Population Genetics in Nonmodel Organisms: II. Natural Selection in Marginal Habitats Revealed by Deep Sequencing on Dual Platforms

Renchao Zhou,†,[1] Shaoping Ling,†,[2] Wenming Zhao,†,[2] Naoki Osada,[3] Sufang Chen,[1] Meng Zhang,[1] Ziwen He,[1] Hua Bao,[1] Cairong Zhong,[4] Bing Zhang,[2] Xuemei Lu,[1,2] David Turissini,[5] Norman C. Duke,[6] Jian Lu,*,[7] Suhua Shi,*,[1] and Chung-I Wu*,[1,5,8]

[1]State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-sen University, Guangzhou, China

[2]Key Laboratory in Genome Science and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[3]Department of Biomedical Resources, National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan

[4]Hainan Dongzhai Harbor National Nature Reserve, Haikou, China

[5]Department of Ecology and Evolution, University of Chicago

[6]School of Biological Sciences, University of Queensland, St Lucia, Queensland, Australia

[7]Department of Molecular Biology and Genetics, Cornell University

[8]Laboratory in Disease Genomics and Individualized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

†These authors contributed equally to this work.

***Corresponding author:** E-mail: JL2434@cornell.edu; lssssh@mail.sysu.edu.cn; ciwu@uchicago.edu.

Associate editor: Hideki Innan

Research article

Abstract

Population genetics of species living in marginal habitats could be particularly informative about the genetics of adaptation, but such analyses have not been readily feasible until recently. *Sonneratia alba*, a mangrove species widely distributed in the Indo-West Pacific, provides a very suitable system for the study of local adaptation. In this study, we analyzed DNA variation by pooling 71 genes from 85–100 individuals for DNA sequencing. For each of the two nearby *S. alba* populations, we obtained ~2,500 × coverage on the Illumina GA platform and for the Sanya population, an additional 5,400 × coverage on the AB SOLiD platform. For the Sanya sample, although each sequencing method called many putative single nucleotide polymorphisms, the two sets of calls did not overlap, suggesting platform-dependent errors. Conventional sequencing corroborated that each population is monomorphic. The two populations differ by 54 bp of 79,000 sites, but 90% of the variants are found in 10% of the genes. Strong local adaptation and high migration may help to explain the extensive monomorphism shared by the two populations in the presence of a small number of highly differentiated loci.

Key words: population genomics, next-generation sequencing, natural selection, *Sonneratia*, mangroves.

Introduction

Evolutionary dynamics of populations thriving in marginal habitats are fascinating in several aspects (Kawecki 2008). First of all, selective pressure for the desirable traits is presumably intense in marginal habitats. Positive Darwinian selection will drive the beneficial mutations (either de novo formed in local populations or migrated from core populations) to high frequency or fixation in the marginal populations rapidly. Second, marginal populations are usually sparse and fragmented, hence fluctuation of population size or extinction–recolonization will significantly reduce genetic diversities harbored in these populations (Glemin et al. 2003). Third, gene flows between core and marginal populations are usually asymmetric (Manier and Arnold 2005; Kawecki 2008). Fourth, marginal populations usually have small effective population size so that deleterious mutations would accumulate in the populations (Nieminen et al. 2001). Therefore, population genetics analysis of marginal populations (or species) could be particularly informative about the genetics of adaptation as well as other evolutionary forces.

Mangroves are a very suitable system for the study of adaptation in marginal habitats. These are woody plants that inhabit the intertidal zones of tropical and subtropical coasts. Consisting of approximately 70 species from about 20 families (Duke 1992), most mangroves tolerate fluctuations in environmental conditions, such as salinity, tidal currents, winds, temperature fluctuations, and muddy anaerobic soils. It was suggested that no other group of woody plants was so highly adapted to such extreme conditions (Kathiresan and Bingham 2001). *Sonneratia* L., comprising 5–6 species, is a typical mangrove genus in the Indo-West Pacific region (Duke and Jackes 1987). *S. alba* (supplementary fig. S1, Supplementary Material online) is the most widespread species of this genus,

ranging from eastern Africa to southern China and from northern Australia to Okinawa, Japan. Growing in the estuary, it is one of the most salt tolerant mangroves (Ball and Pidsley 1995).

Our previous surveys on a small number of genes and individuals revealed that *S. alba* harbors little genetic variation (Zhou et al. 2007). In species known to be lowly polymorphic, including many endangered species and species living in marginal habitats such as *S. alba*, the precise level of low diversity, whether at 10^{-4} or $<10^{-5}$ per nucleotide site, has important implications. For example, if no polymorphism is detected by exhaustive sequencing, the population may be effectively a single clone, resulting from a recent severe bottleneck, extreme selective pressure, or both. Genetic diversity surveys of the marginal populations will also provide insights into conservation strategies.

In this study, we employed next-generation sequencing techniques (G2 for short) to investigate the genetic diversities of *S. alba* at a much larger scale than was done previously (Zhou et al. 2007). Given its unprecedented capacity, a particularly powerful application of G2 may be the accurate determination of genetic diversity at the low limit. However, a great challenge is its relatively high error rate (between 10^{-3} and 10^{-2}), which could confound the assessment of low-level polymorphisms (Margulies et al. 2005; Dohm et al. 2008; Shendure and Ji 2008). Errors may occur in every step of the process starting from library construction to sequence determination. Due to the intrinsic differences in design principles, different G2 sequencing platforms (including Roche 454, Illumina/Solexa Genome Analyzer, and ABI SOLiD) have distinct rates and types of sequencing errors. For example, the Roche 454 platform is systematically biased toward homopolymer errors (Wicker et al. 2006), and Illumina GA has higher single nucleotide error rates at the 3′ end (Rougemont et al. 2008). A recent study comparing the three G2 techniques estimated that the false-positive rate of single nucleotide polymorphism (SNP) calling for 454, Illumina GA, and SOLiD is 2.5%, 6.3%, and 7.8%, respectively, and the false-negative rate is 3.1%, 0%, and 0.9% (Harismendy et al. 2009).

Because the literature on applying the G2 technology to population genomics, especially to the precise measurement of very low levels of genetic diversity at a large scale, is still limited, we devoted the first half of this study to the technical development. We used both Illumina GA and SOLiD platforms to targeted gene regions of one population of *S. alba* to cross validate the sequencing results. We found that the combination of the two G2 platforms could greatly increase the confidence in SNP callings. We then applied the procedure to the specific problems of genetic diversity in mangrove populations. By comparing polymorphisms within and between two populations of *S. alba* (fig. 1), we found interesting biological patterns related to local adaptation of mangroves. The procedure developed in this study should nevertheless be generally applicable to systems where the precise determination of low-level genetic diversity is crucial.

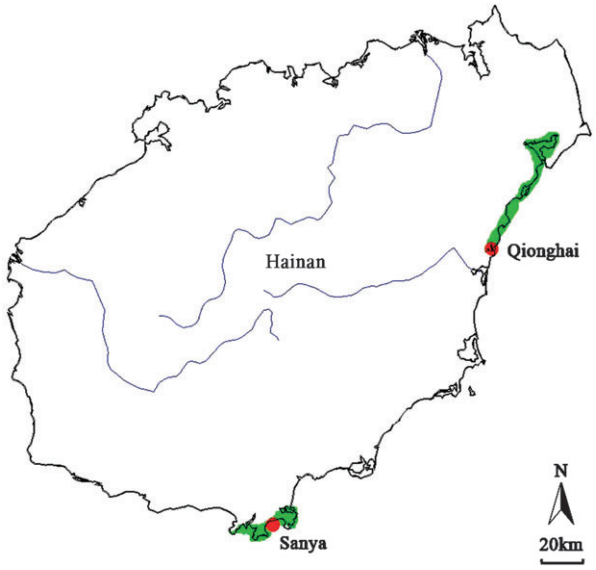

FIG. 1. Map locations of the Sanya (Yalong Bay, 18°22′N, 109°63′E) and Qionghai (Tanmen Harbor, 19°14′N, 110°37′E) populations of *S. alba*. Green colored areas show the distribution of mangrove forests, whereas the red dots indicate sampling sites. The distance between the two populations is about 100 km. The salinity is 32‰ for the Sanya samples and ranges from 11‰ to 22‰ for the Qionghai population.

Materials and Methods

Plant Materials and Salinity Investigation

The two populations of *S. alba* on Hainan Island are located in Tanmen Harbor, Qionghai (19°14′N, 110°37′E) and Yalong Bay, Sanya (18°22′N, 109°63′E), respectively (supplementary fig. S1, Supplementary Material online). We collected 100 individuals from the Qionghai population and 85 from the Sanya population, respectively. Intervals between sampled individuals are at least 15 m. One or two leaves were collected from each individual tree and stored in plastic bags with silica gels for DNA extraction.

We collected six seawater samples in the sampling area of *S. alba* from Qionghai, alongshore from the estuary (Tanmen Harbor) to the end of distribution of *S. alba*, with a minimum distance between two adjacent subpopulations of 500 m (fig. 1). One seawater sample was collected in the sampling site of *S. alba* from Sanya, Yalong Bay (fig. 1). Salinity was measured using a Handheld Refraction Meter (Saltmeter; Chengdu Taihua). For the Qionghai population, salinities ranged from 11‰ to 22‰ with lowest levels upstream and highest levels toward the river mouth. The salinity of Sanya population was 32‰.

Molecular Experiments

The aim of this study is to detect low-frequency SNPs but not to trace the identity of SNPs in the samples, thus for each of the Sanya (85 individuals) and Qionghai (100 individuals) populations, we prepared pooled samples of leaves from all individuals of the same population and isolated total DNAs from the mixed samples. Dry weight of leaves from each individual was carefully measured, and the same amount of leave tissues

(0.02 g) from each individual was used in the pooled samples. We used CTAB method (Stewart and Via 1993) to extract DNA from the pooled samples.

Based on the sequences of over 200 clones from a cDNA library of a congeneric species *S. caseolaris*, we designed primers of 71 genes for *S. alba*. These genes consist of 70 nuclear genes and one chloroplast fragment. All the primers were synthesized at the Invitrogen (Shanghai) company. The 71 loci were respectively amplified for the two mixed samples using corresponding primer pairs. We used high-fidelity PrimeStar HS DNA polymerase (TaKaRa, Dalian, China) to reduce nucleotide mismatches during polymerase chain reaction (PCR) amplification. The PCR products were purified by electrophoresis through 1.2% agarose gels followed by use of the Pearl Gel Extraction Kit (Pearl Bio-tech, Guangzhou).

The purified PCR products for the 71 loci were pooled in equal quantity based on the measures of Ultraviolet Spectrophotometer mass mixtures (70 ng per amplicon, 5 ug in total). Single-end library preparation, sequencing, and base calling for pooled samples were done with Illumina GA I according to the manufacturer's recommendations at Shenzhen Huada Genomics Institute (Shenzhen). Each pooled sample was sequenced with one GA channel.

We used ABI SOLiD sequencing platform to sequence the pooled Sanya sample. One quarter of ABI SOLiD sequencing run was performed. SOLiD fragment library preparation, emulsion PCR, sequencing, and the primary analysis on the SOLiD machine were conducted according to the manufacturer's instructions, and the raw color-space sequences and their Quality Values (QV) were generated at Beijing Institute of Genomics, the Chinese Academy of Sciences.

Double-stranded Sanger sequencing was carried out with amplification primers and internal locus-specific primers (when required) on ABI 3730 DNA automated sequencer with the BigDye chemistry for two purposes. One aim is to obtain reference sequences of each gene for each population, and the other aim is to validate the putative SNPs called by GA and SOLiD sequencing platforms. For either population, we randomly selected and sequenced one individual to obtain the reference sequences of the 71 loci. All sequences have been deposited in GenBank with accession numbers GQ121922–GQ121992 and GQ121993–GQ122063. We sequenced seven genes (18, 32, 48, 51, 58, 59, and 61) in all 85 individuals from the Sanya population and two (18 and 27) in all 100 individuals from the Qionghai population.

Data Analysis

Sequence reads generated by GA and SOLiD platforms were mapped to the reference sequence sets with at most 2 and 3 mismatches for each 35-bp sequence read by using MAQ (Li et al. 2008) and Corona_lite, respectively. We wrote a series of Perl scripts to compute the following seven parameters for the two mapping results.

1) Site Coverage (SC): SC = Number of reads covering one site.
2) Allele Coverage (AC): AC = Number of reads covering one allele of a site.
3) Forward/Reverse Start-Point (FSP/RSP): FSP/RSP = Number of Start-Points covering one site in the forward/reverse direction.
4) Average Read QV (ARQ):

$$\mathrm{ARQ} = -10 * \log\Big(\sum_{i}^{N} 10^{-Q_i^{\mathrm{Phred}}/10}/N\Big)/\log(10),$$

where N is the read length.

5) Average Allele QV at one site (AAQ):

$$\mathrm{AAQ} = -10 * \log\Big(\sum_{i}^{M} 10^{-Q_i^{\mathrm{Phred}}/10}/M\Big)/\log(10),$$

where M is the number of reads mapping to a given allele.

6) Most Abundant Allele Frequency (MOF): MOF = Number of reads for the most abundant allele/SC.
7) Minor Allele Frequency (MAF): MAF = Number of reads for the second abundant allele/SC.

Here Start-Point is the unique starting positions of mapped reads covering one site. There are at most 35 Start-Points for one site in each direction.

Here QV is traditional Phred quality. AB SOLiD QV uses the Phred Quality as its standard. Illumina GA QV′ can be transferred to Phred quality according to the formula:

$$\mathrm{QV} = 10 * \log(1 + 10^{\mathrm{QV}'}/10)/\log 10.$$

SNPs were called when they passed the minimum AAQ, minimum SC, minimum FSP/RSP, and minimum MAF. We set three threshold values (0, 8, and 11) for the minimum ARQ, three threshold values (8, 15, 20) for the minimum AAQ, 10× for the minimum SC, 1 for both the minimum FSP and RSP, and 0.01 for the minimum MAF. Table 2 shows putative SNPs detected in the Sanya population at the three levels of stringency: high (AAQ >= 20), medium (AAQ >= 15), and low (AAQ >= 8), in addition to the common criteria of ARQ >= 11, SC >= 10×, FSP/RSP>= 1, and MAF > 0.01. If we assume each site has 2,500 × coverage and the frequency of the putative SNPs is greater than 0.01, then in theory, the possibility that a SNP is a sequencing error is 10^{-50} under the high stringency and 10^{-20} under the low-stringency criteria.

After SNP calling from each platform, we wrote a series of Python scripts to identify the joint calls from the two platforms. We classified all putative SNPs into three categories: 1) GA_only, 2) SOLiD_only, and 3) Joint call from GA/SOLiD platforms. Here, $\mathrm{MAF_{GA}}$ and $\mathrm{MAF_{SOLiD}}$ were the MAF obtained from GA platform and SOLiD platform, respectively.

1) GA_only—Sites where $\mathrm{MAF_{GA}}/\mathrm{MAF_{SOLiD}} > 2$ or $(\mathrm{MAF_{GA}}-\mathrm{MAF_{SOLiD}}) > 5\%$;
2) SOLiD_only—Sites where $\mathrm{MAF_{SOLiD}}/\mathrm{MAF_{GA}} > 2$ or $(\mathrm{MAF_{SOLiD}}-\mathrm{MAF_{GA}}) > 5\%$; and
3) Joint calls—Sites where the ratio between two MAFs is less than 2-fold, and the difference is less than 5%. A joint call also has to be out of the homopolymer or microsatellite region.

Note that GA calls and SOLiD calls in table 1 are slightly different from GA_only and SOLiD_only defined here. GA

Table 1. A Summary of the Sequencing Data for Sanya and Qionghai Populations of *Sonneratia alba*.

Samples	Sanya	Sanya	Qionghai
Number of individuals sampled	85	85	100
Number of genes	71	71	71
Sequencing platform	ABI SOLiD	Illumina GA	Illumina GA
Number of total reads	29,157,609	6,904,694	8,161,413
Mapping program	Corona	Maq	Maq
Percentage of mapping reads	41.95	82.01	67.11
Sequencing depth	5,423×	2,510×	2,428×

calls includes GA_only, joint calls and those calls meeting the first condition of joint call but being in the homopolymer or microsatellite region. SOLiD calls are defined similarly.

Frequency Spectra

We converted the sequencing frequencies of the SNPs in the Sanya population into frequencies in a sample of 170 chromosomes (85 individuals pooled and sequenced). The expected frequency spectra were calculated according to Fu (1995).

Simulations

Computer simulations were performed to test whether the observed clustering of high F_{st} sites is compatible with the isolation–bottleneck models. Here, we assumed that the effective population size of the two extant populations were extremely small ($4N_e\mu = 0.001$, where μ is mutation rate per locus) and no migration between the populations. We also set $4N_eu = 0.000001$ (which is extremely low) to conduct the simulation and obtained the same conclusion. The two populations have split from the common ancestor with an effective population size of N_A before time τ, which is measured by the unit of $4N_A$. We used ms software to generate random genealogies (Hudson 2002). In each step, genealogies of 100 samples from each population for 71 loci of 1000 bp length were generated with given τ and θ, where $\theta = 4N_A\mu$. Note that $\theta = 1$ per locus is equivalent to $\theta = 0.001$ per base pair. We also incorporated variable mutation rates among loci. For each locus, a mutation rate parameter, r, was drawn from the gamma distribution with a shape parameter $= \alpha$ and a scale parameter $= 1/\alpha$. Because the average of r is 1, we multiplied θ by r for each locus to account the rate variation.

We approximated the posterior distribution of τ and θ by a rejection sampling method (Tavare et al. 1997; Weiss and von Haeseler 1998). Let X be the number of zero-divergence genes in the sample. Here, X was used as a summary statistic for fitting the parameters to data. In each step, τ and θ were randomly selected from the uniform distribution (U[0,1]), and the number of zero-divergence genes in the simulated data, x, is estimated. A set of τ and θ are accepted when $X = x$. When τ and θ are accepted, ith highest number of differentiated sites ($F_{st} \geq 0.95$) per gene, y_i, in the simulated data is simultaneously estimated. After 1 million rounds, we obtained the joint posterior distribution of τ and θ and the distribution of y_i with given posterior distribution. Figure 3 in the main text shows the result when $\alpha = 5$. It is intuitive that the assumption of high mutation rate variation (smaller α) makes the test conservative. We estimated α using the divergence data between *S. alba* and *S. caseolaris*. We estimated that the value of α was 5.7 between *S. alba* and *S. caseolaris*, as implemented in the Mathematica routine FindMinimum and NMaximize, indicating that our choice of $\alpha = 5$ makes the test conservative. Furthermore, the test was robust when we assumed a higher mutation rate variation ($\alpha = 1$).

Results

Determination of Low-Level Genetic Diversity by Dual Sequencing Platforms

Using both Illumina GA and SOLiD platforms, we sequenced pooled DNAs of 71 genes from 85 individuals of *S. alba* collected in Sanya on Hainan Island in southern China (fig. 1). In parallel, one individual was randomly selected for conventional Sanger sequencing in order to obtain full-length reference sequences. Detailed information on experimental procedures is presented in Materials and Methods. The 71 genes were selected on the basis of cDNA sequencing results in our previous study (Zhou et al. 2007). The orthologous sequences of the 71 genes have also been determined in four *Sonneratia* species (Zhou et al. 2007). The total length of the 71 genes is close to 80 kb (with gene length ranging from 530 to 2,092 bp and mean length at 1,112 bp; supplementary table S1, Supplementary Material online).

We obtained 6,904,694 raw reads from the Sanya sample by Illumina GA (table 1). The aligned reads give ~2,500 × coverage or 30 × per individual (table 1). We used three levels of stringencies in SNP calling (see Materials and Methods). Under the low-stringency criteria, 210 SNPs in 57 genes were identified (table 2). After correcting for coverage, the frequency spectra of the minor alleles were plotted in figure 2A (170 chromosomes from 85 diploid individuals). In comparison, we also plotted the expected frequency spectra under neutral equilibrium conditions. (Note that bins are merged for presentation.) It is notable that the observed minor alleles are significantly skewed toward the very low-frequency classes. Although there are biological explanations for the strong skew, errors in SNP callings have to be addressed in the first place. For example, 137 (65.2%) of the minor alleles are Ts and slightly more than 50% of the SNPs are in local homopolymers or oligonucleotide repeats. The number of SNP calls decreases as the stringency increases. We observed 13 and 3 SNPs under the medium and high stringency, respectively (table 2).

With the SOLiD platform, we obtained more than 12 million mapped reads (~5,400 × coverage) and expected to see all SNPs detected by Illumina GA (table 1). However, SOLiD sequencing discovered only 23 SNPs under the low-stringency criteria, most of which are in very low frequencies (fig. 2B). Under the medium and high stringency criteria, we observed 3 and 0 SNPs, respectively (table 2).

Although reducing the rate of false positives, the high stringency calls could have missed many true SNPs (false

Table 2. Numbers of Putative SNPs in Sanya Population Called by the GA, SOLiD, or Both Platforms.

Level of Stringency in SNP Calling[a]	High	Medium	Low
GA calls	3 (0)	13 (9)	210 (80)
SOLiD calls	0 (0)	3 (2)	23 (16)
Joint calls	0	0	2
% reads discarded (% reads discarded among mapped reads)	58.7 (21.2)	44.1 (17.9)	30.0 (13.1)

The numbers in the parentheses in the upper two rows are putative SNPs that fall near locally repetitive regions.
Three levels of stringency of SNP calling are defined in Materials and Methods. Specific criteria include sequencing depth, quality value, and MAF. Joint call is defined in Materials and Methods.

negatives). The results in table 2 highlight the quandary in calling SNPs at low frequency. Between platforms and among levels of stringencies, the number of SNPs could vary from 0 to 210. It is hence necessary to compare and combine the performances of the two sequencing platforms, and the unusual sequencing depth in this study has allowed us to do so. We wish to know the rates of false negatives and positives on either platform and to evaluate the power of combining dual platforms. Figure 3 shows the MAF called by either platform at each site. (Minor allele denotes the second most common variant at a site.) Unfiltered data were used in order to compare the "raw performance" of each platform. It is striking that the MAFs called by the two platforms fall predominantly on the two edges. In other words, most sites appear polymorphic on only one of the two platforms. For comparison, the inset in figure 3 shows the data that have been filtered with low-stringency criteria (see also table 2). Although there are far fewer SNPs in the inset, the scattering is similar.

A salient observation in table 2 is that the number of joint calls (see Materials and Methods) is consistently low even when stringency is relaxed. Only two SNPs were discovered by dual platforms using the low-stringency criteria but neither remains true when the stringency is raised (table 2). Our subsequent validation by the conventional Sanger sequencing method on seven genes (see supplementary table S1, Supplementary Material online) from all 85 individuals confirms the absence of polymorphism. The results suggest that the Sanya population of *S. alba* is indeed extremely low in genetic diversity. It also suggests that the concordance between the two G2 platforms is effective in screening out sequencing errors.

There are two explanations for the discordance between the results of the two platforms. First, true polymorphic SNPs were missed by either of the two platforms (i.e., false negatives). False negatives can happen in two ways—the rare variant is not sampled or, when it is sampled, it is read incorrectly. Because the error rate is rarely >5% for any site

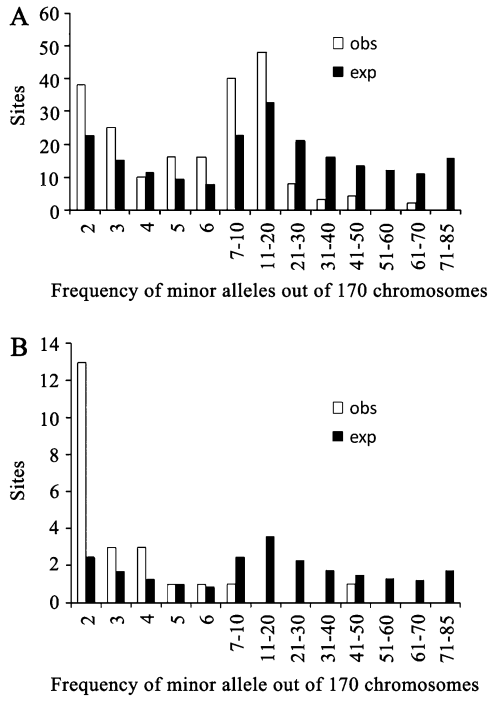

Fig. 2. Frequency spectra of SNPs of the Sanya population by Illumina GA (*A*) and SOLiD (*B*) platforms. The expected frequency (Exp) of the minor allele under neutral evolution is given together with the observed (Obs). Note that the bins are merged to spread the counts more evenly.

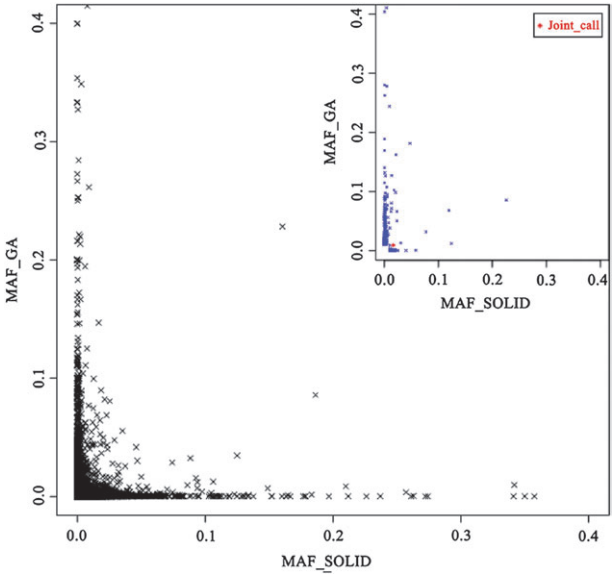

Fig. 3. MAF of putative SNPs called by either SOLiD or Illumina GA. These calls are unfiltered "raw" data. The inset in the upper right corner shows MAFs at low stringency (see Materials and Methods and table 2). The *x* axis is the MAF of a putative SNP by SOLiD, and the *y* axis is the MAF of the corresponding SNP by Illumina GA. Although the number of putative SNPs is markedly lower in the inset, the general patterns of nearly mutually exclusive calls are similar between the main figure and the inset. Sanger sequencing was carried out on selected genes to validate the SNP calls and no SNPs were validated.

Table 3. Distribution of False-Positive Rates in G2 Sequencing for the Sanya Population of *Sonneratia alba*.

	Frequency	
False-Positive Rate	SOLiD	Illumina Genome Analyzer
0–0.001	0.7456	0.5373
0.001–0.01	0.2356	0.4466
0.01–0.02	0.0112	0.0083
0.02–0.03	0.0036	0.0025
0.03–0.04	0.0014	0.0014
0.04–0.05	0.0007	0.0010
0.05–0.06	0.0004	0.0007
0.06–0.07	0.0004	0.0003
0.07–0.08	0.0002	0.0003
0.08–0.09	0.0002	0.0003
0.09–0.10	0.0001	0.0003
>0.10	0.0007	0.0011
Mean	0.0009	0.0014

(table 3), false negatives are negligible when the scheme ensures multiple reads. In this study, thanks to the very deep coverage (2,500–5,400×), even a rare variant of 1% should have been read more than 25 times (the probability of being missed is $<10^{-10}$ by a Poisson distribution). In supplementary figure S2, Supplementary Material online, we plotted the depth of coverage for the sites where SNPs were called by either platform. The coverage for the SNP sites is about the same as non-SNP sites. In short, given the high coverage, it seems improbable that any true polymorphic site could have been missed by either platform in this study.

The results suggest a second explanation; namely, the discordant SNP calls are platform-dependent errors. These errors are not surprising given that the two platforms rely on different sequencing chemistries and associated techniques (Mardis 2008). We may then calculate the false-positive rates for both platforms. Although the average raw error rate per site ranged from 0.6–1.5% in previous applications (Margulies et al. 2005; Dohm et al. 2008; Shendure and Ji 2008), the mean has decreased to 0.1–0.2% in our study, mainly due to technical improvements in the intervening period. Platform-dependent sequencing errors are tallied in table 3. Although the mean has become reasonably low, the distribution of errors among sites appears nonuniform. Given that 0.2–0.3% of sites have an error rate of >5% in our data set (table 3), 160 sites would falsely appear to be polymorphic at the 5% level due to sequencing errors among the 80,000 sites.

In summary, the joint applications of the Illumina GA and SOLiD sequencing platforms can effectively reduce the false-positive rates of SNP identification. Our population genetics result suggests that the Sanya population of *S. alba* indeed harbor zero polymorphisms in the surveyed gene regions.

Low Level of Polymorphism and Strong Differentiation in the Populations of *S. alba*

In addition to the Sanya sample, we also sequenced the same 71 genes of 100 individuals from the nearby Qionghai site (100 km to the north of Sanya, see fig. 1). We should note that salinity is a critical factor for the growth and distribution of *S. alba*, and these two sites differ in this respect. The salinity is 32‰ in the sampling site in Sanya and 11–22‰ in the sampling area in Qionghai. We obtained ~2,500 × coverage for the targeted regions by the GA method (table 1). At the high and medium stringency, 0 and 13 SNPs were called, respectively. The analysis above has shown that the medium stringency allowed for many errors, whereas the high stringency calls are not associated with detectable false negatives. Thus, it is possible that the Qionghai population also has zero polymorphism. For verification, we resequenced two genes (gene 18 and 27, supplementary table S1, Supplementary Material online) from all 100 individuals by the conventional Sanger method. These genes are indeed monomorphic. Therefore, there is no evidence for polymorphism in the targeted genomic regions in both the Sanya and Qionghai populations.

We next examined the distribution of genetically differentiated sites between the two populations across genes. Among the 79,000 sites, 54 have an F_{st} value of 1, and the rest are all 0. (F_{st} is a measure of population differentiation; $F_{st} = 0$ and $F_{st} = 1$ indicate complete overlap and complete nonoverlap in genetic variation, respectively.) If demographic factors are solely responsible for the reduction in genetic diversity, we would expect sites with $F_{st} = 1$ to be randomly distributed across loci. Among the 71 genes, 59 show zero divergence between the two populations, yet the two most differentiated genes have 14 and 11 high F_{st} sites, respectively, accounting for nearly half of the 54 sites (table 4).

The overall population genetic pattern is thus 1) little polymorphism in either population and 2) the concentration of divergent sites in a small number of genes. There are a number of explanations for the low polymorphism, but the skewed distribution of divergent sites may entail more complex forces. The main question is how much migration and selection are necessary. Strong migration may explain the prevalence of monomorphic genes (with $F_{st} = 0$) between populations, but it would also reduce the level of divergence (and increase the level of polymorphism). Hence, to account for the observed pattern, migration would need to be counteracted by local selection. We ask if the overall pattern can be accounted for by simple factors (ancestral population size, time of colonization, mutation rate variation, etc.) without invoking the more complex forces of migration and local selection.

We used computer simulations to determine the relationship between the number of undifferentiated genes (the *x* axis of fig. 4) and the divergence of the most differentiated gene (the *y* axis). A wide range of demographic parameters that include the ancestral population size and the timing of population separation were incorporated in our simulations (see Materials and Methods). We assumed that the two populations split from a common ancestor since time τ before present when the effective population size was N_A. Given zero polymorphism, the effective population size after the split must have been small. The effective gene flow, which elevates within-population

Table 4. Putative Functions and Numbers of Divergent Sites at 12 Genes Between Qionghai and Sanya Populations of *Sonneratia alba*.

GenBank Accession Number (Sanya)	Putative Gene Function	Length (bp)	Total Number of Differentiated Sites	Number of Synonymous Changes	Number of Nonsynonymous Changes	Sequence Divergence Between *S. alba* and *S. caseolaris*
GQ121954	Lecithine cholesterol acyltransferase-like protein	963	14	13	1	0.0168
GQ121977	Acireductone dioxygenase (iron(II)-requiring)/metal ion binding	1411	11	11	0	0.0345
GQ121981	Na^+/H^+ antiporter	1114	7	6	1	0.0310
GQ121982	ACR3; amino acid binding	1477	6	6	0	0.0324
GQ121932	SOUL-like protein	942	4	3	1	0.0271
GQ121965	NADP-dependent malate dehydrogenase	926	4	4	0	0.0319
GQ121930	Cysteine proteinase inhibitor	645	3	3	0	0.0254
GQ121957	Microtubule-associated protein 1 light chain 3	1494	1	1	0	0.0336
GQ121964	SLL1 protein	1133	1	1	0	0.0290
GQ121966	Pathogenesis-related protein PR1	1061	1	1	0	0.0168
GQ121968	Macrophage migration inhibitory factor family protein/MIF family protein	839	1	1	0	0.0189
GQ121972	Transcriptional corepressor LEUNIG	1124	1	1	0	0.0226

polymorphism, should have been close to zero as well. Hence, the two unknown parameters in this model are τ and θ ($\theta = 4N_A\mu$, where μ is mutation rate per locus). We allowed the mutation rate to vary among loci according to a Gamma distribution (with a shape parameter of α and a scale parameter of $1/\alpha$). We estimated the posterior value of α based on the distribution of gene divergence between *S. alba* and a congener *S. caseolaris* as estimated in our previous study (Zhou et al. 2007). We found the best fit for α is 5.7, with a 95% confidence interval (3.7–9.0). (The simulation results were robust when we used different α values [ranging from 1 to 5]. When we set $\alpha = 1$, which makes the gamma distribution more dispersed, we still obtained consistent results. In figures 4 and 5, we set $\alpha = 5$.).

In figure 4, we present the simulation results on the number of zero-divergence genes among 71 genes (x axis) and the highest number of sites with $F_{st} = 1$ in a gene (y axis) under the isolation–bottleneck model with 1,000,000 repetitions. The simulated number of zero-divergence genes is significantly lower than the observed number ($P < 0.001$). Most important, the highest number of sites with $F_{st} = 1$ in any gene is significantly lower than the observed value (marked with asterisk in fig. 4).

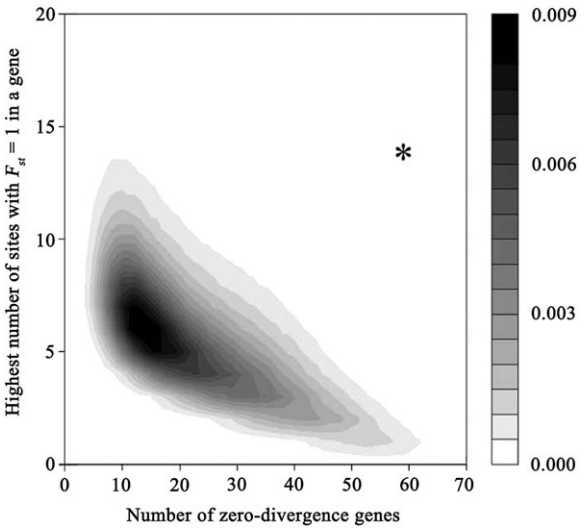

Fig. 4. Simulation results on the number of zero-divergence genes and the highest number of sites with $F_{st} = 1$ in a gene under the isolation–bottleneck model. The contour spectra indicate the probability out of 1,000,000 simulations. Asterisk indicates the observed values in the sample.

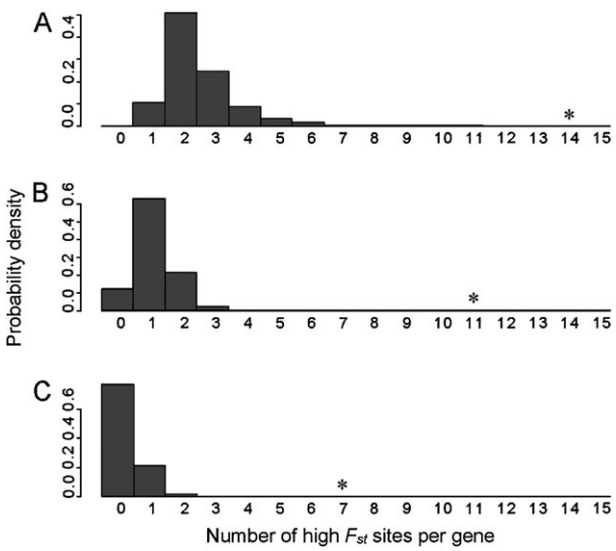

Fig. 5. Probability density of the number of high F_{st} sites in a gene. Distributions of the first (*A*), second (*B*), and third (*C*) highest number of high F_{st} sites among 71 genes are shown. Asterisks indicate the observed values in the sample.

In figure 5, we present the simulation results on the number of high F_{st} (>0.95) sites in a gene when the number of zero-divergence genes is as observed (59 of 71). The simulation was replicated for 1,000,000 times under the isolation–bottleneck model. Figure 5*A* is the number of divergent sites (i.e., sites with $F_{st} > 0.95$) in the most divergent gene in our sample against the number of zero-divergence genes. The observation marked with an asterisk is out of the range of the simulated results. To minimize the outlier effect, we also compared the values for the second and third most divergent genes between observation and simulation (fig. 5*B* and *C*). Thus, the three most divergent genes all have significantly more high F_{st} sites than expected under the isolation–bottleneck model. The disparity between the observation and simulations shown in figure 5 would be even greater if migration is incorporated into the isolation–bottleneck model. Migration would decrease the level of population divergence, and hence, the most divergent gene should have even fewer high F_{st} sites than observed. In summary, the isolation–bottleneck–migration models without invoking selection cannot account for the observation. It is possible that some very complex demographic scenarios may still be compatible with the observation. However, the possibility of divergent selection between local populations should be considered along with the more complex demography (see Discussion).

We also exclude the possibility that mutational bias might account for the observed unevenness across genes. In our previous studies, we sequenced the orthologous sequences of the 71 genes in four mangrove species (Zhou et al. 2007). There is no evidence that the most differentiated genes have higher divergence values or mutation rates (table 4).

Discussion

A main observation here is the very low level of polymorphism in both populations. *Sonneratia alba* is diploid and reproduces sexually with abundant seeds and high rate of germination (Kathiresan and Qasim 2005). Flowers of *S. alba* are pollinated by a wide variety of potential pollinators including insects, birds, and bats (Coupland 2002). Vegetative spread of *S. alba* occurs only on Malaysian rocky coastlines where seedlings do not become established (Holbrook and Putz 1982). Thus, the observed monomorphism cannot be attributed to asexual propagations.

A simple and intuitive explanation for the low polymorphism is that both populations were colonized very recently. (Variations of the same theme include severe reductions in population size in the recent past or fluctuations in local population size.) All these possibilities were frequently observed in other marginal populations (Kawecki 2008). However, the existence of a few highly divergent loci makes these simple explanations untenable.

A possible explanation for the existence of highly divergent genes among many zero-divergence loci is migration counteracted by local adaptation. The two habitats in Sanya and Qionghai are 100 km apart and differ in salinity by about 2-fold. These two populations are connected by ocean currents (fig. 1), and their divergence may have occurred in the presence of gene flow. Natural selection might favor alleles pertaining to physiochemical characters adapted to one of the two environments. At some loci, local selection may be strong enough to overcome the homogenizing effect of gene flow. The pattern echoes a similar view on speciation (Wang et al. 1997; Dieckmann and Doebeli 1999; Via 2001; Wu 2001; Wu and Ting 2004). Interestingly, some genes that are differentiated between these populations are known to be salt or stress responsive (table 4). For example, nonsynonymous changes were observed in a SOUL-like protein (GenBank accession number GQ121932) and an Na^+/H^+ antiporter (GenBank accession number GQ121981). Both are known to play a role in salt response (Apse et al. 1999; Gao et al. 2008). These are candidate genes for local adaptation.

In this migration-local adaptation model, only a fraction of the genome contributes to fitness differences. This fraction became differentiated between populations as gene flow at these loci is selected against. In contrast, gene flow contributes to the homogenization of the rest of the genome, which is not differentially adapted. We estimate that the proportion of the genomes that is differentially adapted may be 10–15% (12 of the 71 genes having at least one high F_{st} site; 7 having more than one such site).

The migration-local selection model can account for the disparity in divergence among loci. The low polymorphism in each population may be attributed to the low effective population size, and a recent bottleneck is a possible cause. Nevertheless, there may be other elements of the mangroves biology that contribute to the very low effective population size. First, mangroves live in marginal habitats for woody plants. It is conceivable that selection for the fittest genotype may be very strong in these habitats. Second, the longevity of *S. alba* may enable them to produce a vast quantity of seeds over their life time, thus increasing selective differentials among genotypes. These two elements may help to further reduce the effective population size, vis-a-vis the very large extant populations.

This study also illustrates the importance of accurately estimating the lower bound of genetic diversity in natural populations. The application of G2 sequencing methods has to distinguish true polymorphisms from sequencing errors for these methods to be widely useful in population studies. Many solutions, including reading the sequence contexts and raising the level of stringencies in SNP calling, have been proposed (Druley et al. 2009). Although the sequence context indeed affects the error pattern, the contexts have not been fully worked out yet. Neither is it practical in a genome scan to examine the sequence context of each SNP individually. Raising the stringency of SNP calls would certainly reduce errors but may also bias the estimation of polymorphisms (He et al. in press). Besides, determining the right level of stringency can sometimes be an arbitrary exercise. In this study, we propose that the joint application of the two G2 platforms provides an efficient means of filtering out sequencing errors over a wide range of stringencies.

In reality, errors come from many sources including every step of the library preparation. Hence, one expects to filter out many false positives even on one single platform, if the entire experiment is repeated. Because the 71 genes in both Sanya and Qionghai populations are monomorphic, comparing GA sequencing results from the two populations is equivalent to carrying out GA sequencing twice, provided that we ignore the fixed divergent sites. In supplementary fig. S3, Supplementary Material online, we plotted the GA SNP calls from the two populations. With raw data, we can indeed see most points falling on the edges. However, unlike the GA-SOLiD plots, there are many points that represent concordant SNP calls between the two GA runs, even though there is no polymorphism. The more meaningful result is that these concordant SNP calls tend to remain when the stringency is raised. In other words, simply increasing the depth of sequencing coverage or running the same platform twice may not reduce the false-discovery rates as effectively as using dual platforms.

Besides surveying genetic variations in nonmodel organisms, there is a growing interest in applying G2 sequencing to expand the scope of population genetic analyses of more familiar subjects, such as the 1,000 Genomes Project, the *Drosophila* Genetics Reference Panel, and the *Drosophila* Population Genomics Project. Hence, our results may have general relevance to many large-scale population genomic studies.

Supplementary Material

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

Acknowledgments

We thank Peter Raven, Richard Abbott, Roman Arguello, and David Boufford for comments and suggestions. This study is supported by grants from the National Natural Science Foundation of China (30730008, 30800060, 31071914, and 40976081), the National Basic Research Program of China (2007CB815701, 2007CB411600), National S&T Major Project of China (2009ZX08010-017B, 2009ZX08009-149B), and the Chinese Academy of Sciences (KSCX1-YW-22).

References

Apse MP, Aharon GS, Snedden WA, Blumwald E. 1999. Salt tolerance conferred by overexpression of a vacuolar Na+/H+ antiport in Arabidopsis. *Science* 285:1256–1258.

Ball MC, Pidsley SM. 1995. Growth responses to salinity in relation to distribution of two mangrove species, Sonneratia alba and S. lanceolata, in Northern Australia. *Funct Ecol*. 9:77–85.

Coupland GT. 2002. The ecological interaction between insects and mangroves in Darwin Harbour, Australia. [PhD Thesis]. [Darwin (Australia)]: Charles Darwin University.

Dieckmann U, Doebeli M. 1999. On the origin of species by sympatric speciation. *Nature* 400:354–357.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*. 36:e105.

Druley TE, Vallania FL, Wegner DJ, et al. (12 co-authors). 2009. Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods*. 6:263–265.

Duke NC. 1992. Mangrove floristics and biogeography. In: Robertson A, Alongi D, editors. Tropical mangrove ecosystems. Washington (DC): American Geophysical Union. p. 63–100.

Duke NC, Jackes BR. 1987. A systematic revision of the mangrove genus Sonneratia (Sonneratiaceae) in Australasia. *Blumea*. 32:277–302.

Fu YX. 1995. Statistical properties of segregating sites. *Theor Popul Biol*. 48:172–197.

Gao F, Zhou YJ, Huang LY, He DC, Zhang GF. 2008. Proteomic analysis of long-term salinity stress-responsive proteins in Thellungiella halophila leaves. *Chin Sci Bull*. 53:3530–3537.

Glemin S, Ronfort J, Bataillon T. 2003. Patterns of inbreeding depression and architecture of the load in subdivided populations. *Genetics* 165:2193–2212.

Harismendy O, Ng PC, Strausberg RL, et al. (11 co-authors). 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*. 10:R32.

He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu CI, Shi S. Forthcoming. Two evolutionary histories in the genome of rice: the roles of domestication genes. *PLoS Genetics*.

Holbrook NM, Putz FE. 1982. Vegetative seaward expansion of *Sonneratia alba* trees in a Malaysian mangrove forest. *Malays For*. 45:278–281.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Kathiresan K, Bingham BL. 2001. Biology of mangroves and mangrove ecosystems. *Adv Mar Biol*. 40:81–251.

Kathiresan K, Qasim SZ. 2005. Biodiversity of mangrove ecosystems. New Delhi (India): Hindustan Publishing Corporation.

Kawecki JT. 2008. Adaptation to marginal habitats. *Annu Rev Ecol Evol Syst*. 39:321–342.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 18:1851–1858.

Manier MK, Arnold SJ. 2005. Population genetic analysis identifies source-sink dynamics for two sympatric garter snake species (Thamnophis elegans and Thamnophis sirtalis). *Mol Ecol*. 14:3965–3976.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 24:133–141.

Margulies M, Egholm M, Altman WE, et al. (63 co-authors). 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.

Nieminen M, Singer M, Fortelius W, Schops K, Hanski I. 2001. Experimental confirmation that inbreeding depression increases extinction risk in butterfly populations. *Am Nat*. 157:237.

Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F. 2008. Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9:431.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol*. 26:1135–1145.

Stewart CN Jr, Via LE. 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *Biotechniques* 14:748–750.

Tavare S, Balding DJ, Griffiths RC, Donnelly P. 1997. Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.

Via S. 2001. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol Evol*. 16:381–390.

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.

Weiss G, von Haeseler A. 1998. Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.

Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7:275.

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

Wu CI, Ting CT. 2004. Genes and speciation. *Nat Rev Genet*. 5:114–122.

Zhou R, Zeng K, Wu W, Chen X, Yang Z, Shi S, Wu CI. 2007. Population genetics of speciation in nonmodel organisms: I. Ancestral polymorphism in mangroves. *Mol Biol Evol*. 24:2746–2754.