Ecological and conservation genomics for the tropical tree species *Metrosideros polymorpha* and *Shorea leprosula*

by
AYAKO IZUNO

A dissertation submitted to
Kyoto University
for the degree of

Doctor of Agriculture

accepted on the recommendation of
Professor Yuji Isagi
Professor Kanehiro Kitayama
Professor Mamoru Kanzaki

March 2016

Contents

Abstract

Tree species, which have developed adaptive traits to survive in diverse environmental conditions, provide excellent opportunities to address the fundamental question of how organisms have adapted and expanded their distributions. The scientific findings regarding adaptive mechanisms of tree species can be applied to predict the responses of forests after climate and/or anthropogenic activities, promote efficiency in the breeding of timber varieties, and determine priorities in forest conservation. Rapid progress in high-throughput sequencing technologies has provided enormous genetic information for various organisms. The genomics is expected to reveal the mechanisms of adaptation and speciation, and thus greatly advance evolutional and conservation biology of tree species. In this thesis, I aimed to: 1) reveal the genetic bases underlying environmental adaptation of tree species at the genome level and; 2) derive the genetic implications for sustainable conservation of tropical forests.

In Chapters 2 and 3, I investigated the genomic divergence in *Metrosideros polymorpha* (Myrtaceae), which occupies a wide range of ecological habitats and shows adaptive radiation within a species in Hawaiian Islands. The biological functions of genetic variations observed within this species could give us valuable insights into the drastic evolution found in a single lineage.

In Chapter 2, I sequenced and assembled *de novo* genome sequences of one *M. polymorpha* tree to provide the basic genomic parameters about this species and to develop our knowledge about ecological divergences. The assembly yielded 304 Mb genome sequences, half of which were covered by 19 scaffolds with >5 Mb, and contained 30 K protein-coding genes. Demographic history inferred from the genome-wide heterozygosity indicated that this species experienced the dramatic rise and fall in the effective population size possibly according to the past geographic or climatic changes in Hawaiian Islands. The present *M. polymorpha* genome assembly represents a high-quality genome resource useful for future functional analyses of genetic variations or comparative genomics among intra- and inter-species.

In Chapter 3, a *de novo* draft genome sequence provided in Chapter 2 and 2,247 single nucleotide polymorphism (SNP) markers were used to reveal the population genetic structure of nine populations across five elevations and two ages of substrates on Mauna Loa, the island of Hawaii. The nine populations were genetically differentiated according to elevation as well as age of lava but largely admixed, particularly in the lower elevations. A genome scan for the 2,247 SNPs revealed that a small fraction of the genome (35 SNPs on 26 scaffolds; 1.56%) was likely under divergent selection, and alleles on these non-neutral SNPs were fixed in one or more populations. Generalized mixed modeling for pairwise population differentiation according to geographic and environmental variables revealed that population differentiation in most of the genome followed the isolation-by-distance model, whereas divergence at non-neutral SNPs followed the isolation-by-environment model.

Accordingly, the current study reveals the genomic mosaic of *M. polymorpha* comprising contrasting divergence patterns. Although the genome was largely mixed among populations, a small fraction of the genome appears to be subject to environmental selection and responsible for the dramatic divergence in phenotype and adaptation to a wide range of environments.

In Chapters 4 and 5, I focused on the *Shorea leprosula* (Dipterocarpaceae) plantation managed by a private-sector forestry company in Central Kalimantan. Because the family Dipterocarpaceae constitutes the core of biodiversity and is faced with heavy exploitation in the Southeast Asian tropical rainforests, planting of native dipterocarp trees is a valuable policy. However, the genetic concerns of the planted trees and species richness of phyllosphere fungi associated with them have not been evaluated.

In Chapter 4, the genetic diversity of *S. leprosula* and *S. parvifolia* in plantations and those in natural populations were compared using microsatellite markers. Genetic diversity in the planted populations was as high as that in the natural populations. No clear genetic differences between each planted population and the natural forest populations were found. The genetic variation present in planted *S. leprosula* and *S. parvifolia* populations did not appear to deteriorate in the planting system adopted in Indonesia, known as Tebang Pilih Tanam Jalur. These results indicate that the current plantation method practiced in the region is suitable for maintaining the original genetic composition and achieving sustainable use of tropical rainforests.

In Chapter 5, I sought to estimate the species diversity and community structure of phyllosphere fungi in the *S. leprosula* plantation. I conducted a massively parallel amplicon sequencing analysis of fungi collected from the leaves of *S. leprosula.* Phyllosphere fungal compositions and spatial variability were investigated for 31 *S. leprosula* trees across four plots within a plantation stand. In total, 488 fungal operational taxonomic units (OTUs) were recognized in 153,194 ribosomal internal transcribed spacer reads at 95 % OTU identity level. Rare OTUs accounted for the majority of fungal diversity detected in the study site; 200 OTUs (41 %) comprised fewer than 10 reads and 465 OTUs (95 %) were found in fewer than half of the leaf samples. Fungal OTU compositions of *S. leprosula* trees were differentiated within a narrow area of the plantation and even between plots that were separated by 15 m. These findings indicate that highly diverse fungal OTUs form spatially structured communities even within a tropical plantation stand of single tree species.

Overall, this thesis has updated our knowledge about the ecological adaptations of *M. polymorpha* and clearly demonstrated the effectiveness and possibility of genome-wide analyses in the field of ecology and evolution. Further, the appropriateness of the dipterocarp planting in Central Kalimantan was evaluated in terms of genetic and species diversity harbored in the plantation stands. The field of ecological and conservation genetics for wild organisms, including forest trees, is becoming an interesting platforms to uncover idiosyncratic ecological phenomena by taking the advantage of massive genomic data. It is expected that we will be able to discuss a more specific evolutional process regarding fitness-related traits and predicting the impact of climate change or artificial management on the fitness and longevity of target populations.

1

General introduction

1.1. Ecological and conservation genetics for tree species

Tree species adapt to various climates on most continents, and constitute forest ecosystems, playing indispensable roles within the global ecosystem and economy. Forests cover 28% of the global land area, representing 82% of the terrestrial biomass (Roy et al. 2001). Trees provide habitat for a large diversity of organisms, and >50% of the global biodiversity is harbored in forests (Neale and Kremer 2011). The raw materials and genetic resources supplied by forests are widely utilized by man for wood or paper products, energy, and foods. Tree species, which have developed adaptive traits to survive in diverse environmental conditions, provide excellent opportunities to address the fundamental question of how organisms have adapted and expanded their distributions. The scientific findings regarding adaptive mechanisms of tree species can be applied to predict the responses of forests after climate and/or anthropogenic activities, promote efficiency in the breeding of timber varieties, and determine priorities in forest conservation. In the current scenario of prevalence of more serious global changes than ever, evolutionary and conservational studies on tree species are very likely key scientific challenges.

The genome keeps tracks of evolutionary processes that a species has experienced as nucleotide variations. Natural and artificial selection can alter the frequency of specific alleles that are strongly related to adaptive or useful traits (e.g., Colosimo et al. 2005; Axelsson et al. 2013). Changes in the effective population size due to habitat shifts caused by climate change (Buckley et al. 2012; Pauls et al. 2013; Forcada and Hoffman 2014), anthropogenic habitat fragmentation (Finger et al. 2012), or other catastrophic events including damage caused by pest species, forest fires, or hunting, induce changes in allele frequencies within populations. Introgression or artificial mating additionally alters the population genetic structure or adaptive capacity of a

population (Ryan et al. 2009; Fitzpatrick et al. 2010; Heliconius Genome Consortium 2012). The current genetic characteristics observed in a species reveals the genetic bases of ecological successes (i. e., ecological genetics). Based on the current genetic variations, which critically affect the capacity of evolution of species (Frankham et al. 2010), we can derive appropriate conservation strategies for the future (i. e., conservation genetics).

1.2. A breakthrough in genetics

While non-neutral genetic variations indicate key genes subject to selection, i.e., major genetic mechanisms underlying adaptations (Jones et al. 2012), neutral genetic variations reveal population genetic structures or temporal changes in population size (Lexer et al. 2014; Liu et al. 2014; Trucchi et al. 2014; de Kort et al. 2015). The drivers and extent of genetic differentiations are variable so that neutral and non-neutral loci adjoin across a genome (i. e., genomic mosaics; Nosil et al. 2009).

For model organisms such as a human, mice, zebrafish, *Arabidopsis thaliana*, and rice, we have been able to access genome-wide divergences and gene function. In ecological genetics in which most target species are non-model organisms growing in the field, a handful of neutral genetic variations have been used as genetic markers to assess genetic characteristics. However, rapid progress in high-throughput sequencing technologies has provided enormous genetic information for various organisms at a lower cost than previous methods. The decrease in sequencing costs has occurred at a much quicker rate than expected according to Moore's law (Fig. 1-1).

An advantage of these new technologies is the ease in generating thousands of genetic markers with a greater number of sequence polymorphisms across a genome (Baird et al. 2008; Davey et al. 2011). Genome-wide markers have successfully revealed spatial genetic structures, phylogenies, and linkage maps that could not have been obtained using a handful of neutral markers (e.g., Emerson et al. 2010; Hohenlohe et al. 2012; Wagner et al. 2012; Catchen et al. 2013). These markers have additionally detected precise or candidate genetic regions subject to selection using genome scanning or comparative genomic analysis (e.g., Hohenlohe et al. 2010; Therkildsen et al. 2013; Baute et al. 2015). A breakthrough in genomics has further facilitated genome sequencing, even in various non-model organisms (Ellegren 2014). Whole-genome sequences have allowed the identification of locations on chromosomes and the biological functions of genome-wide markers (Hohenlohe et al. 2012; Orsini et al. 2012; Baute et al. 2015). Population genomic studies using genome-wide markers and genome sequences are expected to reveal the mechanisms of adaptation and speciation, and thus greatly advance evolutional and conservation biology.

1.3. Metabarcoding of fungal species

DNA barcoding, which can identify a taxon via sequencing standardized DNA regions, is a powerful tool for species identification. Metabarcoding is an extensive DNA barcoding to simultaneously identify multiple species contained in an environmental

sample (Valentini et al. 2009). Ecologists have applied meta-barcoding to reveal the species richness and composition in various environmental samples including soil (Peay et al. 2010; Toju et al. 2013), water (DeLong et al. 2006; Brown et al. 2009), animal feces (Taberlet et al. 2007; Ando et al. 2013), plant tissues (Jumpponen and Jones 2009; Redford et al. 2010; Bulgarelli et al. 2012), and permafrost sediment (Willerslev et al. 2003; Willerslev et al. 2007). By selecting DNA regions to sequence, the taxa and the resolutions of community data can be optimized (Valentini et al. 2009; Toju et al. 2012). The methodology and efficiency of metabarcoding have been rapidly advanced by the introduction of high-throughput sequencers and required bioinformatics tools or DNA databases have been accordingly increased (Edgar et al. 2011; Li et al. 2012; Koljalg et al. 2013).

Fungi, which play key roles in the ecosystems and harbor valuable organic matter, have typically been described based on evidence from culture mediums; thus, species that cannot be cultured or collected are difficult to identify. In other words, unrecognized species, some of which may provide useful bioresources, likely remain within the fungi kingdom. Metabarcoding can be utilized to provide a comprehensively inventory of fungal species in an environmental sample, including visible or uncultivable species. Metabarcoding of fungi has effectively revealed the community structures or the interactions between fungi and host plants that have never been addressed using the conventional methods (Toju et al. 2013).

1.4. Study systems

1.4.1. Ecological genomics of tree populations on oceanic islands

As the so-called "living laboratories of evolution," islands are ideal ecosystems for ecological and evolutionary studies (Whittaker and Fernandez-Palacios 2007). Because of biogeographic isolation, most of the available ecological niches are released, and only a limited number of organisms are able to access these isolated locations. Consequently, a single lineage can acquire a wide range of ecological niches without competitors or enemies that appear on continents. Such an adaptive diversification is well known by the adaptive radiations of Darwin's finches on the Galapagos Islands (Grant and Grant 2002) or Hawaiian honeycreepers (Lerner et al. 2011). The small land area is an additional feature affecting evolution on islands. Small population sizes enhance genetic drift, by which allelic compositions can accidentally alter, even under conditions of random mating in a population, thereby facilitating more rapid evolution on islands than on continents. The obvious and rapid evolution observed on islands is often unique, resulting in high endemism. These characteristics of insular evolution provide advantages for studying adaptive processes of a single lineage.

The Hawaiian archipelago, which is located approximately 4,000 km from the nearest continent, is the most typical of the oceanic islands. This archipelago possesses drastic environmental gradients: elevations rise to >4,000 m above sea level, and the mean annual temperature and precipitation range from 5 to 25°C and from 250 to 11,000 mm, respectively (Price 2004). Because of the continuous activity of volcanoes,

the soil ages within the archipelago also vary (Vitousek 1992; Kitayama & Mueller-Dombois 1995). As a consequence of the nature of island ecosystems, a limited number of species comprise the flora of the Hawaiian Islands, resulting in disharmonic biomes with low diversity (Carlquist 1980). A species occupying a wide range of ecological niches in the Hawaiian Islands is ideal for study of intraspecies genomic diversifications, which possess evidence and trajectories of adaptive evolution of a single lineage.

Metrosideros polymorpha Gaud. (Myrtaceae), a tree species endemic to the Hawaiian Islands, is one such species. It grows from sea level to alpine timberline and comprises early-successional vegetation on new lava flows as well as climax forests on mature substrates (Fig. 1-2a; Stemmermann 1983; Joel et al. 1994; Kitayama & Mueller-Dombois 1995; Cordell et al. 1998). The phenotype, particularly leaf size, the amount of trichomes on the leaf surfaces, and tree height are extremely variable across populations (Fig. 1-2b; Stemmermann 1983; Vitousek 1992; Kitayama et al. 1997). The leaf traits represented by leaf area, leaf mass per area, and trichome mass per area were highly predictable as a function of temperature, precipitation and lava age (Tsujii et al. 2015). Additionally, a similar magnitude of the variation in leaf traits was also observed in the common garden (Cordell et al. 1998; Martin et al. 2007; Tsujii et al. 2015), suggesting that the phenotypic variations observed in the field are based on genetic differentiations. Physiological functions were different in accordance with leaf morphologies (Hoof et al. 2008), suggesting that the variations in leaf traits of this species could be adaptive consequences.

The previous population genetics studies targeting *M. polymorpha* revealed the significant genetic differentiation among populations in accordance with leaf morphologies or ecological niches (Harbaugh et al. 2009; DeBoer & Stacy 2013; Stacy et al. 2014). However, the small number of genetic markers used in previous studies captured insufficient genetic information per individual, and consequently three major knowledge gaps remain: (1) the genetic bases underlying environmental adaptation; (2) the population genetic structures at the individual level, and; (3) the demographic history. A large number of genetic markers extracted throughout the genome are promising to address these questions.

1.4.2. Conservation genetics of tree populations in tropical rain forests

Southeast Asia is recognized as one of the major biodiversity hotspots on Earth (Myers et al. 2000; Fisher et al. 2011b). This region harbors one of the highest species richness and endemism on Earth (Sodhi et al. 2004). In Sundaland, which covers the island of Borneo, Sumatra, Java and the southern Malay peninsula, 25,000 vascular plant species including 15,000 endemic species (5.0% of the 300,000 global vascular plant species) and 1,800 vertebrate species including 701 endemic species (2.6% of the 27,298 global vertebrate species) are reported (Myers et al. 2000). This high biodiversity faces greater threats than any other tropical region (Laurance 2007; Bradshaw et al. 2009), primarily because of unsustainable logging and burning of forests for agricultural land (Kettle et al. 2010). To prevent further loss of biodiversity and resources, the establishment of appropriate and sustainable forest management is an urgent requirement.

Although primary forests are irreplaceable ecosystems (Gibson et al. 2011), degraded forests, such as selectively logged forests, also provide habitats for numerous bird and insect species (Edwards et al. 2011), most of which are lost once forests are converted into intensive monoculture plantations, such as oil palm plantations (Koh and Welcome 2008). These degraded forests additionally contribute to carbon storage (Berry et al. 2010). In Southeast Asia, 70% of the original primary forests have already been lost and secondary forests cover more land area than primary forests (Wilcove et al. 2013), indicating that the success of the conservation of biodiversity and carbon stocks in Southeast Asia largely depends on the management of the secondary forests. Nonetheless, the secondary forests cover has decreased at even a more rapid speed than that of primary forests; the areas of secondary and primary forests have reduced at rate of 0.67% and 0.35% per year, respectively (FAO 2010). Therefore, the sustainable use of secondary forests should be urgently established. Enrichment planting, which introduces timber species into secondary forests, is one of the management strategies adopted in some tropical regions (Paquette et al. 2009). The natural regeneration on existing understory contributes to maintaining native biota, including local plants and associated insects, animals, birds, and microorganisms. This also increases the value of secondary forests as resources. Thus, this strategy could be effective for preventing further deforestation and loss of biodiversity (Dalle et al. 2006).

Several genetic aspects should be taken into account for the sustainability of the forest stands when establishing plantations, including enrichment plantings (Fig. 1-3). One of the concerns is species selection. Native tree species are recognized to adapt to the local biotic and abiotic environment; therefore, they are expected to establish over a long-term and contribute to the maintenance of associated native biodiversity and ecosystem services at a landscape scale (Tang et al. 2007), and should therefore be favored over exotic species. However, in Southeast Asia, non-native tree species including *Acacia*, *Pinus*, and *Eucalyptus* are often used within plantations (Harwood and Nambiar 2014). The sources of planting materials are also important. Ideally, the sources should originate from a forest as near as possible to the target plantation stands. Generally, local adaptations or genetic differentiation occur within a tree species (González-Martínez et al. 2006). By planting appropriate materials that are closely related to the native populations and adapted to the plantation sites, we can increase the survivorship and growth of the planted trees (Thomas et al. 2014) and prevent genetic pollution of surrounding natural forests through gene flow (Aitken et al. 2008; Millar et al. 2012). Moreover, the planted materials should harbor genetic diversity within stands as populations with low genetic diversity may cause inbreeding depression, which reduces fitness and decrease population size (Reed and Frankham 2003). Genetic bottlenecks that lower the original genetic diversity in planted populations should be avoided by choosing appropriate planting practices (Kettle et al. 2008; Li et al. 2012). All of these genetic characteristics could critically influence the productivity and ecosystem functioning of plantation stands.

Recently, an intensive enrichment planting of dipterocarp trees has been adopted in Indonesia (see Chapter 4 for details). Because the family Dipterocarpaceae constitutes the core of biodiversity and is faced with heavy exploitation in the Southeast Asian tropical rainforests, enrichment planting of native dipterocarp trees is

a valuable policy. Although this policy has already adopted optimized planting techniques (Ministry of Forestry 2005), the genetic concerns of the planted trees have not been evaluated. Besides the genetic characteristics of the planted dipterocarp trees, the associating microbes can also affect the sustainability or ecological value of the plantations. Because microbes have significant harmful and beneficial relationships between plant species, the spatial structures of microbe communities in stands would be informative for predicting the expansion of pests and improve plantation management. In addition, the comprehensive investigation of fungal species associating with native dipterocarp trees would contribute to the construction of an inventory of fungal species, which is currently extremely insufficient for the Southeast Asian tropical rainforests (Hawksworth 2012).

1.5. Objectives and outline of this thesis

In this thesis, I aimed to: (1) reveal the genetic bases underlying environmental adaptation of tree species at the genome level and; (2) derive the genetic implications for sustainable conservation of tropical forests.

In Chapter 2, whole genome of *M. polymorpha* was sequenced and assembled to provide a permanent genome resource that can be widely used in ecological and comparative genomic studies. Based on the whole genome sequence, a genome-wide heterozygosity, which reflect the extent of gene flow of a species, was measured and the historical changes in effective population sizes were inferred.

In Chapter 3, genomic differentiations and their drivers were investigated in *M. polymorpha* populations occupying a wide range of ecological niches. Thousands of polymorphic DNA markers and a draft genome sequence provided in Chapter 2 were used to reveal the population genetic structure at an individual level and to detect outlier loci, which show non-neutral differentiation patterns and may have key roles in adaptation to various environments. Additionally, the effect of geography and environment on population differentiation at outlier and genome-wide loci was evaluated.

In Chapter 4, the genetic diversity of planted dipterocarp trees in Central Kalimantan, Indonesia was compared with that of the natural populations in the same region. Here the question whether the current method for planting dipterocarp trees in the region is adequate to maintain genetic diversity and to avoid genetic differentiation from the surrounding natural forest was addressed.

In Chapter 5, using massively parallel high throughput sequencing the diversity and spatial variability of phyllosphere fungal communities in the dipterocarp plantations were characterized.

Finally, in Chapter 6, I have discussed the environmental adaptations of *M. polymorpha* and the conservation of dipterocarp tree populations in the Southeast Asian tropical rainforests in light of population genetics. Moreover, I provided perspectives for ecological and conservation genomics.

Figures

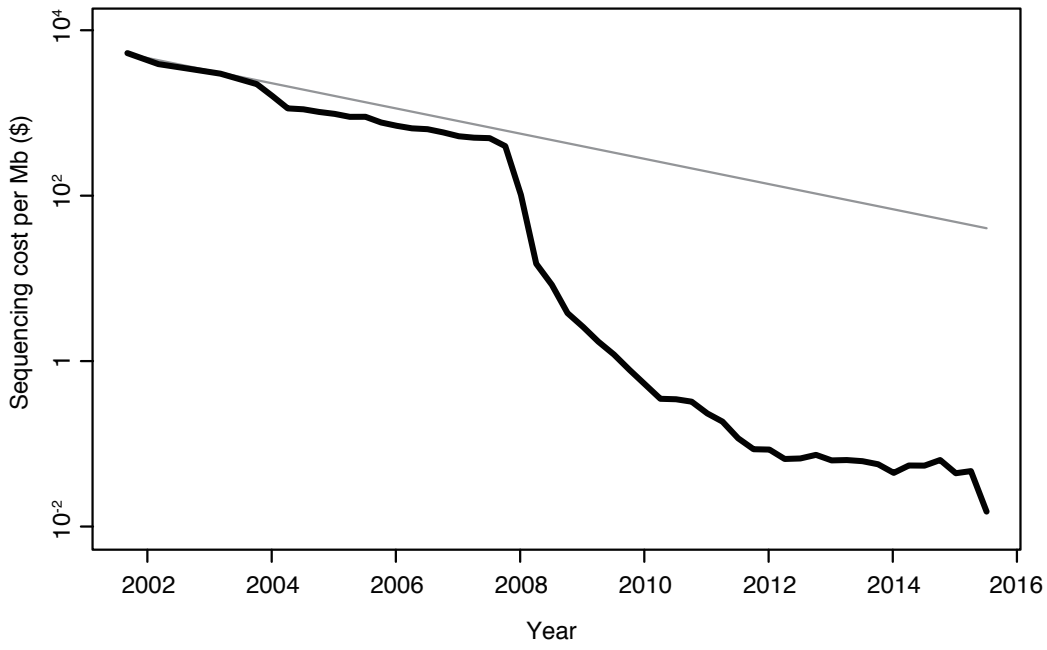

Fig. 1-1

Rapid decline in sequencing costs from 2001 to 2015. Note the y-axis is on a logarithmic scale. The grey line indicates the hypothetical data reflecting Moore's Law, which describes the doubling of central processing unit (CPU) power of computers every approximately 2 years. The original data were downloaded from the National Human Genome Research Institute (NHGRI) (http://www.genome.gov/sequencingcosts/).

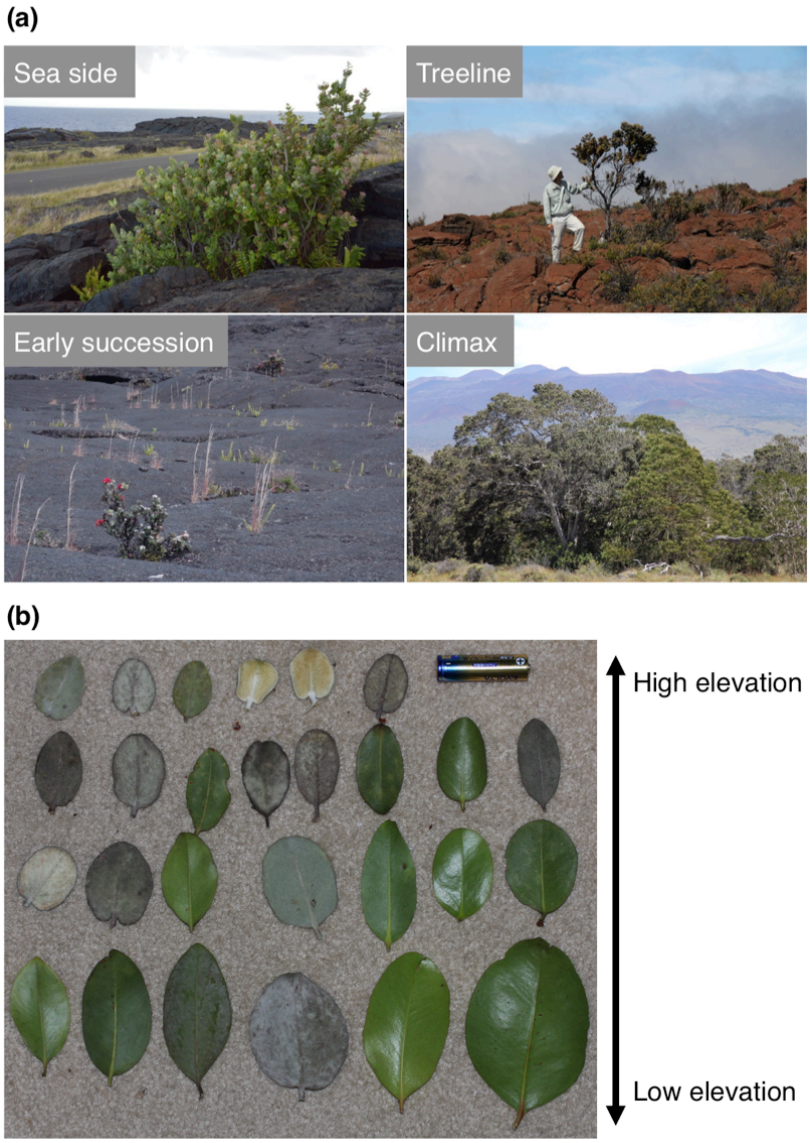

Fig. 1-2
(a) Various ecological habitats of *Metrosideros polymorpha*. (b) Remarkable variations in leaf morphology such as size, shape and color (reflecting presence or absence of trichome) of *M. polymorpha*. The abaxial side is shown for all leaves. An AA battery is included as a scale (June 2013, the island of Hawaii).

Fig. 1-3
Enrichment planting of dipterocarp trees (*Shorea leprosula*) conducted in Indonesia (January 2011, Central Kalimantan).

Fig. 1-4
Diverse fungi observed on the leaves of *Shorea leprosula* (January 2011, Central Kalimantan).

2

Genome sequencing of *Metrosideros polymorpha*

2.1. Introduction

Whole genome sequences of various plant species have provided marked information to biological research (Michael and Jackson 2013). By identifying every gene and its biological function harbored in a species, evolutionary events at the genome scale have been revealed, such as whole genome duplications (D'Hont et al. 2014; The Tomato Genome Consortium 2012; Kim et al. 2014; Akama et al. 2014), diversifications in genome size (Vu et al. 2015), diversification of body plan (Rensing et al. 2008; Banks et al. 2011), and changes in reproductive systems (Slotte et al. 2013). It is also known that genome-wide polymorphism, even in an individual genome, can demonstrate population genetic parameters. Genome-wide heterozygosity (the frequency of heterozygous nucleotide sites out of all nucleotide sites) reflects the degree of outcrossing, which is one of the most important indices in population/conservation genetics (Hartl and Clark 2007). By comparing the time to most recent common ancestor (TMRCA) between two alleles on heterozygous nucleotide sites throughout individual genome sequences, the changes in effective population size in the past can be inferred (Li and Durbin 2011). This fundamental and genome-wide information can be obtained even in non-model organisms using high-throughput sequencing technologies (Ellegren 2014).

Genome resources of *M. polymorpha* can address questions regarding how and when this species acquired various ecological niches. For example, genome-wide association studies (McKown et al. 2014; Slavov et al. 2014) or genome scanning (Foll and Gaggiotti 2008) are useful in detecting genes relating to phenotypic variations and assuming key roles in environmental adaptations. Although these approaches can be performed without reference genome sequences, a reference genome is required to reveal the biological functions of the adaptive loci (Hohenlohe et al. 2012). Besides the

adaptive loci, neutral variants across a genome are also essential means to address these questions. Population genomics with genome-wide variants, even if the variants are detected within individual genome sequences, can estimate the genomic divergence within a species in time and space as well as the required time scale and environmental factors for the ecological divergence (Brandvain et al. 2014; Halley et al. 2014; Wu et al. 2014; Fitak et al. 2015; Meyer et al. 2015). Genome sequences of this species can also be used for studies targeting other *Metrosideros* species, which are broadly distributed around Pacific Ocean (Wright et al. 2000). Phylogeography of the genus *Metrosideros* or ecological genomics of other related species could provide insights not only about the origin of *M. polymorpha* but also the speciation or adaptive dynamics of this genus.

In the present study, a draft genome of *M. polymorpha* as a genome resource widely used in intra- and interspecies genomic studies to facilitate our knowledge about genetic mechanisms of ecological divergences is reported. D*e novo* genome sequences of a wild *M. polymorpha* plant individual were assembled, genome-wide heterozygosity was measured, and historical changes in effective population sizes of this species were estimated.

2.2. Materials and methods

2.2.1. Plant material and estimation of genome size

Mature leaves were collected from one individual tree growing in the eastern frank of Mauna Loa, the island of Hawaii (700 m above sea level on a 3000 year-old lava flow). The chromosome number of this species is $2n$ = 22 (Carr 1978). Before genome sequencing, the genome size was estimated by flow cytometry. Approximately 25 mm^2 of fresh leaves were chopped with a steel razor blade in 300 μl of extraction buffer (Partec, Görlitz, Germany) to extract intact nuclei. After filtering through a CellTrics disposable filter (30 μm; Partec), more than three volumes of staining solution (Partec) was added to the solution and then incubated for 10 min. *Eucalyptus globulus* (1C = 0.62 pg; Praça et al. 2009; http://www.rbgkew.org.uk/cval/homepage.html) was used as a standard. *Eucalyptus* was the closest genus to *M. polymorpha* among genera whose genomes have been completely sequenced (Myburg et al. 2014). A total of three mixtures were prepared, each of which contained equal volumes of *M. polymorpha* and standard *E. globulus* samples, and were independently analyzed using CyStain UV precise P (Partec). The absolute DNA amount of *M. polymorpha* was calculated based on the average relative fluorescence intensity versus *E. globulus*.

2.2.2. Library construction and sequencing

Short read libraries (200-b and 500-b inserts) and mate pair libraries (3-kb, 4-kb, 5-kb, 7-kb, 10-kb, 12-kb, 15-kb, 20 -kb and 24 -kb inserts) were constructed.

For the short read library, DNA was extracted with the CTAB method (Murray and Thompson 1980); briefly, dried leaf tissues (20 mg) were crushed into a powder

with TissueLyser II (QIAGEN, Netherlands); then, 800 μl of 2 × CTAB buffer, which contained 2% CTAB, 20-mM EDTA, 100-mM Tris-HCl (pH 8.0), 2% PVP, 5-mM L-ascorbic acid, 1.42-M NaCl, 4-mM Sodium diethyldithiocarbamate trihydrate and 0.5% 2-mercaptoethanol, was added. The solution was incubated for 15 min at 65°C. After one volume of chloroform/isoamyl alcohol (CIA) was added, the solution was mixed for 5 min and then centrifuged for 3 min at 13,000 × g. The aqueous phase was transferred into a new tube and the nucleic acids were precipitated with 3/4 volume of 100% isopropyl alcohol and washed with cold 70% ethanol. After drying, the pellet was resuspended into TE.

Two short insertion libraries (200-b and 500-b inserts) were constructed using the NEB Next Ultra DNA Library Prep for Illumina (New England Biolabs, MA, USA). Briefly, for each library, 150 ng of DNA was sonicated with the Covaris (Covaris, Inc., MA, USA) using settings specific to the desired fragment size. The fragmented DNA samples were end-repaired and then ligated to adapters with the code (5'—AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT CCGATCT—3'). After the DNA libraries were size selected using Agencourt AMPure XP (Beckman Coulter Inc., CA, USA), fragments containing adapters on both ends were selectively enriched using the polymerase chain reaction (PCR). The quantity and quality of the enriched libraries were validated using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific K.K., MA, USA) with Qubit 1.0 Fluorometer (Thermo Fisher Scientific K.K., MA, USA) and the Caliper GX LabChip GX (Caliper Life Sciences, Inc., MA, USA). Each short insertion library was sequenced in one lane of HiSeq2500 (Illumina) with 100-bp paired-end sequencing technology.

For the mate pair library, high molecular weight DNA was prepared according to the CTAB method described by Murray and Thompson (1980) with minor modifications. Frozen leaf tissues (4 g) were crushed into a powder with Shake Master NEO (BMS). Furthermore, 20 ml of prewarmed 2 × CTAB buffer, which contained 2% CTAB, 1.4-M NaCl, 100-mM Tris-HCl (pH 8.0), 20-mM EDTA, 1% PVP, and 0.5% 2-mercaptoethanol, was added to the tissues, and the solution was incubated for 30 min at 65°C. After one volume of CIA was added, the solution was slowly mixed for 10 min and then centrifuged for 30 min at 9,100 ×g. The aqueous phase was transferred into a new tube and 1/10 volume of 10% CTAB was added. After one volume of CIA was added, the solution was mixed slowly for 10 min and then centrifuged for 30 min at 9,100 xg. The aqueous phase was transferred into a new tube and 3 volume of CTAB precipitation buffer, which contained 1% CTAB, 50 mM Tris-HCl (pH 8.0) and 10 mM EDTA, was added. After 30 min of centrifugation at 9,100 xg, the supernatant was discarded. The precipitate dissolved into 2.5 ml of 1M NaCl-TE was precipitated with 5 ml of 100% ethanol and washed with 2.5 ml of 70% ethanol. The pellet was air-dried and resuspended into TE. The nucleic acids were purified with RNaseA (QIAGEN, Netherlands) and QIAGEN Genomic-tip 20/G (QIAGEN, Netherlands).

Mate pair libraries were constructed mostly according to the instruction of Nextera Mate Pair Sample Preparation Kit Illumina, CA, USA) in combination with KAPA Hyper Prep kit (KAPA Biosystems, MA, USA). The high molecular weight DNA was fragmented with Nextera transposons (Nextera Mate Pair Sample Preparation Kit, Illumina, CA, USA), and then, the resulting fragments were size

selected at nine levels (3–4 kb, 4–5 kb, 5–7.5 kb, 7.5–10 kb, 10–12.5 kb, 12.5–15 kb, 15–20 kb, 20–24.8 kb and 24.8–38.4 kb). The DNA fragments in each library were circularized under 30°C for 17 hours. The circularized fragments were sheared using Covaris S2 (Covaris, Inc., MA, USA) with T6 tubes (70 sec, 6°C), and then the fragments that contained the biotinylated junction adapter were purified with Dynabeads M-280 streptavidin magnetic beads (Thermo Fisher Scientific K.K., MA, USA). The mate pair fragments were end-repaired, A-tailed, and ligated to illumina index adapters (Table 1). The concentration of each index adapter was adjusted in accordance with the molarity of each mate pair fragment library (Table 1). The libraries were quantified with the KAPA Library Quantification Kit (KAPA Biosystems, MA, USA) and enriched with PCR using a KAPA Hyper Prep Kit (KAPA Biosystems, MA, USA), for which the number of PCR cycles was determined according to the determined concentration so that the final concentration would be more than 2 nM. The quantity and size distribution of the libraries were validated with quantitative PCR using KAPA Library Quantification Kit (KAPA Biosystems, MA, USA) and TapeStation 2200 using High Sensitivity D1000 Kit (Agilent Technologies, CA, USA). The barcoded nine mate pair read libraries were pooled and sequenced in one lane of HiSeq2500 (Illumina) with 125-bp paired-end sequencing technology.

2.2.3. *De novo* assembly

2.2.3.1. 1st draft genome

Before conducting the *de novo* assembly, I trimmed low quality bases and adapter sequences from the raw sequence data. For the two short insertion libraries, raw reads were mapped against the PhiX genome sequence (ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/PhiX/Illumina/RTA/PhiX_Illumina_RTA.tar.gz; accessed September 16th, 2015) using Bowtie2 (ver. 2.2.1; Langmead and Salzberg 2012) and the unmapped sequences were used for the subsequent analysis. The adapter sequences and the bases with low quality were trimmed using Trimmomatic (ver. 0.30; Bolger et al. 2014).

The frequency of unique k-mer (k = 21–91 in 10-bp increments) was counted in the quality-filtered reads of the two short insertion libraries using KmerGenie (Chikhi and Medvedev 2013). KmerGenie estimates the frequency distributions of genomic k-mers, which were distinguished from k-mers derived by sequencing errors. Therefore, the k-mer size that found the most abundant unique genomic k-mer fragments was determined as the optimal k-mer length required for de Bruijn graph assembly. The total number of unique genomic k-mer fragments at the optimal k was regarded as the estimated genome size. KmerGenie algorithm further divides the frequency distributions of genomic k-mers into those of homozygous and heterozygous k-mers; therefore, the heterozygosity of this species was estimated with the peak patterns of the frequency distributions.

De novo assembly was conducted using Platanus (ver. 1.2.1; Kajitani *et al*. 2014) with a k-mer of 51, which was inferred by KmerGenie (Chikhi & Medvedev 2013). Gene predictions were conducted on the assembled scaffolds using AUGUSTUS (ver

2.7; Stanke *et al*. 2004) with *Arabidopsis thaliana* genome (TAIR10) as a reference.

2.2.3.2. 2nd draft genome

In the 2nd assembly, for the nine mate pair libraries, the sequences derived from Nextera transposase, adapter sequences, and bases with low quality were trimmed using Trimmomatic (ver. 0.33).

The duplicate reads possibly derived by PCR amplification, which could affect scaffolding results (Xu et al. 2012), were identified and removed with FastUniq (ver. 1.1; Xu et al. 2012) for each library. The retained reads were assembled using Platanus (ver. 1.2.1; Kajitani et al. 2014) with a starting k-mer of 51 and extended to k-mer of 84. The completeness of the assembly was evaluated using CEGMA (ver. 2.5; Parra et al. 2007, 2009), which annotates the highly conserved core eukaryotic genes (CEGs) in assembled sequences. To evaluate the insertion sizes in the 11 libraries, the original reads used for *de novo* assembly were mapped on the assembled sequences using BWA (ver. 0.7.12-r1039; Li and Durbin 2009), and the distances between the paired reads was surveyed using Picard (ver. 1.140; http://broadinstitute.github.io/picard/).

Repeat regions were identified and masked by referring to the *A. thaliana* repeat database using RepeatMasker (ver. 4.0.5; Smit et al. 2013). Gene predictions were performed on the repeat masked scaffolds using Augustus (ver. 3.2.0; Stanke et al. 2004). An *ab initio* gene prediction with the *A. thaliana* genome (TAIR10) was performed as a reference and then improved the annotations with RNA-seq data (obtained in Appendix S1) as a hint. To evaluate the accuracy of the gene predictions, the number of RNA-seq reads (Appendix S1) mapped on exons were counted using HTseq (ver. 0.6.1; Anders et al. 2014). Furthermore, the percentage of CEGs among the annotated transcripts was surveyed using CEGMA (ver. 2.5).

2.2.4. Polymorphism identification and demographic analysis

SNP sites on the assembled scaffolds were identified according to Fitak et al. (2015). The 200-bp paired-end reads to the repeat-masked scaffolds were aligned using the BWA-MEM algorithm (ver. 0.7.12-r1039; Li and Durbin 2009), only properly paired and unambiguously mapped reads were extracted using Samtools (ver. 1.2; Li et al. 2009), and duplicated reads were removed using Picard (ver. 1.140). The depth of the alignment was evaluated using BamTools (ver. 2.4.0; Barnett et al. 2011). SNPs were identified using two independent software tools: Samtools (ver. 1.2; Li et al. 2009) and Platypus (ver. 0.8.1; Rimmer et al. 2014). The consensus sites between the two methods were used for subsequent analysis (Baes et al. 2014). Bi-allelic SNPs were further extracted with a phred-scaled quality score of ≥20 and a depth from 1/3 to twice the mean depth. To determine the genome-wide heterozygosity, SNP densities within non-overlapping 1-kb windows were calculated using Vcftools (ver. 0.1.14; Danecek et al. 2011).

The past changes in effective population size of *M. polymorpha* were inferred from the genome sequences using the pairwise sequentially Markovian coalescent (PSMC) model (ver. 0.6.5-r67; Li and Durbin 2011). PSMC estimates TMRCA of two alleles

throughout individual diploid genome sequences and effective population size at a given time based on the distribution of TMRCAs. The scaffold sequences with SNP sites that were identified and filtered as mentioned above were used, and the presence or absence of heterozygous SNP sites over non-overlapping 100-bp windows were summarized. PSMC runs were performed for 30 iterations with settings as follows: maximum $2N_0$ coalescent time (t) = 15, initial recombination/mutation rate (r) = 4, division of estimation (p) = "1*2+2*1+25*2+1*4+1*6". The variability of the PSMC estimations was evaluated by 100 bootstrapping analyses. The scaled mutation rate (θ) was plotted based on the pairwise sequence differences (d) (Prado-Martinez et al. 2013). The parameters were scaled under the assumptions of a generation time (g) of 30 years and a mutation rate (μ) of $7.1 \times 10^{-9} \pm 0.7 \times 10^{-9}$ per generation per site (*Arabidopsis thaliana*; Ossowski et al. 2010); the effective population size and the time were represented by $\theta/(4\mu)$ and $d/(2\mu/g)$, respectively.

2.3. Results

2.3.1. Genome size estimated by flow cytometry

T he average relative fluorescence intensity in *M. polymorpha* versus *E. globulus* was 0.57 ± 0.01. Because the DNA amount in a haploid chromosome of *E. globulus* is 0.62 pg (Praça et al. 2009), the DNA amount in *M. polymorpha* was calculated to be 0.36 ± 0.01 pg per chromosome. According to the report from Doležel et al. (2003) showing that 1 pg DNA = 0.978×10^9 bp, the genome size of this species was estimated to be 347.3 ± 7.2 Mb.

2.3.2. Sequencing and assembly

2.3.2.1. 1st draft genome

In total, 203,300,721 and 133,575,168 pairs of reads were kept for the 200- and 500-bp libraries, respectively (Table 2-1). *De novo* assembly based on a total of 336,875,889 pairs of reads resulted in a total of 60,163 scaffolds (Fig. 2-1). The total scaffold length, including gaps, was 285.5 M base, and the N50 scaffold size, including gaps, was 43,614 bases. The smallest and largest scaffolds were 500 and 273,848 bases, respectively. On the 15,333 of the 60,163 scaffolds, 54,416 putative genes were predicted.

2.3.2.2. 2nd draft genome

Totally, 356,076,785 and 195,467,842 raw reads for the short insertion and mate pair libraries, respectively, were obtained (Table 1; DDBJ accessions DRA004245 and DRA004246). After quality filtering, 336,875,889 reads (94.6%) from the two short insertion libraries and 143,669,563 reads (73.5%) from the mate pair libraries were retained (Table 1). The k-mer frequency distributions at different k-mer sizes showed

the greatest number of unique k-mers in the trimmed short insertion reads at k = 91 with a local maximum at k = 51 (Figure S1a). At k = 91, the estimated genome size was 300,148,157 bases. Owing to the fact that unique k-mers found at k = 91 may include those derived from sequence errors, a starting k-mer of 51 was set in the subsequent *de novo* assembly. At every k (k = 21–91 in 10-bp increments), similar heights of peaks were found in homozygous- and heterozygous-derived k-mer frequency distributions, indicating a high heterozygosity in this plant individual (Figure S1b for k = 51). The removal of PCR duplicates resulted in a total of >93 giga bases (Gb) composed of 311,767,203 reads (87.6% of the raw reads) and 124,112,642 reads (63.5% of the raw reads) from the short insertion and mate pair libraries, respectively (Table 1).

The assembled genome sequences of *M. polymorpha* represented a total of 304,366,837 bases, with half of the assembly covered by 19 scaffolds more than 5 Mb in length (Table 2; Figure 1; DDBJ accession BCNH01000001–BCNH01036376). Each of the 55,346 contigs on the 36,376 scaffolds was assembled from an average of 176.2-fold coverage of reads (i.e., depth) (Table 2). Of the 248 most highly conserved CEGs, 89.52 % were found in the assembly. An average of 69.0% (range: 56.6–95.1%) of the original reads were mapped in proper pair on the assembled scaffolds (Table 1). In each library, the insertion sizes that were calculated based on the alignments of original reads on the assembled scaffolds mostly agreed with their nominal sizes, except for the library of paired-end 500 bp (Figure S2). The assembled scaffolds contained 4.14% of repetitive regions. The gene prediction on the repeat masked scaffolds based on the *A. thaliana* genome and RNA-seq data (Appendix S1) predicted a total of 39,305 genes containing 41,874 transcripts. Of the 10,454,613 reads counted by HTseq, 9,006,495 reads (86.1%) were mapped on exons. The 41,874 transcripts included 96.37% of the 248 most highly conserved CEGs.

2.3.3. Polymorphism identification and demographic analysis

Heterozygous sites based on the alignment of original short reads to the assembled genome sequences were identified. Of the 395,624,376 reads obtained from the 200-bp paired-end library, 351,782,023 reads (88.9%) were mapped on the repeat masked scaffolds. The mean coverage across the scaffolds was 113.6 folds. In total, 847,078 SNPs with a quality score of ≥20 and a depth between 38 and 226 were identified. The mean heterozygosity across the genome was estimated to be 2.839×10^{-3} using this method.

The historical changes in the effective population size of this species were estimated using the PSMC model (Figure 2). The effective population size slightly decreased around the time of $d = 2 \times 10^{-3}$ and monotonically increased until the time of $d = 2 \times 10^{-4}$ followed by a decline until the time of $d = 5 \times 10^{-5}$ (Figure 2). The bootstrapping analyses showed consistent demographic patterns, although larger variances were found in the estimation after the time of $d = 5 \times 10^{-4}$ (Figure 2).

2.4. Discussion

The genome sequences of *M. polymorpha* have been assembled using a total of >93 Gb obtained from two short insertion and nine mate pair libraries (Table 1). The assembled genome represented 304 Mb in total length (Table 2), which was 101% and 88% of the genome size estimated with k-mer frequency distribution (300 Mb) and flow cytometry (347 Mb), respectively. Of the 248 highly conserved CEGs, 96.4% were identified in the annotated transcripts. Therefore, my assembly almost completely identified the whole genome of this species. The assembly statistics indicated that my genome sequences were of high quality and contained a small fraction of artificial assembly, i.e., an N50 length of ≥5 Mb, an average contig depth of 176 folds, and a mapping rate of original reads of 85.9% (Table 2). Although the 500-bp paired-end library contained shorter fragments than expected, all the other libraries contained artificial fragments (Figure S2). Because an original database of repeat sequences specific to this species was not constructed, the fraction of repetitive regions found in the present genome was relatively smaller than in the other plant species (Michael and Jackson 2013). Based on the gene prediction utilizing the *A. thaliana* genome and RNA-seq data, the *M. polymorpha* genome encoded 39,305 putative protein-coding genes, which were slightly more than those encoded by *Eucalyptus grandis* (36,376 protein-coding genes; Myburg et al. 2014). These characteristics regarding gene components or genome structure of this species need to be validated by additional annotation analysis using alternative annotation software or homology-based annotation (Yandell and Ence 2012).

As expected from a k-mer frequency distribution (Figure S2), a considerable amount of heterozygosity was found in *M. polymorpha*, although the procedure used to identify heterozygosity was likely to be conservative; the mean frequency of heterozygous nucleotide sites across the genome was estimated to be 0.28%. This rate was similar to that of other outcrossing or dioecy plant species such as *Populus trichocarpa* (0.26%; Tuskan et al. 2006) and *Phoenix dactylifera* (0.46%; Al-Dous et al. 2011), contrasting to the rates observed in self-fertilized or inbred plant genomes such as *Carica papaya* undergoing 25 generations of inbreeding (0.06%; Ming et al. 2008). The considerable heterozygosity found in the *M. polymorpha* genome corresponded with the reproductive characteristics. This species is recognized to be outcrossing with partial self-incompatibility having a low inbreeding coefficient (F_{IS}) (Carpenter 1976; DeBoer and Stacy 2013; Shimizu and Tsuchimatsu 2015) and to harbor a high dispersal capacity characterized by wind-dispersed seeds and bird- and insect-pollinated pollens (Carpenter 1976; Drake 1992). Population genetic analyses have revealed that gene flows indeed occurred among populations as well as islands (Harbaugh et al. 2009).

The PSMC model based on the genome-wide polymorphism revealed changes in effective population size of this species (Figure 2). The results indicated three main demographic events in the past: a slight population shrink, following steady growth, and dramatic population decline (Figure 2). In assuming a mutation rate (μ) of 7.1×10^{-9} per generation per site (*Arabidopsis thaliana*; Ossowski et al. 2010) and a generation time (g) of 30 years, the first population shrink likely occurred at 4 million YBP. This could represent a population bottleneck at the first arrival of this species to Kauai Island, which was formed approximately 4.7 million YBP (Price and Clague 2000). The following population growth could have occurred during the subsequent 3.5 million years and corresponds to continual formation of volcanic islands that could have

provided novel habitats to this species. These scenarios about the arrival and one-by-one expansion from older to younger islands of the species agree with the observation by Percy et al. (2008), in which the oldest and most diverse chloroplast haplotypes were found in Kauai and more derivative haplotypes were found in the younger islands. Factors underlying the third event of rapid population decline might include a larger oscillation in climate between glacial and interglacial periods during the last 1 million years, which could result in a the fluctuation in the inversion level or snowline (Gavenda 1992) and/or subsidences that decreased land area in each island since the time of land formation (Price and Clague 2000). Although different assumptions about mutation rate or generation time would date demographic events differently and the precise time for these demographic events is unavailable, the dramatic rise and fall of the population size in the past can be affirmed.

By setting my genome sequences as a starting point, the genome resource of this species should be further enriched to characterize the functions of any genome differences found within or among species (Michael and Jackson 2013). Enhancing insights about the adaptive processes of this species may include re-sequencing of additional individuals that grow in a different habitat from the present individual and thus harbor different ecological features. Comparative genomics among multiple individuals are expected to reveal genetic regions or traits subject to natural selection (Axelsson et al. 2013; Wu et al. 2014; Meyer et al. 2015) or genome-wide patterns of nucleotide variations associated with climates (Fischer et al. 2013; Kubota et al. 2015). These developments would provide valuable resources to reveal the genetic mechanisms underlying an ecological divergence.

Tables and Figures

Table 2-1
Summary for the sequencing libraries of *Metrosideros polymorpha*

Library name	Sequence method	Index sequence	Adapter conc. (μM)	Nominal insert size (nt)	Read count			Mean depth per contig (fold)	Mapped reads in proper pair (%)
					Raw data	After quality filtering	After PCR duplicates removal		
PE200	PE	-	-	200	211,790,654	203,300,721	197,812,188	87.23	95.1
PE500	PE	-	-	500	144,286,131	133,575,168	113,955,015	49.68	92.3
MP03	MP	GCCAAT	0.75	3,000	53,001,052	39,480,821	34,342,702	11.82	67.3
MP04	MP	CTTGTA	0.75	4,000	28,131,255	20,110,668	15,735,145	5.15	64.6
MP05	MP	ACAGTG	1.50	5,000	45,988,172	33,509,763	30,042,658	10.08	65.1
MP07	MP	GTGAAA	1.50	7,000	27,461,166	20,342,232	16,570,918	5.70	64.8
MP10	MP	CGATGT	1.50	10,000	14,539,275	10,512,413	9,585,411	3.17	62.7
MP12	MP	ATGTCA	1.50	12,000	12,918,586	9,604,482	8,925,313	3.10	65.7
MP15	MP	CAGATC	3.00	15,000	8,355,884	6,438,467	5,829,528	2.05	65.3
MP20	MP	TGACCA	1.50	20,000	3,712,677	2,668,468	2,244,634	0.73	59.6
MP24	MP	ACTTGA	1.50	24,000	1,359,775	1,002,249	836,333	0.28	56.6
Total					551,544,627	480,545,452	435,879,845	176.16	85.9
Mean					50,140,421	43,685,950	39,625,440	16.27	69.0

PE, 100 bp paired-end sequencing; MP, 125 bp mate-pair sequencing

Table 2-2
Summary for the *de novo* assembly of *Metrosideros polymorpha* genome sequences

Number of scaffolds	36,376
Number of contigs	55,346
Total number of bases of scaffolds (nt)	304,366,837
Total number of base in contigs (nt)	294,061,642
Scaffold N50 length (nt)	5,051,733
Contig N50 length (nt)	31,875
Number of scaffolds covering half of the assembly	19
Minimum scaffold length (nt)	83
Maximum scaffold length (nt)	15,115,857
Mean scaffold length (nt)	8,367
Percent of gaps (%)	3.4
GC content (%)	40.0
Repeat content (%)	4.1
Number of putative genes	39,305
Proportion of 248 CEGs (%) found in the assembly	89.5

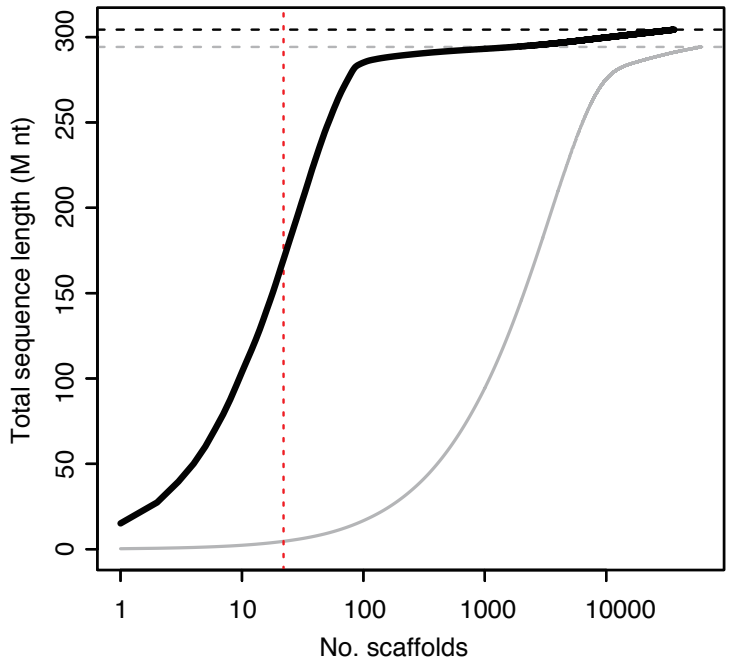

Fig. 2-1
Cumulative length of 1st (gray) and 2nd (bold black) draft genome of *Metrosideros polymorpha*. A red dashed vertical line indicates the number of chromosomes of this species ($2n = 22$).

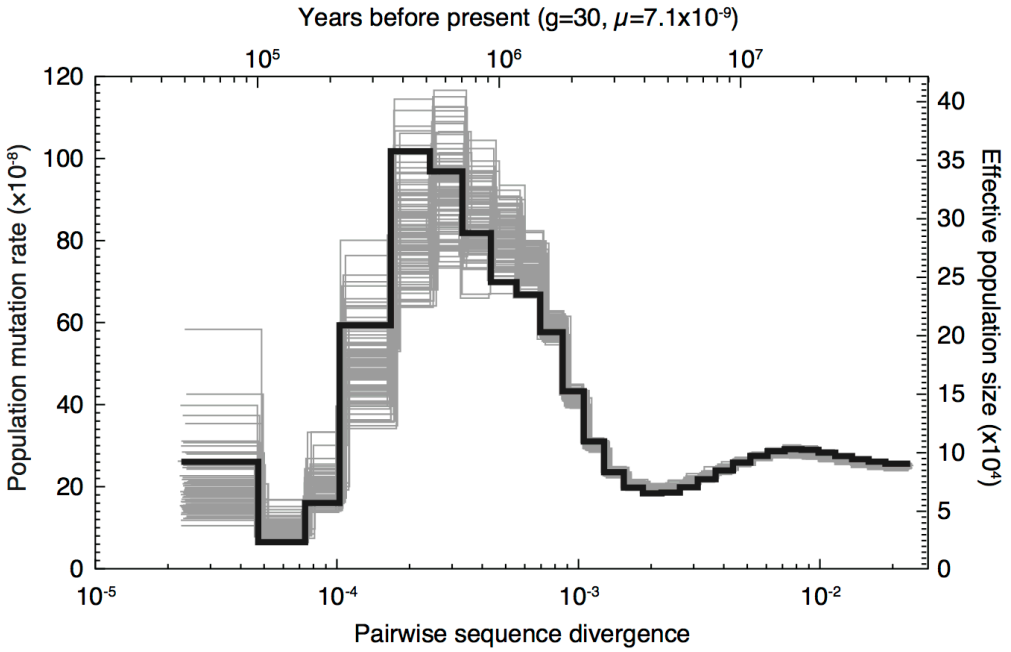

Fig. 2-2

Demographic history of *Metrosideros polymorpha* inferred from the genome-wide polymorphism based on the PSMC model. The lower x-axis indicates the pairwise sequence differences and left y-axis indicates the scaled mutation rate. The upper x-axis indicates the time scaled in years and right y-axis indicates the effective population size. The parameters were scaled using a mutation rate of 7.1×10^{-9} per generation per site (*Arabidopsis thaliana*; Ossowski et al. 2010) and a generation time of 30 years. Thin lines indicate 100 bootstrapping analyses.

Supplemental materials

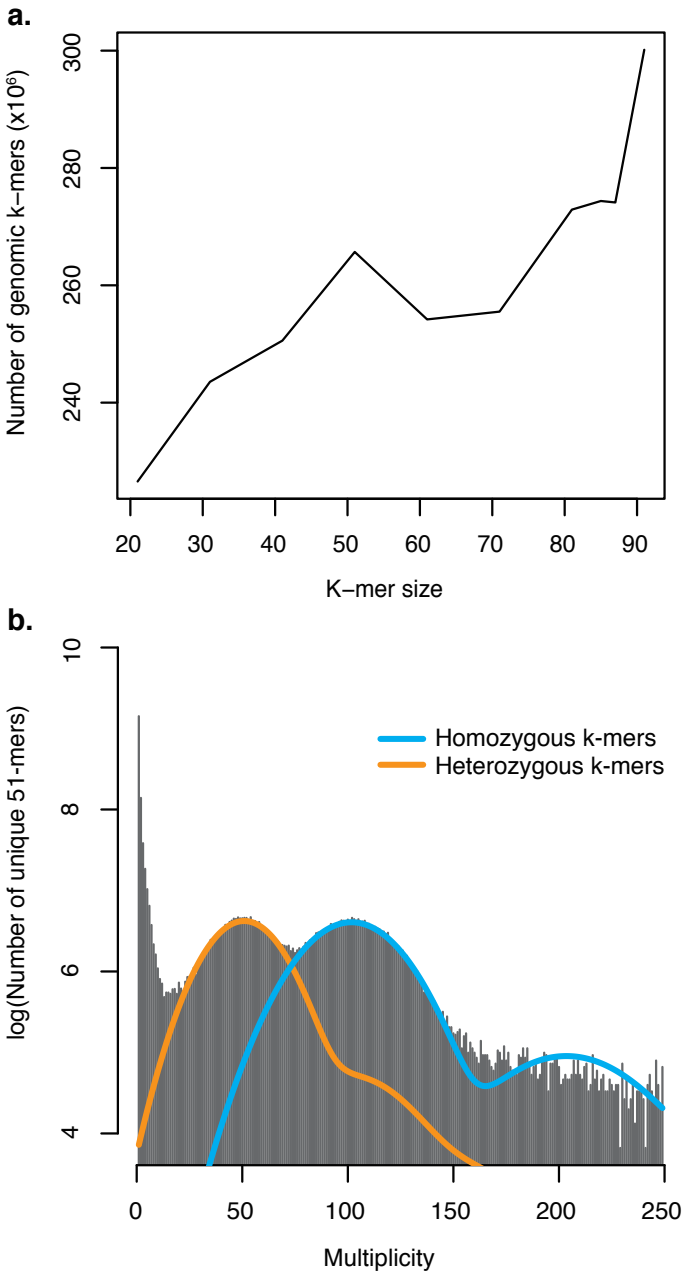

Fig. 2-S1

K-mer frequency distribution in a *Metrosideros polymorpah* genome. (a) The number of unique genomic k-mers counted in each k-mer size (k = 21–91 in 10-bp increments). (b) K-mer frequency distribution at k = 51. Homozygous- and heterozygous-derived k-mer frequency distribution, which was fitted using the KmerGenie algorithm, is indicated with light blue and orange lines, respectively.

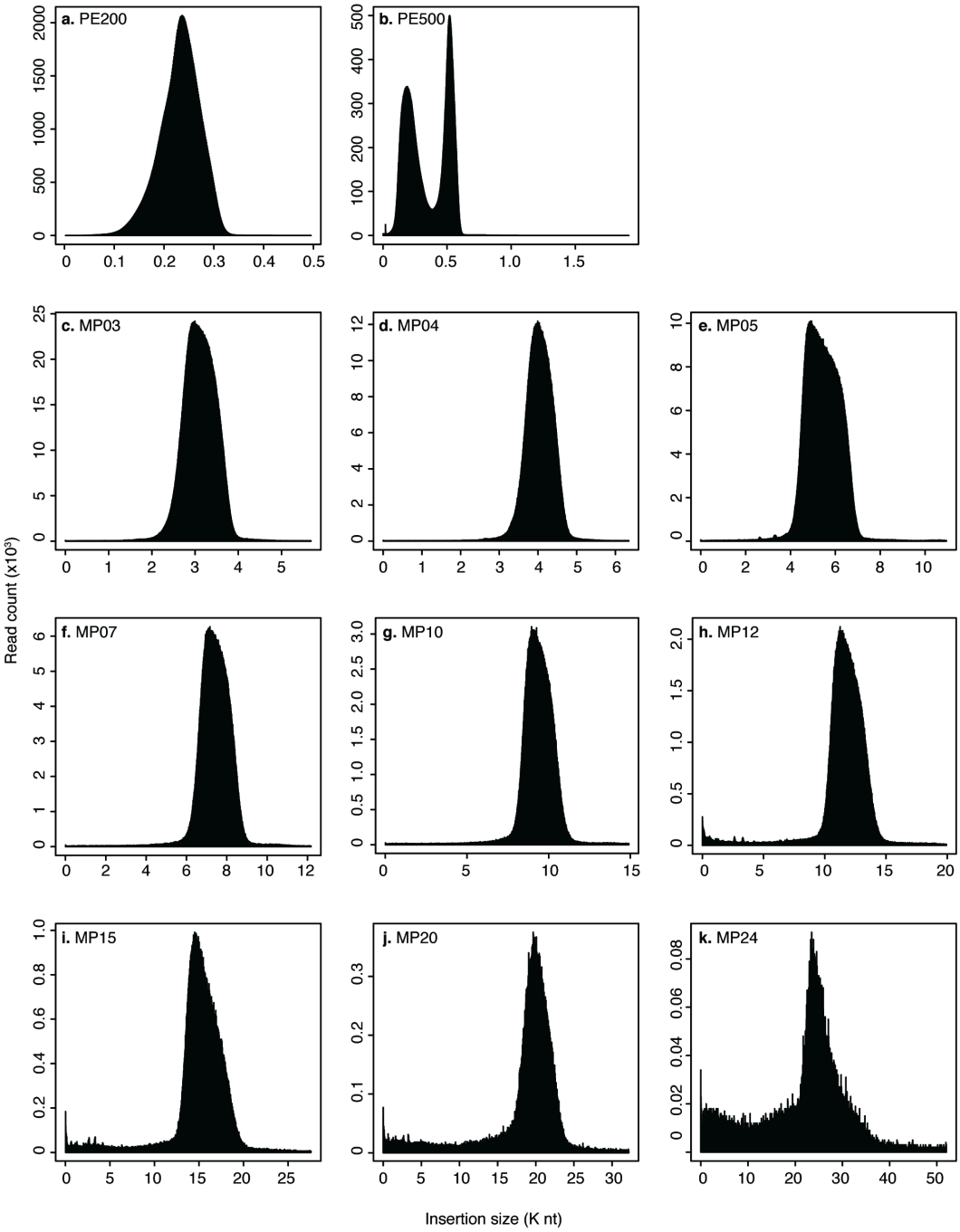

Fig. 2-S2
Distribution of insertion sizes in each library. Except for the 500-bp paired-end library, libraries were constructed with few artifacts. Library names are shown in Table 1.

Appendix 2-S1
Methods and summary results for the RNA sequencing conducted with a *Metrosideros polymorpha* plant individual. The obtained reads were used as a hint in the gene prediction on the assembled genome sequences.

1. Plant materials

Young leaf buds were collected from an *Metrosideros polymorpha* individual at 700 m above sea level on a 150 years old lava flow on the eastern frank of Mauna Loa. This plant individual is different from the one used for genome sequencing. The collected bud tissues were torn into small pieces and immediately stored in RNAlater (Qiagen) under −30°C until RNA extraction.

2. RNA extraction

The tissues were crushed into a powder in a micro tube with stainless beads (diameter = 5 mm and 3 mm). The tube was cooled off for 1 min in a freezer (−30°C) after every 30 s of crushing. 900 μl of 2 × CTAB buffer, which contained 2% CTAB, 2.0-M NaCl, 100-mM Tris-HCl (pH 8.0), 25-mM EDTA, and 2% 2-mercaptoethanol, was added to the tube and the solution was incubated for 10 min at 65°C. After one volume of chloroform/isoamyl alcohol (CIA) was added, the solution was mixed for 5 min and then centrifuged for 10 min at 9,500 ×g. This procedure was repeated for the aqueous phase transferred into a new tube. The aqueous phase was transferred into a new tube, LiCl was added to be a final concentration of 2 M, and incubated at −20 °C for 16 hours. After spinning for 20 min at 9,500 ×g at 4°C, the supernatant was discarded. The precipitate was dissolved into 500 μl of STE buffer (65°C), which contained 10 mM of Tris-HCl (pH 8.0), 1-mM EDTA, and 2-M NaCl. After one volume of CIA was added, the solution was mixed for 5 min and then centrifuged for 15 min at 9,500 × g under 4°C. The supernatant was precipitated with 1/5 volume of 5-M NaCl and 2.5 volumes of 100% ethanol at −20°C for 2 hours. Following spinning for 20 min at 9,500 ×g at 4°C, the supernatant was washed with 400 μl of 70% ethanol. The pellet was air-dried and resuspended into nuclease free water.

3. Library construction and sequencing

The sequencing library was prepared from 500 ng of total RNA using Illumina TruSeq Library preparation kit (Illumina) and normalized using the duplex-specific thermostable Nuclease (DSN) enzyme (Illumina). Poly (A)-containing mRNA molecules were extracted from total RNA using poly (T) oligo-attached magnetic beads; then, mRNA molecules were fragmented into small pieces using fragmentation reagent. From the fragmented mRNA, first-strand cDNA was synthesized using random hexamer-primed reverse transcription. This was followed by a second-strand cDNA synthesis using DNA polymerase I with dNTPs and RNaseH. The synthesized cDNA was purified, end-repaired, and added poly (A) to the 3′-end. The sequence adapters were ligated to the fragments. The ligation products were size selected with

TAE-agarose gel electrophoresis, and the desired fragments were amplified with PCR. Normalization treatment with DSN enzyme was performed for the obtained library. The quantity and quality of the final library was validated using an Agilent 2100 Bioanalyzer and an ABI StepOnePlus Real-Time PCR System. The 200-bp insertion library was sequenced using Illumina HiSeq 2000 (Illumina, CA, USA) with 100-bp paired-end sequencing technology. This library construction and sequencing was conducted at Beijing Genomics Institute
(Hong Kong, China).

4. Quantity and quality of the obtained reads

From the raw reads, adapter sequences were trimmed; reads with more than 10% of ambiguous bases and reads containing more than 50% of bases with a quality score less than 10 were removed using SOAPnuke developed by BGI. As a result, 5,505,887 clean reads with a mean *Phred* score of 38 were obtained (DDBJ accession DRX045362, DRR050356).

3

The population genomics signature of environmental selection with gene flow in *Metrosideros polymorpha*

3.1. Introduction

Spatial patterns of gene flow among populations have been attributed to two primary scenarios: isolation-by-distance (IBD) and isolation-by-environment (IBE). When geographically distant populations differentiate due to a decrease in gene flow, a significant positive correlation is observed between the genetic and geographic distances in populations (IBD model; Wight 1943). Gene flow can also be restricted by specific selective factors independent of geography, such as environment (IBE model; Sexton *et al.* 2014) and resource use (Nosil *et al.* 2008; Edelaar *et al.* 2012). Although geographic and environmental distances often correlate (Shafer & Wang 2013), recent improvements in statistics and easier acquisition of detailed environmental data enable us to discriminate IBE from IBD in spatial genetic structures (Bradburd *et al.* 2013). To date, as pointed out by Shafer and Wolf (2013), most studies testing the effect of geography and environment on spatial genetic structures have used a limited number of neutral loci (Mallet *et al.* 2014; DeWoody *et al.* 2015, but see Lexer *et al.* 2014; Manthey & Moyle 2015). Nonetheless, genetic differentiation can be variable across a genome (Nosil *et al.* 2008), and the factors underlying the differentiation among populations could be different among loci. Thus, the role of IBD and IBE in population differentiation should be evaluated at individual loci across a genome.

Population genetic studies targeting M. polymorpha have revealed significant genetic differentiation among populations in accordance with their ecological niches (Aradhya et al. 1993; Harbaugh et al. 2009; DeBoer & Stacy 2013; Stacy et al. 2014). However, the genetic evidences for these dramatic adaptations are insufficient for the following reasons. First, no selective genetic processes have been addressed in this species because the previous studies have focused on neutral genetic processes (Harbaugh et al. 2009; Stacy et al. 2014). Second, the genetic structures at the individual

level have not been revealed because of the large genetic variations within populations. If a large genetic admixture occurs within a population or a population has a complex colonization history, individual plants within the population can be assigned to different genetic groups; thus, gene flow can be revealed at an individual level rather than at a population level. The previous studies that used 9–10 microsatellite markers faced difficulties in inferring a spatial range of gene flow among individuals; the admixture proportions of the genetic clusters did not differ among individuals or populations such that rather homogenous spatial genetic structures across populations were detected (Harbaugh et al. 2009; Stacy et al. 2014). The small number of genetic markers used in these studies could account for these results, yielding poor genetic information per individual. To resolve these issues, the spatial genome structure considering genome-wide admixtures and adaptive loci should be investigated using genome-wide markers that can be annotated based on whole-genome sequences.

In the present study, I aimed to elucidate genome differentiation and its drivers in *M. polymorpha* occupying a wide range of ecological niches. Thousands of polymorphic DNA markers as well as a *de novo* draft genome sequence were used to (1) reveal the spatial genome structure at an individual level, (2) detect outlier loci, which show non-neutral differentiation patterns and may have key roles in adaptation to various environments, and (3) evaluate the effect of IBD and IBE on population differentiation at outlier and genome-wide loci.

3.2. Materials and methods

3.2.1. Population sampling

In June 2013, leaves were collected from 72 *M. polymorpha* trees growing in the common garden at the Volcano Agriculture Station, University of Hawaii (Kitayama *et al.* 1997; Cordell *et al.* 1998). The 72 individuals were grown from the seeds of different mother trees collected from nine populations distributed at five elevations (150, 700, 1,200, 1,800, and 2,400 m above sea level) and two lava flows (150 and 3,000 years old) on the east flank of Mauna Loa, the island of Hawaii (Table 3-1). The 72 samples covered a wide range of phenotypic variations, especially in leaf area and trichrome weight (Table 3-1). The geographic locations of the nine populations are indicated by the GPS point data of the source forests (Fig. 3-2).

3.2.2. Genotyping by sequencing

For the 72 individuals collected to evaluate the genetic differentiation among the nine populations, total genomic DNA was extracted using the modified CTAB method (Milligan 1992) and digested by the restriction enzymes *Bgl* II and *Eco* RI. The digested DNA fragments and two adapters (*Bgl* II adapter and *Eco* RI adapter) were ligated. Digestion and ligation were performed simultaneously at 37°C for 16 h. The reaction mixture consisted of 20 ng of genomic DNA, 5 units of *Bgl* II (NEB), 5 units of *Eco* RI-HF (NEB), 1× NEB buffer2 (NEB), 1× bovine serum albumin (BSA) (NEB), 0.2

microM *Bgl* II adapter, 0.2 microM *Eco* RI adapter, 1 mM ATP (Takara), 300 units of T4 DNA ligase (Enzymatics). The ligation product was purified using the AMpureXP (Beckman coulter) according to the manufacturer's instructions. One-tenth of the purified DNA was used in the PCR enrichment with the KOD-Plus-Neo (TOYOBO). Sequences of the adaptors and primers used in this study are shown in Table 3-S1. PCR product fragments of approximately 320 bp were selected using E-Gel size select 2% (Life technologies). Each restriction-site–associated DNA sequence (RAD-seq) library was uniquely barcoded and pooled. The first 49 bp of the fragments were sequenced in two lanes on HiSeq2500 (Illumina) using TruSeq v3 chemistry at BGI (Hong Kong, China). A total of 237,218,370 reads and an average of 3,294,700 reads per sample were obtained (DRA accession: DRA003994). The reads were mapped to the draft reference genome sequences using Bowtie2 (ver. 2.2.3; Langmead & Salzberg 2012) with default parameter settings. Single nucleotide polymorphism (SNP) sites were identified in the BAM files using the Stacks pipeline (ver. 1.27; Catchen *et al.* 2011). I then selected SNPs with more than 10 read counts per individual being shared by more than 50% of the individuals in at least five of the nine populations. Only biallelic SNPs with a minor allele frequency of more than 5% were selected. Selection of SNPs was performed using the "populations" command implemented in Stacks (ver. 1.27; Catchen *et al.* 2011) and vcftools (ver. 0.1.11; Danecek *et al.* 2011).

3.2.3. Population genomic analysis

3.2.3.1. Population genetic structure

To evaluate genome-wide genetic diversity within populations, expected heterozygosity (H_E) was calculated for all SNP genotypes using GenoDive (ver. 2.0b27; Meirmans & Tienderen 2004). The difference in expected heterozygosites among elevations or lava ages was tested using analysis of variance (ANOVA) on R (ver. 3.1.2; R Core Team 2014). Spatial genetic structure was investigated using the entire SNP genotypes by two methods, a Bayesian clustering (STRUCTURE ver. 2.3.4; Pritchard *et al.* 2000) and a principal component analysis (PCA). In the STRUCTURE analysis, an admixture model that assumed correlated allele frequencies among populations was used. Ten replicate simulations were run for each K (K = 1–9), with 50,000 burn-in steps followed by 100,000 Markov chain Monte Carlo steps. The optimal K was inferred based on the delta K method (Evanno *et al.* 2005) implemented in STRUCTURE HARVESTER (ver. 0.6.94; Earl & vonHoldt 2012). Admixture proportions from replicate simulations at the optimal K were averaged using CLUMPP (ver. 1.1.2; Jakobsson & Rosenberg 2007). The effect of elevation and/or lava age on the admixture proportion of each genetic cluster was tested using a two-way ANOVA implemented in the "car" package (Fox & Weisberg 2011) on R (ver. 3.1.2; R Core Team 2014). PCA of the genotypes was conducted using GenoDive (ver. 2.0b27; Meirmans & van Tienderen 2004). The correlation between elevation or lava age and the coordinates on PC1 or PC2 axis in each individual was tested using Pearson's product-moment correlation test on R (ver. 3.1.2; R Core Team 2014).

3.2.3.2. Detection and characterization of outlier loci

To detect largely differentiated SNPs that could be subject to divergent selection in the nine populations, a genome scan based on the Bayesian method was carried out using Bayescan (ver. 2.1; Foll & Gaggiotti 2008). Bayescan calculates the Bayes factor, which is the ratio of posterior probabilities of natural selection and neutral model at a given locus, and judges whether a locus was under natural selection (Foll & Gaggiotti 2008; Nielsen *et al.* 2009). The analysis was carried out under the default parameter settings as follows: 20 pilot runs of 5,000 iterations and an additional 50,000 burn-in iterations, followed by 5,000 iterations with a thinning interval of 10. The prior odds were set to 10, indicating that a neutral differentiation is 10 times likely than selection at a locus. Posterior odds (PO) represent Bayes factors and loci with $\log_{10}$ PO > 1, which conversely indicates selection is 10 times more likely than a neutral differentiation at a locus, were identified as outlier SNPs (Nielsen *et al.* 2009; Milano *et al.* 2014). For all outlier SNPs, allele frequencies were compared among populations.

The population differentiation patterns were compared between all 2,247 SNPs and the outlier SNPs using three methods: pairwise F_{ST} values among populations, an analysis of molecular variance (AMOVA), and investigation of a phylogenic network among the 72 samples for each of the entire and outlier SNP datasets. Pairwise F_{ST} values were calculated using GenoDive (ver. 2.0b27; Meirmans & van Tienderen 2004). AMOVA divided the whole genetic variance into two hierarchical categories of "among individuals within populations" and "among populations" using Arlequin (ver. 3.5.2.1; Excoffier & Lischer 2010). Phylogenetic analyses were performed using the Neighbor-net method (Bryant & Moulton 2004) implemented in SplitsTree (ver. 4.13.1; Huson & Bryant 2006) based on uncorrected p-distances among individuals.

To estimate the biological functions of the genes linked to the outlier SNPs, protein sequences were extracted for the putative genes located within 10 kb of the outlier SNPs and then used for homology search analysis. For each protein sequence, BLASTP searching for known proteins was conducted against the NCBI "nr" database (ver. "bg2_may15") and then gene ontology (GO) terms that were searched among the flowering plants (taxid: 3398). Both the BLASTP searching and GO annotations were conducted with an E-value threshold of 1.0×10^{-6} using BLAST2GO (ver. 3.0.10; Conesa *et al.* 2005). To test which GO terms were over expressed in the putative genes located within 10 kb of the outlier SNPs compared to all the putative genes, a GO enrichment analysis was conducted using Fisher's exact test with a threshold of $p < 0.01$ after Bonferroni correction on GOTermFinder (Boyle *et al.* 2004). In the GO enrichment analysis, I used GO terms obtained in the BLASTP searching of protein sequences of all the putative genes against the TAIR10 peptide database (downloaded in December 14th, 2014) with an E-value threshold of 1.0×10^{-6}.

3.2.3.3. Factors underlying genomic differentiation

To identify factors driving population genetic differentiation, I used a generalized linear mixed modeling (GLMM) approach following Lexer *et al.* (2014). Here I intended to explore whether the spatial genetic divergence was derived by IBD (Wight *et al.*

1943) or IBE (Shafer & Wolf 2013). Two matrixes representing pairwise differentiation among the nine populations were used as response variables: F_{ST} matrices calculated for the multilocus genotype data for all and outlier SNPs (described above). Geographic distance and environmental distance were implemented in the GLMMs as explanatory variables. Geographic distances among the nine populations were calculated using the GPS points of the source populations. Environmental distance was determined as the Euclidian distance in the eight-dimensional space composed of eight axes in a PCA of the 20 environmental variables. The 20 environmental variables were composed of 19 WorldClim variables (Hijmans *et al.* 2005) and one variable of lava age (150 or 3,000 years) for the nine populations (Table 3-S2, Supporting information). For each response variable, four models were built: a null model with no explanatory variable (NULL), a model with a single explanatory variable of geographic distance among populations (IBD), a model with a single variable of environmental distance among populations (IBE), and a model with two variables of geographic and environmental distance among populations (IBD + IBE). The best-supported model was then identified based on the deviance information criterion (DIC). The relative significance of the models was evaluated using delta DIC and DIC weight. I used the "MCMCglmm" package (Hadfield 2010) on R (ver. 3.1.2; R Core Team 2014) to calculate the DICs under the parameter set used in Lexer *et al.* (2014).

3.3. Results

3.3.1. Mapping of the RAD-seq reads, and SNP calling

Of a total of 237,218,370 RAD-seq reads obtained from 72 samples, 225,111,711 reads (95%) were successfully mapped on the reference genome sequences (Table 3-S3, Supporting information). After filtering, 2,247 SNPs on the 1,388 scaffolds were used for the following population genomic analysis (Data S4, Supporting information). The number of SNPs recovered from a sample was 1,773 on average (range, 648–2,110), and the number of samples sharing an SNP site was 57 on average (range, 17–72) (Fig. 3-S1, Supporting information).

3.3.2. Population genetic structure

Genetic diversity within populations and population genetic structures were evaluated using the entire 2,247 SNP genotypes. Expected heterozygosity in the nine populations was 0.23 on average (range, 0.19–0.25) (Fig. 3-3). The expected heterozygosity decreased as a function of elevation or lava age but statistically not significant (Fig. 3-3). In a Bayesian clustering of the nine populations, delta K peaked at K = 2 followed by K = 3 (Fig. 3-4a). In a scenario of K = 2, a clear genetic differentiation between high and low elevation was found with partly admixing each other in the middle elevation (1,200 m) (Fig. 3-4b). The genetic cluster found in the lower elevations was further differentiated into two clusters in a scenario of K = 3 (Fig. 3-4b). A two-way ANOVA revealed that the admixture proportion of all the five genetic clusters found in K = 2

and $K = 3$ was influenced by elevation ($p < 0.001$) and that of a genetic cluster in $K = 3$ (shown in blue in Fig. 3-4b) differed between lava ages ($p = 0.01$). PCA of 2,247 SNP genotypes also revealed the prime genetic differentiation among elevations; the PC1 axis, which explained 9.05% of the total genetic variance, showed a significant correlation with elevation ($r = 0.86$; $p < 0.001$) (Fig. 3-4c). PC2 axis with 5.5% of total genetic variance correlated significantly with lava age ($r = 0.25$; $p = 0.03$) (Fig. 3-4c).

3.3.3. Detection and characterization of outlier loci

Of the 2,247 SNPs, 35 (1.56%) on 26 scaffolds showed more than 1 of $\log_{10}$ PO in Bayescan, suggesting divergent differentiation among the nine populations (Fig. 3-5a). The F_{ST} value per each SNP estimated in Bayescan was 0.15 on average (range, 0.06–0.53) (Fig. 3-5b). More than 95% of the SNPs (2,152 of 2,247; 95.8%) showed less than 0.2 of the F_{ST} value (Fig. 3-5b). The change in allele frequencies at the 35 outlier SNPs was 0.91 on average (range, 0.64–1.00), and alleles were fixed in one or more populations at all 35 outlier SNPs (Fig. 3-6). Indeed, populations were greatly differentiated at the outlier SNPs compared with the entire SNPs. The pairwise F_{ST} among the nine populations was 0.10 ± 0.06 (SD) and 0.37 ± 0.26 (SD) for the entire 2,247 SNPs and 35 outlier SNPs, respectively (Fig. 3-7a). AMOVA revealed the proportion of genetic variance among populations and within populations was 45.13% and 54.87%, respectively, at the outlier SNPs, and 10.85% and 89.15% at all SNPs (Fig. 3-7b and c). These patterns were supported by the phylogenetic networks, as that of the outlier SNPs had longer branch lengths between populations than that of the entire SNPs (Fig. 3-7b and c). On the 26 scaffolds (length, 7,320–200,288 bases), 83 putative genes were found within 10 kb (range, 154–9,975 bases) of the outlier SNPs. Sixty-four genes obtained significant hits (E-value < 1.0×10^{-6}) against the seven flowering plant species in BLASTP and GO analysis (Table 3-S4, Supporting information). The GO enrichment analysis showed that no GO terms were significantly over expressed in the 83 putative genes.

3.3.4. Factors underlying genomic differentiation

Whether the population genetic differentiation at all and outlier SNPs were derived by IBD or IBE was tested using a GLMM approach. When the pairwise F_{ST} values calculated for the entire 2,247 SNPs were set as response variables of GLMMs, the best model was composed of only an explanatory variable of geographic distance (IBD model) (Table 3-2). The largest DIC weight was found in the IBD model, followed by the IBE and IBD + IBE model with a slight difference (Table 3-2). In contrast, most F_{ST} values at the 35 outlier SNPs were explained by the IBE model, which incorporated only the variable of environment distance among populations (Table 3-2). The DIC weight of the IBE model was 0.70, a large difference from the other models (Table 3-2).

3.4. Discussion

3.4.1. Population genetic structure revealed by genome-wide SNP markers

Although the traditional genetic markers did not have sufficient resolution to resolve the large genetic variations within populations and infer individual-level genetic structure of this species (Aradhya et al. 1993; Harbaugh et al. 2009; DeBoer & Stacy 2013; Stacy et al. 2014), the present study with genome-wide markers as many as 2,247 SNPs successfully revealed the spatial genetic structure of 72 *M. polymorpha* individuals originating from the eastern flank of Mauna Loa.

The populations were principally differentiated by elevation (Figs. 1b and 2). The Bayesian clustering clearly inferred two genetic clusters, occupying the lower and higher elevations (Fig. 3-4b), and the coordinates on the PC1 axis significantly correlated with elevation (Fig. 3-4c). However, the spatial genetic structure of the nine populations was not likely determined by elevation alone. In the Bayesian clustering, the scenario of K = 3 showed the second largest value of delta K compared with the other scenarios (Fig. 3-4a), indicating the presence of an additional genetic cluster besides the two detected in the scenario of K = 2 (Fig. 3-1b). The PC1 axis explained as low as 9.05% of the genetic variance (Fig. 3-4c), suggesting that factors in addition to elevation also affected the spatial genetic variation. One of the candidate factors is lava age. Statistical tests showed that the admixture proportions of one genetic cluster found in a scenario of K = 3 (shown in blue in Fig. 3-4b) were differentiated between lava age, and PC2 coordinates significantly correlated with lava age. Nonetheless, the genetic variance explained by elevation and lava age was likely small; PC1 plus PC2 explained as low as 14.55% of the total genetic variance (Fig. 3-4c), and AMOVA showed 89.15% of the total genetic variance was contained within populations. Because the gene flow of this species occurred even among island populations (Harbaugh *et al.* 2009), frequent gene flow among the populations may neutralize the population differentiations along elevations and/or lava flows.

In the scenarios of K = 3, a total of three genetic clusters were found in 72 *M. polymorpha* trees. The extent of genetic admixtures of the three genetic clusters varied among the nine populations. In the lower and middle elevations (150, 700, and 1,200 m), multiple genetic clusters were largely admixed (Fig. 3-4b), resulting in higher genetic variations retained in the populations (Fig. 3-3). This admixture could be due to the overlap in preferred habitat among the genetic clusters (DeBoer & Stacy 2013) and frequent hybridizations (Corn & Hiesey 1973). In contrast, the populations at higher elevations (1,800 and 2,400 m) were mostly composed of a single genetic cluster, and the genetic variance within populations was relatively low (Fig. 3-3). This could indicate a strong bottleneck because of the relatively small population sizes (Table 3-1) and/or purifying selection due to the harsh environmental condition as shown in the low temperature and precipitation, limited nitrogen, and strong wind (Vitousek 1992; Stacy *et al.* 2014).

3.4.2. Evidence of adaptive differentiation of *M. polymorpha* populations

A genome scan based on 2,247 SNP markers revealed that 35 SNPs differed significantly among populations. Because the differentiations at these outlier SNPs were mostly explained by the IBE model indicating a significant correlation with environmental distance, they were likely consequences of adaptive differentiation (Salmela 2014).

The 35 outlier SNPs were located on 26 scaffolds, and the scaffolds had 83 putative genes within 10 kb of the outlier SNPs (Table 3-S4, Supporting information). These genes encoded some important biological functions relevant to environmental adaptation to elevation. First, a GO term of "response to (abiotic) stress" was found for four genes, and the gene "g15857" found 6,327–6,353 bases from Outlier04, Outlier05, and Outlier13 is likely a homolog of "*kda* class I heat shock" gene in *Eucalyptus grandis*, which plays a role in the response to oxidative stress (Table 3-S4, Supporting information). Given the extreme environmental conditions at higher elevations, including high-light intensity, strong wind, and oxidative stress, the finding that some outlier SNPs have linkages with the genes involved in stress tolerance is relevant. Second, the gene "g27368" found 4,320–4,343 bases from Outlier01 and Outlier02 was annotated to an *E. grandis* homolog of "probable membrane-associated kinase regulator 4" concerning to response to hormone including auxin and brassinosteroid (Table 3-S4, Supporting information). Auxin as well as brassinosteroid is known to play diverse roles in plant growth, development, and stress responses (Wolters & Jürgens 2009) and in plant fitness (e.g., Keuskamp *et al.* 2010; Lu *et al.* 2014; De Wit *et al.* 2014). Finally, the gene "g15782" found 8,678 bases from Outlier10 functions in the transport of magnesium ions in *E. grandis* (Table 3-S4, Supporting information). Vitousek (1992) reported that magnesium concentrations in soils decrease with increasing elevation; thus, differences in the ability to transport magnesium ions could affect fitness in populations at lower elevations. Further studies such as expression analysis under controlled environmental conditions (Fraser *et al.* 2015) or transgenic experiments using model plant species (Kobayashi *et al.* 2013) are needed to confirm the biological functions of these genes in an ecological context.

The percentage of outlier SNPs among all SNPs was as low as 1.56% (Fig. 3-5). Although comparisons should be made with caution, as geographic scales critically affected the extent of genetic differentiation (De Kort *et al.* 2015), the fraction of outlier SNPs was smaller than that found in other tree species, for example, *Frangula alnus* (2.7%–6.6%, De Kort *et al.* 2015), *Eucalyptus tricarpa* (2.6%; Steane *et al.* 2014), and *Populus trichocarpa* (3.6%; Geraldes *et al.* 2014). This difference could be due to the large gene flow occurring in a narrow geographic scale. In the case of *M. polymorpha*, it is predicted that adaptive traits could be controlled by a relatively small number of genes.

The 35 outlier SNPs showed evident differentiation across the genome as well as among populations. The F_{ST} values of outlier SNPs were approximately three times larger than those of the 2,247 genome-wide SNPs (0.37 ± 0.026 vs. 0.10 ± 0.06, respectively) (Fig. 3-7a). AMOVAs revealed the outlier SNPs explained larger genetic variance among populations (45.13%) than the genome-wide SNPs (10.85%). This pattern was consistent with the phylogenetic networks (Fig. 3-7b and c). Fixations of alleles at all 35 outlier SNPs suggested strong purifying selection in populations at

higher and lower elevations (Fig. 3-6).

3.4.3. Genomic mosaics of *M. polymorpha*

Gene flow and natural selection have antagonistic effects on genetic divergence in sympatric populations (Via 2009; Feder *et al.* 2012; Via 2012), although both have essential effects on spatial genetic structures. The relative strength of each determines the extent of genetic divergence at individual loci and eventually in populations and sometimes causes reproductive isolation or speciation. In the current study, I revealed heterogeneous genetic divergence within a genome of *M. polymorpha*; that is, the extent and drivers of genetic divergence differ among individual loci.

Population differentiation across the entire set of genome-wide markers was better explained by the IBD model than other models (Table 3-2); gene flow among populations was mainly determined by distance. As much as 89.15% of the total genetic variance in all SNPs was contained within populations (Fig. 3-7b), indicating alleles were largely admixed in most of the genome. This species is bird and insect pollinated (Kitayama *et al.* 1997) and produces a large number of seeds dispersed by the wind (Drake 1992). In addition to the high dispersal ability of this species, the absence of geographic barriers (Aradhya *et al.* 1993) could lead to large gene flow among populations across a wide range of environmental conditions. Large gene flow must lead to a high rate of recombination across a genome, which could prevent genetic differentiation among populations (Feder *et al.* 2012). In contrast, a small fraction of the genome diverged among populations under an IBE scenario (Table 3-2). The larger differentiation at the outlier loci compared with the genome-wide loci was indicated by the difference in F_{ST} values of approximately a factor of 3 (Fig. 3-7a). Thus, at the outlier loci, strong selection could drive genetic divergence and overcome the homogenizing effects of gene flow among populations. These differentiation patterns across a genome correspond to the concept of "genomic mosaics," in which natural selection, migration, and genetic drift have different influences on the individual loci (Hohenlohe *et al.* 2010; Feder *et al.* 2012; Gompert *et al.* 2012; Nosil *et al.* 2008). In the case of *M. polymorpha*, gene flow and selection cause divergence at neutral and putative adaptive loci, respectively.

When the genomic mosaics found in this species are considered in the context of speciation, the genome of *M. polymorpha* is likely in the very early stage of speciation. Feder et al. (2012) proposed a scenario describing genomic divergence patterns in face of gene flow (the speciation-with-gene-flow scenario). According to this scenario, the process of genomic divergence from population differentiation to speciation can be divided into four phases: phase 1, natural selection initially acts on and decreases recombination at adaptive loci; phase 2, recombination decreases not only at adaptive loci but also at nearby loci; phase 3, the decrease in recombination expands to loci far from the adaptive loci; and phase 4, a large part of the genome, including adaptive and neutral genes, is largely differentiated (Feder et al. 2012). In the case of *M. polymorpha*, the small proportion of outlier loci (35 of 2,247 SNPs; Fig. 3-5) in addition to the large gene flow across the genome (approximately 90% of the total genomic variance was contained within populations; Fig. 3-7) indicates that the genome of this species is

likely in the very early stage of speciation, corresponding to phase 1 or the beginning of phase 2 in the speciation-with-gene-flow scenario. The relative lower genomic divergence of this species contrasts with the dramatic divergence in phenotype which covers almost the entire range of global phenotypic variation in woody species (Tsujii et al. 2015). However, a limited number of lcoi may involve a reproductive isolation resulting in speciation if the loci have a large effect on functional phenotype (Lexer et al. 2003; Kronforst et al. 2006; Barr & Fishman 2010).

Overall, I found the genomic mosaic of *M. polymorpha*, which consists of loci with contrasting divergence patterns. AMOVA showed that approximately 90% of the total genetic variance was retained within populations (Fig. 3-7b), and each SNP showed an average F_{ST} of 0.15 (Fig. 3-5b), indicating that the genome was largely mixed among populations. Meanwhile, large phenotypic variations were maintained, even in the common garden, and significant divergences in accordance with an environmental gradient were found at a limited number of SNPs (Table 3-2; Fig. 3-6). These observations suggest that a small fraction of the genome could be subject to environmental selection that contributes to this phenotypic divergence. To understand and reconstruct the genetic processes underlying the adaptation of *M. polymorpha* in the Hawaiian Islands, the genomic architecture of this species should be studied to reveal the physical linkages between adaptive genes (Nadeau *et al.* 2011) and the distribution of adaptive genes across chromosomes (Guo *et al.* 2015; Bradbury *et al.* 2010). Such information could tell us how many and what genes were involved in the expansion of ecological niches of this species.

Tables and Figures

Table 3-1
Summary of the 9 source populations of the 72 *M. polymorpha* individuals analyzed.

Population	Elevation (m)	Lava age (years)	*N*	Leaf area (cm^2) (mean ± SD) *	Weight of trichome ($mg \cdot cm^{-2}$) (mean ± SD) *
O150	150	3000	10	13.0 ± 5.5	1.2 ± 1.9
Y150	150	150	7	10.9 ± 2.2	1.6 ± 0.9
O700	700	3000	7	16.2 ± 5.1	1.0 ± 1.5
Y700	700	150	12	12.6 ± 3.4	0.6 ± 2.4
O1200	1200	3000	8	7.3 ± 1.2	1.7 ± 1.8
Y1200	1200	150	12	7.2 ± 3.3	1.5 ± 2.2
O1800	1800	3000	4	3.9 ± 1.1	8.3 ± 2.5
Y1800	1800	150	6	4.9 ± 0.7	5.9 ± 1.5
O2400	2400	3000	6	4.4 ± 1.4	7.6 ± 2.0

N, number of plant individuals analyzed; *Data from Tsujii *et al.* (2015)

Table 3-2

Results of generalized linear mixed models predicting the genetic differentiation between populations of *M. polymorpha* with geographic distance and an environmental distance. The best model was shown in bold.

	F_{ST} of all SNP genotypes			F_{ST} of outlier SNP genotypes		
Model	DIC	Delta DIC	DIC weight	DIC	Delta DIC	DIC weight
NULL	-108.11	23.31	0.00	2.36	28.22	0.00
IBD	**-131.42**	**0.00**	**0.46**	-20.72	5.14	0.05
IBE	-130.68	0.74	0.32	**-25.86**	**0.00**	**0.70**
IBD+IBE	-130.02	1.40	0.23	-23.80	2.06	0.25

NULL, model with no variable; IBD, model with geographic distance; IBE, model with an environmental distance; IBD+IBE, model with geographic distance and an environmental distance; DIC, the deviance information criteria; Delta DIC, Difference of DIC from the best model

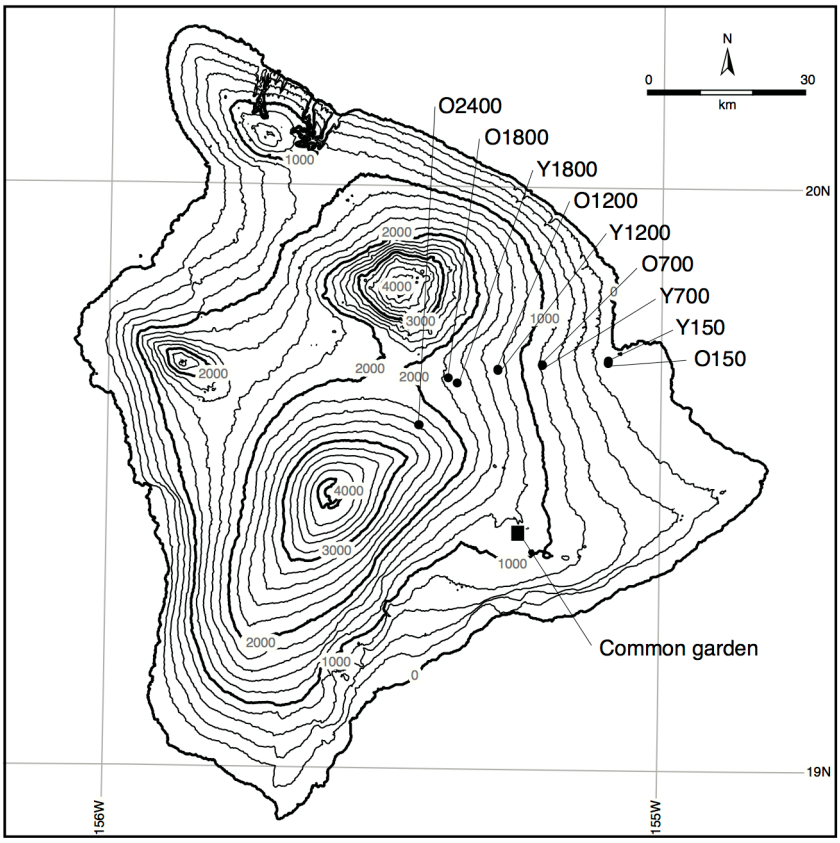

Fig. 3-2

The location of the common garden at the Volcano Agriculture Station, University of Hawaii and the nine original seed sources of the 72 *Metrosideros polymorpha* trees analyzed in this study

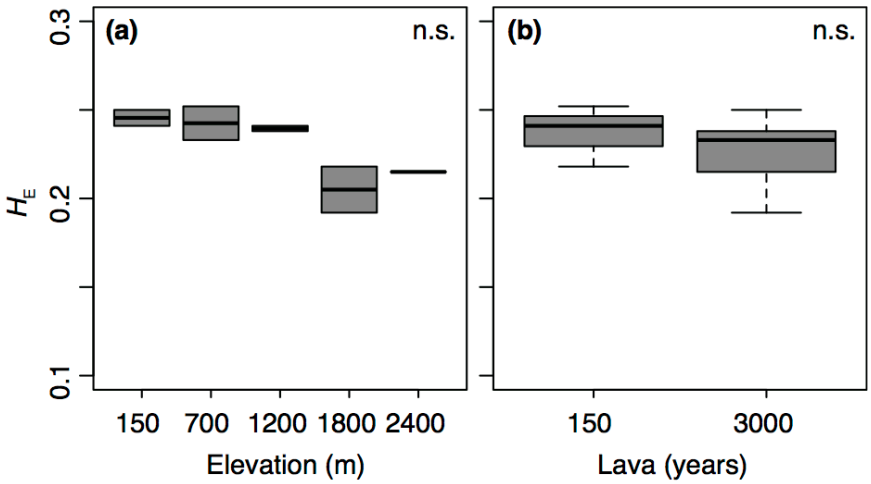

Fig. 3-3
Population level heterozygosity in nine populations of *Metrosideros polymorpha* trees on Mauna Loa. Boxplots (mean, range, and upper/lower quartiles) show the expected heterozygosities based on 2,247 genome-wide single nucleotide polymorphism genotypes against (**a**) elevation and (**b**) lava age.

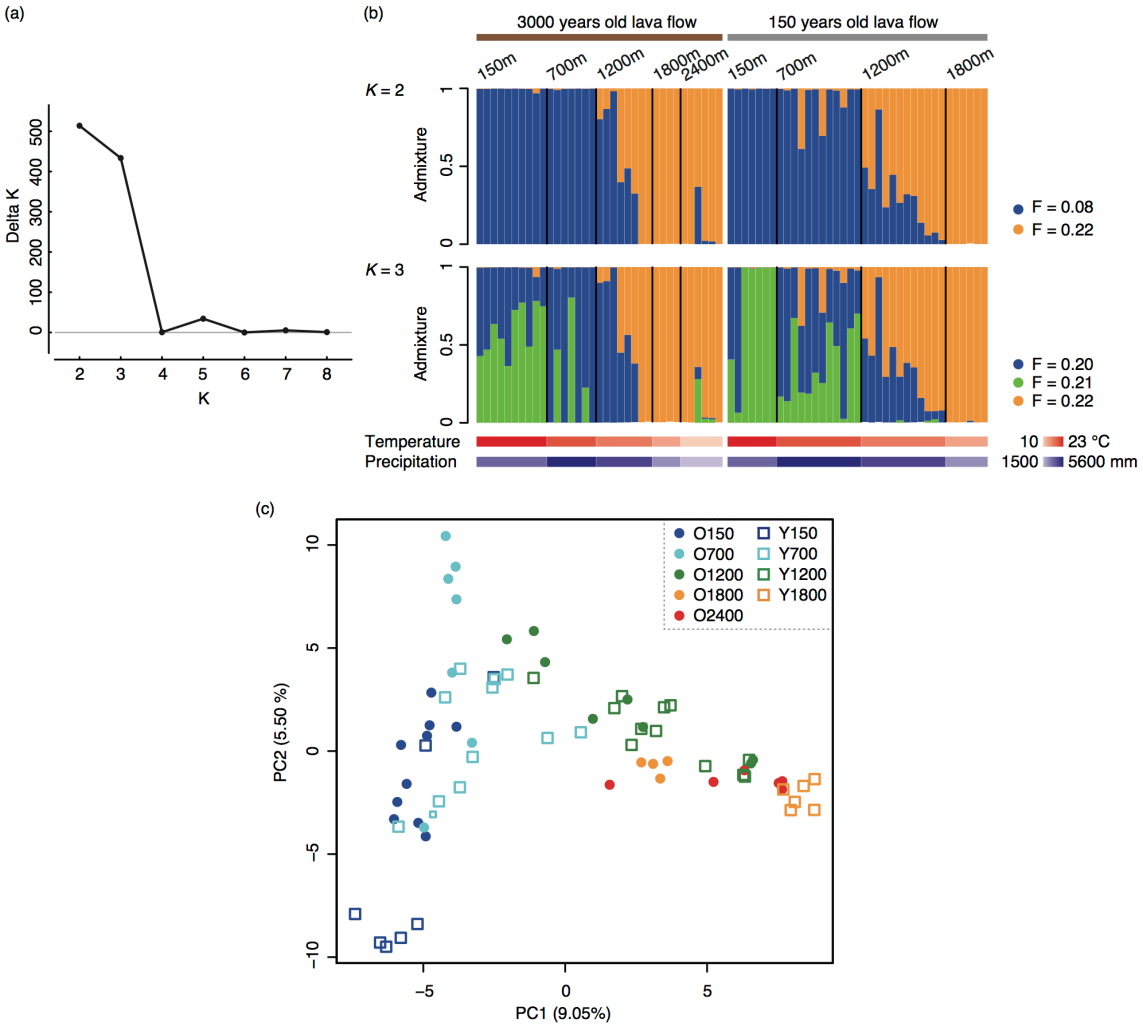

Fig. 3-4

Spatial genetic admixture of 72 *Metrosideros polymorpha* trees across nine populations on Mauna Loa based on genotypes at the 2,247 single nucleotide polymorphisms (SNPs). (**a**) Plot of delta *K* as a function of the number of genetic clusters (*K*) according to Evanno *et al.* (2005). (**b**) Admixture proportions of genetic clusters in individual trees. Bar plots are shown for a scenario of *K* = 2 and *K* = 3. Population profiles for mean annual temperature and mean annual precipitation are also shown. (**c**) Principal component analysis of genotypes at the 2,247 SNPs for the 72 *M. polymorpha* trees. Population profiles are shown in Table 1.

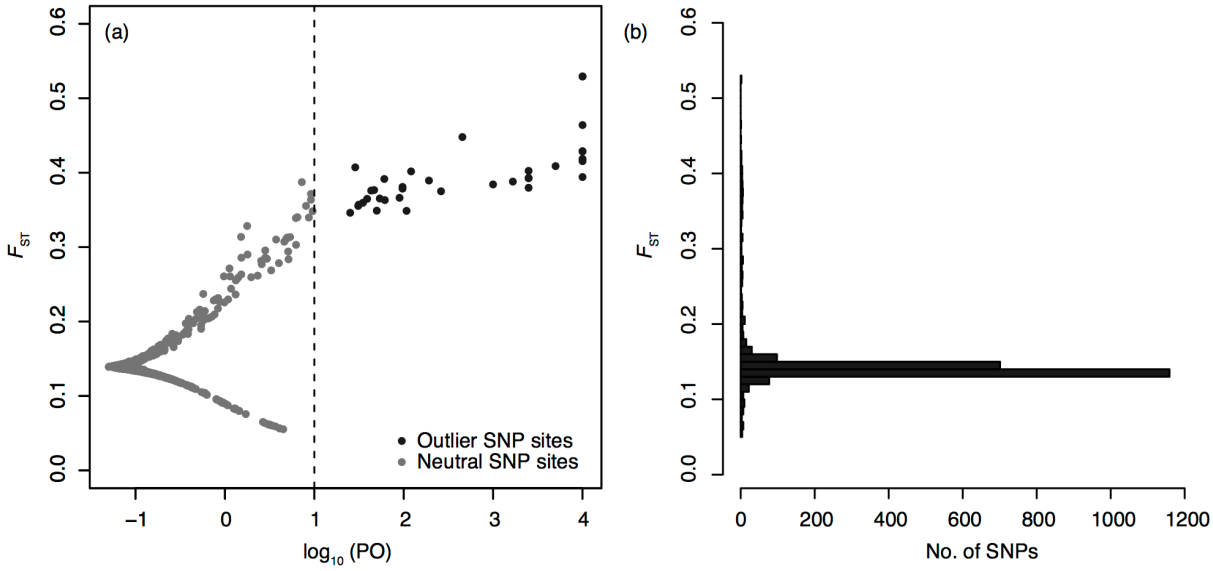

Fig. 3-5

Results of Bayesian outlier analysis for 2,247 single nucleotide polymorphisms (SNPs) in *Metrosideros polymorpha*. (**a**) For each SNP, F_{ST} was plotted against $\log_{10}$ (posterior odds [POs]). SNPs with $\log_{10}$ (PO) > 1 were recognized as outlier SNPs (shown in black). (**b**) Distribution of F_{ST} across the 2,247 SNPs.

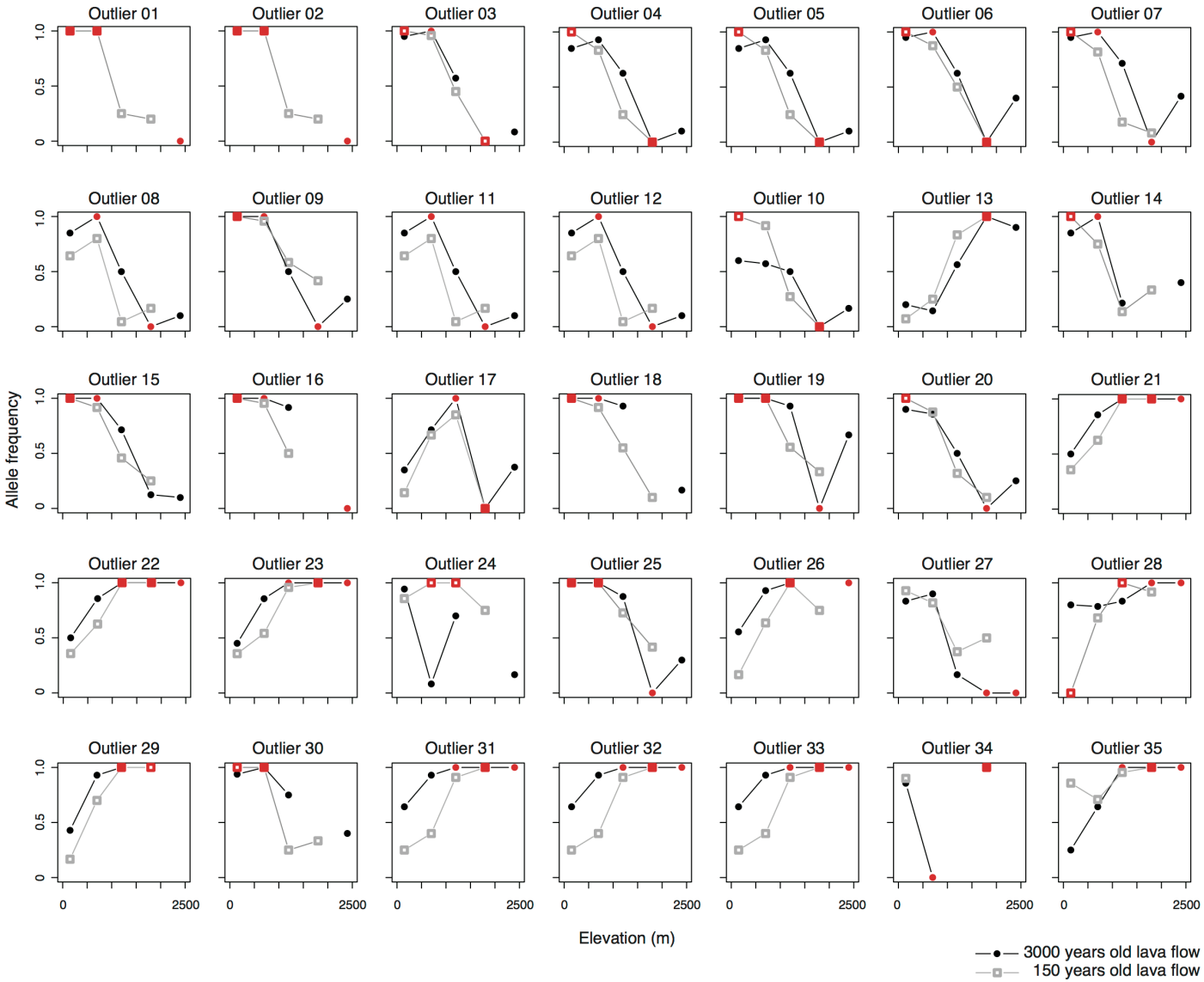
Outlier 01
Outlier 02
Outlier 03
Outlier 04
Outlier 05
Outlier 06
Outlier 07
Outlier 08
Outlier 09
Outlier 11
Outlier 12
Outlier 10
Outlier 13
Outlier 14
Outlier 15
Outlier 16
Outlier 17
Outlier 18
Outlier 19
Outlier 20
Outlier 21
Outlier 22
Outlier 23
Outlier 24
Outlier 25
Outlier 26
Outlier 27
Outlier 28
Outlier 29
Outlier 30
Outlier 31
Outlier 32
Outlier 33
Outlier 34
Outlier 35
Allele frequency
0
0.5
1.0
0
2500
Elevation (m)
3000 years old lava flow
150 years old lava flow

Fig. 3-6

Allele frequencies at the 35 outlier single nucleotide polymorphism loci in each of the nine populations of *Metrosideros polymorpha* on Mauna Loa. Black circles and gray squares indicate the allele frequencies in the populations on 3,000- and 150-year-old lava flows, respectively. Allele frequencies of 0 or 1 are indicated in red.

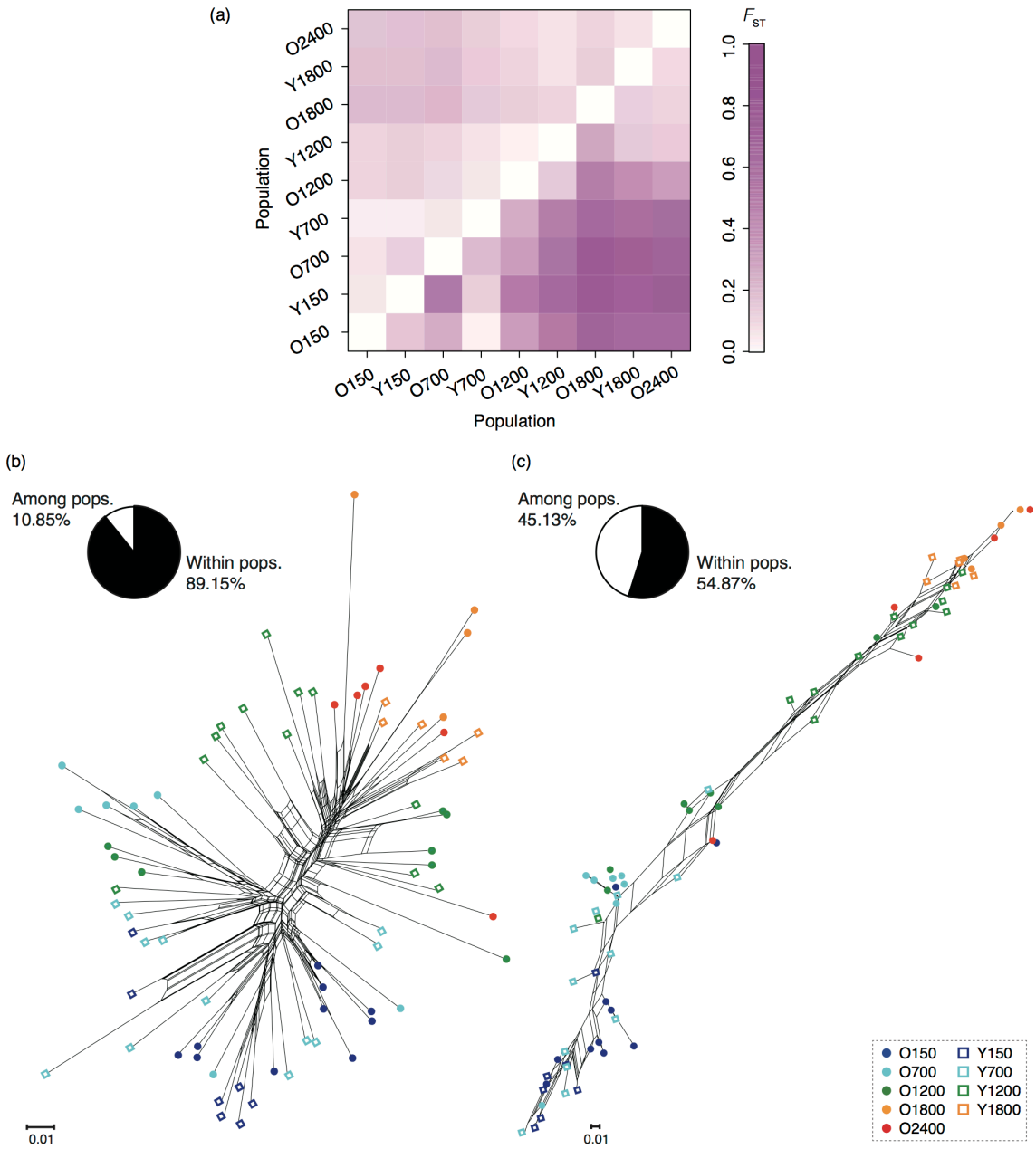

Fig. 3-7

Population and individual differentiation at genome-wide 2,247 and 35 outlier single nucleotide polymorphisms (SNPs) among 72 individuals of *Metrosideros polymorpha* on Mauna Loa. (**a**) Pairwise F_{ST} values between the nine populations based on the genome-wide 2,247 (above the diagonal) and the 35 outlier (below the diagonal) SNP genotypes. Population profiles are shown in Table 1. (**b** and **c**) Phylogenetic network of the 72 individuals from the nine populations based on the *p*-distance at (**b**) genome-wide 2,247 and (**c**) the 35 outlier SNPs. Inset pie charts indicate the proportion of genetic variance among and within populations calculated using analysis of molecular variance.

Supplemental materials

Table 3-S1
Oligonucleotide sequences of adapters and primers used for RAD-seq libraries

Primer / Adapter name	Sequence
TruSeq_EcoRI_adaptor1	/5Phos/A*A*TTGAGATCGGAAGAGCACACG TCTGAACTCCAGTC*A*C
TruSeq_EcoRI_adaptor2	G*T*CAAGTTTCACAGCTCTTCCGATC*T*C
TruSeq_BglII_adaptor1	A*A*TGATACGGCGACCACCGAGATCTACA CTCTTTCCCTACACGACGCTCTT*C*C
TruSeq_BglII_adaptor2	G*A*TCGGAAGAGCTGTGCAGA*C*T
TruSeq_Univ_primer	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGA
TruSeq_IP001_xxxxxx	CAAGCAGAAGACGGCATACGAGATXXXXX XGTGACTGGAGTTCAGACGTGT

*: phosphorothioate bond
XXXXXX: 6mer index sequence

Table 3-S2

Twenty climate variables for the nine populations of *Metrosideros polymorpha* trees on Mauna Loa comprised of 19 WorldClim variables (BIO1-19) plus one variable of lava age

Pop.	Latitude	Longitude	BIO 1	BIO 2	BIO 3	BIO 4	BIO 5	BIO 6	BIO 7	BIO 8	BIO 9	BIO 10	BIO 11	BIO 12	BIO 13	BIO 14	BIO 15	BIO 16	BIO 17	BIO 18	BIO 19	Lava
O150	19.70	-155.10	226	85	72	1063	284	167	117	214	229	239	212	3921	425	197	22	1196	796	857	1049	3000
Y150	19.70	-155.10	226	85	72	1046	285	168	117	214	231	239	213	3862	417	194	22	1177	782	851	1034	150
O700	19.70	-155.22	187	87	71	1128	249	127	122	175	189	202	173	3898	459	168	28	1256	737	768	1113	3000
Y700	19.69	-155.22	187	87	71	1128	249	127	122	175	189	202	173	3898	459	168	28	1256	737	768	1113	150
O1200	19.69	-155.30	157	89	69	1210	223	95	128	144	158	173	142	2219	289	73	36	758	352	390	715	3000
Y1200	19.69	-155.30	157	89	69	1210	223	95	128	144	158	173	142	2219	289	73	36	758	352	390	715	150
O1800	19.67	-155.39	130	91	70	1218	196	67	129	117	131	146	115	1811	223	70	32	587	320	344	513	3000
Y1800	19.66	-155.37	132	91	70	1215	198	69	129	119	142	148	117	1983	247	78	32	644	347	373	560	150
O2400	19.59	-155.44	103	93	75	1187	162	38	124	95	111	116	87	1020	127	32	34	340	154	163	330	3000

BIO1 = Annual Mean Temperature; BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)); BIO3 = Isothermality (BIO2/BIO7) (* 100); BIO4 = Temperature Seasonality (standard deviation *100); BIO5 = Max Temperature of Warmest Month; BIO6 = Min Temperature of Coldest Month; BIO7 = Temperature Annual Range (BIO5-BIO6); BIO8 = Mean Temperature of Wettest Quarter; BIO9 = Mean Temperature of Driest Quarter; BIO10 = Mean Temperature of Warmest Quarter; BIO11 = Mean Temperature of Coldest Quarter; BIO12 = Annual Precipitation; BIO13 = Precipitation of Wettest Month; BIO14 = Precipitation of Driest Month; BIO15 = Precipitation Seasonality (Coefficient of Variation); BIO16 = Precipitation of Wettest Quarter; BIO17 = Precipitation of Driest Quarter; BIO18 = Precipitation of Warmest Quarter; BIO19 = Precipitation of Coldest Quarter

Table 3-S3
The number of raw RAD-seq reads and mapped reads in a draft genome sequence for each individual

Sample ID	Population	Read count	Read count in BAM file	Mappig rate	Average reads per scaffold
Mpo01	O150	1,636,565	1,475,448	0.90	24
Mpo02	O150	5,874,215	5,617,654	0.96	93
Mpo03	O150	2,770,669	2,662,236	0.96	44
Mpo04	O150	5,068,203	4,914,074	0.97	81
Mpo05	O150	5,027,159	4,807,874	0.96	79
Mpo06	O150	2,747,252	2,570,616	0.94	42
Mpo07	O150	2,368,356	2,279,057	0.96	37
Mpo08	O150	10,656,592	10,205,544	0.96	169
Mpo10	O150	1,923,864	1,855,067	0.96	30
Mpo11	O150	2,870,881	2,778,984	0.97	46
Mpo13	O700	12,171,850	11,571,282	0.95	192
Mpo17	O700	1,581,610	1,502,276	0.95	24
Mpo18	O700	2,557,945	2,458,505	0.96	40
Mpo19	O700	1,661,796	1,582,951	0.95	26
Mpo20	O700	2,575,593	2,463,402	0.96	40
Mpo22	O700	3,740,743	3,595,201	0.96	59
Mpo23	O700	1,355,654	1,296,163	0.96	21
Mpo25	O1200	1,116,373	1,063,108	0.95	17
Mpo26	O1200	5,207,756	4,916,126	0.94	81
Mpo27	Y700	5,683,491	5,363,745	0.94	89
Mpo29	O1200	725,505	690,632	0.95	11
Mpo30	O1200	1,211,230	1,162,510	0.96	19
Mpo31	O1200	301,385	276,179	0.92	4
Mpo32	O1200	1,608,138	1,491,265	0.93	24
Mpo33	O1200	3,266,094	3,152,959	0.97	52
Mpo34	O1200	1,366,579	1,300,290	0.95	21
Mpo35	O2400	1,594,670	1,278,126	0.80	21
Mpo38	O1800	274,223	252,535	0.92	4
Mpo39	O1800	576,437	542,909	0.94	9
Mpo40	O1800	484,296	440,887	0.91	7
Mpo41	O1800	237,678	209,137	0.88	3
Mpo42	O2400	3,630,701	2,475,653	0.68	41
Mpo43	O2400	450,383	420,900	0.93	6
Mpo45	O2400	1,533,894	1,426,151	0.93	23
Mpo47	O2400	989,738	871,903	0.88	14
Mpo48	O2400	1,416,940	1,351,829	0.95	22
Mpo51	Y150	6,471,194	6,170,448	0.95	102
Mpo52	Y150	999,635	945,178	0.95	15
Mpo55	Y150	8,597,656	8,309,826	0.97	138

Mpo56	Y150	12,895,885	12,549,349	0.97	208
Mpo59	Y150	11,331,115	10,930,863	0.96	181
Mpo60	Y150	6,422,302	6,194,333	0.96	102
Mpo61	Y150	3,070,875	2,946,100	0.96	48
Mpo62	Y700	2,143,418	2,060,130	0.96	34
Mpo63	Y700	9,727,253	9,185,075	0.94	152
Mpo64	Y700	7,602,905	7,312,787	0.96	121
Mpo65	Y700	6,052,738	5,754,681	0.95	95
Mpo66	Y700	2,950,568	2,799,044	0.95	46
Mpo67	Y700	2,149,200	2,064,076	0.96	34
Mpo68	Y700	1,248,566	1,168,233	0.94	19
Mpo69	Y700	1,120,097	1,081,819	0.97	17
Mpo70	Y700	1,355,422	1,289,719	0.95	21
Mpo73	Y700	1,124,798	1,070,158	0.95	17
Mpo74	Y700	6,039,110	5,845,232	0.97	97
Mpo75	Y1200	1,507,731	1,452,090	0.96	24
Mpo76	Y1200	3,219,785	3,049,967	0.95	50
Mpo78	Y1200	752,355	680,687	0.90	11
Mpo79	Y1200	763,691	708,438	0.93	11
Mpo80	Y1200	779,686	734,733	0.94	12
Mpo81	Y1200	689,531	642,614	0.93	10
Mpo82	Y1200	2,039,482	1,952,402	0.96	32
Mpo83	Y1200	876,874	793,027	0.90	13
Mpo84	Y1200	2,037,047	1,925,187	0.95	31
Mpo85	Y1200	2,093,140	1,997,772	0.95	33
Mpo87	Y1200	6,016,874	5,742,293	0.95	95
Mpo88	Y1200	2,081,449	1,919,116	0.92	31
Mpo89	Y1800	5,629,235	5,323,697	0.95	88
Mpo90	Y1800	5,523,977	5,307,141	0.96	88
Mpo92	Y1800	2,410,689	2,327,252	0.97	38
Mpo93	Y1800	2,696,553	2,501,574	0.93	41
Mpo94	Y1800	1,321,844	1,227,526	0.93	20
Mpo95	Y1800	7,211,232	6,825,966	0.95	113
Total		237,218,370	225,111,711	0.95	
Mean		3,294,700	3,126,552	0.94	51.43

Table 3-S4
Gene ontology results for the 83 putative genes located within 10 kb of the 35 outlier single nucleotide polymorphisms

	Scaffold ID	Outlier SNP ID	Outlier SNP positions	Gene ID	Start position	End position	Distance	Top-Hit Species	Sequence Description	Blast E-Value Min	Blast Similarity Mean	GOs
1	scaffold6837_len146858_cov183	*Outlier01; Outlier02*	118182, 118205	g27367	110198	111431	7,368		---NA---		-	-
				g27368	113226	114498	4,320	*Eucalyptus grandis*	probable membrane-associated kinase regulator 4	4.9.E-115	61.75	P:response to cyclopentenone; C:plasmodesma; P:response to auxin; P:biological_process; P:response to brassinosteroid
				g27369	116665	118751	474	*Eucalyptus grandis*	exocyst complex component exo70a1-like	0.0.E+00	68.80	C:cytoplasm; P:cellular process; P:transport
				g27370	123568	125551	6,378	*Cicer arietinum*	60s ribosomal protein l35	2.0.E-53	79.55	F:structural molecule activity; C:ribosome; P:translation
2	scaffold6535_len66485_cov231	*Outlier03*	47247	g26510	39403	41341	6,875		---NA---		-	-
				g26511	43470	47131	1,947	*Vitis vinifera*	catalytic region zinc cchc-type peptidase catalytic	1.4.E-106	61.55	F:binding
				g26512	51291	54131	5,464	*Eucalyptus grandis*	PREDICTED: uncharacterized protein LOC104443196	4.1.E-91	67.85	-
				g26513	54187	56921	8,307	*Eucalyptus*	PREDICTED:	8.4.E-151	62.90	-

								grandis	uncharacterized protein LOC104445904			
3	scaffold 3384_len 41585_c ov188	*Outlier04; Outlier05; Outlier13*	24673, 24686, 24699	g15853	15246	17400	8,350	*Eucalyptus grandis*	tubulin beta-1 chain	0.0.E+00	91.05	C:Golgi apparatus; P:biosynthetic process; F:nucleotide binding; F:hydrolase activity; C:membrane; P:cellular process; P:carbohydrate metabolic process; F:structural molecule activity; P:catabolic process; C:cytoskeleton; P:protein metabolic process; P:cellular component organization
				g15854	18106	19141	6,050	*Eucalyptus grandis*	par1 protein	8.0.E-105	77.20	P:biological_process; P:transition metal ion transport
				g15855	23076	24127	1,072	*Eucalyptus grandis*	par1 protein	8.2.E-86	80.65	P:transport
				g15856	27006	30025	3,843	*Eucalyptus grandis*	mitochondrial import inner membrane translocase subunit tim10	1.7.E-42	93.35	C:mitochondrion; P:nucleobase-containing compound metabolic process; P:biosynthetic process; P:cellular process; F:binding; P:transport; P:cellular component organization
				g15857	30586	31466	6,353	*Eucalyptus grandis*	kda class i heat shock	2.8.E-76	53.80	P:response to oxidative stress
				g15858	31614	34061	8,165	*Eucalyptus grandis*	udp-glycosyltransferase 76e2-like	0.0.E+00	72.15	P:metabolic process; F:transferase activity
4	scaffold 1062_len 78411_c ov188	*Outlier06*	11282	g5286	1457	8301	6,403	*Eucalyptus grandis*	structural maintenance of chromosomes protein 2-1-like	0.0.E+00	88.80	C:intracellular; P:cell cycle; F:nucleotide binding; C:nucleus; P:cellular component organization
				g5287	8486	9461	2,309		---NA---		-	-

				g5288	10323	16991	2,375	*Eucalyptus grandis*	no-associated protein chloroplastic mitochondrial-like	0.0.E+00	73.85	F:nucleotide binding; P:metabolic process; F:catalytic activity
5	scaffold 23842_len12745_cov196	*Outlier07*	10140	g48354	936	1760	8,792		---NA---		-	-
				g48355	12117	12681	2,259		---NA---		-	-
6	scaffold 12346_len41828_cov184	*Outlier08; Outlier11: Outlier12*	24048, 24051, 24079	g38900	17686	23733	3,339	*Eucalyptus grandis*	protein nynrin-like	0.0.E+00	80.95	-
				g38901	23760	29841	2,753	*Eucalyptus grandis*	low quality protein: beta-galactosidase	0.0.E+00	86.45	F:hydrolase activity; P:carbohydrate metabolic process; F:carbohydrate binding; C:extracellular region
7	scaffold 19372_len48733_cov183	*Outlier09*	16675	g46083	8477	20731	2,071	*Populus euphratica*	e3 ubiquitin-protein ligase keg isoform x2	0.0.E+00	91.75	F:protein binding; P:response to endogenous stimulus; C:Golgi apparatus; P:cellular process; P:response to stress; P:catabolic process; F:catalytic activity; P:signal transduction; P:nucleobase-containing compound metabolic process; F:nucleotide binding; F:kinase activity; F:binding; P:transport; P:growth; F:transferase activity; P:cellular protein modification process
				g46084	21536	23151	5,669	*Eucalyptus grandis*	e3 ubiquitin-protein ligase keg isoform x1	2.2.E-125	77.70	F:binding; F:transferase activity
				g46085	24006	24712	7,684	*Eucalyptus grandis*	PREDICTED: uncharacterized protein	9.1.E-06	69.00	-

									LOC104449813			
				g46086	26229	27071	9,975	*Eucalyptus grandis*	sigma factor binding protein chloroplastic-like	1.5.E-73	63.85	-
8	scaffold 3364_len 126043_cov208	*Outlier10*	101995	g15778	91626	97401	7,482	*Eucalyptus grandis*	PREDICTED: uncharacterized protein LOC104451536	0.0.E+00	66.55	F:nucleic acid binding; P:DNA integration; C:nucleus; P:proteolysis; F:aspartic-type endopeptidase activity; F:zinc ion binding
				g15779	99536	101222	1,616	*Eucalyptus grandis*	cytochrome p450 87a3-like	2.3.E-98	59.20	-
				g15780	102286	106294	2,295	*Eucalyptus grandis*	vacuolar-processing enzyme-like	7.4.E-56	61.85	F:nucleic acid binding; P:oxidation-reduction process; F:oxidoreductase activity; P:DNA integration; F:zinc ion binding
				g15781	106426	107603	5,020	*Eucalyptus grandis*	PREDICTED: uncharacterized protein LOC104426629	1.8.E-110	67.15	C:plastid; P:biological_process
				g15782	109754	111591	8,678	*Eucalyptus grandis*	calcium-dependent lipid-binding family	3.1.E-11	71.50	C:integral component of membrane; P:magnesium ion transport; F:magnesium ion transmembrane transporter activity
9	scaffold 5210_len 53638_cov189	*Outlier14*	6380	g22259	356	3470	4,467	*Eucalyptus grandis*	antifungal protein ginkbilobin-2-like	1.1.E-48	57.80	-
				g22260	9106	10111	3,229		---NA---		-	-
				g22261	10896	11591	4,864	*Eucalyptus grandis*	antifungal protein ginkbilobin-2-like	1.3.E-54	59.00	-
				g22262	12373	15641	7,627		---NA---		-	-
10	scaffold	*Outlier15*	574	g53047	17	1806	338	*Eucalyptus*	protein nynrin-like	0.0.E+00	70.00	-

	52451_len7320_cov204							*grandis*				
				g53048	4536	7130	5,259		---NA---		-	-
11	scaffold8545_len200286_cov200	*Outlier16*	123225	g31723	116046	119440	5,482	*Eucalyptus grandis*	glutamine amidotransferase	0.0.E+00	78.65	F:hydrolase activity; P:cellular process; P:metabolic process; P:response to stress; P:DNA metabolic process; F:transferase activity
				g31724	120196	124984	635	*Eucalyptus grandis*	transcription factor bim2	0.0.E+00	74.55	F:protein binding
				g31725	127724	138151	9,713	*Eucalyptus grandis*	prolyl endopeptidase-like	0.0.E+00	66.10	F:hydrolase activity
12	scaffold3900_len105384_cov222	*Outlier17*	50888	g17905	48446	53638	154	*Glycine max*	gag-pol polyprotein	7.7.E-38	63.05	F:nucleic acid binding; P:DNA integration; F:zinc ion binding; F:metal ion binding
				g17906	54836	57624	5,342		---NA---		-	-
				g17907	57790	60751	8,383	*Eucalyptus grandis*	probable leucine-rich repeat receptor-like serine threonine-protein kinase at3g14840	3.2.E-15	60.00	-
13	scaffold715_len105575_cov227	*Outlier18*	21345	g3512	8218	20481	6,996	*Vitis vinifera*	retrotransposon unclassified	3.3.E-158	68.95	F:binding
				g3513	21996	25089	2,198	*Vitis vinifera*	hypothetical protein VITISV_029829	3.6.E-153	56.80	P:metabolic process; F:nucleic acid binding
14	scaffold3353_len35297_c	*Outlier19*	1656	g15717	2865	5333	2,443		---NA---		-	-

	ov230										
				g15718	7386	7808	5,941	*Eucalyptus grandis*	disease resistance protein rga4	9.9.E-27	80.25	F:nucleotide binding
				g15719	10237	11641	9,283		---NA---		-	-
15	scaffold 19865_len10737_cov182	*Outlier20*	6780	g46389	2434	4531	3,298	*Eucalyptus grandis*	lamin-like protein	5.8.E-56	77.50	F:molecular_function
				g46390	4936	6146	1,239		---NA---		-	-
				g46391	6244	9851	1,268	*Eucalyptus grandis*	chorismate mutase chloroplastic	8.4.E-157	86.30	P:biosynthetic process; P:cellular process; P:carbohydrate metabolic process; P:metabolic process; P:catabolic process; F:catalytic activity; C:cytosol
16	scaffold 23186_len15195_cov161	*Outlier21; Outlier22; Outlier23*	9780, 9908,9915	g48036	1856	5542	6,081	*Eucalyptus grandis*	heat shock protein with tetratricopeptide repeat isoform 1	0.0.E+00	80.50	-
				g48037	13573	15061	4,537	*Phoenix dactylifera*	triosephosphate chloroplastic-like	5.0.E-16	87.60	P:metabolic process; F:catalytic activity
17	scaffold 5055_len44971_cov137	*Outlier24*	23942	g21760	15027	24001	4,428	*Eucalyptus grandis*	probable disease resistance protein at4g27220	0.0.E+00	64.45	F:nucleotide binding
				g21761	25509	31621	4,623	*Eucalyptus grandis*	cytochrome p450 716b1-like	0.0.E+00	84.45	F:binding; P:metabolic process; F:catalytic activity
18	scaffold 6055_len58982_cov200	*Outlier25*	29422	g24957	18781	20371	9,846	*Eucalyptus grandis*	protein mizu-kussei 1	1.9.E-136	86.70	-
				g24958	22448	24841	5,778	*Eucalyptus*	pentatricopeptide	0.0.E+00	72.80	P:multicellular organismal

							grandis	repeat-containing protein at4g28010			development; P:flower development; P:anatomical structure morphogenesis
				g24959	25086	35656	949	*Eucalyptus grandis*	retinoblastoma-related protein	0.0.E+00	78.55	P:cell cycle; P:nucleobase-containing compound metabolic process; P:biosynthetic process; C:nucleus
				g24960	36341	39591	8,544	*Eucalyptus grandis*	3 -5 exoribonuclease 1-like	0.0.E+00	74.00	P:nucleobase-containing compound metabolic process; F:nuclease activity; F:nucleic acid binding
19	scaffold10239_len37826_cov194	*Outlier26*	4876	g35373	529	1361	3,931		---NA---		-	-
				g35374	2336	4060	1,678	*Vitis vinifera*	probable s-adenosylmethionine-dependent methyltransferase at5g37990	3.2.E-94	59.00	F:methyltransferase activity; P:methylation; F:transferase activity
				g35375	4190	4761	401	*Eucalyptus grandis*	mediator of rna polymerase ii transcription subunit	3.1.E-84	67.30	-
				g35376	6840	9011	3,050	*Cucumis sativus*	carboxypeptidase 2	2.5.E-34	78.35	F:hydrolase activity; P:protein metabolic process
				g35377	9166	11675	5,545	*Eucalyptus grandis*	ell-associated factor 1	8.8.E-152	66.55	P:nucleobase-containing compound metabolic process; P:biosynthetic process; C:nucleoplasm
				g35378	12296	15905	9,225	*Eucalyptus grandis*	heavy metal-associated isoprenylated plant protein 26	1.3.E-96	90.35	F:binding; P:transport
20	scaffold1349_len	*Outlier27*	56021	g6766	42918	49321	9,902	*Eucalyptus grandis*	exocyst complex component	0.0.E+00	65.50	-

	175907_cov192							exo84b-like			
				g6767	50406	59806	915	*Eucalyptus grandis*	phospholipid--sterol o-acyltransferase isoform x1	0.0.E+00	85.20	P:lipid metabolic process; P:cellular process; P:catabolic process; P:multicellular organismal development; F:transferase activity
				g6768	59811	70831	9,300	*Eucalyptus grandis*	ribonuclease e g-like chloroplastic isoform x1	0.0.E+00	73.15	P:nucleobase-containing compound metabolic process; F:nuclease activity; F:RNA binding; F:carbohydrate binding
21	scaffold613_len15745_cov148	*Outlier28*	748	g3074	958	6241	2,852		---NA---		-	-
				g3075	6398	7401	6,152		---NA---		-	-
				g3076	8440	11731	9,338	*Eucalyptus grandis*	probable proteasome inhibitor	5.0.E-72	72.55	C:proteasome complex
22	scaffold3856_len124483_cov228	*Outlier29*	107794	g17724	98533	103141	6,957	*Eucalyptus grandis*	dna rna polymerases superfamily protein	0.0.E+00	62.85	P:DNA metabolic process; F:nucleic acid binding
				g17725	103249	105701	3,319		---NA---		-	-
				g17726	109242	118481	6,068	*Eucalyptus grandis*	subtilisin-like protease	0.0.E+00	94.85	F:hydrolase activity; C:membrane; C:cell wall; P:protein metabolic process
23	scaffold13732_len54471_cov185	*Outlier30*	14029	g40749	1	1å4011	7,023	*Eucalyptus grandis*	cell division cycle protein 27 homolog b-like isoform x2	0.0.E+00	86.90	P:response to endogenous stimulus; P:biosynthetic process; P:cellular process; P:response to stress; P:catabolic process; P:multicellular organismal development; P:DNA metabolic process; P:embryo development; P:post-embryonic

											development; P:anatomical structure morphogenesis; P:cell differentiation; C:cytoplasm; P:cell cycle; P:response to abiotic stimulus; P:cell growth; C:cytoskeleton; P:protein metabolic process; P:reproduction; P:cellular component organization
				g40750	14916	15885	1,372	*Eucalyptus grandis*	uncharacterized atp-dependent helicase	2.0.E-12	87.00	F:metal ion binding; F:ATP binding; F:zinc ion binding; F:DNA binding
				g40751	17046	28568	8,778	*Eucalyptus grandis*	uncharacterized atp-dependent helicase	0.0.E+00	79.50	F:hydrolase activity; F:nucleotide binding; F:binding; F:DNA binding
24	scaffold 288_len36095_cov167	*Outlier31; Outlier32; Outlier33*	211, 216, 219	g1347	45	3291	1,457		---NA---		-	-
				g1348	5866	10396	7,920	*Eucalyptus grandis*	alpha beta-hydrolases superfamily protein	0.0.E+00	78.05	P:metabolic process; F:catalytic activity
25	scaffold 7682_len52324_cov195	*Outlier34*	46576	g29533	41765	44731	3,328	*Eucalyptus grandis*	chaperone -domain-containing isoform partial	1.3.E-155	62.95	-
				g29534	49490	50691	3,515		---NA---		-	-
				g29535	51408	52151	5,204		---NA---		-	-
26	scaffold 3764_len46300_cov180	*Outlier35*	383	g17349	1006	4029	2,135	*Eucalyptus grandis*	(-)-germacrene d synthase-like	0.0.E+00	82.10	F:binding; P:metabolic process; F:catalytic activity

[a] Distance between SNP and center of the gene

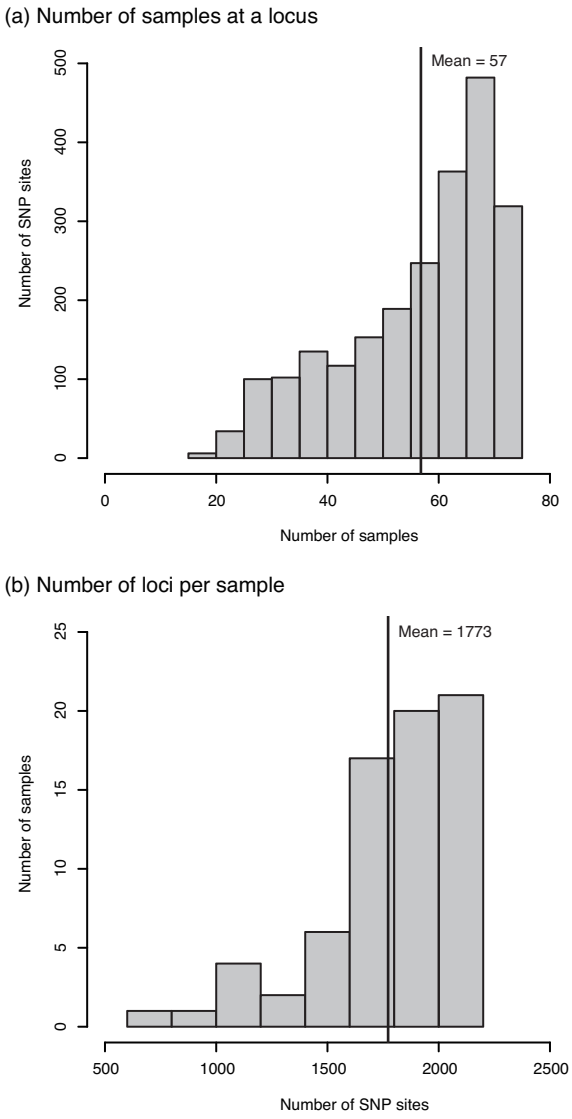

Fig. 3-S1

Genotype data coverage of 2,247 single nucleotide polymorphisms (SNPs) among 72 *Metrosideros polymorpha* samples. (**a**) Distribution of the number of samples at an SNP site. Each SNP covered the genotype data for 57 samples, on average. (**b**) Distribution of the number of SNPs per sample. Each sample obtained the genotype data for 1,773 SNPs, on average.

4

Current plantation practices have negligible genetic effects on planted dipterocarps in the tropical rainforest

4.1. Introduction

Dipterocarp trees are important resources, both ecologically and commercially, in the Southeast Asian tropical rainforest. The family Dipterocarpaceae is classified into 13 genera, approximately 475 species of which can be found in Southeast Asia (Bawa 1998; Ashton 2004). Dipterocarp trees are dominant over substantial areas of the forests in this region and constitute the core elements of biodiversity in tropical ecosystems. Dipterocarp trees are also important timber trees and account for 80% of timber exports from the region (Kettle 2010). Therefore, developing a sustainable system for dipterocarp timber production will contribute not only to the maintenance and improvement of the economic activities of rural residents but also to the conservation of biodiversity.

Until recently, planting and efforts to enhance timber production from dipterocarp trees have been unsuccessful (Appanah 1998). One problem was an inconsistent supply of seeds and seedlings because of irregular flowering and fruiting, a short viability period (Adjers and Otsamo 1996), poor seed quality resulting from frequent insect and fungal attacks (Sakai 1980; Tompsett 1987), and lack of seed storage and handling facilities (Ng 1977). However, forest concession companies have introduced good nursery facilities for storage of wild seedlings (wildlings) collected during mass flowering years and have successfully managed commercial dipterocarp plantations. Determination of optimal plantation conditions such as tree density, light conditions, and soil humidity has also been another problem. For example, in Indonesia, one of the silviculture system used to manage tropical rainforests is selective cutting and strip planting (Tebang Pilih Tanam Jalur, TPTJ). TPTJ was introduced in 1998 (Ministry of forestry 1998) and modified in 2005 and 2009 as Indonesia Selective Cutting and Intensive Planting System (Tebang Pilih Tanam Indonesia Intensif, TPTII) and TPTJ with Silviculture Technique, respectively (Ministry of forestry 2005; Ministry of

forestry 2009). In these systems, enrichment planting is conducted in strips planting in order to increase the density of desired tree species in secondary forests (Sovu *et al.* 2010). Trees are planted at the following varied intervals: 5 × 25 m (80 trees/ha), 2.5 × 20 m (200 trees/ha), and 5 × 20 m (100 trees/ha) for TPTJ 1998, TPTII, and TPTJ 2009, respectively (Ministry of forestry 1998; Ministry of forestry 2005; Hardiansyah *et al.* 2006; Ministry of forestry 2009). As a result of the improved guidelines for growing wild seeds in nurseries and the enforcement of the TPTJ system, the cultivation of planted trees has become a sustainable enterprise.

Although strategies for tree planting in plantations have been considerably optimized, the genetic composition of planted trees has not been specifically studied. In the case of planting individuals of native plant species that have originated from wild rather than genetically improved seeds and seedlings, the following two genetic aspects should be considered: genetic diversity within and genetic differentiation among populations. Genetic diversity is essential for conservation of genetic resources, resistance to various pests in natural ecosystems, avoidance of inbreeding depression (Frankham 1995), and ecosystem-specific processes including community production and maintenance of floral and faunal diversity (Crutsinger *et al.* 2006). These advantages are ecologically and economically important because population variability is needed for long-term timber harvesting. Moreover, genetic differentiation between the plantations and the surrounding natural forests should be considered (O'Brien *et al.* 2010). Even within a single species, genetic differentiation exists between subpopulations because of local adaptations and/or genetic drift (Linhart and Grant 1996). If individuals are introduced to a plantation site from genetically differentiated populations, growth rate and/or population fitness may decrease (Hufford and Mazer 2003). In the event that the resulting generations grow successfully at the plantation site, genetic crossing with naturally grown individuals may lead to genetic pollution or disturbance of the genetic structure and outbreeding depression in the natural forests (Hufford and Mazer 2003; Potts *et al.* 2003). Therefore, planted populations should have a similar genetic composition to those present in the natural forests.

Shorea leprosula Miq. and *S. parvifolia* Dyer (Dipterocarpaceae) are common species in the lowland rainforests of Southeast Asia and are known as useful timber species because of their high growth rates and ease of processing for lumber. In this study, I analyzed the genetic composition of *S. leprosula* and *S. parvifolia* planted in Central Kalimantan, Indonesia. I then compared these characteristics with those of the natural populations in the same region. My objective was to investigate whether the current method for planting dipterocarp trees in the region is adequate to maintain genetic diversity and avoid genetic differentiation from the surrounding natural forest.

4.2. Materials and methods

4.2.1. Plant materials and study site

Plant materials for DNA analysis were collected at the concession area managed by PT Sari Bumi Kusuma (SBK), a private forestry company located in Central Kalimantan

(Fig. 4-1). SBK manages approximately 148 km^2 of the concession area and harvests timbers of 5 *Shorea* species (*S. leprosula*, *S. parvifolia*, *Shorea johonensis*, *Shorea macrophylla*, and *Shorea platyclados*) (Hardiansyah *et al.* 2006). In SBK, wild seedlings (wildlings) are collected in the natural rainforest surrounding the concession area, Bukit Baka Bukit Raya National Park. After being maintained in the nursery for 10 months, the wildlings are planted according to the TPTJ system (Fig. 4-2). For the present analysis, 211 leaves from six plantation stands of *S. leprosula* (Lpl_2000, Lpl_2003, Lpl_2005, Lpl_2006, Lpl_2007, and Lpl_2008) and 139 leaves from four populations of *S. parvifolia* (Ppl_2005, Ppl_2006, Ppl_2008_A, and Ppl_2008_B) (numbers shown in population names indicate planting year) were collected (Table 4-1). To estimate the genetic composition of these two species in natural populations, the leaves of 80 individuals of *S. leprosula* (LNF) and 41 individuals of *S. parvifolia* (PNF) were sampled from the natural forest surrounding SBK.

4.2.2. Microsatellite analysis

Total genomic DNA was extracted from dried leaf tissues using a modified cetyltrimethylammonium bromide (CTAB) method (Milligan 1992). All plant individuals were genotyped using seven expressed sequence tag-linked microsatellite loci (EST-SSR) that were developed previously for *S. leprosula* (Ng *et al.* 2009). Forward primers of each locus were labeled with fluorescent dyes as follows: *SleE05*_VIC, *SleE07*_PET, *SleE08*_6-FAM, *SleE13*_NED, *SleE14*_VIC, *SleE16*_PET, and *SleE21*_6-FAM (Applied Biosystems, Life technologies Corporation, Eugene, OR, USA). Multiplex amplification reactions were performed in a total volume of 10 μl containing 5 ng of genomic DNA, 5 μl of 2× Multiplex PCR Master Mix (QIAGEN Multiplex PCR Kit; Qiagen, Valencia, CA, USA), and 0.2 μM of each primer. Using a GeneAmp PCR System 2700 thermal cycler (Applied Biosystems), I performed the amplification reactions under the following conditions: 1× (95°C for 15 min), 25× (94°C for 30 s, annealing temperature for 90 s, 72°C for 60 s), and 1× (60°C for 30 min). The primer-pair specific annealing temperature was 45°C for *SleE07*, *SleE13*, *SleE14*, and *SleE16*, 45.2°C for *SleE05* and *SleE08*, 50°C for *SleE15*, and 50.1°C for *SleE21*. PCR fragment sizes were determined using an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems), 3130*xl* Genetic Analyzer (Applied Biosystems), Genotyper™ 3.7 software (Applied Biosystems), and GeneMapper™ 3.0 software (Applied Biosystems).

4.2.3. Assessment of Hardy–Weinberg equilibrium and linkage disequilibrium of EST-SSR markers

The assumptions of random mating and Hardy–Weinberg equilibrium for each population were tested by the U test and p-values were estimated with Markov chain algorithm using GenePop 4.2 (Raymond and Rousset 1995). Linkage disequilibrium was tested for each pair of loci by the log likelihood ratio statistic using GenePop 4.2 (Raymond and Rousset 1995).

4.2.4. Assessment of genetic diversity and inbreeding coefficients

Number of alleles per locus (N_A), observed heterozygosity (H_O), and expected (H_E) heterozygosity were calculated with GenAlEx 6.4 (Peakall *et al.* 2006). Allelic richness (A_R; El Mousadik and Petit, 1996) and inbreeding coefficient (F_{IS}) were calculated using FSTAT 2.9.3.2 (Goudet 1995). Differences in A_R, H_O, and H_E between each populations and natural population were tested using ANOVA and Tukey's multiple comparisons test using values for each locus as a replicate in R (R development core team 2010). Deviation of F_{IS} values from zero was evaluated by a significant deficit of heterozygotes for each population at 0.05 of significance levels adjusted by Bonferroni correction for multiple testing using FSTAT 2.9.3.2 (Goudet, 1995).

4.2.5. Assessment of genetic differentiation between populations

To clarify the level of genetic differentiation between populations, Nei's *D* (Nei 1978) and G''_{ST} (Meirmans and Hedrick 2011) were calculated for each pair of populations using GenoDive 2.0 b24 (Meirmans and Tienderer 2004). Pairwise differentiations were tested with the log-likelihood G-statistics by 999 permutations on GenoDive 2.0 b24 (Meirmans and Tienderer 2004) and p-values were adjusted by Bonferroni correction. An analysis of molecular variance (AMOVA; Excoffier *et al.* 1992) was used to partition the entire genetic variance into those within and among populations using GenAlEx v6.4 (Peakall *et al.* 2006). To estimate the number of populations providing genetic sources, a Bayesian cluster analysis was applied to all individuals using STRUCTURE 2.3 (Pritchard *et al.* 2000). An admixture model that assumed correlated allele frequencies among populations was used. Ten runs for each *K* (*K* = 1–10) for each species were carried out. For each run, the number of burn-in and MCMC interactions were 100,000 and 1,000,000, respectively. The most likely number of clusters was selected by comparing the log Pr (X|K) [ln *P*(D)] and the statistic delta *K* (Evanno *et al.* 2005) in each *K*. For *S. parvifolia*, the data did not fulfill the assumptions of Hardy–Weinberg equilibrium; therefore clustering analysis was not conducted.

4.3. Results

4.3.1. Hardy–Weinberg equilibrium and linkage disequilibrium of EST-SSR markers

Significant deviation ($p < 0.01$) from Hardy–Weinberg equilibrium was detected in four populations of *S. leprosula* and all populations of *S. parvifolia* (Table 4-2). Random mating within most populations was found in *S. leprosula* but not in *S. parvifolia*. Linkage disequilibrium was detected in three locus pairs (14%) in *S. leprosula* and 15 locus pairs (74%) in *S. parvifolia* (Table 4-3). Near-complete independence of the seven loci was shown in *S. leprosula* but not in *S. parvifolia*, suggesting a relatively high rate of inbreeding within each population.

4.3.2. Genetic diversity and inbreeding coefficients

The degree of genetic diversity and inbreeding coefficient for each population of *S. leprosula* and *S. parvifolia* is presented in Table 4-1. The degree of genetic diversity (A_R, H_O, and H_E) in each plantation was not significantly different from the natural population in either species except for a difference in A_R between Ppl_2006 and PNF in *S. parvifolia*. Two populations of *S. leprosula* and all populations of *S. parvifolia* presented significantly high F_{IS} values.

4.3.3. Genetic differentiation between populations

In *S. leprosula*, G''_{ST} and Nei's *D* values between each plantation and natural population were 0.057 on average (range: 0.010–0.098) and 0.047 on average (range: 0.007–0.075), respectively (Table 4-4 (a)). In *S. parvifolia*, G''_{ST} and Nei's *D* values between each plantation and natural population were 0.055 on average (range: 0.002–0.101) and 0.054 on average (range: 0.012–0.093), respectively (Table 4-4 (b)). As a result of testing population differentiation, 18 pairs of populations in *S. leprosula* and seven pairs in *S. parvifolia* showed significant differentiations. Ppl_2005 presented relatively large differentiation from the natural population (PNF), as shown in G''_{ST} and *D* values of 0.101 and 0.093, respectively (Table 4-4 (b)). AMOVA of hierarchical genetic diversity revealed that genetic variation within populations accounted for 97% of the total molecular variance, and 3% of the variance occurred among populations in both species (Table 4-5).

In the Bayesian analysis for clustering of all individuals, the ln *P*(D) value was highest at $K = 1$ for *S. leprosula*. At $K > 2$, each individual contained alleles derived from each cluster in a similar proportion (Fig. 4-3). Thus, all *S. leprosula* individuals likely originated from a single gene pool and genetic differentiations between populations were not found.

4.4. Discussion

4.4.1. Genetic status of plantation stands in SBK

In both *Shorea leprosula* and *S. parvifolia*, each plantation contained genetic diversity as large as that in the natural population (Table 4-1, Table 4-5). However, Ppl_2006 showed significantly lower allelic richness ($A_R = 7.86$) than the natural forest (PNF, $A_R = 11.23$), the smallest value of observed heterozygosity ($H_O = 0.52$), and the largest inbreeding coefficient ($F_{IS} = 0.28$) among the plantations of *S. parvifolia* (Table 4-1). Because the study plantations were established by selecting seedlings without genetic information, incidental lack of genetic diversity within a plantation stand could occur. It is likely that *S. parvifolia* presented a relatively higher level of inbreeding coefficients than *S. leprosula*, and all populations of *S. parvifolia* deviated from the Hardy–Weinberg equilibrium (Table 4-1). This difference could have been affected by tree density in the wild populations from which seedlings were taken. In the wild, the density of

conspecific flowering trees affects selfing rate in the population (Fukue *et al.* 2007) and lower tree density drives a higher rate of inbreeding (Tani *et al.* 2009). As a whole, the genetic diversity contained in almost all plantation stands was as high as that in the natural population. This indicates that the current plantation methodology is suitable for maintaining genetic diversity in the planted dipterocarp trees. However, it should be noted that an incidental lack of genetic diversity could occur when the source seedlings are taken from wild populations of low density, or genetically related seedlings are transplanted from a nursery to the same plantation stand.

Significant genetic differentiations between study populations were found, indicating that genetic compositions were nonhomogenous between the plantation stands (Table 4-4). Genetic compositions could be different between plantation stands because the individuals may be only a part of trees in the original natural forest. However, based on the Bayesian clustering of *S. leprosula*, the most likely number of genetic clusters was one (Fig. 4-3), suggesting that the source of trees in each plantation was similar to each other. Therefore, genetic degradations and disturbance of the genetic structure probably had no importance.

4.4.2. Implication for the dipterocarp plantations in the tropical rainforest

Based on TPTJ, the current plantation method incorporates several factors that may contribute to the maintenance of genetic diversity in plantations. One of these factors is the use of wild seeds and seedlings collected from the natural forest surrounding the concession area. Unlike clonal propagation such as cutting, wild seeds and seedlings derived from natural regeneration are likely to maintain a high degree of genetic diversity. Another factor is the presence of a healthy, natural forest surrounding the concession area, which facilitates the easy collection of seeds and seedlings. Because *S. leprosula* and *S. parvifolia* are outcrossing species (Lee *et al.* 2000; Sakai *et al.* 1999) and planted populations may influence the genetic composition of natural populations through gene flow, it is safe to plant seedlings derived from a neighboring natural population. Conservation of the natural forest near the concession area is an advisable policy for the maintenance of the forest landscape and the sustainability of the forestry industry.

To maintain the current genetic status in the plantation stands, appropriate method to select seedlings for plantation is important. The process that may cause the reduction of genetic variation such as selection for specific phenotypic traits or propagation by cutting should be avoided. By mixing the seedlings that originated from different mother trees, I can expect that plantation stands containing high genetic diversity similar to the natural forests may be established and the reduction of the genetic diversity in the process of transplantation of seedlings from nurseries to plantation stands may be prevented.

The sustainability of dipterocarp plantation stands is ecologically and economically important. Ecologically, dipterocarp plantations contribute to ecological services such as the conservation of biodiversity (Barlow *et al.* 2007; Meijaard and Sheil

2007), the storage of carbon (Kettle 2010). The timber production provides an important income for rural people. Although some plantation systems including TPTJ have been established after repeated improvements, their genetic effectiveness has not been recognized. I report the first genetic evaluation of an intensive enrichment planting system practiced in Indonesia, TPTJ and conclude that the current plantation methodology is basically suitable for maintaining the genetic variation present in natural populations. The continuance and improvement of dipterocarp plantation techniques, with regard to genetic effects are expected.

Tables and Figures

Table 4-1
Genetic diversity within each population of *Shorea leprosula* and *S. parvifolia*

Species	Population	*N*	*Na*	A_R	H_O	H_E	F_{IS}
Shorea leprosula	LNF	80	13	9.90^{ab}	0.66	0.76^{ab}	0.13*
	Lpl_2000	33	11	11.14 ^{a}	0.71	0.81 ^{a}	0.15*
	Lpl_2003	34	10	10.21^{ab}	0.76	0.78^{ab}	0.05
	Lpl_2005	35	10	9.55^{ab}	0.71	0.78 ^{b}	0.06
	Lpl_2006	35	10	9.57^{ab}	0.74	0.77^{ab}	0.05
	Lpl_2007	37	10	9.49 ^{b}	0.71	0.77^{ab}	0.09
	Lpl_2008	37	10	9.53 ^{b}	0.75	0.77^{ab}	0.04
Shorea parvifollia	PNF	41	12	11.23 ^{c}	0.60	0.78^{cd}	0.25*
	Ppl_2005	47	11	9.74^{cd}	0.60	0.76^{cd}	0.22*
	Ppl_2006	30	8	7.86 ^{d}	0.52	0.71 ^{c}	0.28*
	Ppl_2008_A	32	11	10.42 ^{c}	0.59	0.78 ^{d}	0.26*
	Ppl_2008_B	30	9	9.14^{cd}	0.67	0.77^{cd}	0.15*

Sample size (*N*); mean number of alleles per locus (*Na*); allelic richness (A_R); observed (H_O) and expected heterozygosity (H_E); inbreeding coefficient (F_{IS}). LNF: Natural forest population of *S. leprosula*, PNF: Natural forest population of *S. parvifolia*, Lpl: Planted populations of *S. leprosula*, Ppl: Planted populations of *S. parvifolia*, and numbers shown in population names indicate plantation year. Different lowercase letters above the values indicate significant $p < 0.05$ differences among populations. Asterisks indicate F_{IS} values are significantly different from 0 ($p < 0.05$).

Table 4-2
Summary of chi-square tests for Hardy-Weinberg equilibrium for each locus in each population

	S. leprosula							*S. parvifolia*				
Locus	LNF	Lpl_2000	Lpl_2003	Lpl_2005	Lpl_2006	Lpl_2007	Lpl_2008	PNF	Ppl_2005	Ppl_2006	Ppl_2008_A	Ppl_2008_B
SleE08	ns	ns	ns	ns	ns	ns	ns	***	ns	ns	ns	ns
SleE05	***	***	***	ns	ns	**	ns	**	***	***	***	***
SleE13	***	ns	ns	ns	ns	***	ns	***	***	**	***	ns
SleE16	ns	**	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
SleE14	ns	ns	ns	ns	ns	ns	ns	**	ns	**	**	ns
SleE07	ns	ns	ns	ns	ns	ns	ns	**	ns	ns	ns	ns
SleE21	***	*	ns	ns	**	***	ns	***	***	***	***	***
Multi-locus	***	***	ns	ns	**	***	ns	***	***	***	***	***

ns=not significant, ** p <0.01, *** p <0.001

LNF: Natural forest population of *S. leprosula*, PNF: Natural forest population of *S. parvifolia*, Lpl: Planted populations of *S. leprosula*, Ppl: Planted populations of *S. parvifolia*, and numbers shown in population names indicate plantation year.

Table 4-3
Summary of log likelihood ratio statistic tests for linkage disequilibrium in *Shorea leprosula* (above diagonal) and *S. parvifolia* (below diagonal)

	SleE08	*SleE05*	*SleE13*	*SleE16*	*SleE14*	*SleE07*	*SleE21*
SleE08	-	ns	ns	ns	ns	*	ns
SleE05	*	-	ns	ns	ns	ns	ns
SleE13	*	*	-	ns	ns	*	*
SleE16	*	*	*	-	ns	ns	ns
SleE14	*	*	ns	*	-	ns	ns
SleE07	*	*	ns	ns	ns	-	ns
SleE21	*	*	*	ns	*	ns	-

ns=not significant, * $p < 0.05$

Table 4-4
G''_{ST} values (above diagonal) and Nei's *D* (below diagonal) for every pairwise populations of (a) *Shorea leprosula* and (b) *S. parvifolia*

(a) *S. leprosula*

	LNF	Lpl_2000	Lpl_2003	Lpl_2005	Lpl_2006	Lpl_2007	Lpl_2008
LNF	-	0.098*	0.010	0.065*	0.053*	0.037*	0.080*
Lpl_2000	0.075	-	0.052	0.078*	0.090*	0.073*	0.119*
Lpl_2003	0.007	0.047	-	0.053*	0.090*	0.049*	0.080*
Lpl_2005	0.055	0.058	0.042	-	0.110*	0.078*	0.136*
Lpl_2006	0.044	0.076	0.077	0.092	-	0.022	0.116*
Lpl_2007	0.032	0.062	0.043	0.066	0.021	-	0.036*
Lpl_2008	0.067	0.101	0.068	0.116	0.100	0.032	-

(b) *S. parvifolia*

	PNF	Ppl_2005	Ppl_2006	Ppl_2008_A	Ppl_2008_B
PNF	-	0.101*	0.071*	0.045*	0.002
Ppl_2005	0.093	-	0.100*	0.067*	0.087*
Ppl_2006	0.058	0.085	-	0.068	0.109*
Ppl_2008_A	0.050	0.065	0.055	-	0.023
Ppl_2008_B	0.012	0.080	0.091	0.030	-

* Significant population differentiation at 0.05 of significance level adjusted by Bonferroni correction for multiple testing

Table 4-5
Summary of analysis of molecular variance for EST-SSR genotypes

Source of variation	*df*	Sum of squares	Variance components	Percentage of variation
(a) *S. leprosula*				
Among populations	6	70.014	0.166	2.7
Within populations	284	1686.776	5.939	97.3
Total	290	1762.790	6.105	100.0
(b) *S. parvifolia*				
Among populations	4	51.712	0.175	3.0
Within populations	175	1168.654	6.678	97.0
Total	179	1220.367	6.853	100.0

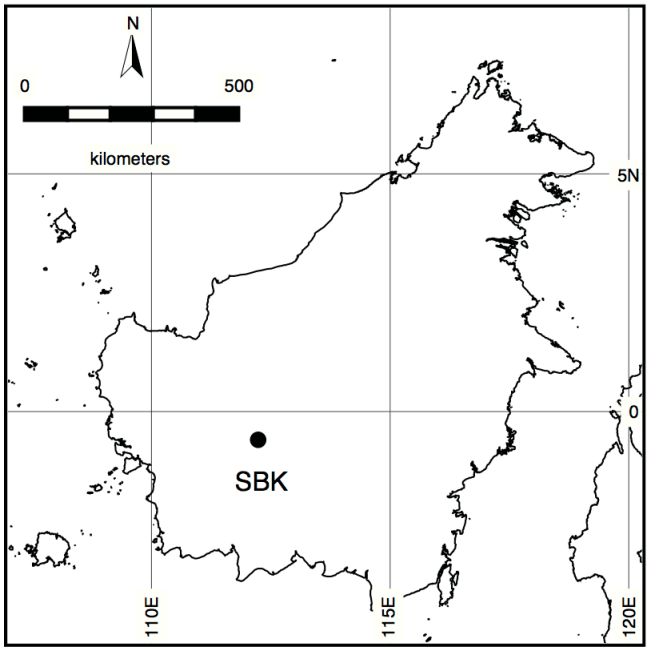

Fig. 4-1
Location of PT. Sari Bumi Kusuma (SBK) in Central Kalimantan, Indonesia

Fig. 4-2
(a) A nursery of *Shorea leprosula* and (b) a plantation stand of *S. leprosula* established in 2006 in SBK.

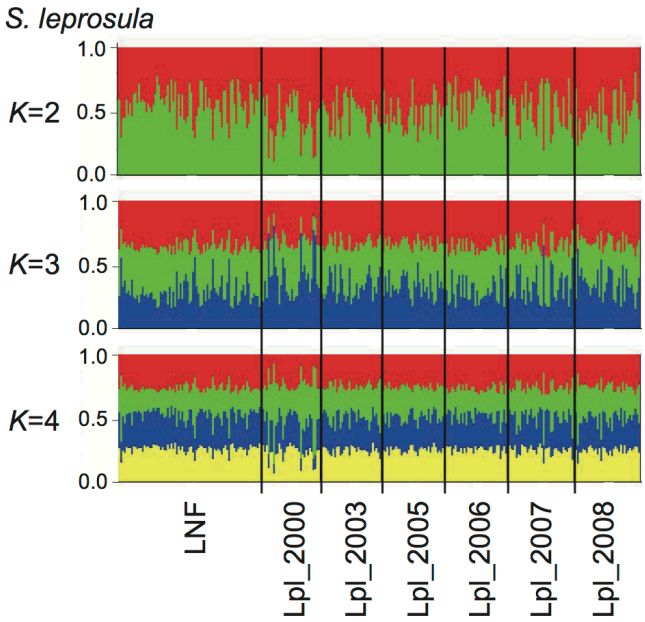

Fig. 4-3
Bayesian structure analysis of *Shorea leprosula* with the STRUCTURE software program from $K = 2$ to $K = 4$ (Pritchard *et al.* 2000). LNF, natural forest population of *S. leprosula*; Lpl, planted populations of *S. leprosula*. Numbers shown in the population names indicate plantation year.

5

Structure of phyllosphere fungal communities in a tropical dipterocarp plantation

5.1. Introduction

Phyllosphere fungi, which include epiphytes on the surfaces of leaves (Lindow and Brandl 2003), endophytes living asymptomatically within leaf tissues (Rodriguez et al. 2009) and pathogens, exert neutral, negative, or positive influences on their host plants and play various ecological and physiological roles in terrestrial ecosystems. For example, they may reduce the photosynthetic rates of host plants (Pinto et al. 2000), pathogenically attack host plants (Newton et al. 2010), enhance the resistance of hosts to pathogens (Arnold et al. 2003), or act as decomposers of leaf litter and drive nutrient cycles (Osono 2006). Phyllosphere fungi have been recognized to be the most species-rich groups of fungi (Unterseher et al. 2011). The habitats of phyllosphere fungi are likely very large and diverse, as the total global leaf surface area exceeds 4 × 108 km^2 (Morris and Kinkel 2002), a value approximately 2.7 times larger than the total land area. This likelihood of diversity is further supported by the fact that phyllosphere fungi inhabit all major land plant lineages from the tropics to the Arctic (Arnold 2007). Recent methodological developments based on next-generation sequencing facilitate the sequencing of DNA extracted directly from environmental materials, thereby permitting the estimation of species richness among various populations of microorganisms. Metagenomic amplicon sequence analyses with high-throughput sequencers have begun to uncover the tremendous diversity of fungi from aquatic (Brown et al. 2009; Stoeck et al. 2010) to terrestrial environments, including plant roots (e.g., Davison et al. 2012; Clemmensen et al. 2013; Toju et al. 2013) and leaves (e.g., Jumpponen and Jones 2009; Cordier et al. 2012; Zimmerman and Vitousek 2012). The investigation of community composition, richness and dynamics of phyllosphere fungi via metagenomic amplicon sequence analyses will contribute to the exploration of new fungal bioresources and understanding of the nature of the

ecological interactions between fungi and plants.

Many factors determine fungal diversity and community structures. Differences in abiotic conditions, along with environmental gradients, such as elevation (Davey et al. 2013), continental-scale climate conditions (U'Ren et al. 2012) and the degree of urbanization (Jumpponen and Jones 2009), generate spatial variation in phyllosphere fungal communities. Although dispersal can counteract this spatial variation, topological or anthropogenic factors that limit dispersal among sites are likely to cause small-scale geographic variations in community structure (Lomolino et al. 2006; Adams et al. 2013). The effects of these factors on fungal community structures can differ both spatially and temporally (Matulich et al. 2015). Next-generation sequencing is an effective and comprehensive method of investigating the phyllosphere fungal community structure (Jumpponen and Jones 2009; Cordier et al. 2012), but to my knowledge few studies have used high-throughput meta-barcoding analysis to reveal the assemblies of phyllosphere fungi in the tropics.

Southeast Asia contains one of the most diverse floras in the world. It comprises 29,332 endemic vascular plant species (about 10% of approximately 300,000 global vascular plant species; Myers et al. 2000). In view of this diversity, fungi that interact with those plants (such as phyllosphere fungi and mycorrhizal fungi) are likely to be highly diverse. Nonetheless, the richness of fungal species in this region remains poorly investigated (Hawksworth and Rossman 1997; Webb et al. 2010; but see Peay et al. 2010). Within this region, plantations of native tree species represent an important land use because they contribute to sustainable timber production without further accelerated deforestation, which would inevitably result in a substantial loss of biodiversity. Indonesia has established commercial timber production by planting trees after cutting (Hardiansyah et al. 2006; Ministry of Forestry 2009). However, pathogens, pests and diseases interfere with commercial plantation forestry (van Staden et al. 2004). Plantation forests contribute to the maintenance of biodiversity in the region as well as in primary forests (Barlow et al. 2007; Meijaard and Sheil 2007). For this reason, the accumulation of knowledge regarding the richness and spatial distribution patterns of fungal species is important for the sustainable management of biodiversity and commercial plantations.

In the present study, I addressed the question "How diverse are the phyllsphere fungi inhabiting in the Southeast Asian tropics?" I used massively parallel next-generation sequencing to characterize the diversity and spatial variability of phyllosphere fungal communities of managed *Shorea leprosula* Miq. (Dipterocarpaceae), which is commonly distributed in the lowland rainforests of Southeast Asia and is an important timber species called red meranti. I investigated the phyllosphere fungal assemblage of trees grown in plantation stands, evaluated total fungal richness, and then compared fungal community compositions among spatially distant tree individuals.

5.2. Materials and methods

5.2.1. Study site and plant materials

In July 2012, leaves were collected at *S. leprosula* plantation stands managed by PT. Sari Bumi Kusuma (SBK), a private forestry company in Central Kalimantan, Indonesia (0°35.5′S, 112°14.2′E; Fig. 5-1A). In SBK-managed plantation stands, young *S. leprosula* trees had been planted at 2.5-m intervals and had grown to approximately 10 m in height and 15 cm in diameter at breast height during a six-year period. To evaluate the effects of geographic locations on fungal assemblage, I established four plots comprising seven to eight adjoining trees that had originated from different mother trees (Fig. 5-1B). The distances separating the four plots ranged from 15 m to 3.5 km (Fig. 5-1B). A total of 31 *S. leprosula* trees were selected as target plant individuals. To minimize the effect of the heterogeneous distribution of fungi within an individual plant on the observed species diversity, 10 leaves from three branches growing in different directions were collected from each plant. The collected leaves were then dried in silica gel until DNA extraction was performed.

5.2.2. DNA extraction

A 25-mm^2 piece was cut from a dried leaf with a disposable knife, and 10 leaf pieces from the 10 leaves originating on the same plant were mixed in a microcentrifuge tube. The knife blade was discarded after each set of 10 leaves. All leaf pieces appeared healthy and showed no symptoms of pathogenic fungal infection. To avoid contamination by fungi that did not originate on *S. leprosula* leaves (such as common airborne fungi), each dried leaf piece was cleaned three times with 0.005% (w/v) Aerosol-OT solution (di-2-ethylhexyl sodium sulfosuccinate; Cytec Industries, West Paterson, NJ, USA) and rinsed twice with sterile distilled water (Osono et al. 2009). After the remaining water was removed from the tube, the 10 pieces collected from an individual plant (approximately 20 mg in total) were homogenized to powder with stainless steel beads (3 mm in diameter), after which total genomic DNA was extracted using the modified cetyl trimethylammonium bromide method (Milligan 1992). A tube was prepared without leaf contents to serve as a negative control for each extraction procedure. Hereafter, a sample is defined as a mixture of leaf pieces originating from an individual plant.

5.2.3. Parallel amplicon sequencing

To identify fungal species inhabiting the leaves, the internal transcribed spacer (ITS) region was amplified and sequenced. ITS1 sequences (approximately 250 base pairs [bp]) were amplified via PCR to allow the use of a sequencer Ion-PGM (Ion Torrent, Life Technologies, Guilford, CT, USA), for which the sequence length is limited to approximately 300 bp at the time of sequencing. For parallel amplicon sequencing, I used two types of fusion primer: a PGM-sequencing primer (Ion A primer), followed by tag sequences for sample identification (Hamady et al. 2008) and ITS1F_KYO2 (Toju et al. 2012), and a DNA capture bead annealing primer for emulsion PCR (trP1 primer), followed by ITS2_KYO2 (Toju et al. 2012). Fungal ITS1 regions were amplified in a total

volume of 50 μl that contained 10 ng of template DNA, a 200-nM concentration of each primer and 45 μl of Platinum PCR SuperMix High Fidelity (Invitrogen/Life Technologies Inc., Burlington, Ontario, Canada). Using a Veriti 96-well thermal cycler (Applied Biosystems/Life Technologies, Foster City, CA, USA), amplification reactions were performed under the following conditions: 94°C for 3 min (initial denaturation) and 35 cycles at 94°C for 30 s, 60°C for 30 s, and 68°C for 2 min and 40 s.

Approximately 0.3 pmol of 200–400-bp PCR products from each sample was mixed. Amplicons with a length of 200–300 bp were further extracted using E-Gel SizeSelect Agarose Gels (Invitrogen). The extracted amplicons were then purified using AgenCourt AMPure XP PCR Purification Reagent (AgenCourt Bioscience, Beverly, MA, USA), quantified with an Agilent 2100 Bioanalyzer DNA High Sensitivity Kit (Agilent Technologies, Santa Clara, CA, USA), and diluted to a final concentration of 26 pM. The amplicon library was amplified using Ion One Touch System (Ion Torrent) with an Ion One Touch 200 Template Kit v2 (Ion Torrent) and sequenced using Ion-PGM with an Ion 200 Sequencing Kit (Ion Torrent). Six of the 31 samples were sequenced in the first run using an Ion 314 Chip (Ion Torrent), whereas the remaining 25 samples were sequenced using an Ion 318 Chip (Ion Torrent). In total, 485,710 and 5,566,576 reads were obtained for the first and second runs, respectively (DDBJ Sequence Read Archive accession: DRA001737).

5.2.4. Bioinformatics

Raw sequencing reads from two Ion-PGM runs were sorted using sample-specific tag sequences and filtered using the "clsplitseq" command in the Claident v0.2.2015.03.11 software package (Tanabe 2012b). A minimum average quality value of 27 over the sequence (Kunin et al. 2010) was required, low-quality 3′-tail reads were removed, and reads shorter than 150 bp were excluded. Potentially chimeric reads and reads that were likely to contain a high proportion of sequencing errors (noisy reads) were detected and removed by the "clcleanseq" command of Claident. This command uses UCHIME v4.2.40 (Edgar et al. 2011) to detect potentially chimeric reads and the algorithm developed by Li et al. (2012) to detect noisy reads.

Filtered reads were clustered to generate operational taxonomic units (OTUs) using the Assams software v0.2.2015.03.11 (Tanabe 2012a; Toju et al. 2014), which was invoked with the "clclassseq," "clclassclass," and "clreclassclass" commands in Claident v0.2.2015.03.11 (Tanabe 2012b). Assams enables highly parallelized processing along with the accurate assembly program, Minimus (Sommer et al. 2007). Filtered reads were clustered within each sample with a cutoff sequence similarity of 99%. Then, sequence clusters within samples were merged across all samples with three cutoff sequence similarities of 95%, 97% and 98.5%. (Jumpponen and Jones 2009; U'Ren et al. 2009; Kemler et al 2013). The resulting clustered sequences were used as OTUs in the subsequent analysis.

Taxonomic identification of the obtained OTUs was based on the query-centric auto-*k*-nearest-neighbor (QCauto) method (Tanabe and Toju 2013) using the "clidentseq" and "classigntax" commands in Claident v0.2.2015.03.11. An intensive benchmark analysis indicated that the QCauto method returned the most conservative

and accurate taxonomic identification results among the existing automated DNA barcoding methods if all potentially observable species had not been completely described and registered in the reference database (Tanabe and Toju 2013). To permit QCauto-method-based identification of the obtained fungal OTUs while using Claident software, I downloaded the "fungi_ITS_species" database from the Claident web site on July 7, 2015. This "fungi_ITS_species" database comprises a subset of the National Center for Biotechnology Information "nt" database and contains fungal reference ITS sequences with species-rank taxonomic information (Tanabe and Toju 2013). Claident uses the lowest-common-ancestor algorithm to return the lowest taxonomic level common to the homologous sequences found in the reference database (Huson et al. 2007). After the taxonomic identification process, OTUs that were not assigned to fungi or were included in negative control samples were excluded from the datasets. Because singletons likely include PCR and sequencing artifacts (Tedersoo et al. 2010; Brown et al. 2015), singletons were also removed from the datasets. Samples containing fewer than 100 reads were also excluded from the subsequent statistical analysis. The command script for Claident v0.2.2015.03.11 is provided in Data 5-S1.

To test the effect of sequence lengths of OTUs on taxonomic identifications, the sequence length differences between identified and unidentified OTUs were tested at each taxonomic level (phylum, class, order, family, genus and species levels) by Wilcoxon rank sum test.

5.2.5. Statistical analysis

Considering that sequence dissimilarity among OTUs is critical in biodiversity analyses, I separately analyzed and compared three datasets differing in the OTU identity level (95%, 97% and 98.5%).

5.2.5.1. Diversity of phyllosphere fungi in the tropical plantation

The effect of OTU clustering level (95%, 97% and 98.5%) on OTU richness was evaluated by rarefaction curves. For each dataset, the differences in sequencing intensities among samples or plots were further evaluated by the rarefaction curves. Rarefaction curves were generated using the "rarecurve" or "specaccum" function in the R vegan 2.0–10 package (Oksanen et al. 2011). To assess the evenness of number of reads among OTUs, rank-abundance plots were drawn. Whether the fungal communities were completely surveyed was tested by fitting the community datasets to log-normal species abundance distributions. The goodness of fit was tested with chi-square tests.

5.2.5.2. Spatial variability of phyllosphere fungal communities

To discriminate the abundant OTUs from rare or occasional OTUs, OTUs represented in more than 50% of samples were defined as core (abundant) OTUs and the rest were defined as satellite (rare or occasional) OTUs (Unterseher et al. 2011). The number and taxonomy were compared between core and satellite OTUs. Spatial variability in the

phyllosphere fungal communities was investigated by comparison of the core or satellite OTU compositions in samples among the four plots (P1–P4). I calculated the Raup–Crick dissimilarity index (Chase et al. 2011) for the presence/absence data of fungal OTUs using the "raupcrick" function in the vegan package. This function accommodates presence/absence data and corrects for differences in observed numbers of OTUs among groups and differences in sampling intensities among OTUs (Chase et al. 2011; Oksanen et al. 2011). The core or satellite OTU composition of each sample was plotted on a two-dimensional surface according to the pairwise Raup–Crick dissimilarity using the nonmetric multidimensional scaling (NMDS) analysis in the vegan package. The effect of plot on core or satellite fungal community dissimilarity was evaluated with a permutation test for multivariate analysis of variance (PERMANOVA; Anderson 2001) and another for homogeneity of multivariate dispersions (PERMDISP; Anderson 2006) in 999 permutations, available in the vegan package. Pairwise dissimilarity of fungal communities was evaluated by the overlap of standard deviation ellipses of the weighted averages of plots on the NMDS ordinations.

5.3. Results

5.3.1. Diversity of phyllosphere fungi in the tropical plantation

Among the total reads obtained by sequencing, 177,777 and 1,607,247 reads contained the tag sequences corresponding to the 6 and 25 samples in the first and second run, respectively (Table 5-S2). After removal of short and low-quality reads, 40,002 (22.5%) reads for the six first-run samples and 856,126 reads (53.3%) for the remaining 25 samples were retained (Table 5-S2). These reads contained 2149 and 98,765 possibly chimeric reads and 24,216 and 274,484 noisy reads for the first and second run, respectively (Table 5-S2). Thus, 26,365 (14.8% of the 177,777 raw reads) and 373,249 (23.2% of the 1,607,247 raw reads) were recognized as artificial reads. After removal of the artificial reads, 13,637 (7.7% of the 177,777 raw reads) and 482,877 (30.0% of the 1,607,247 raw reads) reads from the first and second run were used for OTU clustering (Table 5-S2). Clustering these reads with a 95%, 97% and 98.5% similarity cutoff yielded 888, 1267 and 3046 OTUs, respectively (Table 5-S2). After removal of the OTUs contained in the negative controls and non-fungal OTUs and singletons, 488 OTUs (153,194 reads), 697 OTUs (154,300 reads) and 1562 OTUs (193,381 reads) were obtained for the 95%, 97% and 98.5% OTU similarity datasets, respectively (Table 5-S2).

As a result of OTU identification in the dataset of 95% OTU similarity, 80.9%, 42.2%, 32.6%, 19.3%, 13.5% and 4.9% of total OTUs were identified at the phylum, class, order, family, genus and species levels, respectively. Note that the "unidentified" OTUs indeed obtained some homology hits belonging to the kingdom Fungi in the "fungi_ITS_species" database, but the taxonomies were inconsistent among the hits. In such cases, a taxonomic name was unassigned because of the lowest common ancestor algorithm (Huson et al. 2007) to avoid misidentification (Tanabe and Toju 2013). The Wilcox rank sum tests showed no difference in sequence length between identified and

unidentified OTUs at every taxonomic level except for the phylum-level (W = 21003, p = 0.03 for the phylum-level; W = 31063, p = 0.19 for the class-level; W = 28512, p = 0.12 for the order-level; W = 19865, p = 0.27 for the family-level; W = 15252, p = 0.21 for the genus-level; W = 5933, p = 0.59 for the species-level).

The observed OTU richness decreased along with decreasing thresholds of sequence similarity for OTU clustering (Fig. 5-2). The OTU accumulation curves drawn for the three datasets were mostly saturated but did not reach stable plateaus (Fig. 5-2). In each dataset, the OTU accumulations were different remarkably among samples as well as plots (Fig. 5-S3). Even in the dataset of 95% OTU similarity, the OTU richness of most samples did not reach saturation against the number of reads (Fig. 5-5A).

The rank abundance plots showed that most OTUs comprised small numbers of reads (Figs. 3A, S6A, B). The numbers of OTUs comprising fewer than 10 reads were 200 (41.0%), 326 (46.8%) and 824 (52.8%) in the datasets of 95%, 97% and 98.5% OTU similarity, respectively (Figs. 3A, S6A, B). Chi-square tests and the probability plots revealed that no datasets followed log-normal species abundance distributions (Figs. 3B, S6C, D).

5.3.2. Spatial variability of phyllosphere fungal communities

OTUs were divided into core and satellite OTUs based on the number of representative samples. In the dataset of 95% OTU similarity, 23 OTUs (72,990 reads) inhabiting 16–28 samples were recognized as core OTUs and the remaining 465 OTUs (80,204 reads) represented by 1–15 samples belonged to satellite OTUs (Table 5-1; Fig. 5-4). The number of satellite OTUs was 20.2 times greater than that of core OTUs (Table 5-1; Fig. 5-4). Among the 465 satellite OTUs, 146 (29.9%) were distributed in a single sample (Fig. 5-4). Similar results for the relative abundance and distribution patterns between core and satellite OTUs were observed in the other two datasets (97% and 98.5% OTU similarity) (Table 5-1; Fig. 5-S5A, B).

Taxonomic compositions were compared between core and satellite OTUs. At the phylum level, relative abundances between Ascomycota and Basidiomycota were similar between core and satellite OTUs; the ratio of Ascomycota to Basidiomycota was 6:1 in core OTUs and 5:1 in satellite OTUs in the dataset of 95% OTU similarity (Fig. 5-6A). At the class and order levels, more diverse taxonomy was found in satellite OTUs than in core OTUs. The 23 core OTUs found in the dataset of 95% OTU similarity belonged to only four ascomycote orders: Xylariales, Sordariales, Diaporthales (Sordariomycetes) and Pleosporales (Dothideomycetes) (Fig. 5-6B, C) and 465 satellite OTUs were represented in 19 orders in 11 classes (Fig. 5-6B, C). The other two datasets (97% and 98.5% OTU similarity) showed qualitatively similar taxonomic compositions.

Spatial variability in fungal community compositions was separately tested for core and satellite OTUs. NMDS ordinations and PERMANOVA showed that both the core and satellite fungal communities on *S. leprosula* differed among geographically distant plots (Table 5-1; Fig. 5-6). PERMDISP showed that community dissimilarity variances within each plot did not differ among the four plots; thus, the significant differences in fungal communities among plots found in PERMANOVA were attributed to those among means rather than variances of fungal communities (Table

5-1). The proportions of variation explained by plot were 51–56% and 74–80% in core and satellite OTU communities, respectively (Table 5-1). In NMDS ordinations of satellite OTUs compositions, standard deviations did not overlap among the four plots (Figs. 6B, S7E, F).

5.4. Discussion

Using high-throughput DNA sequencing, I investigated and identified species diversity among tropical phyllosphere fungi. Although a previous study using a conventional isolation method identified 42 endophyte OTUs from six woody plant species growing within a 1500- ha area in a tropical forest (Arnold and Lutzoni 2007), the present meta-barcoding analysis detected 488 phyllosphere fungal OTUs from a single plant species growing within an approximately 400 m^2 area of a managed tropical plantation. The present study showed the effectiveness of high-throughput sequencing for detecting phyllosphere fungi, as also shown in previous studies. For example, 1599 OTUs were detected on nine trees 300 m apart (Cordier et al. 2012), and 360 OTUs were found on 12 trees growing across two regions 8.5 km apart (Jumpoonen and Jones 2009). Meta-barcoding analyses with high throughput sequencers have enabled us to detect various fungal OTUs, of which many are inconspicuous and unculturable (Ekblom and Galindo 2011; Davey et al. 2013; Unterseher et al. 2013).

The phyllosphere fungal diversity revealed in the present study likely represents only part of the entire fungal flora for the following reasons. First, rank-abundance plots based on the read count per OTU showed that most fungal OTUs comprised small numbers of reads (Figs. 3A, S6A, B) and the distribution patterns of the read count per OTU did not follow log-normal species abundance distributions (Figs. 3B, S6C, D). Rarefaction curves for samples or plots did not reach stable plateaus (Fig. 5-S3). Under the assumption that a speceis-abundance distribution derived by a complete sampling follows log-normal distribution (Ulrich et al. 2010), these results indicated that the phyllosphere fungal communities were sampled incompletely (Unterseher et al. 2011), suggesting that more diverse and comprehensive phyllosphere fungal species compositions would be detected with additional sampling and sequencing efforts per sample. Second, silica gel tissue storage may reduce the DNA quality and yield limited fungal compositions (U'Ren et al. 2014). Third, the plant samples analyzed represent only a fraction of the diverse flora inhabiting the tropical rainforest. Although planted *S. leprosula* trees were analyzed in this study, 10,000–12,000 species of flowering plant species are likely present on Borneo island (Soepadmo and Wong 1995). The fungal species diversity would thus be many times higher than that of associated plant communities (Bruns 1995; Dickie 2007). More diverse and complex fungal communities may be found in larger areas, including natural forests. Furthermore, seasonality may affect fungi associated with plant phyllospheres (Albrectsen et al. 2010; Jumpponen and Jones 2010; Davey et al. 2012). The leaf sampling in the present study was conducted during the dry season in Borneo; however, greater species richness or more alternative patterns of diversity may be

observed in phyllosphere fungal communities during the rainy season.

As shown in previous studies of phyllosphere endophytic fungi in various types of habitats (Arnold 2007), Ascomycota was dominant among the fungal OTUs observed in this study (Fig. 5-5A). Within Ascomycota, Sordariomycetes was more prevalent than Dothideomycetes (Fig. 5-5B), consistent with the patterns observed in tropical forests (Arnold and Lutzoni 2007). At the order level, Xylariales, Diaporthales (Sordariomycetes) and Pleosporales (Dothideomycetes), which were found in core OTUs, were consistent with fungi previously isolated from living leaf tissues of Dipterocarpaceae trees in northern Thailand (Osono et al. 2009; Sutjaritvorakul et al. 2011). Dominance of Sordariomycetes, especially Xylariales and Diaporthales, has also been reported in another tropical region (Arnold and Lutzoni 2007). Thus, the phyllosphere fungal flora of *S. leprosula* in Central Kalimantan was partly consistent with those in other tropical regions.

These characteristics in the taxonomic composition, however, were distinct from those in temperate forests. In a temperate forest in the state of Kansas, USA, the most prevalent fungal order found to inhabit *Quercus macrocarpa* (N = 18) was Pleosporales (Dothideomycetes), followed by Capnodiales (Dothideomycetes), Erysiphales (Leotiomycetes), Tremellales (Tremellomycetes) and Taphrinales (Taphrinomycetes) (Jumpponen and Jones 2009). In another temperate forest in Laveyron, France, Taphrinales (Taphrinomycetes) was detected at a high frequency on *Fagus silvatica* (N = 27), along with ubiquitous detection of *Aureobasidium pullulans* (Dothideales; Dothideomycetes) and *Mycosphaerella punctiformis* (Capnodiales; Dothideomycetes) (Cordier et al. 2012). The proportions of identified OTUs, which were not influenced by sequence lengths, were low. The difficulty of correct identifications of the obtained reads could be due to the absence of matching sequences in the "nt" database, which is likely because of the absence of previous comprehensive investigations of phyllosphere fungal flora in the Southeast Asian tropical rainforest.

Among the total 488 fungal OTUs in the dataset of 95% OTU similarity, 23 OTUs (4.7%) and the remaining 465 OTUs (95.3%) were analyzed as core and satellite OTUs, respectively (Fig. 5-4). The relative abundance of core and satellite OTUs revealed that tropical phyllosphere fungal communities were composed of more satellite OTUs than temperate forests. The ratio of core to satellite OTUs was 1:20.2 in the present study (Fig. 5-4), in comparison with 1:13.3 in Cordier et al. (2012), and 1:7.6 in Jumpponen and Jones (2009), both of which performed meta-barcoding analysis of phyllosphere fungal communities on a single woody plant species in temperate regions. The distribution of the 23 core OTUs were spatially structured (Table 5-1; Fig. 5-6A), presumably because of the limitations of fungal dispersal. The geographic effects on fungal community dissimilarity were previously known (Adams et al. 2013). Clear spatial structures in satellite OTUs' communities were also found (Table 5-1; Fig. 5-6B). Compositions of satellite OTUs were different even between plots 15 m apart resulting in the high heterogeneity of fungal communities within a plantation stand. These satellite OTUs may include transiently settling fungi, which do not necessarily have specific interaction with host trees (Unterseher et al. 2011), but likely accounted for most of the fungal diversity found in the study site.

Overall, the planted *Shorea* trees in SBK harbored unknown species diversity of

phyllosphere fungi: most of the fungal OTUs were found in low frequency and were unidentified at the species level. The high species diversity found in the target plantation stands could be attributed to the management method, which reduces the disturbance of understory vegetation (i.e., TPTJ; Ministry of Forestry 1998; Hardiansyah et al. 2006), and/or the retained intact forests surrounding SBK. However, to detect the effect of forest management on species diversity of phyllosphere fungi, fungal flora need to be compared between different plantation stands managed by different methods.

Tables and Figures

Table 5-1
Summary of three datasets and results of statistical tests for spatial variability of fungal communities.

Dataset[a]	Number of OTUs	Number of reads	PERMANOVA[c]			PERMDISP[d]	
			$F_{df=3}$	R^2	p	$F_{df=3}$	p
95% OTU similarity							
Core OTUs[b]	23	72990	9.8	0.52	<0.001	1.8	0.18
Satellite OTUs	465	80204	25.4	0.74	<0.001	1.3	0.31
97% OTU similarity							
Core OTUs[b]	20	71640	11.6	0.56	<0.001	2.2	0.11
Satellite OTUs	677	82660	35.0	0.80	<0.001	3.5	0.03
98.5% OTU similarity							
Core OTUs[b]	26	64174	9.3	0.51	<0.001	2.0	0.13
Satellite OTUs	1536	129207	29.1	0.76	<0.001	2.7	0.07

[a] The three datasets differ in the sequence similarity within operational taxonomic units (OTUs) (95%, 97% and 98.5%).
[b] Core OTUs represented by more than 50% of samples were distinguished from the remaining satellite OTUs.
[c] Permutation tests for multivariate analysis of variance
[d] Permutation tests for homogeneity of multivariate dispersions

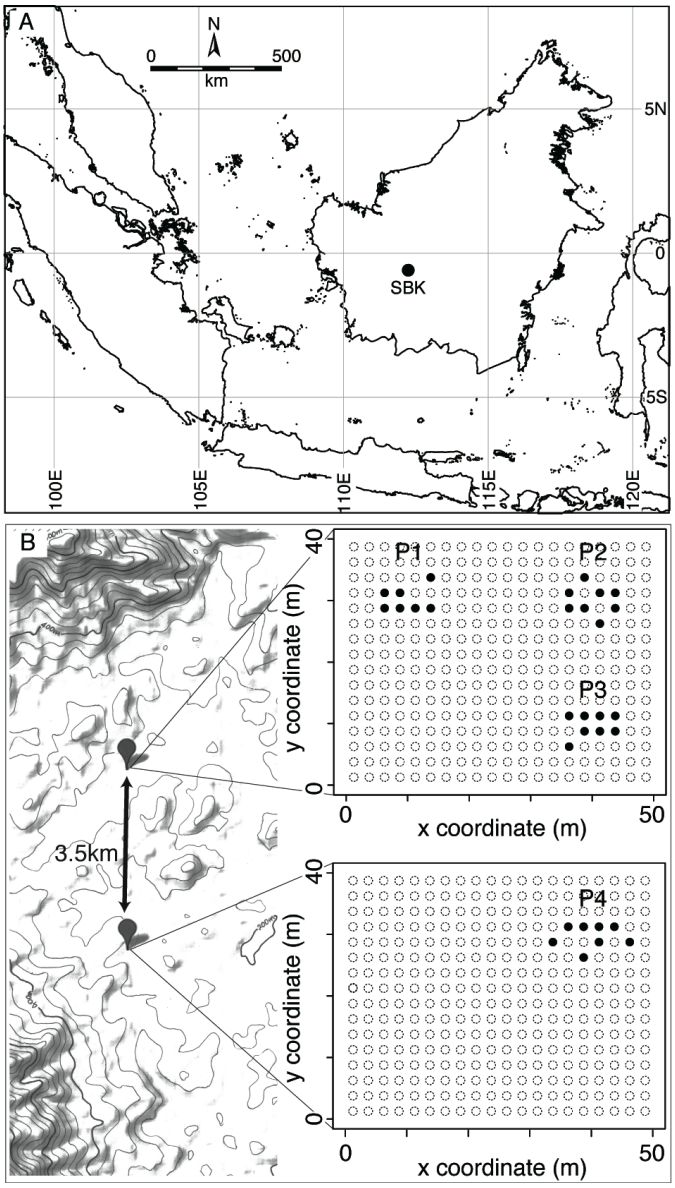

Fig. 5-1

Study site and plots where leaf samples were collected. A: Location of PT. Sari Bumi Kusuma (SBK) in Central Kalimantan, Indonesia. B: Map of the four study plots established in the SBK concession area. The four plots were distributed across two plantation stands of *Shorea leprosula* located 3.5 km apart. In each plantation stand, 320 trees were growing within a 200-m^2 area (white circles). *Shorea leprosula* leaves were collected from seven to eight trees per plot (black circles).

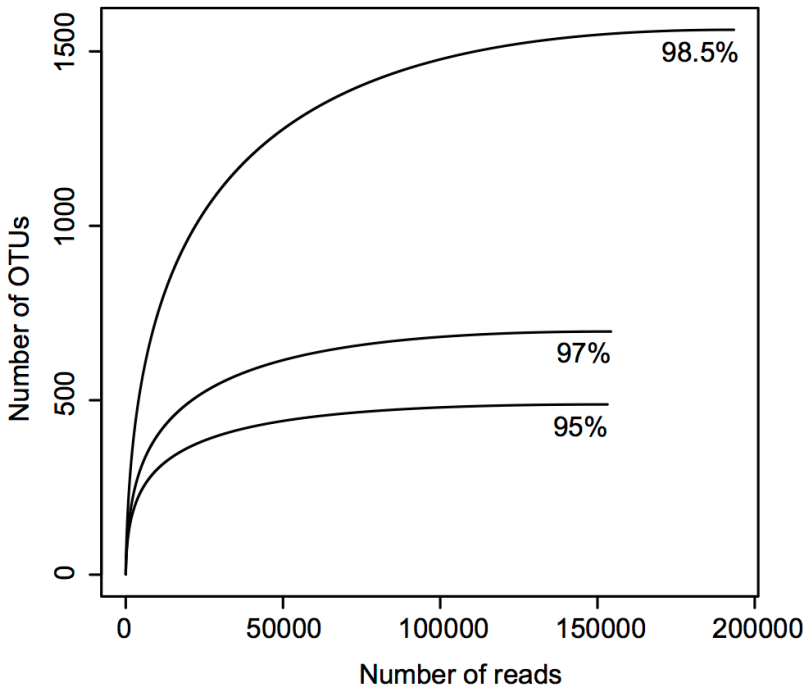

Fig. 5-2
Accumulation curves of fungal OTUs detected from *Shorea leprosula* leaves of the three datasets differing in sequence similarity within OTUs (95%, 97% and 98.5%).

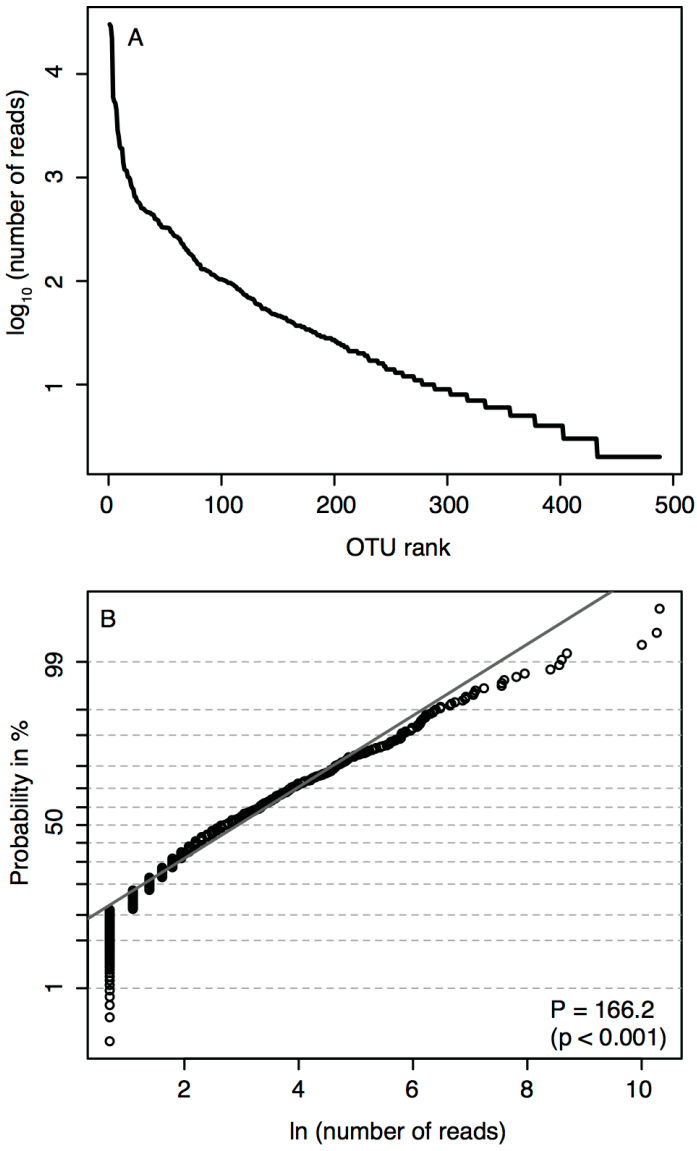

Fig. 5-3

Rank-abundance plot (A) and probability plot (B) of 488 OTUs in the dataset of 95% OTU similarity. In the probability plot (B), a gray line indicates a log-normal species abundance distribution. A chi-square test indicated significant deviation from a log-normal species abundance distribution of the observed OTU abundance (P = 166.2, $p < 0.001$). See Fig. 5-S4 for the datasets of 97% and 98.5% OTU similarity.

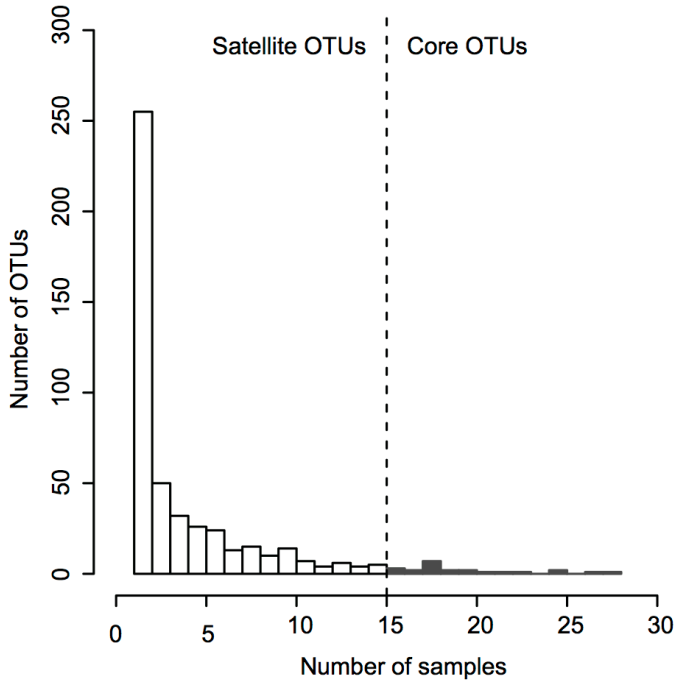

Fig. 5-4
Number of representative samples of 488 OTUs in the dataset of 95% OTU similarity. Twenty-three OTUs represented by more than 16 samples were classified as core OTUs and the remaining 465 OTUs as satellite OTUs.

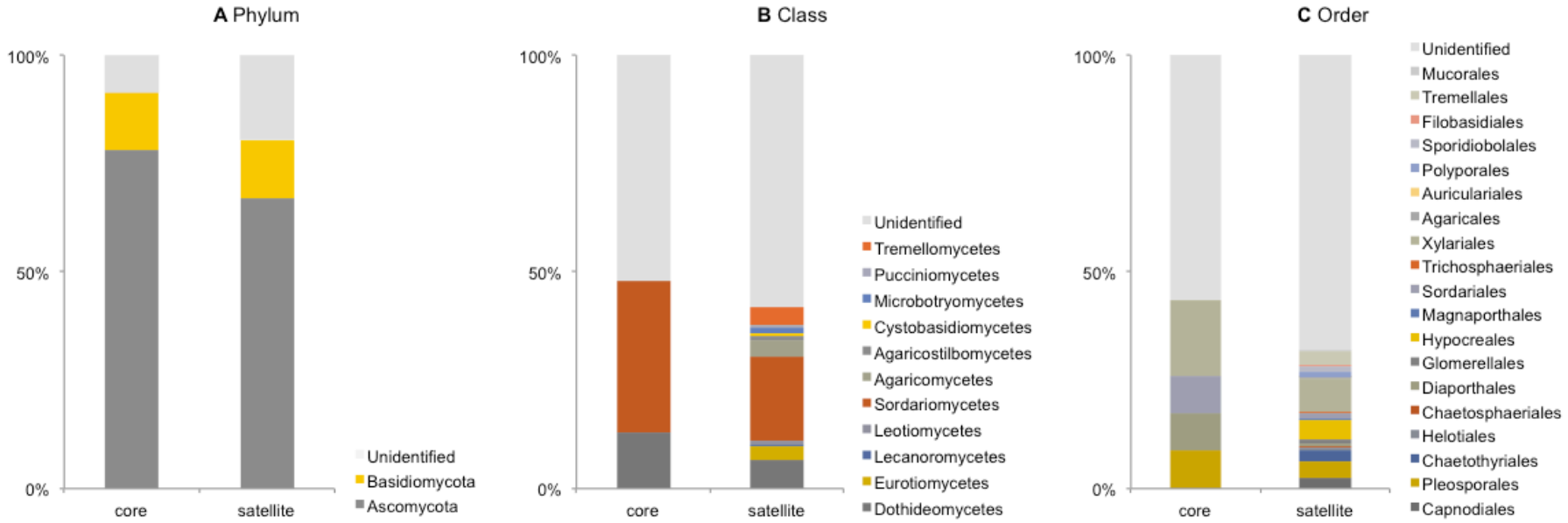

Fig. 5-5

Phylum (A), class (B) and order (C) level taxonomic composition of 23 core and 465 satellite fungal OTUs detected on *Shorea leprosula* leaves in the dataset of 95% OTU similarity.

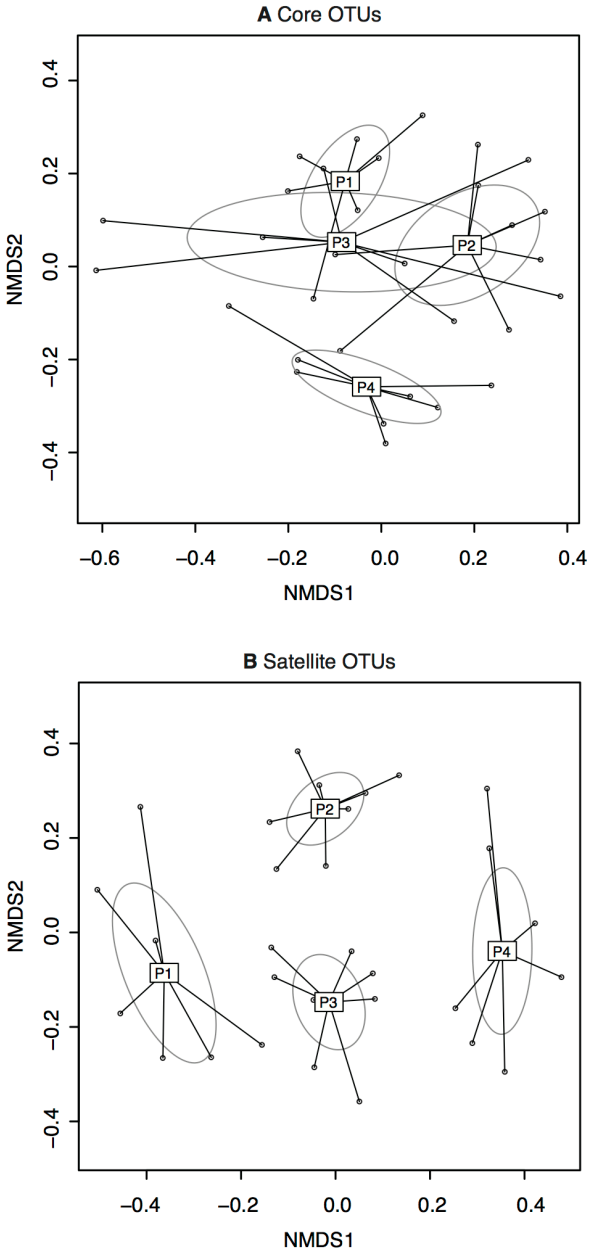

Fig. 5-6
Two-dimensional NMDS plots of fungal communities based on pairwise Raup–Crick dissimilarities in the fungal OTU compositions of 31 *Shorea leprosula* trees sampled from four plots (P1, P2, P3 and P4). NMDS ordinations were performed for 23 core (A) and 465 satellite (B) OTUs in the dataset of 95% OTU similarity. Ellipses indicate the standard deviations of the weighted averages of plots (see Fig. 5-1 for the detailed locations of plots).

Supplemental materials

Data 5-S1
Command script for Claident v0.2.2015.03.11 used for filtering, clustering and taxonomy identification of OTUs.

```
# Split the entire sequences according to tag sequences

clsplitseq ¥
   --runname=R1 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_1.fasta ¥
   --numthreads=10 ¥
   120913.fastq R1_samples

clsplitseq ¥
   --runname=R1 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_2.fasta ¥
   --numthreads=10 ¥
   120913.fastq --append R1_samples

clsplitseq ¥
   --runname=R1 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_3.fasta ¥
   --numthreads=10 ¥
   120913.fastq --append R1_samples

clsplitseq ¥
   --runname=R2 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_1.fasta ¥
   --numthreads=10 ¥
   121012.fastq R2_samples

clsplitseq ¥
   --runname=R2 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_2.fasta ¥
   --numthreads=10 ¥
   121012.fastq --append R2_samples

clsplitseq ¥
   --runname=R2 ¥
   --primerfile=barcodelist.fasta ¥
   --tagfile=taglist120_3.fasta ¥
   --numthreads=10 ¥
   121012.fastq --append R2_samples

# Filter reads

mkdir R1_filtered

gunzip R1_samples/R1__A*__ITS1F.fastq.gz

for fname in R1_samples/*.fastq
do
   clfilterseq ¥
       --minqual=27 ¥
       --minlen=150 ¥
```

```
      --numthreads=10 ¥
      --output=folder ¥
      --append ¥
      $fname R1_filtered
done

gzip R1_samples/R1__A*.fastq

mkdir R2_filtered

gunzip R2_samples/R2__A*__ITS1F.fastq.gz

for fname in R2_samples/*.fastq
do
  clfilterseq ¥
      --minqual=27 ¥
      --minlen=150 ¥
      --numthreads=10 ¥
      --output=folder ¥
      --append ¥
      $fname R2_filtered
done

gzip R2_samples/R2__A*.fastq

clcleanseq ¥
  uchime --minh 0.1 --mindiv 0.8 end ¥
  --numthreads=10 ¥
  R1_filtered/R1__A*.fastq R1_clean

clcleanseq ¥
  uchime --minh 0.1 --mindiv 0.8 end ¥
  --numthreads=10 ¥
  R2_filtered/R2__A*.fastq R2_clean

# Clustering within samples (OTU similarity = 0.99)

mkdir withinsamples

clclassseq ¥
  --minident=0.99 ¥
  --strand=plus ¥
  --numthreads=10 ¥
  --append ¥
  R1_clean/R1__A*.cleaned.dereplicated.fastq.gz ¥
  withinsamples

clclassseq ¥
  --minident=0.99 ¥
  --strand=plus ¥
  --numthreads=10 ¥
  --append ¥
  R2_clean/R2__A*.cleaned.dereplicated.fastq.gz ¥
  withinsamples

# Clustering among samples (OTU similarity = 0.99)

clclassclass ¥
  --minident=0.99 ¥
  --strand=plus ¥
```

```
  --numthreads=10 ¥
  withinsamples/*.assembled.fastq.gz ¥
  amongsamples

# Re-clustering among samples (OTU similarity = 0.95)

clreclassclass ¥
  --minident=0.95 ¥
  --strand=plus ¥
  --numthreads=10 ¥
  amongsamples ¥
  clustering_95

clsumclass ¥
  --output=matrix ¥
  clustering_95/assembled.contigmembers.gz ¥
  clustering_95/clustering_95.txt

clfiltersum ¥
  --samplelist=SBKsamplelist.txt ¥
  --minntotalseqotu=1 ¥
  clustering_95/clustering_95.txt ¥
  clustering_95/clustering_95_lepro.txt

clfiltersum ¥
  --minntotalseqotu=2 ¥
  clustering_95/clustering_95_lepro32.txt ¥
  clustering_95/clustering_95_lepro32_rmsingleton.txt

# Re-clustering among samples (OTU similarity = 0.97)

clreclassclass ¥
  --minident=0.97 ¥
  --strand=plus ¥
  --numthreads=10 ¥
  amongsamples ¥
  clustering_97

clsumclass ¥
  --output=matrix ¥
  clustering_97/assembled.contigmembers.gz ¥
  clustering_97/clustering_97.txt

clfiltersum ¥
  --samplelist=SBKsamplelist.txt ¥
  --minntotalseqotu=1 ¥
  clustering_97/clustering_97.txt ¥
  clustering_97/clustering_97_lepro.txt

clfiltersum ¥
  --minntotalseqotu=2 ¥
  clustering_97/clustering_97_lepro32.txt ¥
  clustering_97/clustering_97_lepro32_rmsingleton.txt

# Re-clustering among samples (OTU similarity = 0.985)

clreclassclass ¥
  --minident=0.985 ¥
  --strand=plus ¥
  --numthreads=10 ¥
```

```
  amongsamples ¥
  clustering_985

clsumclass ¥
  --output=matrix ¥
  clustering_985/assembled.contigmembers.gz ¥
  clustering_985/clustering_985.txt

clfiltersum ¥
  --samplelist=SBKsamplelist.txt ¥
  --minntotalseqotu=1 ¥
  clustering_985/clustering_985.txt ¥
  clustering_985/clustering_985_lepro.txt

clfiltersum ¥
  --minntotalseqotu=2 ¥
  clustering_985/clustering_985_lepro32.txt ¥
  clustering_985/clustering_985_lepro32_rmsingleton.txt

# Taxonomic identification of OTUs (OTU similarity = 0.95)

clidentseq ¥
  blastn -strand plus end ¥
  --blastdb=fungi_ITS_species ¥
  --numthreads=10 ¥
  clustering_95/assembled.fasta ¥
  clustering_95/blastID_95.txt

classigntax ¥
  --taxdb=fungi_ITS_species ¥
  clustering_95/blastID_95.txt ¥
  clustering_95/blastName_95.txt

clidentseq ¥
  blastn -strand plus end ¥
  --blastdb=fungi_ITS_species ¥
  --numthreads=10 ¥
  clustering_95/clustering_95_lepro32.fasta ¥
  clustering_95/blastID_95_lepro32.txt

classigntax ¥
  --taxdb=fungi_ITS_species ¥
  clustering_95/blastID_95_lepro32.txt ¥
  clustering_95/blastName_95_lepro32.txt
```

Table 5-S2

Number of raw reads, filtered and retained reads and OTUs for each sample.

Sample ID	Plot	Sequencing run [a]	Raw reads	Trimmed reads [b]	Chimeric reads	Noisy reads	Used for clustering	Clutering 95%					
								OTUs	OTUs contained in negative control	OTUs not assigned as fungi	Singleton	Retained OTUs	Retained reads
Sle01	P1	R1	41950	32278	461	5749	3462	65	7	15	0	43	2762
Sle02	P1	R1	23969	20019	209	2621	1120	75	7	16	0	52	316
Sle03	P1	R1	29131	23051	316	3635	2129	60	9	11	0	40	670
Sle04	P1	R1	23307	17205	380	3727	1995	44	9	8	0	27	395
Sle05	P1	R1	49329	37532	577	7326	3894	89	7	13	0	69	2586
Sle06	P1	R2	18373	9219	1097	3297	4760	92	6	19	1	66	626
Sle07	P1	R1	10091	7690	206	1158	1037	36	5	1	1	29	186
Sle08	P2	R2	21963	11761	378	5442	4382	39	3	4	1	31	4069
Sle09	P2	R2	121085	52562	5990	21327	41206	49	9	4	3	33	8256
Sle10	P2	R2	37398	16862	1279	7653	11604	53	6	4	3	40	11323
Sle11	P2	R2	21622	12361	850	4735	3676	78	7	11	2	58	3185
Sle12	P2	R2	25885	14091	1036	5420	5338	49	7	4	1	37	3748
Sle13	P2	R2	48127	27097	1923	7929	11178	81	7	12	1	61	9773
Sle14	P2	R2	4315	2138	218	647	1312	20	5	1	0	14	1293
Sle15	P2	R2	46089	23942	1712	10156	10279	48	7	5	1	35	5931
Sle16	P3	R2	118359	54079	8931	16627	38722	161	9	24	3	125	4256
Sle17	P3	R2	121785	52613	10468	18161	40543	150	9	25	1	115	6479
Sle18	P3	R2	44033	18783	3384	5644	16222	61	7	5	0	49	832
Sle19	P3	R2	96813	43257	7008	18260	28288	187	9	44	4	130	7300
Sle20	P3	R2	22519	8938	2057	3946	7578	80	7	11	1	61	798
Sle21	P3	R2	119380	57084	8464	17784	36048	178	9	32	3	134	5530
Sle22	P3	R2	50205	23665	3469	9693	13378	173	9	25	4	135	3048
Sle23	P3	R2	57017	23640	4907	9107	19363	91	7	13	3	68	2199

Sle24	P4	R2	64377	35111	3462	9845	15959	102	8	26	0	68	1151
Sle25	P4	R2	102319	51878	5203	14343	30895	101	8	17	1	75	3751
Sle26	P4	R2	100754	40556	2377	19547	38274	130	7	29	1	93	36378
Sle27	P4	R2	68160	35325	4021	13702	15112	131	7	27	0	97	6996
Sle28	P4	R2	41983	21638	2603	6802	10940	102	8	15	2	77	2675
Sle29	P4	R2	54925	23885	3897	11716	15427	132	7	25	3	97	5945
Sle30	P4	R2	135721	58287	10756	20705	45973	98	9	7	3	79	6744
Sle31	P4	R2	64040	32349	3275	11996	16420	153	8	32	3	110	3993
Total			1785024	888896	100914	298700	496514	888	11	343	46	488	153194
Average												69	4942

[a] R1 and R2 indicates the first and second Ion-PGM sequencing run, respectively

[b] Short reads (<150 bp) + reads with low quality

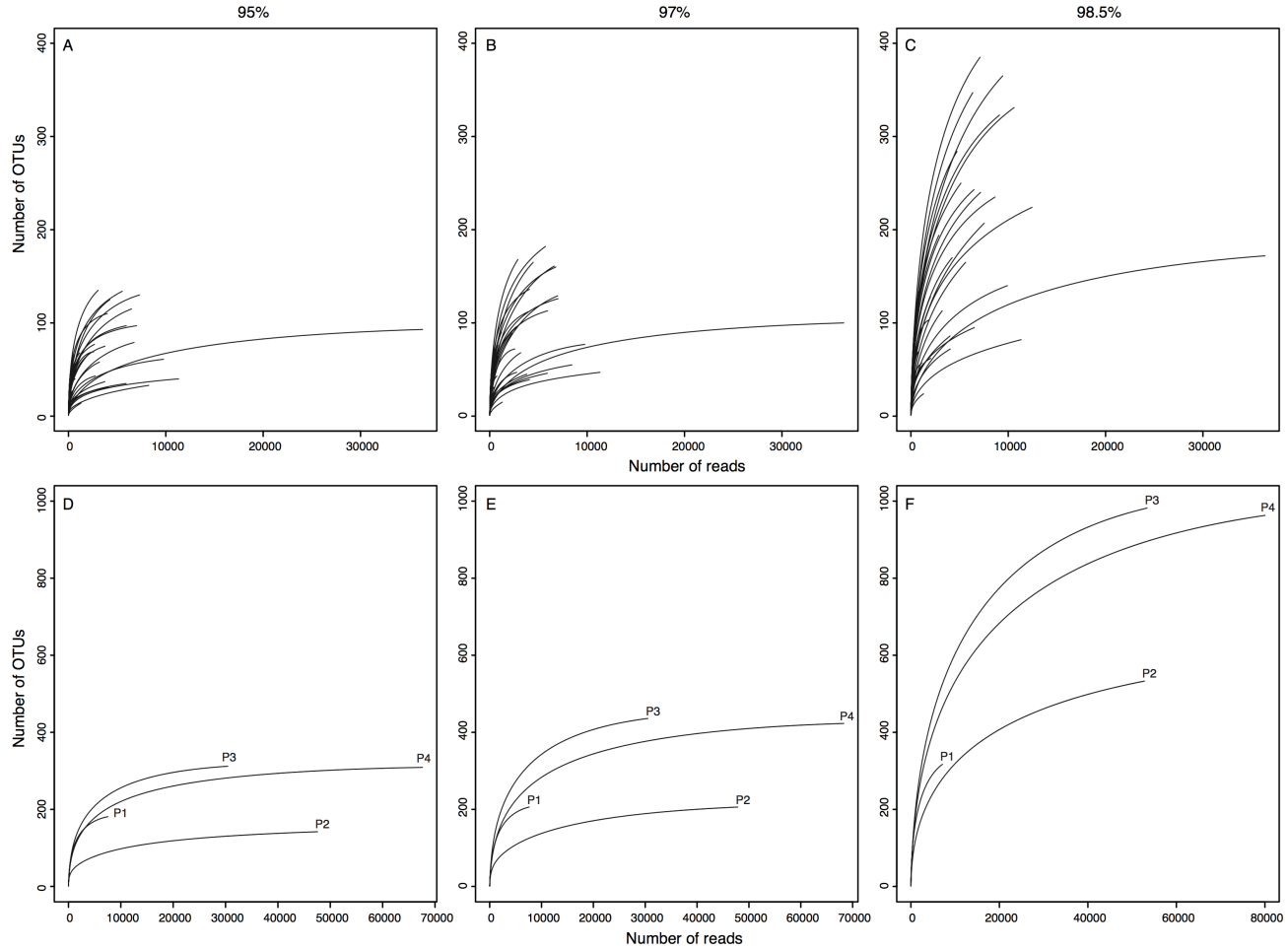

Fig. 5-S3

Accumulation curves of fungal OTUs per sample (A–C) and per plot (D–F) of three datasets differing in sequence similarity within OTUs (95%, 97% and 98.5%).

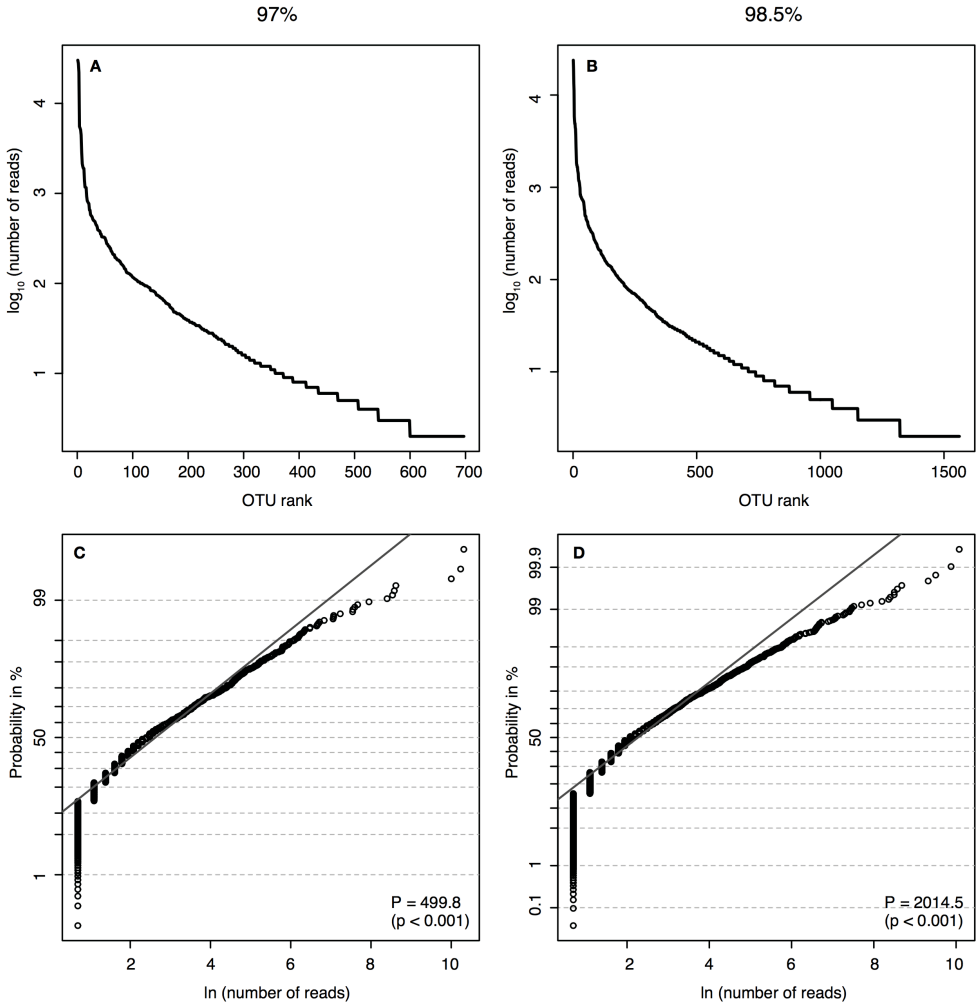

Fig. 5-S4

Rank-abundance plots (A, B) and probability plots (C, D) of 697 and 1562 OTUs in the datasets of 97% and 98.5% OTU similarity, respectively. In the probability plot (C, D), gray lines indicate a log-normal species abundance distribution. Chi-square tests indicated significant deviation from a log-normal species abundance distribution of the observed OTU abundance (P = 499.8, $p < 0.001$ for the dataset of 97% OTU similarity; P = 2014.5, $p < 0.001$ for the dataset of 98.5% OTU similarity). See Fig. 5-3 for the dataset of 95% OTU similarity.

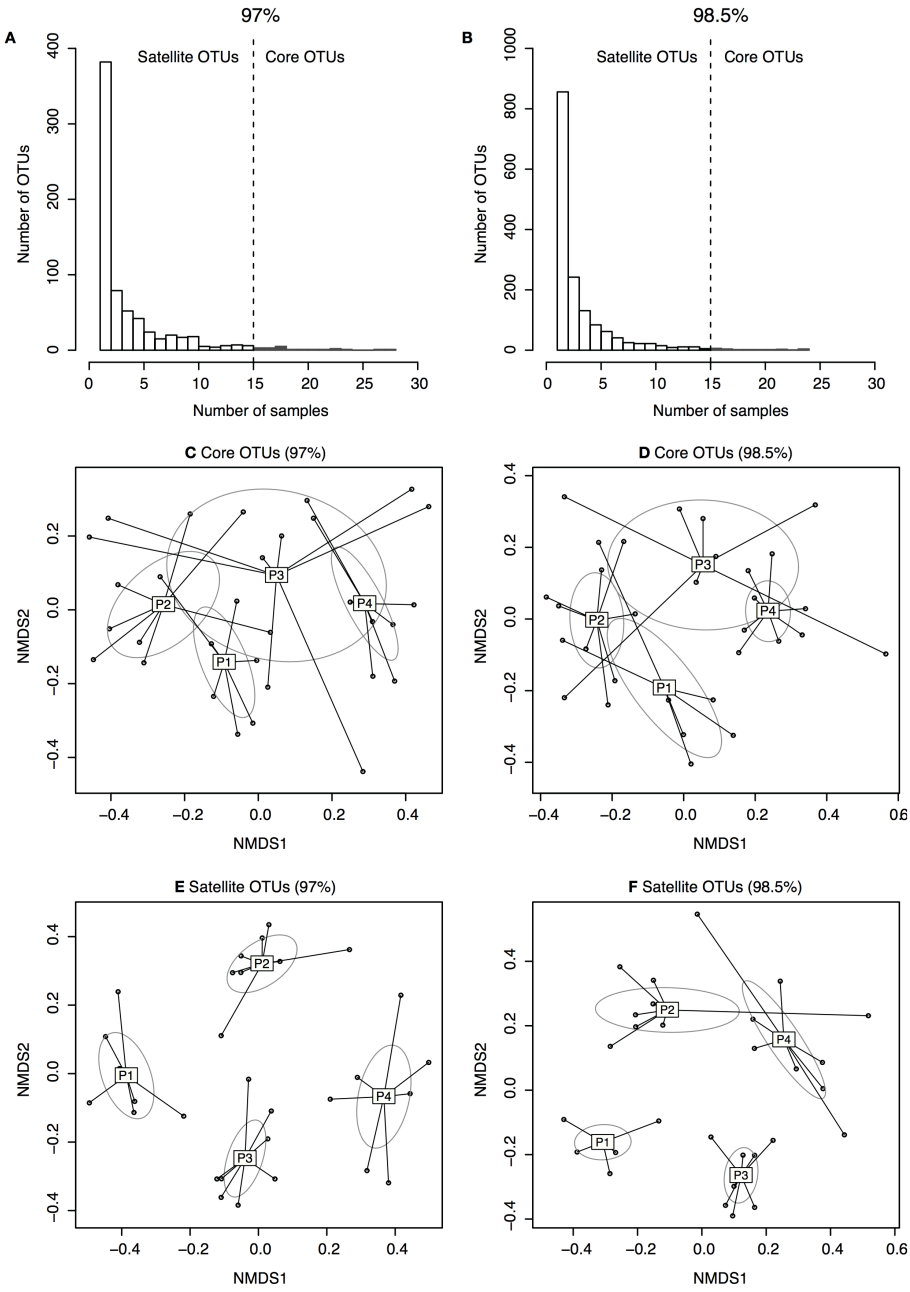

Fig. 5-S5

(A, B) Number of representative samples of 697 and 1562 OTUs in the dataset of 97% and 98.5% OTU similarity, respectively. See Fig. 5-4 for the dataset of 95% OTU similarity. (C–F) Two-dimensional NMDS plots of each fungal community based on pairwise Raup-Crick dissimilarities in the fungal OTU compositions of 31 *Shorea leprosula* trees sampled from four plots (P1, P2, P3 and P4). NMDS ordinations were performed for 20 core (C) and 677 satellite (D) OTUs in the dataset of 97% OTU similarity and 26 core (E) and 1536 satellite (F) OTUs in the dataset of 98.5% OTU similarity. Ellipses indicate the standard deviations of the weighted averages of plots (see Fig. 5-1 for the detailed locations of plots). See Fig. 5-6 for the dataset of 95% OTU similarity.

6

General discussion

6.1. Ecological genomics of *Metrosideros polymorpha*

6.1.1. Environmental adaptations of *M. polymorpha*

In Chapters 2 and 3, I provided genetic evidence underlying the environmental adaptations of *M. polymorpha*, an ecologically and morphologically divergent tree species. In the population genomic analyses, genomic mosaics on the genome of this species were found: each locus showed specific divergence derived by a different factor. Most of the genome was admixed through gene flows, depending on the geographic distance (Table 3-2) and a large degree of genetic variation was retained within populations (Fig. 3-7). On the other hand, a small fraction of the genome showed non-neutral divergence (Fig. 3-4) and correlated to environmental differences between populations (Table 3-2); thus, the largely diverged loci likely played key roles in the adaptations of this species to diverse environments. The genes with relevant biological functions were detected within 10 kb from the non-neutral genetic variations (Table 3-S4). Therefore, it is concluded that *M. polymorpha* had effectively developed specific functional traits that strongly affect environmental adaptations and attained various ecological niches in the Hawaiian Islands. Strong environmental selections that overcome the effects of large gene flows act on the divergence in a small fraction of the genome and maintain the large variations in the phenotype. The demographic analyses based on the genome-wide polymorphism indicated that this species had experienced large population growth, possibly according to the continual formations of volcanic islands and a subsequent rapid bottleneck that may have been caused by geographic or climatic factors (Fig. 2-2). An additional outcome of this thesis is the high-quality genomic resource of *M. polymorpha*; half of the genome was successfully reconstructed as the actual chromosomes (Table 2-2; Fig. 2-1). This genomic resource can be widely

used for ecological genomics in *Metrosideros* species or comparative genomics focusing on evolutionary events across higher taxa.

These findings are brought about by the genome-wide analysis, i.e., the use of thousands of SNP markers extracted throughout the genome and whole genome sequencing. Three advances have been made by the use of genome-wide SNP markers. First, unlike the conventional genetic markers (Harbaugh et al. 2009; DeBoer and Stacy 2013; Stacy et al. 2014), thousands of genome-wide SNP markers captured sufficient genetic variations in each individual to distinguish whether the sources of the variation are within and among populations (Fig. 3-2), and resolved the complex population genetic structures in which individuals are largely admixed (Wagner et al. 2010 focusing on African cichlids) Second, by applying a genomic scanning (Foll and Gaggiotti 2008), non-neutrally diverged SNP sites were found. Finally, genomic mosaics and the genomic status in the context of general speciation were demonstrated. These broad views throughout a genome developed our understanding about genetic differentiations, in other words, the genetic basis of adaptation of this species.

Whole genome sequences also contributed to the population genomics of this species. Only the whole genome sequence information can annotate the non-neutral SNPs. A homology search (e.g., BLAST) of the short nucleotide fragments (49 bp) on which the non-neutral variants exist did not detect homologous sequences in the public nucleotide databases (Izuno A. pers. observ.), because of insufficient length to specify the homology between the query and database sequences. Therefore, in non-model organisms, utilizing precise genome information or closely related to the target species should be required to detect biological functions of any nucleotide variants. Alternatively, RAD-seq with long sequences (≥90–100 bp) should be performed. Parameters such as heterozygosity or effective population size, which are conventionally obtained by sampling multiple individuals, are also effectively obtained based on the whole genome information of a single individual.

Thus, by conducting *de novo* genome sequencing of *M. polymorpha*, I have updated our knowledge about the ecological adaptations of this species and clearly demonstrated the effectiveness and possibility of genome-wide analyses in the field of ecology and evolution.

6.1.2. Further questions to be answered regarding the ecological genomics of *M. polymorpha*

To achieve the goal of revealing the genetic mechanisms and environmental factors that affected the dramatic ecological divergence of this species, further questions still need to be addressed (Fig. 6-1).

The first issue is to identify the adaptation genes that play roles in the development of adaptive traits and contribute to the ecological divergence of this species. Although several candidate loci have been identified by genome scanning in Chapter 3, further candidates can be comprehensively identified by alternative methodologies such as the RNA-seq that focuses on the differentially expressed genes between individuals with different morphologies or the genome-wide association

analysis (GWAS) that detects the loci significantly correlated to trait values. Ecological functions and the effect on fitness of the adaptive candidate genes also require validation (see also 6.3.1.). Moreover, the origin of the variation at adaptive genes can be useful in understanding the evolution of this species. A molecular phylogeny for an adaptive gene of *M. polymorpha* individuals growing in the various habitats and other *Metrosideros* species growing out of the Hawaiian Islands would answer the questions of when and how the genetic divergence was generated and selected during the adaptive radiations within the species.

Another issue is the historical demography of this species. The PSMC model (Chapter 2) is considered to be effective for the whole genome sequence data obtained from even one individual, but its accuracy in the more recent past is suspicious because few genetic variations were contained in an individual coalesced at the periods (Sheehan et al. 2013). By incorporating genetic variations from multiple genomes, the PSMC model would reconstruct the recent demography (Sheehan et al. 2013). Moreover, the effects of leaf morphologies or habitats on the population histories should be examined. Inferring the divergence time between the populations differing in leaf morphologies could promote our understanding about the historical processes of ecological divergence in this species.

6.2. Understanding and conserving the biodiversity in dipterocarp tree plantations in Southeast Asia

In Chapters 4 and 5, I evaluated the appropriateness of the enrichment planting of dipterocarp trees conducted in PT. Sari Bumi Kusuma (SBK) in terms of population genetics of planted trees and species diversity of the phyllosphere fungi associated with them. The planted trees harbored genetic diversity as high as natural populations (Table 4-1), indicating that the original genetic diversity of the species was not deteriorated during the process of planting. As the genetic diversity in the host plants could contribute to the species diversity at the ecosystem level (Crutsinger et al. 2006; Whitham et al. 2006), the high genetic diversity maintained in the planted populations is desirable in terms of ecosystem functioning. No significant genetic differentiation was found between planted and natural populations (Table 4-4). Thus, genetic pollution of surrounding natural forests through gene flows does not likely occur, and the genetic compositions of the planted dipterocarp trees were also favorable for the genetic diversity in the surrounding primary forests. In the plantation stands, diverse fungi including unknown or invisible species inhabited the leaves (Fig. 5-3, 5-4, 5-5). The local compositions of fungal assemblages differed within a stand (Fig. 5-6). At present, the planted dipterocarp trees in the studied forest management unit of SBK represent the subset of the local forests allowing the association with diverse phyllosphere fungi.

Although the ecological values in the enrichment planting were partially evaluated in this thesis, further studies, particularly focusing on the long-term significances, are required to accomplish the sustainable enrichment plantations. To sustain the supply of appropriate planting materials, which should be collected from

the surrounding natural forests, the collected wild seedlings should be stored in a healthy state until used for planting. The wild seeds or seedlings of dipterocarp trees can be obtained during limited years because of their nature of masting behaviors; thus, the maintenance of nurseries or the establishment of cutting techniques, which have already been developed by the local concessions, are effective to conserve the planting sources collected. In addition, because the natural regeneration in the surrounding natural forests cannot supply sufficient sources for the intensive enrichment plantations, seed orchards in which trees are grown to reproduce and supply novel genetic diversity should be established using the local dipterocarp trees. This may prevent the overexploitation of wild seeds or seedlings, which are to constitute the natural forests, for the use of plantations. The genetic monitoring through harvesting cycles could reveal the cumulative genetic differences between planted and wild trees and provide ways to improve the yields and genetic diversity of planted trees. If the genetic tagging on the planting sources is realized, the optimal plantation designs can be suggested through simulation studies which predict the resulting genetic characteristics under scheduled planting or harvesting strategies (Sebbenn et al. 2008; Ng et al. 2009).

The values regarding biodiversity conservation in plantation stands have been unclear (Wilcove et al. 2013). One time rotation probably induced little loss of biodiversity, as shown by the fact that the species richness found in primary forests was retained in selectively logged forests, although the abundances were decreased in selectively logged forests (Cannon et al. 1998; Berry et al. 2010). However, more than two rotations of logging are likely to induce a serious loss of biodiversity (Ghazoul 2002; Shahabuddin et al. 2010). The lower damages of one-time logging also contribute to retain the original biodiversity; the reduced impact logging succeeded in the prevention of the loss of biodiversity as compared with the conventional destructive logging methods (Pinard et al. 2000). Therefore, the interval time between harvests and the amount of logging per harvest should be optimized to balance between the commercial yields and the conservation of biodiversity. Alternatively, connecting the plantation stands that are managed separately could enlarge the habitats and population sizes of wild organisms and thus mitigate the damage to the native biodiversity.

In addition to the ecological aspects, the economical evaluations on plantation or forest management strategies are also required. A policy that enlarges the economic values of secondary forests other than oil palm plantations could improve the circumstances. To induce such a policy, ecologists need to develop measurements to evaluate the ecological functioning of secondary forests and suggest effective management to prevent further loss of biodiversity.

6.3. Perspectives for ecological and conservation genetics

The focuses of ecological and conservation genetics have been dramatically expanded through remarkable advances in high-throughput sequencing technologies. As whole genome or genome-wide information can be obtained for any organism, the gaps

between model and non-model organisms have been reduced (Ellegren 2014). This trend is also true in genetic studies of trees. Challenges experienced such as large genome sizes and long generation times, which prevented whole genome sequencing of many tree species, can be overcome by the further development of the sequencing technologies (Neale and Kremer 2011). In this section, I highlight the prospects in ecological and conservation genetics based on genome-wide information and the expectations for forest tree genomics.

6.3.1. Genetic bases of adaptations

Non-neutral genetic variations can be detected with genome-wide information by genome scanning (Foll and Gaggiotti 2008; Chapter 3) or neutrality tests (Tajima 1989). A genome-wide association analysis (GWAS) is also a powerful tool for identifying loci significantly related to trait values (Dalman et al. 2013; Campbell et al. 2014; McKnown et al. 2014), climate conditions (Hancock et al. 2011), and associated microbe assemblages (Horton et al. 2014). These methodologies facilitate the identification of genomic traces of adaptations by any species, including undetectable traces as expressed traits. Therefore, in the near future, we can address the various adaptive phenomena that classical model organisms do not experience (Feder and Mitchell-Olds 2003) or that are controlled by novel genetic mechanisms such as epigenetics or non-coding RNA, which have not been fully studied yet.

Although the detection of the candidates for adaptive genes has become relatively easy, the confirmation of candidates as truly adaptive remains a challenge in non-model organisms. To argue the ecological relevance of candidate genes found in non-model organisms, we still have to relay on mutants of model organisms (Fig. 6-2 (A)). We need to identify genes homologous to the target candidate genes, conduct transgenic experiments, and eventually examine the expressions of the homologous genes (Fig. 6-2 (B); Kobayashi et al. 2012). In this scheme, the precise identification of homologous loci with the focal genes critically affects the results. In addition, conventional ecological studies to prove the links between traits and fitness are essential as well (Fig. 6-2 (C)).

Once adaptive genes are identified, the extent of natural selection acting on the adaptive loci can be quantified by comparisons with the extent of gene flows at neutral loci. Interspecific comparisons of the genetic mechanisms underlying conserved functional traits, such as flower color (Bradshaw and Schemske 2003; Rausher 2008; Ortiz-Barrientos 2013), could provide novel insights about the origin and evolution of the traits. Genotypes at adaptive loci could also provide significant information for conservation. Conservational decisions including the determination of evolutionary significant units, suitable individuals for *ex situ* conservation, and the strategies for restoration could be improved by considering local adaptations. However, genetic markers should be appropriately selected. As long as the conservation strategies of a species demand information that does not directly involve fitness, the limited numbers of neutral loci can sufficiently characterize the genetic status (Kaneko et al. 2013). These views on conservation based on the adaptive genetic information can be applied to global forest management. Because the climatic tolerance of tree populations are often

narrower than the apparent distributions (Kremer et al. 2012), the genes responsible for the environmental adaptations and the spatial distribution of adaptive genotypes would be informative to facilitate appropriate forest management. Moreover, information on the genetic bases of specific traits could be useful in tree breeding programs. The detection of the genes responsible for the target traits such as tree height, wood density, and resistance to pests could enable the realization of marker-assisted selection, which is recognized as an effective selection scheme, but has remained difficult to implement for forest trees because of the rapid decay of linkage disequilibrium in tree species or the variable expressions of trait-associated genes under different environmental conditions (Neale and Kremer 2011).

6.3.2. Increased reliability of population genetic data via numerous neutral loci

Both the number of genetic markers and samples can increase the reliability of population genetics data. Previously, when the characterization of even a handful of neutral markers was expensive in terms of time and money for most wild organisms, we had to increase the number of samples to obtain reliable data to solve population genetic structures. The breakthrough in genomics has altered the options of genetic markers and thus sampling designs. Genome-wide SNP markers are now available at lower cost in any species, and consequently, we can improve the reliability of population genetic data by merely increasing the number of loci rather than the number of samples. With genome-wide genetic variations, even a single individual contains sufficient has enough information to characterize the population (Chapter 2; Li and Durbin 2011).

This remarkable change has set ecological genetics on a new course. The increased number of neutral loci can clarify the population genetics structures that were not solved by conventional genetic markers because of large gene flows among populations (Chapter 3; Wagner et al. 2012) or introgression (Twyford and Ennos 2012; Hipp et al. 2014; Eaton et al. 2015). The increased number of neutral loci can also be used in cases of polyploid or rare species with small population sizes, to which general assumptions in population genetics are difficult to apply (Allendorf et al. 2010; Defresne et al. 2013; Zhou et al. 2013). A large number of neutral genome-wide markers are expected to reveal population the genetic structures of tree populations across contrasting environments or population histories along the fluctuating climates of the past. Although tree populations are likely to be tolerant to climate to some extent, they are also tolerant to unfavorable environmental conditions (Kremer et al. 2012). Therefore, we need to evaluate not only the spatial distributions of adaptive genetic variations but also the spatial ranges of gene flows to predict the adaptive potential of forest tree populations after climate changes in the near future.

Overall, the field of ecological and conservation genetics for wild organisms, including forest trees, is becoming an interesting platforms to uncover idiosyncratic ecological phenomena by taking the advantage of massive genomic data. It is expected that we will be able to discuss a more specific evolutional process regarding

fitness-related traits and predicting the impact of climate change or artificial management on the fitness and longevity of target populations.

Figures

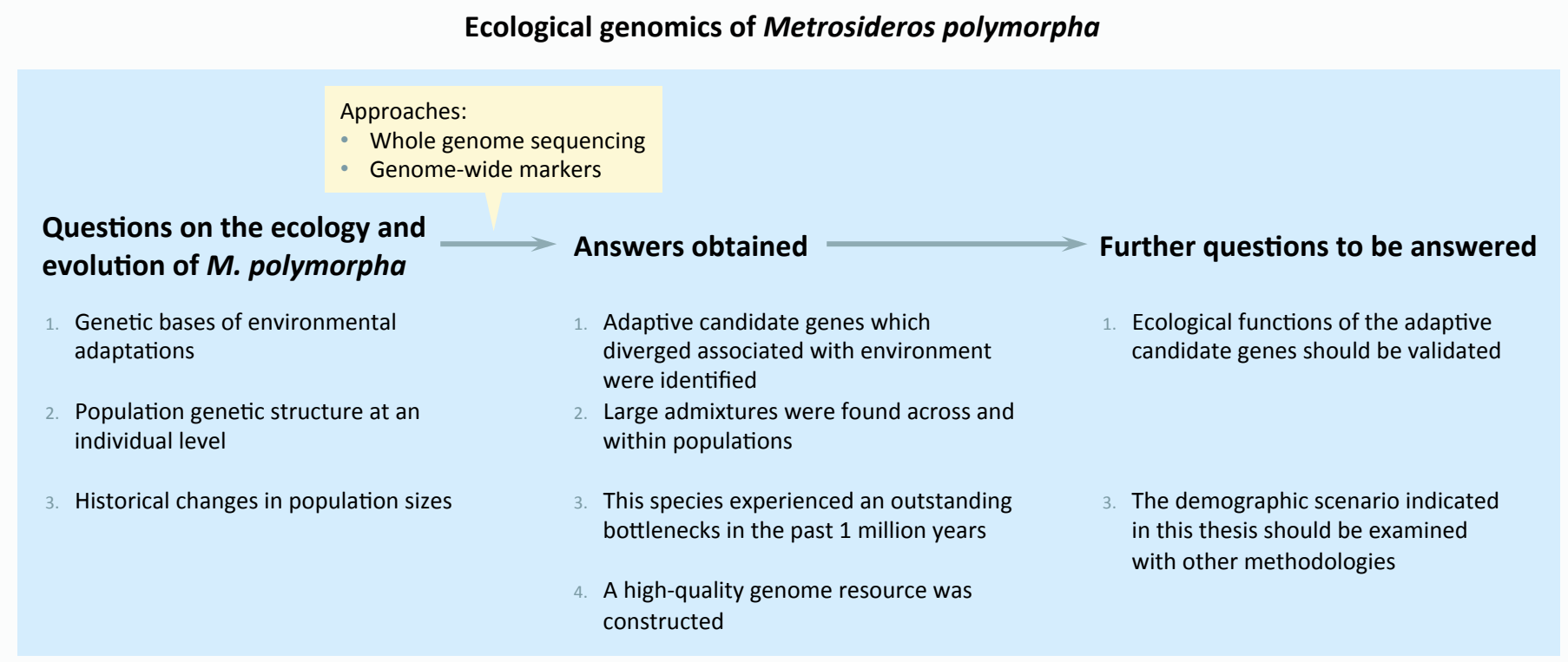

Fig. 6-1
Study scheme of the ecological genomics of *Metrosideros polymorpha*.

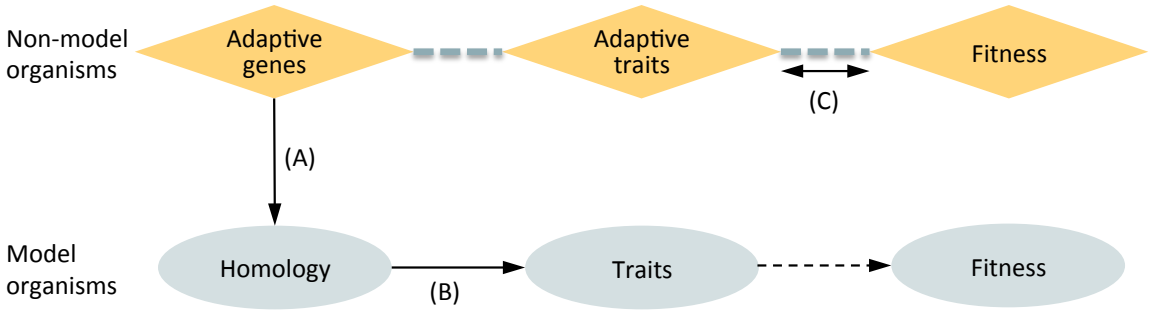

Fig. 6-2
Strategies to prove the ecological relevance of adaptive candidate genes detected in non-model organisms. (A) To argue the ecological relevance of candidate genes found in non-model organisms, we still have to rely on mutants of model organisms. (B) We need to conduct transgenic experiments and examine the expressions of the homologous genes. (C) Ecological studies to prove the links between traits and fitness are also essential.

References

Adams RI, Miletto M, Taylor JW, Bruns TD (2013) Dispersal in microbes: fungi in indoor air are dominated by outdoor air and show dispersal limitation at short distances. *The ISME Journal*, **7**, 1262–1273.

Adjers G, Otsamo A (1996) Seedling production methods of dipterocarps. In: *Dipterocarp forest ecosystems: towards sustainable management* (eds Schulte A, Schone D), pp. 391–410. World Publishing Co. Pte. Ltd., Singapore.

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, **1**, 95–111.

Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J (2014) Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Research*, **42**, e46.

Al-Dous EK, George B, Al-Mahmoud ME *et al.* (2011) *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology*, **29**, 521–527.

Albrectsen BR, Bjorken L, Varad A *et al.* (2010) Endophytic fungi in European aspen (*Populus tremula*) leaves—diversity, detection, and a suggested correlation with herbivory resistance. *Fungal Diversity*, **41**, 17–28.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews Genetics*, **11**, 697–709.

Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Anderson MJ (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, **62**, 245–253.

Anderson MJ (2008) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.

Ando H, Setsuko S, Horikoshi K *et al.* (2013) Diet analysis by next-generation sequencing indicates the frequent consumption of introduced plants by the critically endangered red-headed wood pigeon (*Columba janthinanitens*) in oceanic island habitats. *Ecology and Evolution*, **3**, 4057–4069.

Appanah S (1998) Management of natural forests. In: *A Review of Dipterocarps: Taxonomy, Ecology, and Silviculture* (eds Appanah S, Turnbull JM), pp. 133–149. Center for International Forestry Research, Bogor, Indonesia.

Aradhya KM, Mueller-Dombois D, Ranker TA (1993) Genetic structure and differentiation in *Metrosideros polymorpha* (Myrtaceae) along altitudinal gradients in Maui, Hawaii. *Genetical Research*, **61**, 159–170.

Arnold AE (2007) Understanding the diversity of foliar endophytic fungi: progress, challenges, and frontiers. *Fungal Biology Reviews*, **21**, 51–66.

Arnold AE, Lutzoni F (2007) Diversity and host range of foliar fungal endophytes: Are tropical leaves biodiversity hotspots? *Ecology*, **88**, 541–549.

Arnold AE, Mejía LC, Kyllo D *et al.* (2003) Fungal endophytes limit pathogen damage in a tropical tree. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15649–15654.

Ashton PS (2004) Dipterocarpaceae. In: *Tree flora of Sabah and Sarawak: volume one.* (eds Soepadmo E, Saw LG, Chung RCK), pp. 63–388. Forest Research Institute Malaysia, Kuala Lumpur, Malaysia.

Axelsson E, Ratnakumar A, Arendt M-L *et al.* (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, **495**, 360–364.

Baes CF, Dolezal MA, Koltes JE *et al.* (2014) Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics*, **15**, 948.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One*, **3**, e3376.

Banks JA, Nishiyama T, Hasebe M *et al.* (2011) The compact *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.

Barlow J, Gardner TA, Araujo IS *et al.* (2007a) Quantifying the biodiversity value of tropical primary, secondary, and plantation forests. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 18555–18560.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692.

Barr CM, Fishman L (2010) The nuclear component of a cytonuclear hybrid incompatibility in *Mimulus* maps to a cluster of pentatricopeptide repeat genes. *Genetics*, **184**, 455–465.

Baute GJ, Kane NC, Grassa CJ, Lai Z, Rieseberg LH (2015) Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytologist*, **206**, 830–838.

Berry NJ, Phillips OL, Lewis SL *et al.* (2010) The high value of logged tropical forests: lessons from northern Borneo. *Biodiversity and Conservation*, **19**, 985–997.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Boyle EI, Weng S, Gollub J *et al.* (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, **67**, 3258–3273.

Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 3725–3734.

Bradshaw CJA, Sodhi NS, Brook BW (2009) Tropical turmoil: a biodiversity tragedy in progress. *Frontiers in Ecology and the Environment*, **7**, 79–87.

Bradshaw HD, Schemske DW (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, **426**, 176–178.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and Introgression between *Mimulus nasutus* and *Mimulus guttatus*. *Plos One*, **10**, e1004410.

Brown MV, Philip GK, Bunge JA *et al.* (2009) Microbial community structure in the North Pacific Ocean. *The ISME Journal*, **3**, 1374–1386.

Brown SP, Veach AM, Rigdon-Huss AR *et al.* (2015) Scraping the bottom of the barrel: are rare high throughput sequences artifacts? *Fungal Ecology*, **13**, 221–225.

Bruns TD (1995) Thoughts on the processes that maintain local species diversity of ectomycorrhizal fungi. *Plant and Soil*, **170**, 63–73.

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**, 255–265.

Buckley J, Butlin RK, Bridle JR (2011) Evidence for evolutionary change associated with the recent range expansion of the British butterfly, *Aricia agestis*, in response to climate change. *Molecular Ecology*, **21**, 267–280.

Bulgarelli D, Rott M, Schlaeppi K *et al.* (2012) Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature*, **488**, 91–95.

Campbell NR, LaPatra SE, Overturf K (2014) Association Mapping of Disease Resistance Traits in Rainbow Trout Using Restriction Site Associated DNA Sequencing. *G3*, **4**, 2473–2481.

Cannon CH, Peart DR, Leighton M (1998) Tree species diversity in commercially logged Bornean rainforest. *Science*, **281**, 1366–1368.

Carlquist S (1980) *Hawaii: A Natural History.* Pacific Tropical Botanical Garden, Lawai, Hawaii.

Carpenter FL (1976) Plant-pollinator interactions in Hawaii - pollination energetics of *Metrosideros collina* (Myrtaceae). *Ecology*, **57**, 1125–1144.

Carr GD (1978) Chromosome Numbers of Hawaiian flowering plants and the significance of cytology in selected taxa. *American Journal of Botany*, **65**, 236–242.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3*, **1**, 171–182.

Catchen J, Bassham S, Wilson T *et al.* (2013) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

Chase JM, Kraft NJB, Smith KG, Vellend M, Inouye BD (2011) Using null models to

disentangle variation in community dissimilarity from variation in α-diversity. *Ecosphere*, **2**, art24.

Chikhi R, Medvedev P (2013) Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, **30**, 31–37.

Clemmensen KE, Bahr A, Ovaskainen O *et al.* (2013) Roots and associated fungi drive long-term carbon sequestration in boreal forest. *Science*, **339**, 1615–1618.

Colosimo PF, Hosemann KE, Balabhadra S (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science*, **307**, 1928-1933.

Conesa A, Götz S, García-Gómez JM *et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Cordell S, Goldstein G, Mueller-Dombois D, Webb D, Vitousek PM (1998) Physiological and morphological variation in *Metrosideros polymorpha*, a dominant Hawaiian tree species, along an altitudinal gradient: the role of phenotypic plasticity. *Oecologia*, **113**, 188–196.

Cordier T, Robin C, Capdevielle X, Desprez-Loustau M-L, Vacher C (2012) Spatial variability of phyllosphere fungal assemblages: genetic distance predominates over geographic distance in a European beech stand (*Fagus sylvatica*). *Fungal Ecology*, **5**, 509–520.

Corn CA, Hiesey WM (1973) Altitudinal Variation in Hawaiian *Metrosideros*. *American Journal of Botany*, **60**, 991–1002.

Crutsinger GM, Collins MD, Fordyce JA *et al.* (2006) Plant genotypic diversity predicts community structure and governs an ecosystem process. *Science*, **313**, 966–968.

Dalle SP, de Blois S, Caballero J, Johns T (2006) Integrating analyses of local land-use regulations, cultural perceptions and land-use/land cover data for assessing the success of community-based conservation. *Forest Ecology and Management*, **222**, 370–383.

Dalman K, Himmelstrand K, Olson A *et al.* (2013) A Genome-wide association study identifies genomic regions for virulence in the non-model organism *Heterobasidion annosum* s.s. *Plos One*, **8**, e53525.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davey ML, Heegaard E, Halvorsen R, Kauserud H, Ohlson M (2013) Amplicon-pyrosequencing-based detection of compositional shifts in bryophyte-associated fungal communities along an elevation gradient. *Molecular Ecology*, **22**, 368–383.

Davey ML, Heegaard E, Halvorsen R, Ohlson M, Kauserud H (2012) Seasonal trends in the biomass and structure of bryophyte-associated fungal communities explored

by 454 pyrosequencing. *New Phytologist*, **195**, 844–856.

Davison J, Öpik M, Zobel M *et al.* (2012) Communities of arbuscular mycorrhizal fungi detected in forest soil are spatially heterogeneous but do not vary throughout the growing season. *Plos One*, **7**, e41938.

De Kort H, Vandepitte K, Mergeay J, Mijnsbrugge KV, Honnay O (2015) The population genomic signature of environmental selection in the widespread insect-pollinated tree species *Frangula alnus* at different geographical scales. *Heredity*.

de Wit M, Lorrain S, Fankhauser C (2014) Auxin-mediated plant architectural changes in response to shade and high temperature. *Physiologia plantarum*, **151**, 13–24.

DeBoer N, Stacy EA (2013) Divergence within and among 3 varieties of the endemic tree, 'Ohi'a Lehua (*Metrosideros polymorpha*) on the eastern slope of Hawai'i Island. *The Journal of Heredity*, **104**, 449–458.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, **311**, 496–503.

DeWoody J, Trewin H, Taylor G (2015) Genetic and morphological differentiation in *Populus nigra* L.: isolation by colonization or isolation by adaptation? *Molecular Ecology*, **24**, 2641–2655.

Dickie IA (2007) Host preference, niches and fungal diversity. *New Phytologist*, **174**, 230–233.

Dolezel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233–2244.

Drake DR (1992) Seed dispersal of *Metrosideros polymorpha* (Myrtaceae): a pioneer tree of Hawaiian lava flows. *American Journal of Botany*, **79**, 1224–1228.

Dufresne F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, **23**, 40–69.

D'Hont A, Denoeud F, Aury J-M *et al.* (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, **488**, 213–217.

Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J (2015) Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution*, **69**, 2587–2601.

Edelaar P, Alonso D, Lagerveld S, Senar JC, Björklund M (2012) Population differentiation and restricted gene flow in Spanish crossbills: not isolation-by-distance but isolation-by-ecology. *Journal of Evolutionary Biology*, **25**, 417–430.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves

sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Edwards DP, Larsen TH, Docherty TDS *et al.* (2011) Degraded lands worth protecting: the biological importance of Southeast Asia's repeatedly logged forests. *Proceedings of the Royal Society B: Biological Sciences*, **278**, 82–90.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

ElMousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree *Argania spinosa* (L) Skeels endemic to Morocco. *Theoretical and Applied Genetics*, **92**, 832–839.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources,* **10**, 564–567.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes - application to human mitochondrial-DNA restriction data. *Genetics*, **131**, 479–491.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics*, **4**, 649–655.

Finger A, Kettle CJ, Kaiser-bunbury CN *et al.* (2012) Forest fragmentation genetics in a formerly widespread island endemic tree: *Vateriopsis seychellarum* (Dipterocarpaceae). *Molecular Ecology*, **21**, 2369–2382.

Fischer MC, Rellstab C, Tedder A *et al.* (2013) Population genomic footprints of selection and associations with climate in natural populations of Arabidopsis hallerifrom the Alps. *Molecular Ecology*, **22**, 5594–5607.

Fisher B, Edwards DP, Giam XL, Wilcove DS (2011a) The high costs of conserving Southeast Asia's lowland rainforests. *Frontiers in Ecology and the Environment*, **9**, 329–334.

Fisher B, Edwards DP, Larsen TH *et al.* (2011b) Cost-effective conservation: calculating biodiversity and logging trade-offs in Southeast Asia. *Conservation Letters*, **4**, 443–450.

Fitak RR, Mohandesan E, Corander J, Burger PA (2015) The *de novo* genome assembly and annotation of a female domestic dromedary of North African origin. *Molecular Ecology Resources*.

Fitzpatrick BM, Johnson JR, Kump DK *et al.* (2010) Rapid spread of invasive genes into a threatened native species. *Proceedings of the National Academy of Sciences*, **107**, 3606–3610.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Food and Agriculture Organization (FAO) (2010) Global Forest Resources Assessment 2010. *FAO* (http://www.fao.org/forestry/fra/fra2010/en/)

Forcada J, Hoffman JI (2014) Climate change selects for heterozygosity in a declining fur seal population. *Nature*, **511**, 462–465.

Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.

Frankham R, Ballou JD, Briscoe DA (2010) *Introduction to Conservation Genetics*. Cambridge University Press.

Fraser BA, Künstner A, Reznick DN, Dreyer C, Weigel D (2015) Population genomics of natural and experimental populations of guppies (*Poecilia reticulata*). *Molecular Ecology*, **24**, 389–408.

Gavenda RT (1992) Hawaiian Quaternary paleoenvironments: a review of geological, pedological, and botanical evidence. *Pacific Science*, **46**, 295–307.

Geraldes A, Farzaneh N, Grassa CJ *et al.* (2014) Landscape genomics of *Populus trichocarpa*: the role of hybridization, limited gene flow, and natural selection in shaping patterns of population structure. *Evolution*, **68**, 3260–3280.

Ghazoul J (2002) Impact of logging on the richness and diversity of forest butterflies in a tropical dry forest in Thailand. *Biodiversity and Conservation*, **11**, 521–541.

Gibson L, Lee TM, Koh LP *et al.* (2011) Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature*, **478**, 378–381.

Gompert Z, Lucas LK, Nice CC *et al.* (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

González-Martínez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. *New Phytologist*, **170**, 227–238.

Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F-statistics. *The Journal of Heredity*, **86**, 485–486.

Grant PR, Grant BR (2002) Unpredictable evolution in a 30-year study of Darwin's finches. *Science*, **296**, 707–711.

Guo B, DeFaveri J, Sotelo G, Nair A, Merilä J (2015) Population genomic evidence for adaptive differentiation in Baltic Sea three-spined sticklebacks. *BMC Biology*, **13**, 19.

Hadfield JD (2010) MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, **33**, 1–22.

Halley YA, Dowd SE, Decker JE *et al.* (2014) A draft *de novo* genome assembly for the

northern bobwhite (*Colinus virginianus*) reveals evidence for a rapid decline in effective population size beginning in the late Pleistocene. *Plos One*, **9**, e90240.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, **5**, 235–237.

Hancock AM, Brachi B, Faure N *et al.* (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Harbaugh DT, Wagner WL, Percy DM, James HF, Fleischer RC (2009) Genetic structure of the polymorphic *Metrosideros* (Myrtaceae) complex in the Hawaiian Islands using nuclear microsatellite data. *Plos One*, **4**, e4698.

Hardiansyah G, Hardjanto T, Mulyana M (2006) A brief note on TPTJ (Modified Indonesia Selective Cutting System) from experience of PT Sari Bumi Kusuma (PT SBK) timber concessionair. In: *Permanent Sample Plots, More Than Just Forest Data: Proceedings of International Workshop on Promoting Permanent Sample Plots in Asia and the Pacific Region* (eds Priyadi H, Gunarso P, Kanninen M), pp. 23–31. Center for International Forestry Research, Bogor, Indonesia.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. Sinauer Associates Incorporated.

Harwood CE, Nambiar KS (2014) *Sustainable plantation forestry in South-East Asia*.

Hawksworth DL (2012) Global species numbers of fungi: are tropical studies and molecular approaches contributing to a more robust estimate? *Biodiversity and Conservation*, **21**, 2425–2433.

Hawksworth DL, Rossman AY (1997) Where Are All the Undescribed Fungi? *Phytopathology*, **87**, 888–891.

Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hipp AL, Eaton DAR, Cavender-Bares J *et al.* (2014) A framework phylogeny of the American oak clade based on sequenced RAD data. *Plos One*, **9**, e93975.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 395–408.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *Plos Genetics*, **6**, e1000862.

Hoof J, Sack L, Webb DT, Nilsen ET (2008) Contrasting structure and function of pubescent and glabrous varieties of Hawaiian *Metrosideros polymorpha* (Myrtaceae) at high elevation. *Biotropica*, **40**, 113–118.

Horton MW, Hancock AM, Huang YS *et al.* (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics*, **44**, 212–216.

Hufford KM, Mazer SJ (2003) Plant ecotypes: genetic differentiation in the age of ecological restoration. *Trends in Ecology & Evolution*, **18**, 147–155.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.

ITTO (2005) *Status of Tropical Forest Management*. International Tropical Timber Organization, Yokohama, Japan.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Joel G, Aplet G, Vitousek PM (1994) Leaf morphology along environmental gradients in Hawaiian *Metrosideros polymorpha*. **26**, 17–22.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Jumpponen A, Jones KL (2009) Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist*, **184**, 438–448.

Jumpponen A, Jones KL (2010) Seasonally dynamic fungal communities in the *Quercus macrocarpa* phyllosphere differ between urban and nonurban environments. *New Phytologist*, **186**, 496–513.

Kajitani R, Toshimoto K, Noguchi H *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.

Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. *Molecular Ecology Notes*, **5**, 187–189.

Kaneko S, Abe T, Isagi Y (2013) Complete genotyping in conservation genetics, a case study of a critically endangered shrub, *Stachyurus macrocarpus* var. prunifolius (Stachyuraceae) in the Ogasawara Islands, Japan. *Journal of Plant Research*, **126**, 635–642.

Kemler M, Garnas J, Wingfield MJ *et al.* (2013) Ion Torrent PGM as tool for fungal community analysis: a case study of endophytes in *Eucalyptus grandis* reveals high taxonomic diversity. *Plos One*, **8**, e81718.

Kettle CJ (2010) Ecological considerations for using dipterocarps for restoration of lowland rainforest in Southeast Asia. *Biodiversity and Conservation*, **19**, 1137–1151.

Kettle CJ, Ennos RA, Jaffré T, Gardner M, Hollingsworth PM (2008) Cryptic genetic bottlenecks during restoration of an endangered tropical conifer. *Biological Conservation*, **141**, 1953–1961.

Kettle CJ, Ghazoul J, Ashton P *et al.* (2011) Seeing the fruit for the trees in Borneo. *Conservation Letters*, **4**, 184–191.

Keuskamp DH, Pollmann S, Voesenek LACJ, Peeters AJM, Pierik R (2010) Auxin transport through PIN-FORMED 3 (PIN3) controls shade avoidance and fitness during competition. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 22740–22744.

Kim C, Wang X, Lee TH *et al.* (2014) Comparative analysis of *Miscanthus* and *Saccharum* reveals a shared whole-genome duplication but different evolutionary fates. *The Plant Cell*, **26**, 2420–2429.

Kitayama K, Mueller-Dombois D (1995) Vegetation changes along gradients of long-term soil development in the Hawaiian montane rainforest zone. *Vegetatio*, **120**, 1–20.

Kitayama K, Pattison R, Cordell S, Webb D, Mueller-Dombois D (1997) Ecological and genetic implications of foliar polymorphism in *Metrosideros polymorpha* Gaud. (Myrtaceae) in a habitat matrix on Mauna Loa, Hawaii. *Annals of Botany*, **80**, 491–497.

Kobayashi MJ, Takeuchi Y, Kenta T *et al.* (2013) Mass flowering of the tropical tree *Shorea beccarianawas* preceded by expression changes in flowering and drought-responsive genes. *Molecular Ecology*, **22**, 4767–4782.

Koh LP, Wilcove DS (2008) Is oil palm agriculture really destroying tropical biodiversity? *Conservation Letters*, **1**, 60–64.

Koljalg U, Nilsson RH, Abarenkov K *et al.* (2013) Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, **22**, 5271–5277.

Kremer A, Ronce O, Robledo-Arnuncio JJ *et al.* (2012) Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters*, **15**, 378–392.

Kronforst MR, Young LG, Kapan DD *et al.* (2006) Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences*, **103**, 6575–6580.

Kubota S, Iwasaki T, Hanada K *et al.* (2015) A genome scan for genes underlying microgeographic-scale local adaptation in a wild *Arabidopsis* species. *Plos Genetics*, **11**, e1005361.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, **12**, 118–123.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Laurance WF (2007) Have we overstated the tropical biodiversity crisis? *Trends in Ecology & Evolution*, **22**, 65–70.

Lee SL, Wickneswari R, Mahani MC, Zakri AH (2000) Mating system parameters in a tropical tree species, *Shorea leprosula* Miq. (Dipterocarpaceae), from Malaysian lowland dipterocarp forest. *Biotropica*, **32**, 693–702.

Lerner HRL, Meyer M, James HF, Hofreiter M, Fleischer RC (2011) Multilocus resolution of phylogeny and timescale in the extant adaptive radiation of Hawaiian honeycreepers. *Current Biology*, **21**, 1–7.

Lexer C, Lai Z, Rieseberg LH (2003) Candidate gene polymorphisms associated with salt tolerance in wild sunflower hybrids: implications for the origin of Helianthus paradoxus, a diploid hybrid species. *New Phytologist*, **161**, 225–233.

Lexer C, Wüest RO, Mangili S *et al.* (2014) Genomics of the divergence continuum in an African plant biodiversity hotspot, I: drivers of population divergence in *Restio capensis* (Restionaceae). *Molecular Ecology*, **23**, 4373–4386.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li W, Fu L, Niu B, Wu S, Wooley J (2012a) Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, **13**, 656–668.

Li Y-Y, Tsang EPK, Cui M-Y, Chen X-Y (2012b) Too early to call it success: An evaluation of the natural regeneration of the endangered *Metasequoia glyptostroboides*. *Biological Conservation*, **150**, 1–4.

Lindow SE, Brandl MT (2003) Microbiology of the phyllosphere. *Applied and environmental microbiology*, **69**, 1875–1883.

Linhart YB, Grant MC (1996) Evolutionary significance of local genetic differentiation in plants. *Annual review of ecology and systematics*, **27**, 237–277.

Liu S, Lorenzen ED, Fumagalli M *et al.* (2014) Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, **157**, 785–794.

Lomolino M, Riddle B, Brown J (2006) *Biogeography*. Sinauer Associates, Incorporated, Sunderland, MA, USA.

Lu G, Wu F-Q, Wu W *et al.* (2014) Rice LTG1is involved in adaptive growth and fitness under low ambient temperature. *The Plant Journal*, **78**, 468–480.

Mallet B, Martos F, Blambert L, Pailler T, Humeau L (2014) Evidence for isolation-by-habitat among populations of an epiphytic orchid species on a small oceanic island. *Plos One*, **9**, e87469–12.

Manthey JD, Moyle RG (2015) Isolation by environment in white-breasted nuthatches (*Sitta carolinensis*) of the Madrean Archipelago sky islands: a landscape genomics approach. *Molecular Ecology*, **24**, 3628–3638.

Martin RE, Asner GP, Sack L (2007) Genetic variation in leaf pigment, optical and photosynthetic function among diverse phenotypes of *Metrosideros polymorpha* grown in a common garden. *Oecologia*, **151**, 387–400.

Matulich KL, Weihe C, Allison SD *et al.* (2015) Temporal variation overshadows the

response of leaf litter microbial communities to simulated global change. *The ISME journal*, **9**, 2477–2489.

McKown AD, Klápště J, Guy RD *et al.* (2014) Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist*, **203**, 535–553.

Meijaard E, Sheil D (2007a) A logged forest in Borneo is better than none at all. *Nature*, **446**, 974–974.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*, **11**, 5–18.

Meirmans PG, van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Meyer WK, Venkat A, Kermany AR *et al.* (2015) Evolutionary history inferred from the *de novo* assembly of a nonmodel organism, the blue-eyed black lemur. *Molecular Ecology*, **24**, 4392–4405.

Michael TP, Jackson S (2013) The first 50 plant genomes. *The Plant Genome*, **6**, 1–7.

Milano I, Babbucci M, Cariani A *et al.* (2014) Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*). *Molecular Ecology*, **23**, 118–135.

Millar MA, Byrne M, Nuberg IK (2012) High levels of genetic contamination in remnant populations of *Acacia saligna* from a genetically divergent planted stand. *Restoration Ecology*, **20**, 260–267.

Milligan B (1992) Plant DNA isolation. In: *Molecular-Genetic Analysis of Populations - a Practical Approach - Hoelzel,Ar* (ed Hoelzel AR), pp. 59–88. IRL Press, Oxford.

Ming R, Hou S, Feng Y *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, **452**, 991–996.

Ministry of Forestry (1998) Decree no: 625/Kpts-II/1998: The silvicultural system of selective cutting and strip planting (TPTJ) silviculture system in the management of natural production forests.

Ministry of Forestry (2005) Decree no: 226/VI-BPHA/2005: The guideline of Indonesia selective cutting and intensive planting system (Tebang Pilih Tanam Indonesia Intensif / TPTII); Directorate General of BPK.

Ministry of Forestry (2009) Decree no: 11/Menhut-II/2009: Silvicultural system in the natural forest productions.

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, **8**, 4321–4325.

Myburg AA, Grattapaglia D, Tuskan GA *et al.* (2014) The genome of Eucalyptus grandis. *Nature*, **510**, 356–362.

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Nadeau NJ, Whibley A, Jones RT *et al.* (2011) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 343–353.

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nature Genetics*, **12**, 111–122.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583–590.

Newton AC, Fitt BDL, Atkins SD, Walters DR, Daniell TJ (2010) Pathogenesis, parasitism and mutualism in the trophic space of microbe–plant interactions. *Trends in Microbiology* **18**, 365–373.

Ng FSP (1977) Gregarious flowering of dipterocarps in Kepong 1976. *Malaysian Forester*, **40**, 126–137.

Ng KKS, Lee SL, Ueno S (2009a) Impact of selective logging on genetic diversity of two tropical tree species with contrasting breeding systems using direct comparison and simulation methods. *Forest Ecology and Management*, **257**, 107–116.

Ng KKS, Lee SL, Tsumura Y *et al.* (2009b) Expressed sequence tag-simple sequence repeats isolated from *Shorea leprosula* and their transferability to 36 species within the Dipterocarpaceae. *Molecular Ecology Resources*, **9**, 393–398.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276–11.

Nosil P, Egan SP, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: "isolation by adaptation" and multiple roles for divergent selection. *Evolution*, **62**, 316–336.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

O'Brien EK, Krauss SL (2010) Testing the home-site advantage in forest trees on disturbed and undisturbed sites. *Restoration Ecology*, **18**, 359–372.

Oksanen J, Blanchet FG, Kindt R *et al.* (2011) Vegan: community ecology package. R package version 2.0-2. (http://CRAN.R-project.org/package=vegan)

Orsini L, Spanier KI, De Meester L (2012) Genomic signature of natural and anthropogenic stress in wild populations of the waterflea *Daphnia magna*: validation in space, time and experimental evolution. *Molecular Ecology*, **21**, 2160–2175.

Ortiz-Barrientos D (2013) The color genes of speciation in plants. *Genetics*, **194**, 39–42.

Osono T (2006) Role of phyllosphere fungi of forest trees in the development of decomposer fungal communities and decomposition processes of leaf litter. *Canadian journal of microbiology*, **52**, 701–716.

Osono T, Ishii Y, Takeda H, Seramethakun T (2009) Fungal succession and lignin decomposition on *Shorea obtusa* leaves in a tropical seasonal forest in northern

Thailand. *Fungal Diversity*, **36**, 101–119.

Ossowski S, Schneeberger K, Lucas-Lledó JI (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

Öpik M, Tedersoo L, Schnittler M (2011) Species abundance distributions and richness estimations in fungal metagenomics--lessons learned from community ecology. *Molecular Ecology*, **20**, 275–285.

Paquette A, Hawryshyn J, Senikas AV, Potvin C (2009) Enrichment planting in secondary forests: a promising clean development mechanism to increase terrestrial carbon sinks. *Ecology and Society*, **14**, 31.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Parra G, Bradnam K, Ning Z, Keane T, Korf I (2009) Assessing the gene space in draft genomes. *Nucleic Acids Research*, **37**, 289–297.

Pauls SU, Nowak C, Bálint M, Pfenninger M (2012) The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology*, **22**, 925–946.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Peay KG, Kennedy PG, Davies SJ, Tan S, Bruns TD (2010) Potential link between plant and fungal distributions in a dipterocarp rainforest: community and phylogenetic structure of tropical ectomycorrhizal fungi across a plant and soil ecotone. *New Phytologist*, **185**, 529–542.

Percy DM, Garver AM, Wagner WL *et al.* (2008) Progressive island colonization and ancient origin of Hawaiian *Metrosideros* (Myrtaceae). *Proceedings of the Royal Society B: Biological Sciences*, **275**, 1479–1490.

Pinard MA, Barker MG, Tay J (2000) Soil disturbance and post-logging forest recovery on bulldozer paths in Sabah, Malaysia. *Forest Ecology and Management*, **130**, 213–225.

Pinto L, Azevedo JL, Pereira JO, Vieira M, Labate CA (2000) Symptomless infection of banana and maize by endophytic fungi impairs photosynthetic efficiency. *New Phytologist*, **147**, 609–615.

Potts BM, Barbour RC, Hingston AB, Vaillancourt RE (2003) Genetic pollution of native eucalypt gene pools - identifying the risks. *Australian Journal of Botany*, **51**, 1–25.

Praça MM, Carvalho CR, Novaes CRDB (2009) Nuclear DNA content of three Eucalyptusspecies estimated by flow and image cytometry. *Australian Journal of Botany*, **57**, 524–8.

Prado-Martinez J, Sudmant PH, Kidd JM *et al.* (2013) Great ape genetic diversity and population history. *Nature*, **499**, 471–475.

Price JP (2004) Floristic biogeography of the Hawaiian Islands: influences of area, environment and paleogeography. *Journal of Biogeography*, **31**, 487–500.

Price JP, Clague DA (2002) How old is the Hawaiian biota? Geology and phylogeny suggest recent divergence. *Proceedings of the Royal Society B: Biological Sciences*, **269**, 2429–2435.

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

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/)

Rausher MD (2008) Evolutionary transitions in floral color. *International Journal of Plant Sciences*, **169**, 7–21.

Raymond M, Rousset F (1995) GENEPOP (version-1.2) - population-genetics software for exact tests and ecumenicism. The *Journal of Heredity*, **86**, 248–249.

Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N (2010) The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environmental Microbiology*, **12**, 2885–2893.

Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conservation Biology*, **17**, 230–237.

Rensing SA, Lang D, Zimmer AD *et al.* (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. **319**, 64–69.

Rimmer A, Phan H, Mathieson I *et al.* (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, **46**, 912–918.

Rodriguez RJ, White JF, Arnold AE, Redman RS (2009) Fungal endophytes: diversity and functional roles. *New Phytologist*, **182**, 314–330.

Roy J, Mooney HA, Saugier B (2001) *Terrestrial Global Productivity*. Elsevier Science.

Ryan ME, Johnson JR, Fitzpatrick BM (2009) Invasive hybrid tiger salamander genotypes impact native amphibians. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 11166–11171.

Sakai S, Momose K, Yumoto T, Kato M, Inoue T (1999) Beetle pollination of *Shorea parvifolia* (section Mutica, Dipterocarpaceae) in a general flowering period in Sarawak, Malaysia. American Journal of Botany, **86**, 62–69.

Salmela MJ (2014) Rethinking local adaptation: Mind the environment! *Forest Ecology and Management*, **312**, 271–281.

Sasaki S (1980) Storage and germination of dipterocarp seeds. *Malaysian Forester*, **43**, 290–308.

Sebbenn AM, Degen B, Azevedo VCR *et al.* (2008) Modelling the long-term impacts of selective logging on genetic diversity and demographic structure of four tropical tree species in the Amazon forest. *Forest Ecology and Management*, **254**, 335–349.

Sexton JP, Hangartner SB, Hoffmann AA (2013) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, **68**, 1–15.

Shafer ABA, Wolf JBW (2013) Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecology Letters*, **16**, 940–950.

Shahabuddin, Hidayat P, Manuwoto S *et al.* (2009) Diversity and body size of dung beetles attracted to different dung types along a tropical land-use gradient in Sulawesi, Indonesia. *Journal of Tropical Ecology*, **26**, 53–14.

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, **194**, 647–662.

Shimizu KK, Tsuchimatsu T (2014) Evolution of selfing: recurrent patterns in molecular adaptation. *Annual Review of Ecology, Evolution, and Systematics*, **46**, 593–622.

Slavov GT, Nipper R, Robson P *et al.* (2013) Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass *Miscanthus sinensis*. *New Phytologist*, **201**, 1227–1239.

Slotte T, Hazzouri KM, Ågren JA *et al.* (2013) The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics*, **45**, 831–835.

Smit AFA, Hubley R, Green P (2013) RepeatMasker Open-4.0. (http://www.repeatmasker.org)

Sodhi NS, Koh LP, Brook BW, Ng PKL (2004) Southeast Asian biodiversity: an impending disaster. *Trends in Ecology & Evolution*, **19**, 654–660.

Soepadmo E, Wong KM (1995) *Tree Flora of Sabah and Sarawak*. Forest Research Institute Malaysia, Kuala Lumpur, Malaysia.

Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) *Minimus*: a fast, lightweight genome assembler. *BMC bioinformatics*, **8**, 64.

Sovu, Tigabu M, Savadogo P, Odén PC, Xayvongsa L (2010) Enrichment planting in a logged-over tropical mixed deciduous forest of Laos. *Journal of Forestry Research*, **21**, 273–280.

Stacy EA, Johansen JB, Sakishima T, Price DK, Pillon Y (2014) Incipient radiation within the dominant Hawaiian tree *Metrosideros polymorpha*. *Heredity*, **113**, 334–342.

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, **32**, W309–W312.

Steane DA, Potts BM, McLean E *et al.* (2014) Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology*, **23**, 2500–2513.

Stemmermann L (1983) Ecological-studies of Hawaiian *Metrosideros* in a successional context. *Pacific Science*, **37**, 361–373.

Stoeck T, Bass D, Nebel M *et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21–31.

Sutjaritvorakul T, Whalley A, Sihanonth P (2011) Antimicrobial activity from endophytic fungi isolated from plant leaves in Dipterocarpous forest at Viengsa

district Nan province, Thailand. *Journal of Agricultural Technology*, **7**, 115–121.

Taberlet P, Coissac E, Pompanon F *et al.* (2007) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, **35**, e14.

Tajima F (1989) Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tanabe AS, Toju H (2013) Two new computational methods for universal DNA barcoding: a benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. *Plos One*, **8**, e76910.

Tang CQ, Hou X, Gao K *et al.* (2007) Man-made versus natural forests in mid-Yunnan, Southwestern China. *Mountain Research and Development*, **27**, 242–249.

Tedersoo L, Nilsson RH, Abarenkov K *et al.* (2010) 454 pyrosequencing and sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, **188**, 291–301.

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Therkildsen NO, Hemmer-Hansen J, Als TD *et al.* (2013) Microevolution in time and space: SNP analysis of historical DNA reveals dynamic signatures of selection in Atlantic cod. *Molecular Ecology*, **22**, 2424–2440.

Thomas E, Jalonen R, Loo J *et al.* (2014) Genetic considerations in ecosystem restoration using native tree species. *Forest Ecology and Management*, **333**, 66–75.

Toju H, Sato H, Tanabe AS (2014) Diversity and spatial structure of belowground plant–fungal symbiosis in a mixed subtropical forest of ectomycorrhizal and arbuscular mycorrhizal plants. *Plos One*, **9**, e86566.

Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-coverage ITS primers for the DNA-based identification of Ascomycetes and Basidiomycetes in environmental samples. *Plos One*, **7**, e40863.

Toju H, Yamamoto S, Sato H *et al.* (2013) Community composition of root-associated fungi in a *Quercus*-dominated temperate forest: "codominance" of mycorrhizal and root-endophytic fungi. *Ecology and Evolution*, **3**, 1281–1293.

Tompsett PB (1987) Desiccation and storage studies on Dipterocarpus seeds. *Annals of Applied Biology*, **110**, 371–379.

Trucchi E, Gratton P, Whittington JD *et al.* (2014) King penguin demography since the last glaciation inferred from genome-wide data. *Proceedings of the Royal Society B*, **281**, 20140528.

Tsujii Y, Onoda Y, Izuno A, Isagi Y, Kitayama K (2015) A quantitative analysis of phenotypic variations of *Metrosideros polymorpha* within and across populations along environmental gradients on Mauna Loa, Hawaii. *Oecologia*.

Tuskan GA, DiFazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression.

Heredity, **108**, 179–189.

U'Ren JM, Dalling JW, Gallery RE *et al.* (2009) Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples. *Mycological Research*, **113**, 432–449.

U'Ren JM, Lutzoni F, Miadlikowska J, Laetsch AD, Arnold AE (2012) Host and geographic structure of endophytic and endolichenic fungi at a continental scale. *American Journal of Botany*, **99**, 898–914.

U'Ren JM, Riddle JM, Monacell JT *et al.* (2014) Tissue storage and primer selection influence pyrosequencing-based inferences of diversity and community composition of endolichenic and endophytic fungi. *Molecular Ecology Resources*, **14**, 1032–1048.

Ulrich W, Ollik M, Ugland KI (2010) A meta-analysis of species-abundance distributions. *Oikos*, **119**, 1149–1155.

Unterseher M, Peršoh D, Schnittler M (2013) Leaf-inhabiting endophytic fungi of European Beech (*Fagus sylvatica* L.) co-occur in leaf litter but are rare on decaying wood of the same host. *Fungal Diversity*, **60**, 43–54.

Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.

van Staden V, Erasmus BFN, Roux J, Wingfield MJ, van Jaarsveld AS (2004) Modelling the spatial distribution of two important South African plantation forestry pathogens. *Forest Ecology and Management*, **187**, 61–73.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences*, **106**, 9939–9946.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 451–460.

Vitousek PM, Aplet G, Turner D, Lockwood JJ (1992) The Mauna Loa environmental matrix - foliar and soil nutrients. *Oecologia*, **89**, 372–382.

Vu GTH, Schmutzer T, Bull F *et al.* (2015) Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. *The Plant Genome*, **8**, 1–14.

Wagner CE, Keller I, Wittwer S *et al.* (2012) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22**, 787–798.

Webb CO, Slik JWF, Triono T (2010) Biodiversity inventory and informatics in Southeast Asia. *Biodiversity and Conservation*, **19**, 955–972.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population-structure. *Evolution*, **38**, 1358–1370.

Whitham TG, Bailey JK, Schweitzer JA *et al.* (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nature Reviews Genetics*, **7**, 510–523.

Whittaker RJ, Fernandez-Palacios JM (2007) *Island Biogeography: Ecology, Evolution, and Conservation*. OUP Oxford.

Wilcove DS, Giam X, Edwards DP, Fisher B, Koh LP (2013) Navjot's nightmare revisited: logging, agriculture, and biodiversity in Southeast Asia. *Trends in Ecology & Evolution*, **28**, 531–540.

Willerslev E, Cappellini E, Boomsma W, Nielsen R (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science*, **317**, 111–114.

Willerslev E, Hansen AJ, Binladen J, Brand TB (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.

Wolters H, Jürgens G (2009) Survival of the flexible: hormonal growth control and adaptation in plant development. *Nature Reviews Genetics*, **10**, 305–317.

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

Wright SD, Yong CG, Dawson JW, Whittaker DJ, Gardner RC (2000) Riding the Ice Age El Nino? Pacific biogeography and evolution of *Metrosideros* subg. *Metrosideros* (Myrtaceae) inferred from nuclear ribosomal DNA. *Proceedings of the National Academy of Sciences*, **97**, 4118–4123.

Wu H, Guang X, Al-Fageeh MB *et al.* (2014) Camelid genomes reveal evolution and adaptation to desert environments. *Nature communications*, **5**, 1–9.

Xu H, Luo X, Qian J *et al.* (2012) FastUniq: A fast *de novo* duplicates removal tool for paired short reads. *Plos One*, **7**, e52249.

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, **13**, 1–14.

Zhou X, Sun F, Xu S *et al.* (2013) Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. *Nature communications*, **4**, 2708.

Zimmerman NB, Vitousek PM (2012) Fungal endophyte communities reflect environmental structuring across a Hawaiian landscape. *Proceedings of the National Academy of Sciences*, **109**, 13022–13027.

Acknowledgments

First and foremost I would like to express my sincere gratitude to my supervisor Prof. Yuji Isagi for the continuous support of my Ph.D. study and related researches, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Additionally, I appreciate Prof. Kanehiro Kitayama and Prof. Mamoru Kanzaki for giving me the opportunities to conduct researches in Hawaiian Islands and Central Kalimantan and for reviewing my thesis and providing me valuable comments that widen my research from various perspectives. I also thank Dr. Atsushi Takayanagi, Dr. Michimasa Yamasaki for their insightful comments and encouragement on my study.

Dr. Yusuke Onoda, Yuki Tsujii, and Gaku Amada helped me with the field works in the island of Hawaii. Prof. Kentaro K. Shimizu provided me the opportunities to conduct the whole genome sequencing of *M. polymorpha* and the insightful comments on my study. Dr. Masaomi Hatakeyama kindly helped me conduct bioinformatics for the genome sequencing. Dr. Rie Shimizu-Inatsugi gave me advices for preparing the library for genome sequencing. Dr. Tomoaki Nishiyama and Keiko Yamada instructed and assisted me in the experiments for mate-pair sequencing. Dr. Ichiro Tamaki kindly provided me advices on coalescent simulation analyses. Dr. Ryuta Sasaki helped me with the flow cytometry experiments. Dr. Nagano J. Atsushi, Dr. Mie N. Honjo, and Prof. Hiroshi Kudoh conducted the experiments for RAD-seq. Catharine Aquino Fournier and the Functional Genomics Center Zurich supported me with Illumina sequencing.

Dr. Sapto Indrioko, Widiyatno, Eko Prasetyo, and Dr. Hatma Suryatmojo helped me with the field works in Central Kalimantan. Yudhi Hendro, Kasmujiono, and the staffs in PT. Sari Bumi Kusuma allowed me to conduct field works in their concession area. Dr. Ruliyana Susanti and Dr. Tomoya Inada guided helped me receive the research permits in Indonesia. Dr. Akifumi S. Tanabe and Dr. Hirokazu Toju instructed me in conducting bioinformatics and provided me insightful comments on the statistical analysis for metagenomics.

Lastly, I thank all the present and former members of Laboratory of Forest Biology, Kyoto University.

The researches in this thesis were financially supported by JSPS KAKENHI grant (22255002 and 22128008), a Strategic Funds for the Promotion of Science and Technology from the Japan Science and Technology Agency, and grant-in-aid for JSPS fellows (13J02516).