of the ((A, B), C) phylogeny, whereas the genes sampled from them show the genealogy of (A, (B, C)). The conflict between the 2 trees results from the segregation of ancient polymorphism, which depends on the effective population size of the ancestral species. Some of the 3-species methods (Nei 1987; Pamilo and Nei 1988; Wu 1991) use only the conflict in tree topology to infer the ancestral polymorphism, whereas others (Yang 2002; Rannala and Yang 2003) use both the conflict in topology and sequence divergence. The latter usually outperform the former because they take the full probabilistic approach, and the parameters are estimated by either maximum likelihood method or Bayesian method. In particular, the Bayesian method can deal with many species simultaneously (Rannala and Yang 2003).

For either category of methods, an outgroup can be added to calibrate the difference in mutation rate across loci (Yang 2002; Osada and Wu 2005). Thus, sequence data from 4 closely related species, with one being a clear outgroup, are most versatile for the type of analysis presented here.

These methods should be generally applicable in studying nonmodel organisms as they require only DNA sequences from a number of genomic regions with no need for "hard-core" genetics. Despite that, these methods have been applied primarily to the human–chimpanzee split (Takahata et al. 1995; Takahata and Satta 1997; Yang 1997; Chen and Li 2001; Yang 2002; Rannala and Yang 2003; Wall 2003; Satta et al. 2004; Osada and Wu 2005) and, occasionally, *Drosophila melanogaster* and its relatives (Li et al. 1999; Zeng K, unpublished data). A noteworthy conclusion from these studies is that the effective population size at the time of speciation appears to be larger than the extant ones. This conclusion is intriguing in that the actual population sizes of both humans and *D. melanogaster* at present are large. This seemingly paradoxical observation has motivated many hypotheses on speciation and natural selection (Osada and Wu 2005; Patterson et al. 2006). It may, therefore, be useful if the same approach is applied to a wide range of species that are not "model organisms."

Our analysis focuses on closely related species between which gene flow has ceased completely since some time ago, as between humans and chimpanzees. These "good species" leave no doubt that we are studying speciation events, rather than differentiation within a polytypic species. An alternative is to study diverging populations that are still exchanging genes. Such an approach has been well explored by other investigators (Wang et al. 1997; Nielsen and Wakeley 2001; Hey and Nielsen 2004).

[1] These authors contributed equally to this study.

Key words: ancestral polymorphism, extant polymorphism, population subdivision, *Sonneratia*, mode of speciation.

E-mail: lssssh@mail.sysu.edu.cn.

*Mol. Biol. Evol.* 24(12):2746–2754. 2007
doi:10.1093/molbev/msm209
Advance Access publication September 28, 2007

In this study, we focus on the paleo-population genetics of *Sonneratia*, a genus of mangroves widely distributed in the Indo-West Pacific region (Tomlinson 1986). It consists of about 5–6 species, among which *Sonneratia alba*, *Sonneratia caseolaris*, *Sonneratia ovata*, and *Sonneratia apetala* are the better known ones and are the focus of this study. *Sonneratia alba*, the most widespread species in this genus, extends from tropical eastern Africa through Indo-Malay to Australia and islands of the West Pacific. The second most common species, *S. caseolaris*, has a similar geographic distribution as *S. alba*, except that it does not occur in eastern Africa. *Sonneratia ovata* and *S. apetala* are only found in Southeast Asia and the Bengal Bay region, respectively. Among the 4 species in question, *S. alba*, *S. caseolaris*, and *S. ovata* have overlapping species ranges, whereas *S. apetala* is allopatric from the other 3 species. All species in this genus grow in the intertidal zones of estuary systems, and the salinity gradient determines the fine-scale habitats of these species—both *S. alba* and *S. ovata* are downstream species, and *S. caseolaris* and *S. apetala* are found upstream. The phylogeny of the 4 species has been inferred previously using nr internal transcribed spacer sequence data (Shi et al. 2000; Zhou et al. 2005).

In what follows, we will analyze sequence data of about 60 homologous genes from the 4 species of *Sonneratia*. We will use 3 different methods (Takahata and Satta 1997; Yang 2002; Rannala and Yang 2003) to estimate levels of ancestral polymorphism. The contrast between the patterns of ancestral and extant polymorphisms could shed some light on the speciation history of these species. We suggest that this population genomic approach may be applicable to many nonmodel organisms.

## Materials and Methods

### Sampling

We collected samples from 11 subpopulations of *S. alba* and 7 subpopulations of *S. caseolaris* in the Indo-West Pacific region, including East Asia, Southeast Asia, South Asia, and Australia. For *S. ovata* and *S. apetala*, we sampled only one population in China because these 2 species are narrowly distributed and, where they can be found are in relatively low abundance. In all, 2–12 individuals were sampled from each subpopulation. Supplementary tables S1 and S2 (Supplementary Material online) summarize the sampling details.

### DNA Sequencing

#### *Sequence Data for Estimating Ancestral Polymorphism*

We constructed a cDNA library from the leaves of *S. caseolaris* using Creator Smart cDNA Library Construction Kit (Clontech, Mountain View, CA) and sequenced 200 clones from the library. Based on these sequences, we designed 67 pairs of primers to amplify the targeted genes from the genomic DNA of the 4 species. The polymerase chain reaction (PCR) products were purified and directly sequenced. For 57 of the 67 sequenced loci, we obtained sequences from all 4 species. For the remaining 10 loci, sequences from 1 or 2 species were missing. The average sequence length of these genes is 1,019 bp. All these sequences have been deposited in GenBank with accession numbers EF585716–EF585975. The positions of exons and introns of each gene were determined according to the corresponding cDNA sequences. Because the coding regions in the amplified fragments are usually very short, we used only noncoding regions (introns) in our analyses. The average length of intron regions over these genes is 707 bp.

#### *Sequence Data for Estimating Extant Polymorphism*

We obtained polymorphism data of 6 nuclear loci. These loci, all having long introns, are genes encoding large subunit 9 of ribosomal protein (*rpl9*), cytochrome B6-F complex iron–sulfur subunit (*cci*), iron-deficiency–responsive protein (*idr*), cysteine proteinase inhibitor (*cpi*), peptidyl–prolyl *cis–trans* isomerase (*ppc*), and phosphatase inhibitor (*phi*). DNA sequences were obtained by PCR amplification and direct sequencing (GenBank accession numbers: EF593405–EF593948 and EU035558–EU035590). Coding and intronic regions were determined by aligning our sequences against the original cDNA sequences. Because the mangrove species in question are diploid, we counted the 2 alleles of each sampled individual and treated a site as polymorphic if there was a "double peak" in the chromatogram. Haplotypes were inferred via "haplotype subtraction" (Clark 1990; Olsen and Schaal 1999). Due to the low level of diversity, such inference is likely to be very accurate. Furthermore, when using only intronic sequences for the analysis in which haplotype information was not required, we obtained very similar results. Sequences were aligned using ClustalW (Thompson et al. 1994). For the first 3 loci, samples were taken from the subpopulations listed in supplementary table S1 (Supplementary Material online), and for the last 3 loci, samples were taken from the subpopulations listed in supplementary table S2 (Supplementary Material online). In general, the sampling schemes for these 2 sets of loci are very similar (supplementary tables S1 and S2, Supplementary Material online). Analyzing these 2 sets of genes separately, we obtained very similar results. Thus, we present them together for clarity.

### Estimating Divergence between Species

We used Kimura's 2-parameter model (Kimura 1980) to estimate genetic divergence. For phylogenetic reconstruction, we concatenated the sequences and used the Neighbor-Joining method (Saitou and Nei 1987) implemented in the PHYLIP package (Felsenstein 2005).

### Estimating Levels of Polymorphism of the Ancestral Species

We first used the Takahata and Satta (1997) method to estimate levels of ancestral polymorphism. This method uses data from 2 species. It assumes neutrality and allopatric speciation. Let $t$ be the divergence time between the 2 species, $\mu$ be the rate of neutral mutation per base pair per generation, $N$ be the effective population size of the ancestral species, and $l_i$ and $k_i$ be the length and the number of

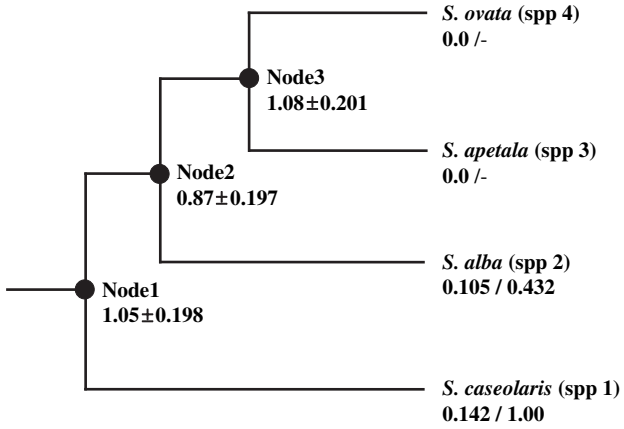


FIG. 1.—Patterns of extant and ancestral polymorphisms. Ancestral species on the phylogeny are denoted as Node 1, Node 2, and Node 3. The numbers beneath the node names are the estimated nucleotide diversity of the ancestral species (table 2). For *S. caseolaris* and *S. alba*, 2 different measures of extant polymorphism are shown ($\overline{\pi_R} \times 10^2$ and $\overline{\pi_T} \times 10^2$, separated by a slash sign; see text and table 4).

difference at the *i*th locus, respectively. If we have $n$ independent orthologous loci, the log-likelihood function can be expressed in terms of the scaled divergence time ($\tau = 2t\mu$) and the population mutation rate ($\theta = 4N\mu$):

$$L(\tau,\theta)=\sum_{i=1}^{n}\left[-l_i\tau-\ln(1+l_i\theta)+\ln\sum_{d=0}^{k_i}\frac{(l_i\tau)^d}{d!}\left(\frac{l_i\theta}{1+l_i\theta}\right)^{k_i-d}\right] \quad (1)$$

(Takahata and Satta 1997). In our implementation, we estimated the number of differences at a locus by multiplying its length by the estimated divergence. No loci show extreme levels of divergence between species (supplementary fig. S1, Supplementary Material online), and the lengths and positions of exons and introns of a gene in different species are quite conserved. Moreover, we selected 5 loci randomly and used additional primers that anchored different locations of the coding regions and obtained the same sequences. Thus, these loci are likely to be truly orthologous.

The model assumes that all loci have the same mutation rate. It has been shown that violation of this assumption can affect the accuracy of the estimate (Yang 1997). Taking advantage of our 4 species system, we calibrated the mutation rate by using outgroup sequences (*S. caseolaris*; see fig. 1). Following the treatment of Osada and Wu (2005), let $\tau_{i,1}$ be the scaled divergence time between species *A* and the outgroup at the *i*th locus and, similarly, let $\tau_{i,2}$ be that between species *B* and the outgroup at the same locus. Further, let $T^* = T + 2N_0$, where $T$ is the time when the outgroup species and the ancestral species leading to species *A* and *B* diverged and $N_0$ is the effective population size of the common ancestor of the 3 species. Then $\tau_{i,0} = 2T^*\mu_i$ can be approximated by $\tau_{i,0} \approx (\tau_{i,1} + \tau_{i,2})/2$. In practice, $\tau_{i1}$ and $\tau_{i2}$ can be estimated by the divergence between the 2 ingroup species (i.e., species *A* and *B*) and the outgroup species. By defining $\alpha = t/T^*$ and $\beta = 2N/T^*$, the local scaled divergence time ($\tau_i$) and the population mutation rate ($\theta_i$) for the *i*th locus can be written as $\tau_i = \alpha\tau_{i,0}$ and $\theta_i = \beta\tau_{i,0}$. Then equation (1) can be rewritten as:

$$L(\alpha,\beta)=\sum_{i=1}^{n}\left[-l_i\tau_{i,o}\alpha-\ln(1+l_i\tau_{i,o}\beta)+\ln\sum_{d=0}^{k_i}\frac{(l_i\tau_{i,o}\alpha)^d}{d!}\left(\frac{l_i\tau_{i,o}\beta}{1+l_i\tau_{i,o}\beta}\right)^{k_i-d}\right]. \quad (2)$$

Simulation studies suggest that this method can effectively account for the variation in mutation across loci, even when $2N_0$ is relatively large (Zeng K, unpublished data).

Note that the Takahata and Satta method (as well as the 3-species method described below) also assumes that every locus evolves in a clock-like manner. It seems that this assumption is well supported by our data (see Results). Furthermore, simulation studies suggest that moderate violation of this assumption does not have much effect on the results (Zeng K, unpublished data).

The two 3-species methods derived using the full probabilistic approach (Yang 2002; Rannala and Yang 2003) can be seen as an extension of the 2-species method (referred to as Yang's method and Rannala and Yang's method, respectively). They combine merits of the 2-species method and the *tree-mismatch* method (Pamilo and Nei 1988; Wu 1991) and models substitution using the Jukes and Cantor model (1969). Having to take into account all these factors, the 3-species models are more complex than the 2-species methods; therefore, we omit the details here.

## Estimating the Patterns of Extant Polymorphism

We used $\pi$ (nucleotide diversity; Hartl and Clark 1997) to measure extant polymorphism. For *S. alba* and *S. caseolaris*, samples were collected from hierarchically structured populations. We referred to each collection from the same location as a subpopulation and those subpopulations from the same geographical area/country (usually within one degree latitude and longitude) as a region (supplementary tables S1 and S2, Supplementary Material online). The entire collection of each species is then designated as "Total." We calculated $\pi_S$, $\pi_R$, and $\pi_T$ for each locus, where $\pi_S$ is the nucleotide diversity averaged across subpopulations; $\pi_R$ is the nucleotide diversity averaged across regional samples; and $\pi_T$ is the nucleotide diversity calculated using all sampled sequences. To measure the level of differentiation, the following common definitions are used:

$$F_{SR} = 1 - \overline{\pi_S}/\overline{\pi_R},\qquad F_{RT} = 1 - \overline{\pi_R}/\overline{\pi_T}, \quad (3)$$

where $\overline{\pi_S}$, $\overline{\pi_R}$, and $\overline{\pi_T}$ are, respectively, mean values of $\pi_S$, $\pi_R$, and $\pi_T$ averaged across the 6 loci (e.g., Hudson et al. 1992; Nagylaki 1998). When calculating the *F*-indices using other definitions (Hedrick 1999, 2005), very similar results were obtained (data not shown).

**Table 1**
**Mean Levels of Divergence between the 4 *Sonneratia* Species**

| | *S. ovata* (%) | *S. apetala* (%) | *S. alba* (%) |
|---|---|---|---|
| *S. apetala* | 2.06 (0.14) | | |
| *S. alba* | 2.44 (0.15) | 2.65 (0.15) | |
| *S. caseolaris* | 3.61 (0.15) | 3.65 (0.16) | 3.75 (0.18) |

NOTE.—The mean and standard error of the estimated divergence across loci are shown. Divergence was estimated using Kimura's 2-parameter method.

## Results

### Divergence and Phylogeny of the *Sonneratia* Species

Table 1 summarizes the levels of divergence among the 4 species based on 57–67 loci. *Sonneratia ovata* and *S. apetala* are the most closely related species; the mean divergence between these 2 species is 2.06%. *Sonneratia caseolaris* is an unambiguous outgroup; the genetic distance between *S. caseolaris* and the other 3 species fluctuates around 3.65%. These results are supported by phylogenetic analysis (fig. 1). The phylogeny shown in figure 1 agrees well with that reported in a previous study (Zhou et al. 2005) and has very high bootstrapping support: 97.7% for Node 3 and 100% for Node 2. The pattern of divergence between species also suggests a clock-like rate of substitution. For example, the level of divergence between *S. alba* and *S. apetala* (i.e., 2.65%) is close to that between *S. alba* and *S. ovata* (i.e., 2.44%). In fact, using Tajima's relative rate test (1993) and *S. caseolaris* as the outgroup, very few loci show significant deviation from the molecular clock hypothesis. For example, between *S. apetala* and *S. ovata*, 2 out of the 59 genes deviate from a clock-like pattern at the significance level of 5%, but none is significant after Bonferroni correction.

### Estimating Levels of Ancestral Polymorphisms, $\pi_A$

Table 2 summarizes the estimates of ancestral polymorphism using the 2 maximum likelihood methods (Takahata and Satta 1997; Yang 2002). Despite the differences between the 2 methods, they produced very similar results. Indeed, the 95% confidence intervals for the estimates overlap extensively. Furthermore, variation in mutation rate across loci does not seem important in our data set because the results obtained by using an outgroup are very close to those obtained without an outgroup. For each ancestral species, we calculated the weighted average of different estimates of its level of polymorphism (weighting different estimates according to the inverse of their variance) and obtained an approximate standard error for this weighted average (table 2 and fig. 1). Overall, Node 1 and Node 3 appear to have similar levels of polymorphism, and Node 2 is slightly less polymorphic (table 2 and fig. 1).

To further investigate the reliability of the estimates presented above, we analyzed our data by the Bayesian method (Rannala and Yang 2003; performed using the program MCMCcoal). Here data from all 4 species were used, and the results are shown in table 3. We used 2 kinds of prior distributions: in one, the 95% threshold is about 4 times of the mean (high-variance priors), and in the other, the 95% threshold is twice of the mean (low-variance priors). In most cases, the results given in table 2 fall in the 95% credible intervals of the Bayesian estimates (table 3). Thus, the estimates obtained by different methods are consistent. Similar results were obtained when very different prior distributions were used (supplementary table S3, Supplementary Material online).

### Estimating Patterns of Extant Polymorphism

Genetic diversity of the extant species is measured by nucleotide diversity $\pi$ (Hartl and Clark 1997). Because the *Sonneratia* species have fragmented distributions, we calculated $\pi$ in 3 levels of hierarchical structure: subpopulation (denoted as $\pi_S$), region ($\pi_R$), and the entire species ($\pi_T$). In general, subpopulations refer to mangrove stands and are of less relevance to the modes of speciation. Genetic variations in a geographical region ($\pi_R$) and across the entire species range ($\pi_T$) are of greater interest here.

For both *S. alba* and *S. caseolaris*, the level of diversity increases significantly each step going up the hierarchy (table 4). For instance, the mean level of regional diversity

**Table 2**
**Ancestral Polymorphisms of *Sonneratia* Estimated by Various Methods**

| | $[\pi_A \pm SE] \times 10^2$ (species compared) | | |
|---|---|---|---|
| Method | Node 1 | Node 2 | Node 3 |
| Without outgroup | | | |
| Takahata and Satta (1997) | 1.42 ± 0.26 (1, 2) | 0.82 ± 0.18 (2, 3) | 1.02 ± 0.15 (3, 4) |
| | 1.09 ± 0.20 (1, 3) | 0.99 ± 0.19 (2, 4) | — |
| | 1.03 ± 0.17 (1, 4) | — | — |
| Yang (2002) | 1.11 ± 0.21 (1, 2, 3) | 0.40 ± 0.35 (1, 2, 3) | 1.17 ± 0.28 (1, 3, 4) |
| | 1.11 ± 0.21 (1, 2, 4) | 0.90 ± 0.30 (1, 2, 4) | 1.92 ± 0.85 (2, 3, 4) |
| | 0.80 ± 0.17 (1, 3, 4) | 0.92 ± 0.16 (2, 3, 4) | — |
| Weighted average ± weighted SE | 1.05 ± 0.198 | 0.87 ± 0.197 | 1.08 ± 0.201 |
| With outgroup | | | |
| Takahata and Satta (1997) | — | 0.97 ± 0.16 (2, 4) | 1.05 ± 0.17 (3, 4) |
| | — | 0.69 ± 0.19 (2, 3) | — |
| Yang (2002) | — | 0.92 ± 0.17 (2, 3, 4) | 2.35 ± 1.15 (2, 3, 4) |
| Weighted average ± weighted SE | — | 0.87 ± 0.174 | 1.08 ± 0.191 |

NOTE.—SE, standard error; $\pi_A$ is the estimated level of ancestral nucleotide diversity. Species used for estimating the ancestral polymorphism are given in parentheses. Nodes 1–3 and species 1–4 are labeled in figure 1. *Sonneratia caseolaris* was used as the outgroup. The calculation of the weighted average is described in text.

**Table 3**
**Estimates of the Levels of Ancestral Polymorphism Obtained by Bayesian Analyses**

| | [$\pi_A$ (2.5 percentile, 97.5 percentile)] $\times 10^2$ | |
|---|---|---|
| Node | Prior | Posterior |
| High-variance priors | | |
| Node 1 | 1.05 (0.031, 3.793) | 1.13 (0.789, 1.546) |
| Node 2 | 0.87 (0.083, 2.557) | 1.27 (0.825, 1.864) |
| Node 3 | 1.08 (0.035, 3.855) | 1.87 (0.385, 4.243) |
| Low-variance priors | | |
| Node 1 | 1.05 (0.514, 1.774) | 1.14 (0.831, 1.489) |
| Node 2 | 0.87 (0.391, 1.536) | 1.15 (0.802, 1.576) |
| Node 3 | 1.08 (0.514, 1.774) | 1.14 (0.831, 1.489) |

NOTE.—$\pi_A$ is the nucleotide diversity at a particular node in the phylogenetic tree (fig. 1). We used 2 different sets of prior distributions, labeled "high-variance priors" and "low-variance priors," respectively. The means of the prior distributions are equal to the MLEs given in table 2. Results obtained by using very different priors are similar and can be found in supplementary table S3 (Supplementary Material online). When running the Markov Chain Monte Carlo (MCMC) algorithm, we discarded the first 200,000 samples (burn-in), and then sampled every 20 iterations until 100,000 data points were obtained. Furthermore, for each set of prior distributions, 2 (or more) independent runs of the MCMC algorithm were performed with different random seeds. In all cases, for a given set of priors, the results are very similar across different runs. For each of the prior and posterior distributions, the mean, 2.5% percentile, and 97.5% percentile were shown.

**Table 4**
**Extant Polymorphisms of the *Sonneratia* Species**

| Species | $\overline{\pi_S} \times 10^2$ | $\overline{\pi_R} \times 10^2$ | $\overline{\pi_T} \times 10^2$ | $F_{SR}$ | $F_{RT}$ |
|---|---|---|---|---|---|
| *S. alba* | 0.038 | 0.105 | 0.432 | 0.639 | 0.757 |
| *S. caseolaris* | 0.093 | 0.142 | 1.003 | 0.345 | 0.859 |
| *S. ovata* | — | 0.000 | — | — | — |
| *S. apetala* | — | 0.000 | — | — | — |

NOTE.—$\pi_S$ is the nucleotide diversity of a locus averaged across subpopulations, and $\overline{\pi_S}$ is the mean $\pi_S$ averaged across loci. $\overline{\pi_R}$ and $\overline{\pi_T}$ are similarly defined but are calculated using the regional samples and samples from the entire species, respectively. $F_{SR} = 1 - \overline{\pi_S}/\overline{\pi_R}$ and $F_{RT} = 1 - \overline{\pi_R}/\overline{\pi_T}$. The *F*-indices for *S. ovata* and *S. apetala* are not available because neither species is widely distributed and, hence, only one sample was obtained for each of them. See supplementary tables S1 and S2 (Supplementary Material online) for detailed sampling information.

of *S. alba* (0.105%) is 2.8 times higher than that of the subpopulation level (0.038%), whereas the mean level of the entire species (0.432%) is 4.1 times higher than that of the regional level. This hierarchical structure is reflected in the large *F*-indices (i.e., $F_{SR}$ and $F_{RT}$) as shown in table 4. Very similar results were obtained when we analyzed the 6 loci individually (data not shown). Thus, in both *S. alba* and *S. caseolaris*, there exists a hierarchy of population structure. Each species is divided into geographical regions connected by very limited gene flow. In fact, we cannot rule out zero migration across geographical regions. Within each region, the species is further subdivided into many subpopulations that are loosely connected by gene flow. Such population structure has also been observed in other mangrove species (e.g., Duke et al. 1998; Maguire et al. 2000; Dodd et al. 2002; Su et al. 2006).

*Sonneratia ovata* and *S. apetala* show no variation in our collections. For these 2 species, we only have samples from one location in China. These 2 species are narrowly distributed and, where they can be found, are in relatively low abundance.

### Contrasting Ancestral and Extant Polymorphisms

The results of this study are summarized in figure 1, which shows the values of $\pi_R$ and $\pi_T$ in contrast with the estimated ancestral polymorphism, $\pi_A$. It seems clear that 1) $\pi_A$ values are generally similar and are much larger than $\pi_R$ and 2) $\pi_A$ is as large as, or larger than, $\pi_T$.

## Discussion

### Accuracy of the Estimation

The methods we used to estimate levels of ancestral polymorphism are based on a number of assumptions (see Materials and Methods). Violation of these assumptions can affect the accuracy of the estimation, often resulting in an overestimation of ancestral polymorphism (Yang 1997; Arbogast et al. 2002; Takahata and Satta 2002). Some of the assumptions such as the molecular clock, homogeneous mutation rate across loci, and the loci being orthologous appear to be well supported by our data (tables 1 and 2 and supplementary fig. S1, Supplementary Material online). The methods also assume the absence of intragenic recombination. Although recombination does not affect the mean number of polymorphic sites at a locus, it nevertheless reduces the variance (Hudson 1983). The methods thus incorrectly fit the data with a smaller effective population size, resulting in an underestimation of the level of ancestral polymorphism (Takahata and Satta 2002; Wall 2003; Zeng K, unpublished data). Because it is likely that recombination have occurred during the course of evolution, violation of this assumption tends to make our results conservative.

Although the methods used in this study are different from each other, the main conclusions reached by different methods are similar. The assumptions on which the methods are based are compatible with our data. Therefore, the conclusion on the contrast between the ancestral and extant polymorphisms appears well supported.

### Allopatric Model Without Gene Flow

A main observation of this study is that the levels of ancestral polymorphism ($\pi_A$) are much higher than the levels of regional polymorphism ($\pi_R$). In the standard vicariant model of allopatric speciation (Mayr 1954, 1963), the ancestral population was split into 2 geographically isolated populations (or regional populations) with no gene flow. At the moment of the split, each of the regional population should resemble the ancestral one in the level of polymorphism, hence $\pi_A \sim \pi_R^* \sim \pi_T^*$. We shall use "*" to denote the level of polymorphism right after the time of geographical isolation. These regional populations eventually become reproductively isolated and evolve to become new species. When that happens, each new species may undergo range expansion and, again, experience geographical isolation. This process may also be accompanied by secondary contact with the other new species. (The scenario can be compared with the model portrayed in fig. 2 except that, in the allopatric model, there is no gene flow between regions.) The process repeats itself in cycles of speciation. In this

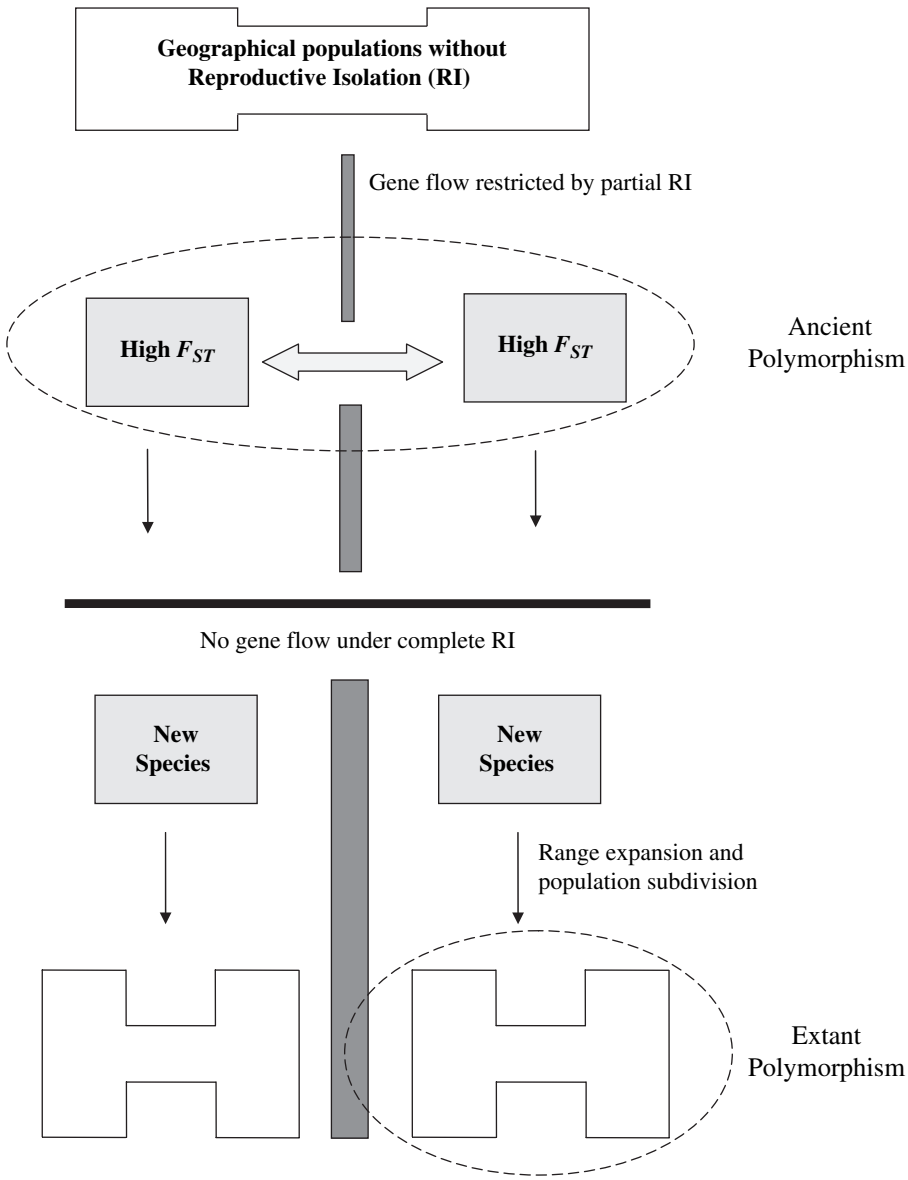


FIG. 2.—A model that explains the high level of polymorphism at the time of speciation. In this model, subdivided ancestral population with continual gene flow that ceases only after reproductive isolation has evolved to completion. The vertical bar indicates geographical isolation which can be either "porous" or complete. The horizontal dark bars denote the time after which there is no gene flow between populations. The cycle of population subdivision and speciation continues as new species form.

ongoing process, we may consider comparable stages in different cycles as equivalent and, hence, $\pi_R \sim \pi_R^*$. In the allopatric model, $\pi_T$ is less relevant as it is a measure of the total diversity across current species and is generally much larger than $\pi_R$.

In the simplest allopatric model described above, the expected pattern may be summarized as $\pi_A (\sim\pi_R^*) \sim \pi_R < \pi_T$. Our observation that $\pi_A$ is often 5–10 times greater than $\pi_R$ suggests that, if simple allopatry is indeed true, these regional populations must have become smaller after geographical isolation and have remained so for a long time. It has often been suggested that severe bottlenecks might be conducive, or even necessary, for the formation of species (Mayr 1954; Carson 1970, 1971; Nei et al. 1975; Templeton 1980; Harrison 1991). Specifically, successive glacial periods during the Pleistocene (Chappell and Shackleton 1986; Wang et al. 1995; Saenger 1998; Voris 2000) may have induced such population bottlenecks.

The hypothesis of severe population bottleneck under strict allopatry is plausible but is not fully compatible with ancillary observations. *Sonneratia alba* and *S. caseolaris* are both very abundant species in each geographical region of the present day. Furthermore, the main effect of past glaciations should be the fragmentation of populations, but this effect should have abated after the retreat of glaciers. Indeed, *S. alba* and *S. caseolaris* are not particularly low in total genetic variation among woody plants. For example, the level of polymorphism is 0.38% in *Cryptomeria japonica* (Kado et al. 2003), 0.49% in *Pinus sylvestris* (Dvornyk et al. 2002), 0.64% in *Pinus taeda* (Brown

et al. 2004), 0.39% in *Picea abies* (Heuertz et al. 2006), and 1.6% in the European aspen, *Populus tremula* (Ingvarsson 2005). At 0.43% and 1.00%, respectively, *S. alba* and *S. caseolaris* are not particularly low in polymorphism. (*Sonneratia ovata* and *S. apetala* are narrowly distributed and have the typical low diversity of endemic species.)

### An Alternative Model of Gene Flow between Ancestral Populations

An alternative explanation for the contrasting patterns of polymorphism of figure 1 is depicted in figure 2. In this model, the ancestral species were composed of interconnected geographical populations, which would differentiate into separate species. During this process, gene flow continues for an extended period of time (Wu 2001; Osada and Wu 2005; Patterson et al. 2006). Gene flow is depicted as the 2-headed arrow in figure 2 between diverging populations with high $F_{ST}$.

The estimated ancestral diversity would reflect the level of polymorphism at the time when the 2 interconnected populations were about to become reproductively isolated species ("ancestral polymorphism"). The total genetic diversity of interconnected regional populations (i.e., $\pi_T$) can be expressed approximately as $\pi_R/(1 - F_{RT})$ (Nei 1973; Slatkin 1991), where $\pi_R$ is regional polymorphism and $F_{RT}$ is the level of differentiation between regions. To account for the large ancestral polymorphism, $\pi_R$ does not have to be large because the total diversity is augmented by a factor of $1/(1 - F_{RT})$ over the local diversity. In the model of figure 2, the population structure at the time of speciation mirrored that of the extant species. The species-level polymorphism is a reflection of the level of subdivision among populations, and this population structure leads to large estimates of ancestral polymorphism (Takahata and Satta 2002; Teshima and Tajima 2002).

Following the notation above, we use "*" to denote the ancestral level of polymorphism (i.e., at the time when the estimated $\pi_A$ applies). The expected pattern under the model of figure 2 is thus $\pi_A \sim \pi_T^* \sim \left[\pi_R^*/\left(1 - F_{RT}^*\right)\right] \geq \pi_T \sim [\pi_R/(1 - F_{RT})]$. (The pattern of $\pi_A \sim \pi_T^* \sim \pi_T$ can be easily understood and we shall explain $\pi_A \sim \pi_T^* > \pi_T$ in the next paragraph.)

### Comparisons of the 2 Models and the Implications

In the strict allopatric model without gene flow, the estimated $\pi_A$ should reflect regional polymorphism, present or past. In the model figure 2, $\pi_A$ should be closer to species-level polymorphism at the time of speciation. Because $\pi_A$ is closer to $\pi_T$ of the extant population (see fig. 1), a structured population with gene flow shown in figure 2 should be a reasonable approximation for the populations both at the time of speciation and at present. This model of figure 2 can also account for the pattern of $\pi_A > \pi_T$ as follows: Structured populations at the time of speciation may be more strongly divided than at any other time; hence, $F_{RT}^* \geq F_{RT}$. As a result, $\pi_A$ may be as large as, or larger than, the observed $\pi_T$.

Although the model of figure 2 appears to be compatible with the parapatric model of speciation, which posits continual (but restricted) gene flow during speciation, it is not by itself a rejection of the allopatric model of speciation. To reject allopatry, it is necessary to rule out a period of geographical isolation. The model of figure 2 suggests that complete geographical isolation may not be necessary (i.e., the genic view of speciation; see Wu 2001) but does not prove that it is indeed absent. Proponents of the allopatric model may still argue that, without such a period, these populations cannot make the transition to fully reproductively isolated species. The presence or absence of a period of complete geographical isolation has to be determined by other means.

In summary, the allopatric model requires complete geographic barriers between populations to stop gene flow. In the parapatric model, gene flow is gradually reduced as more and more loci of local adaptation become differentiated. Genetic exchanges near these loci are restricted by linkage to them. Differential adaptation of geographical populations of mangroves is possible, for example, by salt tolerance. Indeed, substantial genetic differentiation between adjacent inland and littoral populations of a mangrove species has been shown by amplified fragment length polymorphism markers (Tang et al. 2003). Given the proximity of these populations, the observed genetic differentiation might be driven by adaptation to different environments. Finally, the approach outlined in this study can potentially be widely applicable to nonmodel organisms. It would be most interesting to see if the differences between patterns of ancestral and extant polymorphism reported in this study are generally true.

## Supplementary Material

Supplementary tables S1–S3 and fig. S1 are available at *Molecular Biology and Evolution online* (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

We thank Cairong Zhong for his help in collecting plant samples. Thanks are also due to Dr Saitou and 2 anonymous reviewers for their insightful suggestions. R.Z. is supported by the Young Teacher Foundation of Sun Yat-Sen University (2006-33000-1131357). K.Z. is supported by Sun Yat-sen University and the Kaisi Fund. C.-I.W. is supported by National Institutes of Health grants and an grant from the Chinese Academy of Sciences. S.S. is supported by grants from the National Natural Science Foundation of China (30730008, 30470119) and 973 program (2007CB815708).

## Literature Cited

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst. 33:707–740.

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA. 101:15255–15260.

Carson HL. 1970. Chromosome tracers of the origin of species. Science. 168:1414–1418.
Carson HL. 1971. Speciation and the founder principle. Stadler Genet Symp. 3:51–70.
Chappell J, Shackleton NJ. 1986. Oxygen isotopes and sea level. Nature. 324:137–140.
Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet. 68:444–456.
Clark AG. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol. 7:111–122.
Dodd RS, Rafii ZA, Kashani N, Budrick J. 2002. Land barriers and open oceans: effects on gene diversity and population structure in *Avicennia germinans* L. (Avicenniaceae). Mol Ecol. 11:1327–1338.
Duke NC, Benzie JAH, Goodall JA, Ballment ER. 1998. Genetic structure and evolution of species in the mangrove genus *Avicennia* (Avicenniaceae) in the Indo-West Pacific. Evolution. 52:1612–1626.
Dvornyk V, Sirviö A, Mikkonen M, Savolainen O. 2002. Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*. Mol Biol Evol. 19:179–188.
Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.
Harrison RG. 1991. Molecular changes at speciation. Ann Rev Ecol Evol. 22:281–308.
Hartl DL, Clark AG. 1997. Principles of population genetics. Sunderland (MA): Sinauer Associates. p. 59–60.
Hedrick PW. 1999. Highly variable loci and their interpretation in evolution and conservation. Evolution. 53:313–318.
Hedrick PW. 2005. A standardized genetic differentiation measure. Evolution. 59:1633–1638.
Heuertz M, De Paoli E, Kallman T, Larsson H, Jurman I, Morgante M, Lascoux M, Gyllenstrand N. 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. Genetics. 174:2095–2105.
Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics. 167:747–760.
Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 23:183–201.
Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics. 132:583–589.
Ingvarsson PK. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics. 169:945–953.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.
Kado T, Yoshimaru H, Tsumura Y, Tachida H. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). Genetics. 164:1547–1559.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.
Li YJ, Satta Y, Takahata N. 1999. Paleo-demography of the Drosophila melanogaster subgroup: application of the maximum likelihood method. Genes Genet Syst. 74:117–127.
Maguire TL, Saenger P, Baverstock P, Henry R. 2000. Microsatellite analysis of genetic structure in the mangrove species *Avicennia marina* (Forsk.) Vierh. (Avicenniaceae). Mol Ecol. 9:1853–1862.
Mayr E. 1954. Change of genetic environment and evolution. In: Huxley J, Hardy AC, Ford HB, editors. Evolution as a process. London. p. 157–180.
Mayr E. 1963. Animal species and evolution. Cambridge (MA): Belknap Press.
Nagylaki T. 1998. Fixation indices in subdivided populations. Genetics. 148:1325–1332.
Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA. 70:3321–3323.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press. p. 288–289.
Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. Evolution. 29:1–10.
Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics. 158:885–896.
Olsen KM, Schaal BA. 1999. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. Proc Natl Acad Sci USA. 96:5586–5591.
Osada N, Wu CI. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. Genetics. 169: 259–264.
Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol. 5:568–583.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature. 441:1103–1108.
Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.
Saenger P. 1998. Mangrove vegetation: an evolutionary perspective. Mar Freshw Res. 49:277–286.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Satta Y, Hickerson M, Watanabe H, O'hUigin C, Klein J. 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J Mol Evol. 59:478–487.
Shi S, Huang Y, Tan F, He X, Boufford DE. 2000. Phylogenetic analysis of the Sonneratiaceae and its relationship to Lythraceae based on its sequences of nr DNA. J Plant Res. 113:253–258.
Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet Res. 58:167–175.
Su GH, Huang YL, Tan FX, Ni XW, Tang T, Shi SH. 2006. Genetic variation in Lumnitzera racemosa, a mangrove species from the Indo-West Pacific. Aquat Bot. 81:175–788.
Tajima F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics. 135:599–607.
Takahata N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. Genet Res. 48:187–190.
Takahata N. 1993. Mechanisms of molecular evolution: introduction to molecular paleopopulation biology. Sunderland (MA): Sinauer Associates.
Takahata N, Satta Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci USA. 94:4811–4815.
Takahata N, Satta Y. 2002. Pre-speciation coalescence and the effective size of ancestral populations. In: Slatkin M, Veuille M, editors. Modern developments in theoretical population genetics. Oxford: Oxford University Press. p. 52–71.
Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol. 48:198–221.

Tang T, Zhong Y, Jian S, Shi S. 2003. Genetic diversity of *Hibiscus tiliaceus* (Malvaceae) in China assessed using AFLP markers. Ann Bot. 92:409–414.

Templeton AR. 1980. The theory of speciation via the founder principle. Genetics. 94:1011–1038.

Teshima KM, Tajima F. 2002. The effect of migration during the divergence. Theor Popul Biol. 62:81–95.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tomlinson PB. 1986. The botany of mangroves. Cambridge: Cambridge University Press.

Voris HK. 2000. Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. J Biogeogr. 27:1153–1167.

Wall JD. 2003. Estimating ancestral population sizes and divergence times. Genetics. 163:395–404.

Wang P, Wang L, Bian Y, Jian Z. 1995. Late quaternary paleoceanography of the South China Sea: surface circulation and carbonate cycles. Mar Geol. 127:145–166.

Wang RL, Wakeley J, Hey J. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. Genetics. 147:1091–1106.

Wu CI. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics. 127: 429–435.

Wu CI. 2001. The genic view of the process of speciation. J Evol Biol. 14:851–865.

Yang Z. 1997. On the estimation of ancestral population sizes of modern humans. Genet Res. 69:111–116.

Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics. 162:1811–1823.

Zhou R, Shi S, Wu C-I. 2005. Molecular criteria for determining new hybrid species-an application to the *Sonneratia* hybrids. Mol Phylogenet Evol. 35:595–601.

Naruya Saitou, Associate Editor

Accepted September 24, 2007