GENE DISPERSAL IN TROPICAL TREES: ECOLOGICAL PROCESSES AND GENETIC CONSEQUENCES

by

Na Wei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Christopher W. Dick, Chair
Professor Deborah E. Goldberg
Associate Professor Inés Ibáñez
Assistant Professor Timothy Y. James

ACKNOWLEDGMENTS

Many people need to be acknowledged here. First, I would like to thank my advisor, Chris Dick, for his support all the way through my doctoral dissertation work. I would also like to thank my dissertation committee of Deborah Goldberg, Inés Ibáñez and Tim James for their intellectual input into my dissertation research. In addition, I must emphasize that most of work in this dissertation would not have been possible without my collaborators Matteo Detto, Marjolein Bruijning, Jordan Bemmels, Andrew Lowe and Michael Gardner.

I am very grateful to the Panamanian field technicians, Yuriza Guerrero, Anayansi Cerezo and Oldemar Valdez, for assisting me in surveying, mapping and collecting plant samples in the tropical rainforests on Barro Colorado Island, Panama. I thank my husband, Yubo Xia, for helping me with the molecular laboratory work.

Lastly, I would like to acknowledge with gratitude, the funding support provided by a CTFS-ForestGEO Grant from the Center for Tropical Forest Science and Smithsonian Tropical Research Institute, an International Research Award and a Graduate Student Research Grant from Rackham Graduate School, a Winifred B. Chase Fellowship from Matthaei Botanical Gardens and Nichols Arboretum, an Angeline B. Whittier Fellowship and an Emma J. Cole Fellowship from the Department of Ecology and Evolutionary Biology, as well as the financial support provided by a Barbour Scholarship from Rackham Graduate School and fellowships from EEB department.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Tropical trees constitute an ecologically important functional group in terrestrial ecosystems because of the essential roles that they play in sustaining biodiversity and carbon storage. The persistence and evolutionary potentials of tropical trees are, however, increasingly threatened by human-induced rapid changes in abiotic and biotic environments. For long-lived forest trees, gene dispersal by seeds and pollen is critical for tracking shifting climatic niches and for maintaining genetic variation needed to adapt to changing environments. Understanding the potential responses of tropical trees to environmental changes depends in part upon quantifying the rates of seed and pollen dispersal. This dissertation aims to quantify the spatial extent and magnitude of seed and pollen dispersal and their respective genetic impacts in a comparative context, by focusing on four Neotropical tree species that have distinct dispersal and pollination syndromes and life-history strategies. By using parentage inference and inverse modeling, I found that long-distance gene dispersal by seeds is common in these vertebrate-dispersed tropical trees, in which models predicted 1–18% of dispersal events exceeding 1 km. This fraction of pollen dispersal >1 km could reach 10–20% in these species. Furthermore, simulations with gene dispersal distances realistically represented suggest that seed and pollen dispersal limitation can lead to genetic diversity loss in tropical tree populations. By examining the respective genetic impacts of seed vs. pollen dispersal, I found that seed dispersal is the primary force driving spatial genetic patterns in these species. It suggests that the functional

loss of seed-dispersing vertebrates, as a result of anthropogenic disturbance in tropical forests, could alter not only tree population spatial structure and ecological dynamics, but also genetic structure and evolutionary dynamics.

CHAPTER I

Introduction

Two paradoxes exist in the study of gene dispersal via seeds and pollen in tree species, and in plants more broadly. One paradox pertains to the conflict between rapid plant migration (typically 100–1000 m/yr) following postglacial warming inferred from fossil records, and localized seed dispersal inferred from direct observations and plant life-history traits (Reid 1899; Davis & Zabinski 1992; Clark 1998; Clark *et al.* 1998). This paradox is referred to as Reid's Paradox; Reid (1899) first described this dilemma in interpreting the postglacial distribution of oaks in Great Britain, presumably requiring the species to move 1000 km within several thousand years. The second paradox–Slatkin's Paradox–concerns the conflict between indirect inferences of high gene dispersal from low genetic differentiation among populations and direct inferences of gene dispersal limitation from local observations (Slatkin 1987; Mallet 2001), particularly potential pollen dispersal limitation mediated by small insect pollinators in plants (Ashley 2010; Jones 2010). These paradoxes raise important questions about gene dispersal in ecological and evolutionary investigations: Can we reconcile the inconsistent inferences of gene dispersal using alternative lines of evidence?; How far can seeds and pollen move in natural populations?; How do populations respond to potential changes in seed and pollen dispersal processes? Answers to these questions are essential for an improved understanding of biodiversity maintenance at gene and species levels especially in the face of rapid environmental changes.

Seed and pollen-mediated gene dispersal influences the ecological and evolutionary dynamics of forest trees (Levin *et al.* 2003; Kremer *et al.* 2012). Selection for seed dispersal is expected when it facilitates escaping from spatially non-random mortality, colonizing new favorable habitats and maintaining species coexistence (Howe & Smallwood 1982; Hubbell 2001; Levin *et al.* 2003; Muller-Landau *et al.* 2003; Jansen *et al.* 2008). Pollen dispersal, on the other hand, plays a critical role in maintaining genetic connectivity at landscape scales (Adams 1992; Ellstrand 1992; Hamrick & Nason 2000) and sustaining population genetic variation for adaptation to changing environments (Hamrick 2004; Aitken *et al.* 2008; Kremer *et al.* 2012). Our empirical knowledge of the ecological processes of seed and pollen dispersal and their relative genetic importance in tree populations draws primarily from temperate species. In low-diversity, often leafless, temperate forests, airborne pollen travels substantially longer distances than do seeds by wind or animals, and thus pollen dispersal is broadly recognized as the principle avenue of gene movement in temperate zone trees (Ouborg *et al.* 1999; Hamrick 2004; Petit *et al.* 2005; but see Bacles *et al.* 2006). However, in species-rich tropical forests, which harbor 25 times as many tree species as temperate forests at the global scale (Fine & Ree 2006), gene dispersal has been studied in disproportionally few taxa, relative to the total species diversity of tropical trees.

Extrapolating what we have known about gene dispersal in temperate trees to tropical species is challenging, due to the latitudinal differences in pollination and seed dispersal syndromes. Wind pollination that contributes to long-distance gene dispersal in temperate forests declines in frequency to tropical forests (Regal 1982; Bawa 1990), whereas the opposite tendency is found in seed dispersal mediated by vertebrate frugivores

(Jordano 2000; Moles *et al.* 2007). Approximately, 98% of tropical trees are animal pollinated (Bawa 1990) and 70–90% are animal dispersed (Howe & Smallwood 1982). Animal dispersers introduce much variability to seed deposition due to stochastic foraging behaviors, which potentially increase the odds of long-distance seed dispersal events. An appreciable amount of seed intake in conjunction with long retention time and high mobility of large-sized terrestrial and avian frugivores contribute to broad spatial extent of seed deposition (Nathan & Muller-Landau 2000; Nathan 2006; Nathan *et al.* 2008). In addition, assemblages of diverse generalist frugivores may collectively remove, carry and deposit seeds to remote areas away from source fruiting trees (Nathan *et al.* 2008). In light of the sparse yet growing evidence of long-distance seed dispersal by vertebrate frugivores in the tropics (e.g. Sezen *et al.* 2005; Hardesty *et al.* 2006; Russo *et al.* 2006), we may anticipate an increased importance of seed-mediated gene dispersal in many animal-dispersed tropical trees. Such line of research is important because it provides insights into latitudinal patterns of gene dispersal by seeds vs. pollen in tree species and provides fundamental information to guide our forecasts of the adaptive responses of tropical trees to climate change (Kremer *et al.* 2012; Corlett & Westcott 2013) and anthropogenic disturbance (e.g. overhunting, Wright 2003; Harrison *et al.* 2013).

My dissertation quantitatively examines the processes of seed and pollen dispersal and their respective genetic impacts in four tropical tree species with distinct seed dispersal and pollination syndromes and life-history strategies, growing in the mature moist forests of Barro Colorado Island (BCI), Panama. Recognizing the challenges in measuring animal-mediated seed and pollen dispersal in natural populations, especially the long-distance events, I use genetic approaches to match dispersed seedlings to their mother and father

trees to unambiguously measure seed and pollen dispersal distance; I use modeling approaches to integrate immigrant seed and pollen flow that cannot be estimated using parentage inferences. The second chapter focuses on genetic marker discovery for these non-model plant species and discusses the general principles of using next-generation sequencing for marker development. Species-specific microsatellite marker selection and validation are briefly described in Appendix A–C. The third chapter examines the genetic consequences of seed and pollen dispersal, with respect to how the two processes determine the distribution of genetic variation within populations. The fourth chapter quantifies the spatial scale and magnitude of seed and pollen dispersal, particularly the long-distance events, and simulates how populations respond to potential disruptions in seed and pollen dispersal processes. Toward the end, I synthesize the findings of this dissertation and discuss the management implications for tropical trees in the fifth chapter.

Chapter II The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio

Microsatellite markers have been employed to estimate important population parameters in ecological and evolutionary investigations (Schlotterer 2004; Selkoe & Toonen 2006; Guichoux *et al.* 2011), such as dispersal distance, individual reproductive success, inbreeding and genetic structure. Because they can provide sufficient inter-individual variation, microsatellites are particularly powerful in measuring seed and pollen dispersal based on parentage assignments, relative to field observations of the idiosyncratic foraging behaviors of dispersers and pollinators. Long-distance dispersal of seeds and pollen have been discovered on the basis of microsatellite markers (reviewed in Ashley

2010), suggesting the potential of this approach in resolving the paradoxes in gene dispersal inference.

Traditionally, microsatellite development was labor and cost inefficient (Zane *et al.* 2002; Selkoe & Toonen 2006). Next-generation sequencing has revolutionized microsatellite detection in non-model organisms, with substantial reductions in time and capital investment (Guichoux *et al.* 2011; Zalapa *et al.* 2012). Different NGS platforms have been used to generate genomic sequences, from which microsatellite makers can be isolated; however, these platforms differ in the trade-offs between sequencing capacity, read length and sequencing error rate, which may result in their different efficacy for microsatellite maker development. No study has quantitatively examined the effects of read length, read quality and read quantity of NGS on microsatellite development. In this chapter, I performed simulations to assess (1) whether read length is positively correlated with primer design success, microsatellite throughput and the effectiveness of genomic redundancy detection, (2) whether and how read quality affects microsatellite amplification, and (3) whether sequence quality control is necessary for microsatellite development. Based on the findings from these simulations, I compared the performance of different NGS platforms for microsatellite isolation and highlighted some key considerations for projects involving NGS-based microsatellite marker development.

APPENDIX A Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing

One of the study species in my dissertation is insect-pollinated and vertebrate-dispersed Neotropical nutmeg *Virola sebifera* (Myristicaceae). *Virola sebifera* is a shade-

tolerant canopy tree, distributed broadly in mature tropical forests from Central America to the Amazon Basin and Guiana Shield (Croat 1978). The brownish, small-sized flowers are pollinated by various insect pollinators, such as small bees, beetles and wasps (Bawa & Opler 1975; Bawa *et al.* 1985). Nutrient-rich fruits of *V. sebifera* are consumed and dispersed primarily by large-sized birds such as toucans on BCI (Howe 1981). In this appendix, I developed a set of polymorphic microsatellite markers for *V. sebifera* based on genomic sequences obtained from French Guiana samples using shotgun 454 pyrosequencing. These novel markers will be used as part of the dissertation goal of understanding the ecological processes and genetic impacts of gene dispersal in tropical forests, which involves tree species of distinct pollination and dispersal syndromes.

APPENDIX B Characterization of twenty-six microsatellite markers for the tropical pioneer tree species *Cecropia insignis* Liebm. (Urticaceae)

Cecropia insignis (Urticaceae) is a wide-ranging pioneer canopy tree found in lowland moist forests of Central and northern South America (Croat 1978). It attains a height of 40 m and a dbh of 70 cm. *Cecropia insignis* is among the few plant taxa that are wind pollinated in tropical rain forests (Bawa & Opler 1975; Croat 1978). Flowering occurs during the dry season between January and April on BCI. The fruits of *C. insignis* are dispersed by a diverse assemblage of frugivores, including birds, bats and mammals (Brokaw 1986). As a gap specialist, long-distance seed dispersal is expected for *C. insignis* as a means to establish in ephemeral critical environments (Brokaw 1986). Twenty-six microsatellite markers were characterized for this species, of which eleven loci of high polymorphism will be used to study gene dispersal in *C. insignis*.

APPENDIX C Polymorphic microsatellite markers for a wind-dispersed tropical tree species, *Triplaris cumingiana* (Polygonaceae)

Triplaris cumingiana is an insect-pollinated and wind-dispersed midstory tree species (Croat 1978), 10–20 m tall and 12–30 cm in dbh at maturity. Compared to the other study species, *T. cumingiana* is less abundant and more spatially aggregated on BCI. The distribution of this species is associated with high soil phosphorous (Condit *et al.* 2013). Unlike most dioecious tree species that have inconspicuous unisexual flowers (Bawa & Opler 1975), flower sexual dimorphism is pronounced in *T. cumingiana*, which produces bright red bracts signaling flowers on female trees during the dry season of Panama. The large calyx of female flowers in *T. cumingiana* facilitates seed dispersal by wind (Croat 1978). Twelve microsatellite makers were screened and validated, of which nine markers conforming to Hardy-Weinberg expectations will be used.

Chapter III Seed dispersal drives spatial genetic patterns in tropical trees

Seed dispersal is broadly recognized for its ecological importance (Webb & Peart 2001; Levin *et al.* 2003), but its population genetic impacts relative to pollen dispersal as a factor governing spatial genetic patterns are poorly understood in many tropical tree species. Such information is important for our understanding of the potential genetic consequences of increasingly intensive anthropogenic disturbance (e.g. overhunting) in tropical rain forests. The distribution of genetic variation within populations influences short-term evolutionary dynamics of forest trees, such as the level of assortative mating and inbreeding (Epperson 1992). Seed and pollen dispersal collectively determine spatial

genetic structure (SGS). In principle, strong SGS is associated with seed and pollen dispersal limitation.

Several other factors can also influence the strength of SGS in nature populations, such as variation in individual reproductive success, non-random spatial distribution and mortality (Hamrick *et al.* 1993; Doligez *et al.* 1998; Degen *et al.* 2001; Sagnard *et al.* 2011). Specifically, large reproductive variation among individuals intensifies SGS due to increased genetic relatedness resulting from disproportionate reproductive contributions of a few individuals. Spatial aggregation could lead to the overlapping of seed shadows, which lessens SGS. Non-random mortality, if acting in a density-dependent manner, could preferentially target aggregated individuals with higher than average genetic affinity, and thereby reduces SGS intensity. Because different ecological processes and demographic characteristics can be involved, quantifying the extent to which seed and pollen dispersal affect SGS independent from other confounding factors is challenging. It is thus difficult to attribute differences in SGS intensity between taxa to their differences in seed and pollen dispersal distance.

Even if the effects of other factors could be teased apart, separating the respective contributions of seed vs. pollen dispersal to SGS remains difficult. In the cases where theoretical population genetic models, such as island model (Wright 1965) or isolation by distance (Slatkin 1991; Rousset 1997, 2000) at evolutionary equilibrium, are assumed to hold, the relative magnitude of seed vs. pollen dispersal can be retrieved and distinguished from resulting SGS patterns, based on the combined use of biparentally and uniparentally inherited genetic markers (Ouborg *et al.* 1999; Oddou-Muratorio *et al.* 2001). However, whether natural populations conform to theoretical models is highly contentious. In

addition, uniparental markers often do not provide sufficient resolution at fine spatial scales in angiosperms.

To overcome these empirical and theoretical constraints, I developed a novel analytic framework to quantify the respective roles of seed and pollen dispersal in governing SGS in the four study species. This approach has several advantages: (1) it does not depend on theoretical population models or uniparental markers; (2) it takes into account confounding factors, such as spatial structure, population density and reproductive variance; (3) it separates the contributions of seed and pollen dispersal to SGS. Overall, the results suggest that seed dispersal is the primary mechanism driving SGS in these tropical trees, at the least during early life-history stages.

Chapter IV Frequent long-distance seed and pollen dispersal and their genetic impacts in tropical trees

How far seeds and pollen can move affects the responses of forest trees to changing environments (Kremer *et al.* 2012; Corlett & Westcott 2013). Despite their fundamental importance in tree populations, measuring seed and pollen dispersal in nature is notoriously difficult, especially for tropical trees that are primarily animal dispersed and pollinated. This may in part explain our limited empirical knowledge of seed and pollen dispersal in tropical tree species.

Another bias in the studies of gene dispersal by seeds and pollen in tropical trees comes from the disproportionate empirical efforts on pollen dispersal relative to seed dispersal. Despite the unifying nature of seed and pollen dispersal from the plant perspective, most of the previous gene dispersal research in tropical tree species has

focused on pollen movement (e.g. Hamrick & Murawski 1990; Stacy *et al.* 1996; Dawson *et al.* 1997; Nason *et al.* 1998; White *et al.* 2002; Degen *et al.* 2004; Hanson *et al.* 2007; Hufford *et al.* 2009; Tani *et al.* 2009; Collevatti *et al.* 2010; Manoel *et al.* 2012). As a result, there are very few species in which both seed and pollen dispersal have been quantified (Hardesty *et al.* 2006; Ashley 2010).

In this chapter, I quantified the spatial scale and magnitude of seed and pollen dispersal through unambiguous maternal and paternal inferences of established seedlings in the four study species. Then I combined parentage inferences and inverse modeling (Jones & Muller-Landau 2008) to estimate the probability distributions of seed and pollen dispersal (i.e. dispersal kernels) and to assess whether they are light or heavy tailed, the latter suggesting the potential of long-distance gene dispersal. Lastly, I used dynamic simulations to quantify the impacts of short vs. long-distance seed and pollen dispersal.

References

Adams WT (1992) Gene dispersal within forest tree populations. *Population Genetics of Forest Trees* **42**, 217-240.

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.

Ashley MV (2010) Plant parentage, pollination, and dispersal: how DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences* **29**, 148-161.

Bacles CFE, Lowe AJ, Ennos RA (2006) Effective seed dispersal across a fragmented landscape. *Science* **311**, 628-628.

Bawa KS (1990) Plant-pollinator interactions in tropical rain forests. *Annual Review of Ecology and Systematics* **21**, 399-422.

Bawa KS, Bullock SH, Perry DR, Coville RE, Grayum MH (1985) Reproductive biology of tropical lowland rain forest trees. II. Pollination systems. *American Journal of Botany* **72**, 346-356.

Bawa KS, Opler PA (1975) Dioecism in tropical forest trees. *Evolution* **29**, 167-179.

Brokaw NL (1986) Seed dispersal, gap colonization, and the case of *Cecropia insignis*. In: *Frugivores and seed dispersal* (eds. Estrada A, Fleming T), pp. 323-331. Kluwer Academic Publishers, Dordrecht, Netherlands.

Clark JS (1998) Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *American Naturalist* **152**, 204-224.

Clark JS, Fastie C, Hurtt G*, et al.* (1998) Reid's paradox of rapid plant migration. *BioScience* **48**, 13-24.

Collevatti RG, Estolano R, Garcia SF, Hay JD (2010) Short-distance pollen dispersal and high self-pollination in a bat-pollinated neotropical tree. *Tree Genetics & Genomes* **6**, 555-564.

Condit R, Engelbrecht BMJ, Pino D, Perez R, Turner BL (2013) Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5064-5068.

Corlett RT, Westcott DA (2013) Will plant movements keep up with climate change? *Trends in Ecology & Evolution* **28**, 482-488.

Croat TB (1978) *Flora of Barro Colorado Island* Stanford University Press, Stanford, California, USA.

Davis MB, Zabinski C (1992) Changes in geographical range resulting from greenhouse warming: effects on biodiversity in forests. In: *Global warming and biological diversity* (eds. Peters RL, Lovejoy TE), pp. 297-308. Yale University Press, New Haven, Connecticut, USA.

Dawson IK, Waugh R, Simons AJ, Powell W (1997) Simple sequence repeats provide a direct estimate of pollen-mediated gene dispersal in the tropical tree *Gliricidia sepium*. *Molecular Ecology* **6**, 179-183.

Degen B, Bandou E, Caron H (2004) Limited pollen dispersal and biparental inbreeding in *Symphonia globulifera* in French Guiana. *Heredity* **93**, 585-591.

Degen B, Caron H, Bandou E*, et al.* (2001) Fine-scale spatial genetic structure of eight tropical tree species as analysed by RAPDs. *Heredity* **87**, 497-507.
Doligez A, Baril C, Joly HI (1998) Fine-scale spatial genetic structure with nonuniform distribution of individuals. *Genetics* **148**, 905-919.
Ellstrand NC (1992) Gene flow by pollen: implications for plant conservation genetics. *Oikos* **63**, 77-86.
Epperson BK (1992) Spatial structure of genetic variation within populations of forest trees. *New Forests* **6**, 257-278.
Fine PVA, Ree RH (2006) Evidence for a time-integrated species-area effect on the latitudinal gradient in tree diversity. *American Naturalist* **168**, 796-804.
Guichoux E, Lagache L, Wagner S*, et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**, 591-611.
Hamrick JL (2004) Response of forest trees to global environmental changes. *Forest Ecology and Management* **197**, 323-335.
Hamrick JL, Murawski D, Nason J (1993) The influence of seed dispersal mechanisms on the genetic structure of tropical tree populations. In: *Frugivory and seed dispersal: ecological and evolutionary aspects* (eds. Fleming TH, Estrada A), pp. 281-297. Kluwer Academic Publishers, Dordrecht, Netherlands.
Hamrick JL, Murawski DA (1990) The breeding structure of tropical tree populations. *Plant Species Biology* **5**, 157-165.
Hamrick JL, Nason JD (2000) Gene flow in forest trees. In: *Forest conservation genetics : principles and practice* (eds. Young AG, Boshier D, Boyle TJB), pp. 81-90. CSIRO Publishing, Collingwood, Australia.
Hanson T, Brunsfeld S, Finegan B, Waits L (2007) Conventional and genetic measures of seed dispersal for *Dipteryx panamensis* (Fabaceae) in continuous and fragmented Costa Rican rain forest. *Journal of Tropical Ecology* **23**, 635-642.
Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters* **9**, 516-525.
Harrison RD, Tan S, Plotkin JB*, et al.* (2013) Consequences of defaunation for a tropical tree community. *Ecology Letters* **16**, 687-694.
Howe HF (1981) Dispersal of a Neotropical nutmeg (*Virola sebifera*) by birds. *The Auk* **98**, 88-98.
Howe HF, Smallwood J (1982) Ecology of seed dispersal. *Annual Review of Ecology and Systematics* **13**, 201-228.
Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography* Princeton University Press, Princeton, NJ.
Hufford KM, Hamrick JL, Rathbun SL (2009) Male reproductive success at three early life stages in the tropical tree *Platypodium elegans*. *International Journal of Plant Sciences* **170**, 724-734.
Jansen PA, Bongers F, van der Meer PJ (2008) Is farther seed dispersal better? Spatial patterns of offspring mortality in three rainforest tree species with different dispersal abilities. *Ecography* **31**, 43-52.
Jones A (2010) Reconciling field observations of dispersal with estimates of gene flow. *Molecular Ecology* **19**, 4379-4382.

Jones FA, Muller-Landau HC (2008) Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. *Journal of Ecology* **96**, 642-652.
Jordano P (2000) Fruits and frugivory. In: *Seeds: The Ecology of Regeneration in Natural Plant Communities* (ed. Fenner M), pp. 125-166. CABI Publ., Oxon, UK.
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.
Levin SA, Muller-Landau HC, Nathan R, Chave J (2003) The ecology and evolution of seed dispersal: A theoretical perspective. *Annual Review of Ecology Evolution and Systematics* **34**, 575-604.
Mallet JB (2001) Gene flow. In: *Insect movement: mechanisms and consequence* (eds. Woiwod IP, Reynolds DR, Thomas CD), pp. 337-360. CABI Publishing, CAB International, Oxon, UK.
Manoel RO, Alves PF, Dourado CL*, et al.* (2012) Contemporary pollen flow, mating patterns and effective population size inferred from paternity analysis in a small fragmented population of the Neotropical tree *Copaifera langsdorffii* Desf. (Leguminosae-Caesalpinioideae). *Conservation Genetics* **13**, 613-623.
Moles AT, Ackerly DD, Tweddle JC*, et al.* (2007) Global patterns in seed size. *Global Ecology and Biogeography* **16**, 109-116.
Muller-Landau HC, Levin SA, Keymer JE (2003) Theoretical perspectives on evolution of long-distance dispersal and the example of specialized pests. *Ecology* **84**, 1957-1967.
Nason JD, Herre EA, Hamrick JL (1998) The breeding structure of a tropical keystone plant resource. *Nature* **391**, 685-687.
Nathan R (2006) Long-distance dispersal of plants. *Science* **313**, 786-788.
Nathan R, Muller-Landau HC (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* **15**, 278-285.
Nathan R, Schurr FM, Spiegel O*, et al.* (2008) Mechanisms of long-distance seed dispersal. *Trends in Ecology & Evolution* **23**, 638-647.
Oddou-Muratorio S, Petit RJ, Le Guerroue B, Guesnet D, Demesure B (2001) Pollen- versus seed-mediated gene flow in a scattered forest tree species. *Evolution* **55**, 1123-1135.
Ouborg NJ, Piquot Y, Van Groenendael JM (1999) Population genetics, molecular markers and the study of dispersal in plants. *Journal of Ecology* **87**, 551-568.
Petit RJ, Duminil J, Fineschi S*, et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* **14**, 689-701.
Regal PJ (1982) Pollination by wind and animals: ecology of geographic patterns. *Annual Review of Ecology and Systematics* **13**, 497-524.
Reid C (1899) *The origin of the British flora* Dulau & Company, London, UK.
Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* **145**, 1219-1228.
Rousset F (2000) Genetic differentiation between individuals. *Journal of evolutionary biology* **13**, 58-62.

Russo SE, Portnoy S, Augspurger CK (2006) Incorporating animal behavior into seed dispersal models: implications for seed shadows. *Ecology* **87**, 3160-3174.
Sagnard F, Oddou-Muratorio S, Pichot C, Vendramin GG, Fady B (2011) Effects of seed dispersal, adult tree and seedling density on the spatial genetic structure of regeneration at fine temporal and spatial scales. *Tree Genetics & Genomes* **7**, 37-48.
Schlotterer C (2004) The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics* **5**, 63-69.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.
Sezen UU, Chazdon RL, Holsinger KE (2005) Genetic consequences of tropical second-growth forest regeneration. *Science* **307**, 891-891.
Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* **236**, 787-792.
Slatkin M (1991) F_{ST} in a hierarchical island model. *Genetics* **127**, 627.
Stacy EA, Hamrick JL, Nason JD*, et al.* (1996) Pollen dispersal in low-density populations of three Neotropical tree species. *American Naturalist* **148**, 275-298.
Tani N, Tsumura Y, Kado T*, et al.* (2009) Paternity analysis-based inference of pollen dispersal patterns, male fecundity variation, and influence of flowering tree density and general flowering magnitude in two dipterocarp species. *Annals of Botany* **104**, 1421-1434.
Webb CO, Peart DR (2001) High seed dispersal rates in faunally intact tropical rain forest: theoretical and conservation implications. *Ecology Letters* **4**, 491-499.
White GM, Boshier DH, Powell W (2002) Increased pollen flow counteracts fragmentation in a tropical dry forest: An example from *Swietenia humilis* Zuccarini. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 2038-2042.
Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**, 395-420.
Wright SJ (2003) The myriad consequences of hunting for vertebrates and plants in tropical forests. *Perspectives in Plant Ecology Evolution and Systematics* **6**, 73-86.
Zalapa JE, Cuevas H, Zhu HY*, et al.* (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* **99**, 193-208.
Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology* **11**, 1-16.

CHAPTER II

The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio

Abstract

The advent of next-generation sequencing (NGS) technologies has transformed the way microsatellites are isolated for ecological and evolutionary investigations. Recent attempts to employ NGS for microsatellite discovery have used the 454, Illumina, and Ion Torrent platforms, but other methods including single-molecule real-time DNA sequencing (Pacific Biosciences, or PacBio) remain viable alternatives. We outline a workflow from sequence quality control to microsatellite marker validation in three plant species using PacBio circular consensus sequencing (CCS). We then evaluate the performance of PacBio CCS in comparison to other NGS platforms for microsatellite isolation, through simulations that focus on variations in read length, read quantity, and sequencing error rate. Although quality control of CCS reads reduced microsatellite yield by around 50%, hundreds of microsatellite loci that are expected to have improved conversion efficiency to functional markers were retrieved for each species. The simulations quantitatively validate the advantages of long reads, and emphasize the detrimental effects of sequencing errors on NGS-enabled microsatellite development. In view of the continuing improvement in read length on NGS platforms, sequence quality and the corresponding strategies of quality control will become the primary factors to consider for effective microsatellite isolation.

Among current options, PacBio CCS may be optimal for rapid, small-scale microsatellite development due to its flexibility in scaling sequencing effort, while platforms such as Illumina MiSeq will provide cost-efficient solutions for multi-species microsatellite projects.

Keywords: microsatellites, circular consensus sequencing, quality control, read length simulation, sequencing error simulation, error trimming simulation

Introduction

Microsatellites, also referred to as simple sequence repeats (SSRs) or short tandem repeats (STRs), are repetitive short DNA sequences that are scattered throughout the genomes of prokaryotes and eukaryotes (Morgante *et al.* 2002; Ellegren 2004). These molecular markers have seen extensive use in ecology and evolutionary biology (Provan *et al.* 2001; Schlotterer 2004; Selkoe & Toonen 2006). The dominance of microsatellites as the marker of choice for many applications in molecular ecology is, nevertheless, facing new challenges from large genomic datasets generated by next-generation sequencing (NGS) technologies (Ouborg *et al.* 2010). Yet, due to their hypervariability (Schlotterer 2000; Ellegren 2004), microsatellites remain invaluable for investigations of fine-scaled spatial demographic and genetic processes where individuals of interest are closely related, such as dispersal, parentage inference, pedigree reconstruction, linkage mapping, and population structure (Selkoe & Toonen 2006; Guichoux *et al.* 2011; Haasl & Payseur 2011).

Interestingly, the advent of NGS may bolster microsatellite use because acquiring adequate genomic sequences from which microsatellites are retrieved is no longer technically and monetarily difficult. Instead the bottleneck in microsatellite development is now the laborious and costly process of marker validation. Many researchers have advocated NGS-based microsatellite detection in non-model organisms (e.g. Abdelkrim *et al.* 2009; Gardner *et al.* 2011; Jennings *et al.* 2011), with the 454 and Illumina platforms dominating such efforts (reviewed in Zalapa *et al.* 2012). More recently, microsatellite detection has employed other NGS platforms, including Ion Torrent PGM (e.g. Huey *et al.* 2013; Elliott *et al.* 2014), Illumina MiSeq (e.g. McCracken *et al.* 2014; Nowak *et al.* 2014), and Pacific Biosciences RS (e.g. this study; Grohme *et al.* 2013). All these platforms can

deliver hundreds to thousands of microsatellite loci per species, many more than identified using traditional methods (Zane *et al.* 2002), and with substantial reductions in time and capital investment.

The popularity of the 454 platform for microsatellite isolation owes primarily to its long read-length sequencing (Zalapa *et al.* 2012). Long read length is advantageous in that it could benefit primer design by providing sufficient flanking regions (Guichoux *et al.* 2011; Zalapa *et al.* 2012). In addition, longer reads are suggested to allow better detection of genomic redundant sequences that contain low-complexity regions unfavorable for microsatellite amplification and interpretation (Elliott *et al.* 2014). However, the 454 platform is economically inefficient (i.e. high cost per megabases; Glenn 2013), and involves laborious titration steps required in emulsion PCR to precisely link one DNA template to a single bead (Margulies *et al.* 2005). These aspects of the 454 platform eventually translate into a high total cost for microsatellite isolation. In terms of cost reduction, the most dramatic drop has been seen using the Illumina platform due to its high sequence throughput (e.g. Jennings *et al.* 2011; Castoe *et al.* 2012). Although the Illumina platform has much higher sequencing capacity relative to other platforms (Glenn 2013), it produces short reads (single-end up to 150 bp, paired-end up to 300 bp in GAIIx and HiSeq), except for the Illumina MiSeq sequencer, which can generate paired-end reads up to 600 bp (Illumina Incorporation 2013). The Ion Torrent platform represents an intermediate solution regarding the trade-off between (single-end) read length and read quantity (Glenn 2013), as well as the sequencing cost for microsatellite development (Jennings *et al.* 2011; Castoe *et al.* 2012; Elliott *et al.* 2014).

Compared with the above NGS platforms, single-molecule real-time sequencing (SMRT; Eid *et al.* 2009) implemented on the Pacific Biosciences (PacBio) RS system has the longest sequencing capability (Glenn 2013; Pacific Biosciences 2013), which offers potential advantages for microsatellite detection. The PacBio platform differs fundamentally from other platforms in that sequencing is performed on individual molecules without involving DNA amplification (e.g. emulsion PCR on 454 and Ion Torrent; Bridge PCR on Illumina; Glenn 2011), thereby resulting in a more uniform representation of genomic regions (Pacific Biosciences 2013). Although the long read-length sequencing of PacBio comes with a high single-pass error rate (~11%; Pacific Biosciences 2013), improved base-calling accuracy is achieved by circular consensus sequencing (CCS); that is, reading through the same circular template DNA fragment multiple times (Travers *et al.* 2010). In addition, the insensitivity to various types of sequence context biases, such as homopolymers, GC-biased DNA regions, and highly repetitive sequences (Eid *et al.* 2009; Quail *et al.* 2012; Zhang *et al.* 2012), makes PacBio a compelling alternative sequencing platform in this context. Success in microsatellite marker development using CCS has recently been reported on this platform (Grohme *et al.* 2013; Wainwright *et al.* 2013). However, an in-depth evaluation is not yet available regarding sequence characteristics of CCS and corresponding strategies of quality control for microsatellite development; therefore, providing this evaluation is the first objective of this study.

Independently from specific platforms and organisms, the development of microsatellite markers is in general influenced by read length, read quantity, and read quality. Although the commonly agreed-upon benefits of long reads are conceptually

straightforward, robust quantitative evidence for this consensus has been lacking. In addition, sequencing errors can undermine the efficiency of converting *in silico* loci into working markers, because unambiguous and unique sequences are crucial to the construction of amplifiable primers. However, most NGS-based microsatellite development work has been carried out in the absence of the inspection and control of sequence quality (for a counterexample, see Fernandez-Silva *et al.* 2013). It remains unclear the extent to which read quality inflicts a measurable effect on microsatellite marker development. Therefore, the second objective of this work is to provide a quantitative investigation of microsatellite development effectiveness in relation to read length, read quality, and sequence quality control.

For the purpose of assessing the applicability of PacBio CCS in microsatellite isolation, we outline the process of (i) performing quality control (QC) on CCS reads, (ii) identifying microsatellite loci from post-QC CCS reads, and (iii) validating microsatellite markers for three plant species for which no prior genomic information was available. For the second objective of quantifying how sequence characteristics limit microsatellite development, (iv) we conduct read length simulations to test whether increases in sequence length are associated with improvements in primer design success, microsatellite throughput, and genomic redundancy detection; (v) we use sequencing error simulations to examine whether and how read quality affects microsatellite amplification; and (vi) we perform error trimming simulations to validate the need for sequence quality control in microsatellite development. Then, we use the findings from these simulations to guide the performance evaluation of PacBio CCS in comparison with other NGS platforms, and highlight some key considerations for NGS use in microsatellite isolation.

Materials and Methods

DNA sources and PacBio library preparation

We collected leaf tissues of three non-model tropical tree species from the 50-ha Forest Dynamics Plot (FDP) on Barro Colorado Island, Panama: *Alchornea costaricensis* Pax & K. Hoffm. (Euphorbiaceae), *Cecropia insignis* Liebm. (Cecropiaceae), and *Triplaris cumingiana* Fisch. & C.A. Mey. ex C.A. Mey. (Polygonaceae). Genomic DNA was isolated from freeze-dried leaves using DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA). DNA quality was checked using NanoDrop 2000 (Thermo Scientific, Wilmington, Delaware, USA), and dsDNA concentration was measured using Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, California, USA). Double-stranded DNA of at least 30 ng/μL in a 50-μL volume from one tree of each species was sent to the DNA Sequencing Core Laboratory at the University of Michigan for PacBio 500-bp DNA library preparation and circular consensus sequencing (CCS).

In brief, genomic DNA was first sheared to fragments averaging 500 bp in length, and quantified with 2200 TapeStation using DNA 1k Tape (Agilent, Santa Clara, California, USA). Sheared dsDNA was end repaired and ligated with hairpin adapters that contain a sequencing-primer binding site to form the SMRTbell™ structure (i.e. two 55-nt single-stranded hairpin loops plus a dsDNA fragment). Unsuccessful ligation products were removed afterwards by exonuclease (ExoIII and ExoVII). Post-ligation products were quantified a second time with 2200 TapeStation, showing a mean fragment size of 363 bp for *A. costaricensis*, 487 bp for *C. insignis*, and 445 bp for *T. cumingiana*. Then, the SMRTbell™ templates were annealed with sequencing primers and bound to biotinylated

phi29 DNA polymerase mounted at the base of individual reaction chambers in SMRT cells. Nucleotide incorporation in a SMRT cell was monitored using 2 × 45-minute collection mode. Four SMRT cells were run for each species on a PacBio RS sequencer using C2 chemistry. Fragments inserted between adapters of ≥3 × sequencing depths (including the sense and antisense strand) were retained for generating highly accurate adapter-free consensus sequences from CCS (referred to as CCS reads).

Quality control of CCS reads

Species-specific ccs.fastq files from four SMRT cells were combined to fetch CCS reads and the corresponding Phred +33 quality scores. The mean quality score of a CCS read was typically higher than 30 (median = 64, *A. costaricensis*; 62, *C. insignis*; 60, *T. cumingiana*; solid lines, Fig. 2S.1), suggesting that trimming sequences based upon average read quality would be ineffective (also see simulation results below). Thus, we removed terminal low-quality portions of each CCS read using a sliding-window approach implemented in mothur v1.29.2 (Schloss *et al.* 2009). The window size was set to 10 bases, moving one base per step. The minimum window-wide mean quality score was set to 30, equivalent to an error-tolerance rate of 0.1%. If a window below this threshold was encountered, the CCS read was truncated from the last base in the window until the end of the read. We also filtered sequences according to homopolymer length. Some CCS reads contained homopolymers of 30 to 40 bases long, but more than 75% of CCS reads had homopolymers of ≤8 bases (Fig. 2S.2). To retain adequate sequence numbers and eliminate long homopolymers, we omitted from further analyses CCS reads bearing a homopolymer

longer than 8 bases. Comparisons of pre-QC and post-QC base quality were visualized using the qrqc package (Buffalo 2012) in R v2.15.0 (R Development Core Team 2012).

Microsatellite identification and primer design

CCS reads that passed the preceding quality control (referred to as trimmed CCS reads) were used to retrieve microsatellite loci. Perl pipelines in QDD v2.1 (Meglécz *et al.* 2010) were employed to automate the process of detecting microsatellites and designing primers. An initial purging step removed reads either too short (<80 bp) for successful primer design or holding microsatellite motifs of less than five repeats. The resulting microsatellite-containing sequences were screened for genomic redundancy (i.e. low-complexity regions and interspersed repeats) and sequencing redundancy (i.e. multiple copies of the same sequence) based on sequence similarities using BLAST v2.2.25 (Altschul *et al.* 1990) all-against-all pairwise alignments, in which microsatellites were soft masked. Once significant BLAST hits were discerned, the sequences with flanking region similarity less than 95%, likely resulting from genomic redundancy, were eliminated; those of ≥95% flanking region similarity were re-aligned by ClustalW2 (Larkin *et al.* 2007) to generate consensus sequences. The resulting non-redundant microsatellite-containing reads (i.e. singletons and unique consensus sequences) were used to locate appropriate priming regions using Primer3 (Rozen & Skaletsky 2000). Stringent primer-designing criteria (A+B design; definition *sensu* QDD program) were utilized as follows: i) the absence of tandem repeats in priming regions and no homopolymers more than 3 bases; ii) no multiple microsatellites in the target region; iii) primer size between 18 and 22 bases; iv) PCR product of 100–500 bp; v) optimal GC content of 50% (range 40–60%); vi) 57–63 °C

melting temperature with a maximum intra-pair difference of 5 °C; vii) maximum self-complementarity score of 3; viii) presence of one GC clamp.

Microsatellite marker validation

We prioritized the test array of microsatellite markers as follows: for tri- to hexa-nucleotide repeat motifs, the number of repeat units is >7; for a di-nucleotide repeat motif, there are at least 7 or 8 repeats depending on species; no compound repeat motif is allowed. In total, we synthesized 59 primer pairs for *A. costaricensis*, 69 for *C. insignis*, and 62 for *T. cumingiana*. For each species we first screened the markers in 3 individuals. If more than one allele was present at the focal locus, we then assessed marker polymorphism in 9 more individuals collected from the same population in the 50-ha FDP. We defined successful amplification as consistently resulting in easily interpretable allelic patterns, and polymorphism as possessing at least two alleles. We used a fluorescently labeled M13 primer coupled with M13-tagged microsatellite primers in individual PCRs as detailed previously (Wei *et al.* 2013). PCRs were performed using a touchdown protocol of an initial denaturation at 94 °C for 4 min; 28 cycles of 94 °C for 30 s, 59 °C (a decrement of 0.2 °C per cycle) for 40 s and 72 °C for 60 s; 10 cycles of 94 °C for 30 s, 53 °C for 40 s and 72 °C for 60 s; and a final extension at 72 °C for 10 min. Amplicons were sized in ABI 3730 DNA Analyzer (Applied Biosystems, Carlsbad, California, USA), and scored using GeneMarker v1.7 (Softgenetics, State College, Pennsylvania, USA).

Read length simulations

In the simulations of read length effect on microsatellite detection, reads of uniform length at each of 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 1000 and 1200 bp were drawn at random from *Populus trichocarpa* chromosomes 1–19 (v3.0, DOE-JGI, http://www.phytozome.net/poplar; Tuskan *et al.* 2006) using 0.1× coverage (~40 Mb). This equal genome coverage ensures that the ability to locate microsatellites, eliminate genomic redundancy, and design suitable primers for individual loci depends only on how long the reads are, rather than how much sequencing effort was exerted in individual simulations. We conducted platform-independent read length simulations by allowing no sequencing errors in reads using Grinder v0.5.3 (Angly *et al.* 2012). In addition, we incorporated sequencing errors into read length simulations using PBSIM v1.0.3 (Ono *et al.* 2013) with built-in PacBio CCS error profiles (substitutions/insertions/deletions ratio of 6:21:73; read accuracy of 98 ± 2%). Both error-free and error-embedded simulated reads were used directly (i.e. no quality control) to detect microsatellites and design primers, following the above-described procedure except using a relaxed primer GC content of 30–70% (also for the following simulations).

In the situation of equal genome coverage, there exists a balance between read quantity and read length to maintain the total sequence bases; that is, libraries of longer reads contain fewer reads (Tables 2S.1 and 2S.2). To relax the equal genome coverage assumption, we further used equal read quantity in each simulation. To do so, microsatellite parameters (e.g. microsatellite-containing reads, microsatellite loci; Tables 2S.1 and 2S.2) were converted to relative estimates by dividing by the corresponding total read numbers in individual simulations, and then we multiplied these relative estimates by the same read quantity of 160 000.

Sequencing error simulations

Sequencing errors of substitutions and indels (insertions and deletions) were introduced to reads of uniformly 350 bp simulated from the reference genome of *P. trichocarpa* at 0.1× coverage. Taking into account potential effects of sequencing error types on simulated results, we considered both substitution-biased (substitutions/indels ratio of 90:10) and indel-biased (substitutions/indels ratio of 10:90) sequencing errors. In terms of sequencing error rate distribution, we assumed that sequencing errors occurred either uniformly or linearly from the 5′ end to 3′ end of each read. With uniformly distributed sequencing errors, reads were simulated with an error rate of 0, 0.01%, 0.1%, 0.5%, 1%, 2%, 3%, and 5%. With linearly distributed errors, the error rate doubled from the 5′ end to 3′ end: 0, 0.01–0.02%, 0.1–0.2%, 0.2–0.4%, 0.5–1%, 1–2%, and 2–4%. To check the extent to which sequencing errors impair the amplification of microsatellite markers, designed primer pairs (A + B design) from the simulated error-containing reads were aligned back to the reference genome of *P. trichocarpa* using iPCRess implemented in Exonerate v2.2 (Slater & Birney 2005). Successful *in silico* locus amplification was defined conservatively as having unique and perfect (i.e. zero mismatch) alignment between the reference genome and the forward and reverse primer.

Error trimming simulations

To investigate whether sequence quality control is essential for microsatellite development, we compared the rate of *in silico* locus amplification between simulated sequence libraries that were treated with different quality control criteria. We simulated

reads of uniform length of 350 bp from the *P. trichocarpa* genome (0.1× coverage) using PBSIM, following the observed read length distribution and quality profiles of *T. cumingiana* CCS reads in this study. Then two types of quality control were used to filter the simulated sequences. The first QC method was based on mean read quality, requiring a minimum average read quality score of 30, as well as no reads containing homopolymers longer than 8 bases. The second QC method was based on the sliding window approach as described above, including the control of homopolymers. Microsatellite detection and primer design were conducted on both post-QC reads and raw reads. Designed microsatellite primers (A + B type) were tested *in silico* for locus amplification.

Results

Sequencing capacity of PacBio 500-bp CCS

PacBio CCS of 500-bp genomic DNA inserts returned on average 161 000 CCS reads using 4 SMRT cells (Table 2.1). Among species, the number of CCS reads varied (one-way ANOVA, $F_{2,9} = 7.273$, $P < 0.05$; Table 2.1); *A. costaricensis* yielded fewer CCS reads ($n = 105\ 881$; Table 2.1) relative to *C. insignis* (198 989; Holm's adjusted $P < 0.05$ for pairwise *t*-tests) and *T. cumingiana* (178 122, Holm's adjusted $P < 0.05$), whereas the difference between the latter two was negligible (Holm's adjusted $P = 0.436$). At a per-SMRT-cell scale, the number of CCS reads ranged from 19 801 to 57 046 (mean = 40 249; Fig. 2S.3), species identity notwithstanding. The frequency distribution of CCS read lengths revealed a wide size range (11–1391 bp, *A. costaricensis*; 9–1751 bp, *C. insignis*; 12–1917 bp, *T. cumingiana*; Fig. 2.1a), but on average only 0.01% (0.03%, *A.*

costaricensis; 0.003%, *C. insignis*; 0.001%, *T. cumingiana*) of CCS reads were shorter than 80 bp, the minimum read length required in the QDD program for microsatellite detection.

Quality control of CCS reads

Mean sequence quality of CCS reads was greatly augmented after QC (*A. costaricensis*, one-sided Wilcoxon rank sum test, $W = 4.75 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 1.62 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 1.62 \times 10^{10}$, adjusted $P < 0.001$). The minimum mean quality score of post-QC CCS reads (20, *A. costaricensis*; 25, *C. insignis*; 23, *T. cumingiana*; dotted lines, Fig. 2S.1) was nearly double that of raw CCS reads (14, 14, and 13 respectively; solid lines, Fig. 2S.1). CCS read quality was negatively correlated on a log-log scale with read length before QC (Fig. 2S.4), but was positively correlated after QC (Fig. 2S.5).

In addition, QC improved base accuracy, as the percentiles and the mean of base quality scores were elevated (Fig. 2.2). Although base accuracy declined along the length of a CCS read both prior to QC (*A. costaricensis*, $F_{1,1389} = 1.40 \times 10^4$, $P < 0.001$, adjusted $R^2 = 0.910$; *C. insignis*, $F_{1,1749} = 2.24 \times 10^3$, $P < 0.001$, adjusted $R^2 = 0.561$; *T. cumingiana*, $F_{1,1915} = 3.11 \times 10^4$, $P < 0.001$, adjusted $R^2 = 0.942$) and after QC (*A. costaricensis*, $F_{1,879} = 383.6$, $P < 0.001$, adjusted $R^2 = 0.303$; *C. insignis*, $F_{1,1042} = 491.5$, $P < 0.001$, adjusted $R^2 = 0.320$; *T. cumingiana*, $F_{1,979} = 773.6$, $P < 0.001$, adjusted $R^2 = 0.441$), the fitted slope of post-QC base quality with read base position was significantly smaller than the pre-QC fitted slope (*A. costaricensis*, slope $\beta_{\text{post-QC}} = -0.010$, $\beta_{\text{pre-QC}} = -0.036$, one-sided Welch's *t*-test, $t = 1333.8$, df = 1264, $P < 0.001$; *C. insignis*, $\beta_{\text{post-QC}} = -0.010$, $\beta_{\text{pre-QC}} = -0.018$, $t =$

508.1, df = 1958, $P < 0.001$; *T. cumingiana*, $\beta_{post\text{-}QC} = -0.013$, $\beta_{pre\text{-}QC} = -0.025$, $t = 821.8$, df = 1078, $P < 0.001$).

A positive correlation was found between homopolymer length and square root-transformed raw CCS read length (Fig. 2S.6). But homopolymer lengths were appreciably reduced after QC (*A. costaricensis*, Wilcoxon rank sum test, $W = 6.50 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 2.28 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 1.86 \times 10^{10}$, adjusted $P < 0.001$). Furthermore, we found no effect of QC on position GC content of CCS reads; pre-QC sequence position GC content averaged between 36.6% and 39.8%, and post-QC between 38.2% and 38.7% (Fig. 2S.7).

QC filtered out approximately 10% of CCS reads (Table 2.1). Remaining CCS reads were significantly shortened (*A. costaricensis*, Wilcoxon rank sum test, $W = 7.28 \times 10^9$, Holm's adjusted $P < 0.001$; *C. insignis*, $W = 2.60 \times 10^{10}$, adjusted $P < 0.001$; *T. cumingiana*, $W = 2.10 \times 10^{10}$, adjusted $P < 0.001$; Table 2.1), with a mean read length reduction by 32–39%. Post-QC CCS reads were bimodally distributed (Fig. 2.1b), in which on average 76% were longer than 80 bp.

Microsatellite detection

Without quality control, approximately 5400 to 6200 non-redundant microsatellite-containing sequences were retrieved in individual species (Table 2.1), corresponding to 3.1–5.1% of raw CCS reads. With quality control, the non-redundant microsatellite-containing sequences decreased to around 3000 (Table 2.1), accounting for 1.7–3.3% of raw CCS reads. Selected from the non-redundant microsatellite-containing trimmed CCS reads, microsatellite loci (A + B design) varied from 390 in *A. costaricensis* to 512 in *C.*

insignis and 795 in *T. cumingiana* (Table 2.1). These loci accounted for 12.4–26.5% of the non-redundant microsatellite-containing trimmed CCS reads, and 0.3–0.5% of raw CCS reads. With respect to repeat motifs, di-nucleotide motifs were most abundant (69–77% of all repeat motifs), followed by tri-nucleotide motifs (21–26%; Fig. 2.3). Other repeat motifs collectively constituted less than 5%.

We also checked the extent of microsatellite throughput reduction resulting from QC, by comparing the number of microsatellite loci retrieved from trimmed CCS reads with those retrieved from raw CCS reads. The ratio of post-QC microsatellite loci to pre-QC microsatellite loci (pre-QC n = 663, *A. costaricensis*; 1024, *C. insignis*; 1534, *T. cumingiana*) averaged 54% (range 52–59%).

Microsatellite marker validation

For *A. costaricensis*, 59 microsatellite markers were inspected for locus amplification and polymorphism, of which 62.7% (n = 37) were amplifiable, and 42.4% (n = 25) were polymorphic. Likewise, the amplification success in *C. insignis* reached 73.9% (51 of 69); polymorphic loci accounted for 39.1% (27 of 69). In *T. cumingiana*, the rate of locus amplification and polymorphism was 62.9% (39 of 62) and 45.2% (28 of 62) respectively. On average, irrespective of species identity, 66.8% of the screened microsatellite markers were successfully amplified, whereas 42.1% exhibited polymorphism. Further details about the informativeness of species-specific microsatellite markers will be provided elsewhere.

Read length simulations

Read length simulations examined the relationship between read length and microsatellite isolation effectiveness, regarding (1) the likelihood of finding shotgun reads that carry microsatellites, (2) the ability to detect genomic redundancy from microsatellite-containing reads, (3) the success of designing primers and (4) the amount of putative microsatellite loci. First, the percentage of microsatellite-containing reads, relative to total simulated reads, increased in proportion to read length (Pearson correlation coefficient r = 0.998 for both error-free and error-bearing reads; Tables 2S.1 and 2S.2). For instance, a two-fold increase in read length, such as from 200 bp to 400 bp, resulted in a nearly two-fold increase in the proportion of reads containing microsatellites (from 5.1% to 10.1%, error-free reads; from 4.8% to 9.2%, error-bearing reads). Second, with respect to genomic redundancy detection under equal genome coverage, the proportion of grouped sequences that have low levels of similarity (<95%) and thus are not regarded as from the same locus (Meglécz *et al.* 2010), increased by two orders of magnitude with read length from 0.14% at 100 bp to 24.2% at 1200 bp, relative to all microsatellite-containing reads; multihit sequences that contain interspersed repetitive regions increased from 0% at 100 bp to 0.7% at 300 bp, and to 15.6% at 1200 bp, in the situation of no sequencing errors (Fig. 2.4c). A similar magnitude of increase in detectable genomic redundancy was observed when sequencing errors were considered (from 0.1% to 25.1%, grouped sequences; from 0% to 12.9%, multihit sequences; Fig. 2.4c). Third, the rate of primer design (A + B type) for non-redundant microsatellite-containing reads increased fiftyfold from 100 bp (0.3%, error-free reads; 0.4%, error-bearing reads) to 400 bp (18.4%, error-free reads; 17.2%, error-bearing reads), and eightyfold to 1200 bp (27.4% and 26.4% in error-free and error-bearing reads respectively; Fig. 2.4d). Lastly, with respect to microsatellite throughput, the number

of microsatellite loci (≥5 repeats and ≥10 repeats; A + B design) responded positively to read length until approximately 400 bp, after which relationships became nearly asymptotic, for both error-free and error-bearing reads under equal genome coverage (Fig. 2.4a). But with equal read quantity rather than equal genome coverage, the measures of microsatellite yield increased monotonically with read length (e.g. in error-free reads, loci of ≥ 5 repeats and A+B design, linear regression slope $\beta = 5.76$, $F_{1,11} = 583.7$, $P < 0.001$, adjusted $R^2 = 0.980$; loci of ≥ 10 repeats and A+B design, $\beta = 1.18$, $F_{1,11} = 364.1$, $P < 0.001$, adjusted $R^2 = 0.968$; Fig. 2.4b).

Sequencing error simulations

Reductions in microsatellite amplification success were associated with elevated sequencing error rate, irrespective of sequencing error type and error rate distribution (Fig. 2.5). When error rate was 0 (i.e. no sequencing errors), 93.6% of microsatellite loci (primer design A + B) recovered from simulated shotgun sequences amplified *in silico*. Compared against this baseline, reductions in locus amplification became detectable when error rate increased to 0.1% given uniformly distributed base accuracy (indel bias, amplification rate = 89.4%, one-sided Proportion test, $\chi^2 = 13.80$, df =1, Holm's adjusted $P < 0.001$; substitution bias, 88.9%, $\chi^2 = 16.96$, df =1, adjusted $P < 0.001$), and to 0.1–0.2% given linearly distributed base accuracy (indel bias, 87.9%, $\chi^2 = 23.54$, df =1, adjusted $P < 0.001$; substitution bias, 88.1%, $\chi^2 = 22.94$, df =1, adjusted $P < 0.001$). For reads of uniform 1% per-base error rate, approximately 60% of microsatellite loci were amplifiable (61.3%, indel bias; 58.5%, substitution bias), and 40% (41.9%, indel bias; 37.9%, substitution bias) for reads of 2% per-base error rate (Fig. 2.5a). With a linearly distributed sequencing error

rate, approximately 68% of microsatellite loci amplified *in silico* at 0.5–1% error rate and 49% at 1–2% error rate (Fig. 2.5b).

Error trimming simulations

The *in silico* experiment of the effect of error trimming on microsatellite amplification was conducted on simulated reads that closely mimicked the characteristics of observed CCS reads (of *T. cumingiana*) in this study. Quality control based on the sliding window method reduced the test array of microsatellite loci (primer design A+B) by 57% from 1259 to 539, whereas QC based on mean read quality reduced microsatellite loci by 23% to 970. The amplification rate of microsatellite loci with sliding window-based QC was 60.7% (Fig. 2.6), significantly higher than that without QC (53.9%; one-sided Proportion test, $\chi^2 = 6.838$, df = 1, Holm's adjusted $P < 0.05$) and that with mean read quality-based QC (55.3%; $\chi^2 = 3.924$, df = 1, adjusted $P < 0.05$). When comparing the locus amplification rate of simulated reads (60.7%) with that of observed reads in *T. cumingiana* (62.9%) based on the same method of QC, no significant difference was detected ($\chi^2 = 0.042$, df = 1, $P = 0.419$).

Discussion

By elaborating the workflow from sequence quality control to marker validation, we demonstrate the effectiveness of shotgun genome circular consensus sequencing for isolating microsatellites in non-model plant species. On average, approximately 160 000 CCS reads were acquired using four SMRT cells for each species. Quality control reduced microsatellite throughput by *ca*. 50%, but several hundred microsatellite loci were still

obtained per species. These loci are also expected to have higher amplification success than the total pool of microsatellite loci before quality control, as indicated by the error trimming simulations. The initial marker screening revealed that two-thirds of the loci consistently resulted in easily interpretable amplicons, and two-fifths of the loci were polymorphic. Here we discuss the performance of PacBio CCS in comparison with other NGS platforms for microsatellite isolation, in the context of read length, read quality and sequence quality control.

Our read length simulations substantiate the postulated importance of read length in microsatellite development. Long reads increase the likelihood that a sequence contains microsatellites (Tables 2S.1 and 2S.2) by searching through more bases in a genomic location. They also increase the probability that a sequence contains intact microsatellites with sufficient flanking regions (Fig. 2.4d), because a microsatellite is less likely to be located in proximity to either of the two ends in a longer read (Abdelkrim *et al.* 2009). A primer-design success rate of up to 33% was empirically predicted for reads averaging 200 bp (reviewed in Guichoux *et al.* 2011), which coincides with that predicted by simulations here, i.e. *ca*. 25% primer design rate for non-redundant microsatellite-containing reads under relaxed criteria (primer design A–G; data not shown). Furthermore, our read length simulations provide platform-independent evidence of improved genomic redundancy detection with increases in read length under equal genome coverage (Fig. 2.4c). This finding compliments previous inferences by Elliott *et al.* (2014) based on the comparison between Ion Torrent- and 454-specific read lengths.

Notwithstanding the aforementioned benefits of longer reads, increases in read length do not result in a continuing increase in microsatellite yield, when total sequencing

effort is held constant. The relationship between the number of microsatellite loci and read length predicts a threshold read length of approximately 400 bp (Fig. 2.4a; based on visual inspection). Above the threshold, read length is not a limiting factor for microsatellite throughput: any potential gains in the number of microsatellite loci, due to elevated probability of reads containing microsatellites and increased primer design success, are offset by the losses resulting from decreased read numbers and an increased portion of unused redundant reads (e.g. grouped and multihit sequences). Nonetheless, microsatellite loci recovered from longer reads (e.g. 1000 bp) may have a higher chance of successful amplification than loci from reads of 400 bp, because of more effective genomic redundancy removal. An additional comparison of *in silico* microsatellite amplification indeed revealed a small yet significant increase by *ca*. 3% when read length was increased from 400 bp to 1000 bp (data not shown).

Microsatellite yield behaves as a threshold response to read length in the context of equal sequencing depth. However, NGS platforms differ in sequencing throughput; therefore, the estimation of platform-dependent microsatellite yield needs to consider both read length and read numbers without the constraint of equal genome coverage or equal read quantity. Despite amassed NGS-based microsatellite datasets, cross-platform comparisons of the number of microsatellite loci, based on these empirical investigations, are complicated by heterogeneities in genomic microsatellite frequency and genome size among taxa (Toth *et al.* 2000; Morgante *et al.* 2002; Ellegren 2004), and in microsatellite searching and primer design criteria. Therefore, we base this assessment on our simulations, taking into account platform-dependent read numbers and mean read length (Table 2.2). Specifically, we multiplied platform-specific read numbers by the estimated

proportion of reads containing microsatellite loci (primer design A+B; Table 2S.1) for the corresponding mean read length of individual NGS platforms. Despite the use of uniform read length rather than platform-dependent read length distributions, our simulations provide good predictions of microsatellite throughput, as evidenced by the concordance between simulated and observed number of microsatellite loci on PacBio and Illumina MiSeq (Table 2.2). The discrepancy between simulations and empirical findings on Ion Torrent and 454 may result primarily from low microsatellite density in targeted organism genomes (Elliott *et al.* 2014), as noted by the authors. In general, PacBio and Ion Torrent produce a comparable number of microsatellite loci relative to 454 but with a *ca*. 50% reduction in cost; but Illumina MiSeq generates approximately 30 times more microsatellite loci than PacBio and Ion Torrent at the same total cost (Table 2.2). Nevertheless, all the NGS platforms are able to deliver thousands of microsatellite loci, far more than a project could practically screen and genotype.

As *in silico* locus acquisition is no longer the bottleneck for microsatellite development, the efficiency of converting these loci into functional makers is of equivalent, if not greater, importance relative to the initial sequencing step. This consideration is particularly salient in light of the distinct effects that sequencing errors have on microsatellite yield and microsatellite amplification. Microsatellite yield is little affected by the presence of sequencing errors, as error-bearing (read accuracy of 98 ± 2%) and error-free reads of the same read length were able to retrieve a similar number of microsatellite loci (Fig. 2.4a). However, the amplification rate of these microsatellite loci plummets when sequencing errors are present (Fig. 2.5). In practice, all NGS platforms produce sequencing errors but to varying degrees, such as 1.07% reported for 454 GS-FLX Titanium (Gilles *et*

al. 2011), and 1.71% for Ion Torrent PGM and 0.80% for Illumina MiSeq (Quail *et al.* 2012). Given these error rates, roughly 50–68% of microsatellite loci are predicted to be able to amplify unique and interpretable PCR products, according to our simulations; this estimate is consistent with empirical findings (e.g. Castoe *et al.* 2012; Fernandez-Silva *et al.* 2013; Wei *et al.* 2013).

One important implication of the adverse effects of sequencing errors on microsatellite amplification is that quality control is essential for the consideration of cost- and labor-effective marker validation. In this study, significant improvement in microsatellite amplification was achieved by quality control using a sliding-window approach (Fig. 2.6). Although this finding is based on PacBio CCS-dependent error trimming simulations, it can also apply to sequences from other NGS platforms, as base-calling accuracy in general declines with base position (Gilles *et al.* 2011; Loman *et al.* 2012). One question regarding quality control concerns the possible negative effect of shortened read lengths after error trimming. By comparing the *in silico* amplification rate of microsatellite loci retrieved from error-bearing reads of 1000 bp with that of loci from error-free reads of 200 bp (Fig. 2.4a), we found the amplification rate was two-fold higher when errors were absent, despite a five-fold decrease in read length (data not shown). This finding suggests that the importance of read quality outweighs that of read length in terms of obtaining successfully amplifiable microsatellite loci. The methods of quality control may vary between platforms that emphasize long read length (e.g. 454 and PacBio) and those that emphasize high throughput (e.g. Illumina MiSeq and Ion Torrent). The sliding window-based quality control described in this study can be used for 454 and PacBio, because this approach shortens read lengths but is unlikely to result in a substantial

reduction in sequence quantity. Meanwhile, more stringent quality control, such as removing reads of per-base quality score below 30, can be afforded for Illumina MiSeq or Ion Torrent because of high sequence throughput.

Conclusion

This study provides a quantitative demonstration of microsatellite development in relation to sequence attributes, based on which the performance of PacBio CCS is evaluated in comparison to other NGS platforms. PacBio CCS is suitable for fast, small-scale microsatellite development due to its flexibility in scaling sequencing effort, in terms of the number of SMRT cells utilized per project. A single SMRT cell can potentially deliver enough functional microsatellite markers (e.g. Grohme *et al.* 2013; Wainwright *et al.* 2013), at a sequencing cost of ~$200 (not including library preparation). On the other hand, Illumina MiSeq paired-end sequencing can be particularly cost-efficient when greater sequencing effort is required, such as for multi-species microsatellite projects, as well as for organisms that have low genomic microsatellite density. In light of the continuing advances in sequence length on all the platforms, read length may not be the primary concern for NGS use in microsatellite isolation. Instead, sequencing accuracy and the corresponding strategies of quality control are essential for time- and cost-effective microsatellite isolation.

Acknowledgements

This chapter was coauthored with Jordan Bemmels and Christopher Dick and published in Molecular Ecology Resources in 2014. We are grateful to two anonymous reviewers and the editor Travis Glenn for providing invaluable feedback on the manuscript. We thank the Smithsonian Tropical Research Institute and Center for Tropical Forest Science for facilitating fieldwork on Barro Colorado Island. This work was supported by a Rackham Graduate Student Research Grant to N.W. and a grant to C.W.D. from the College of Literature, Science and the Arts at the University of Michigan. N.W. was supported by a Barbour Scholarship from the University of Michigan, and J.B.B. was supported by an NSF Graduate Research Fellowship.

References

Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *Biotechniques* **46**, 185-191.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research* **40**, e94.

Buffalo V (2012) qrqc: Quick Read Quality Control. R package version 1.10.0. http://github.com/vsbuffalo/qrqc.

Castoe TA, Poole AW, de Koning APJ*, et al.* (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PloS one* **7**, e30953.

Eid J, Fehr A, Gray J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**, 435-445.

Elliott CP, Enright NJ, Allcock RJ*, et al.* (2014) Microsatellite markers from the Ion Torrent: a multi-species contrast to 454 shotgun sequencing. *Molecular Ecology Resources* **14**, 554-568.

Fernandez-Silva I, Whitney J, Wainwright B*, et al.* (2013) Microsatellites for next-generation ecologists: a post-sequencing bioinformatics pipeline. *PloS one* **8**, e55990.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines – recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources* **11**, 1093-1101.

Gilles A, Meglecz E, Pech N*, et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics* **12**, 245.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Glenn TC (2013) 2013 NGS Field Guide: Overview. http://www.molecularecologist.com/next-gen-fieldguide-2013/.

Grohme MA, Soler RF, Wink M, Frohme M (2013) Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS. *Biotechniques* **55**, 253-256.

Guichoux E, Lagache L, Wagner S*, et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**, 591-611.

Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* **106**, 158-171.

Huey J, Real K, Mather P*, et al.* (2013) Isolation and characterization of 21 polymorphic microsatellite loci in the iconic Australian lungfish, *Neoceratodus forsteri*, using the Ion Torrent next-generation sequencing platform. *Conservation Genetics Resources* **5**, 737-740.

Illumina Incorporation (2013) Illumina systems. http://www.illumina.com/systems.ilmn.
Jennings TN, Knaus BJ, Mullins TD, Haig SM, Cronn RC (2011) Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources* **11**, 1060-1067.
Larkin MA, Blackshields G, Brown NP*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.
Loman NJ, Misra RV, Dallman TJ*, et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**, 434-439.
Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.
McCracken GR, Wilson KL, Paterson I*, et al.* (2014) Development of 17 novel microsatellite markers for the longnose sucker (*Catostomus catostomus*) and successful cross-specific amplification of 14 previously developed markers from congeneric species. *Conservation Genetics Resources* **6**, 329-332.
Meglécz E, Costedoat C, Dubut V*, et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404.
Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**, 194-200.
Nowak C, Zuther S, Leontyev SV, Geismar J (2014) Rapid development of microsatellite markers for the critically endangered Saiga (*Saiga tatarica*) using Illumina® Miseq next generation sequencing technology. *Conservation Genetics Resources* **6**, 159-162.
Ono Y, Asai K, Hamada M (2013) PBSIM: PacBio reads simulator–toward accurate genome assembly. *Bioinformatics* **29**, 119-121.
Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma R, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. *Trends in Genetics* **26**, 177-187.
Pacific Biosciences (2013) SMRT Technology. http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/.
Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution* **16**, 142-147.
Quail MA, Smith M, Coupland P*, et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* **13**, 341.
R Development Core Team (2012) *R: A language and environment for statistical computing, Version 2.15.0.* R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**, 365-386.
Schloss PD, Westcott SL, Ryabin T*, et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537-7541.
Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365-371.

Schlotterer C (2004) The evolution of molecular markers – just a matter of fashion? *Nature Reviews Genetics* **5**, 63-69.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**, 615-629.
Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *Bmc Bioinformatics* **6**, 31.
Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research* **10**, 967-981.
Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* **38**, e159.
Tuskan GA, DiFazio S, Jansson S*, et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604.
Wainwright B, Arlyza I, Karl S (2013) Isolation and characterization of twenty-one polymorphic microsatellite loci for *Polycarpa aurata* using third generation sequencing. *Conservation Genetics Resources* **5**, 671-673.
Wei N, Dick CW, Lowe AJ, Gardner MG (2013) Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing. *Applications in Plant Sciences* **1**, 1200295.
Zalapa JE, Cuevas H, Zhu HY*, et al.* (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* **99**, 193-208.
Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology* **11**, 1-16.
Zhang XJ, Davenport KW, Gu W*, et al.* (2012) Improving genome assemblies by sequencing PCR products with PacBio. *Biotechniques* **53**, 61-62.

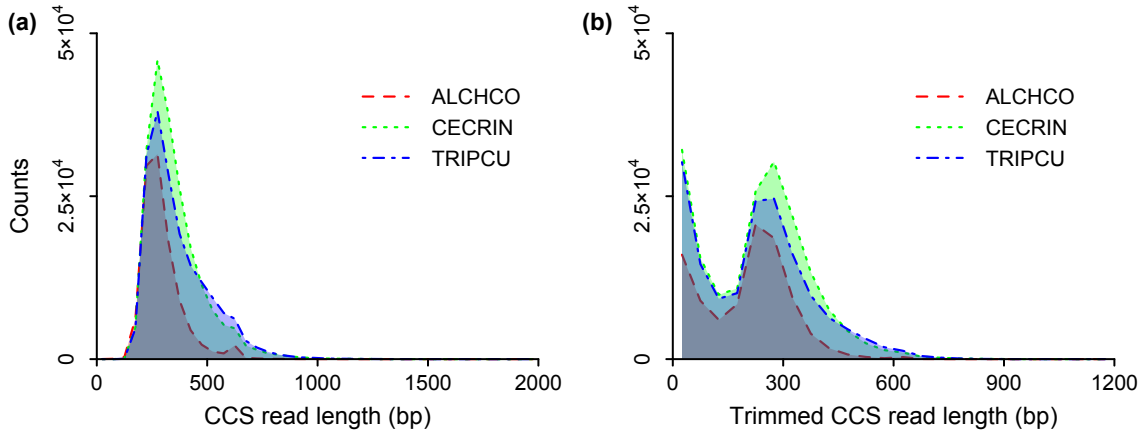

Figure 2.1 Frequency distribution of CCS read lengths generated by 500-bp genomic shotgun circular consensus sequencing. CCS, circular consensus sequencing; CCS reads, adapter-free consensus sequences generated by CCS; trimmed CCS reads, CCS reads passing quality control. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

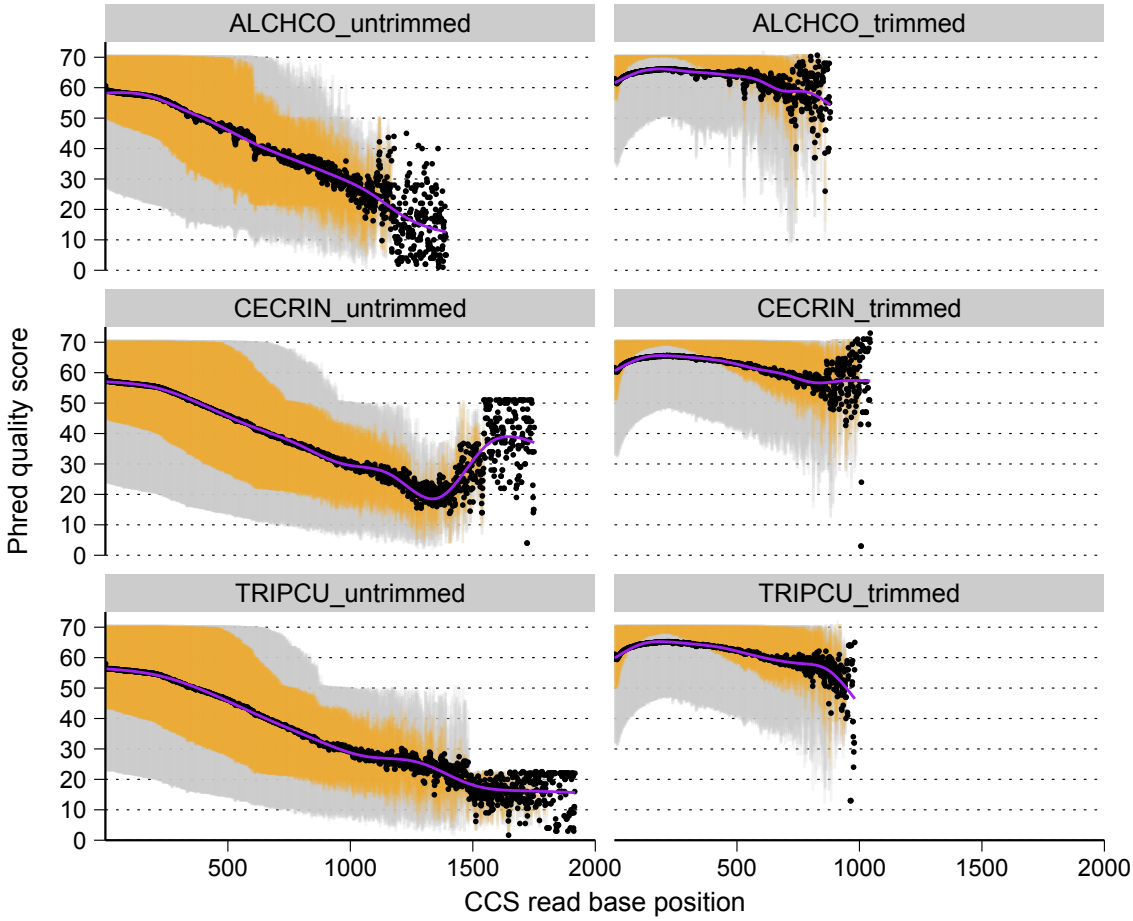

Figure 2.2 Base quality scores of CCS reads before (untrimmed) and after (trimmed) quality control. Outer whiskers (grey regions) represent the 10th to the 90th percentile of position quality scores; inner whiskers (orange regions) represent the 25th to the 75th percentile; dots are the mean quality score at each base position; lines are fitted GAM (generalized additive model) smooth lines. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

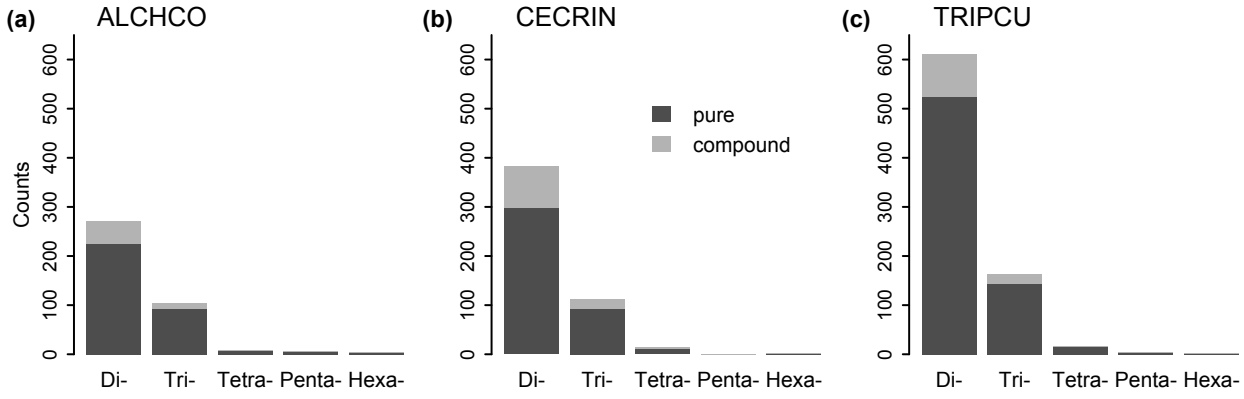

Figure 2.3 Motif length-specific microsatellite loci identified from quality-controlled CCS reads. See text for microsatellite searching and primer design criteria. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

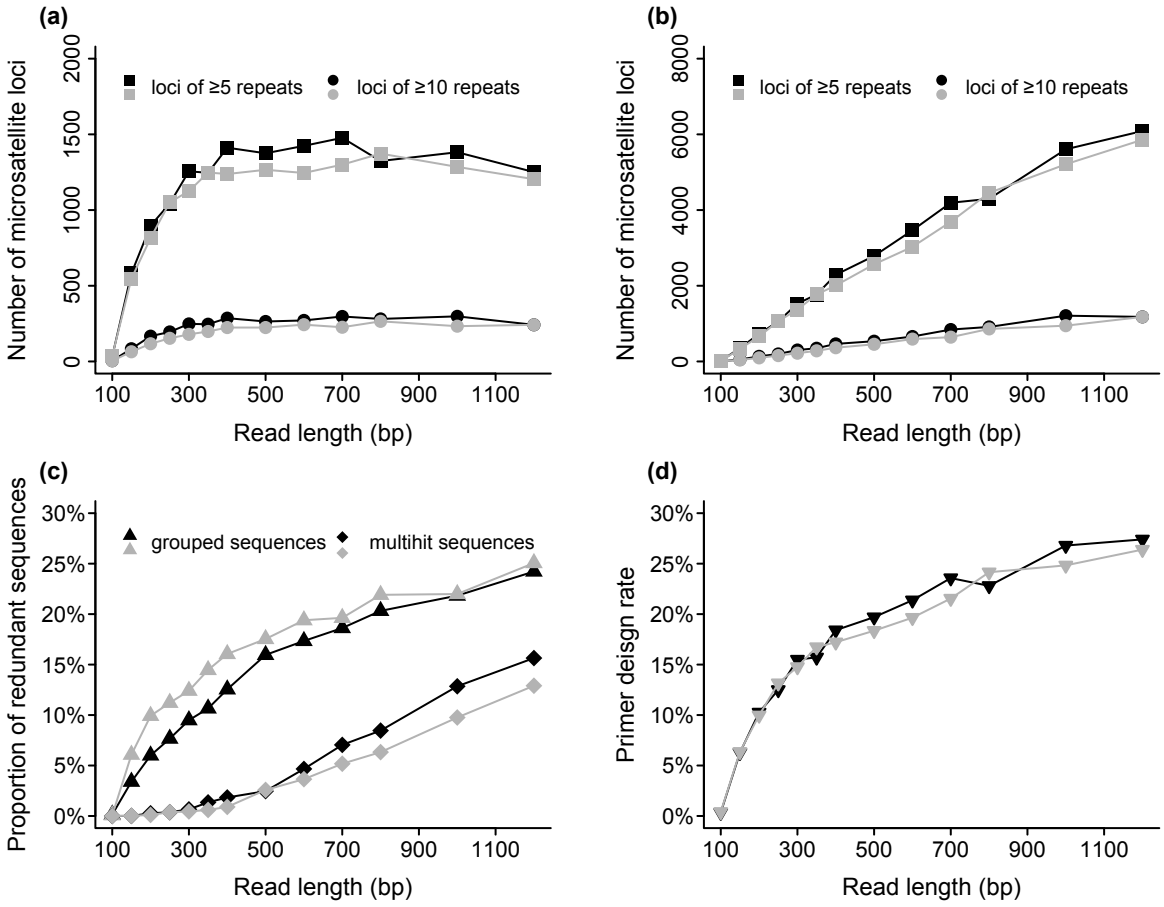

Figure 2.4 The effects of read length on microsatellite yield (a–b), genomic redundancy detection (c), and primer design success rate (d). (a–b) Microsatellite loci (primer design A + B, see text) were retrieved from simulated sequences assuming (a) equal genome coverage of 0.1× and (b) equal read numbers of 160 000 for each read length size from the *Populus trichocarpa* genome. (c) The proportions of grouped and multihit sequences are relative to total microsatellite-containing reads. (d) Primers (A + B type) were designed for non-redundant microsatellite-containing reads. Black symbols indicate the absence of sequencing errors in simulations; grey symbols indicate the inclusion of sequencing errors (read accuracy of 98 ± 2%).

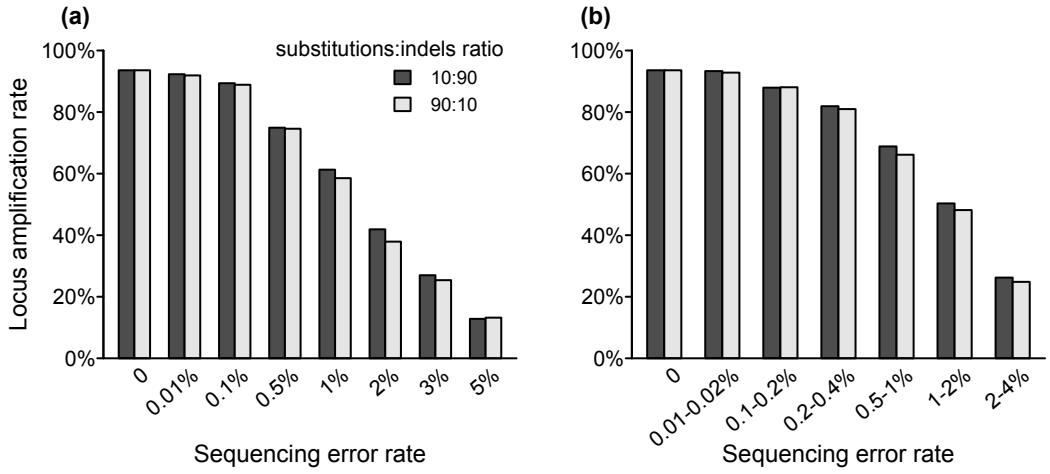

Figure 2.5 Simulations of microsatellite amplification rate in relation to sequencing errors. Sequencing errors of substitutions and indels (insertions and deletions) were introduced to simulated reads of 350 bp (a) uniformly or (b) linearly increasing from the 5′ end to 3′ end from the *Populus trichocarpa* genome with 0.1× coverage. Substitution- and indel-dominated sequencing errors are represented by light bars and dark bars respectively.

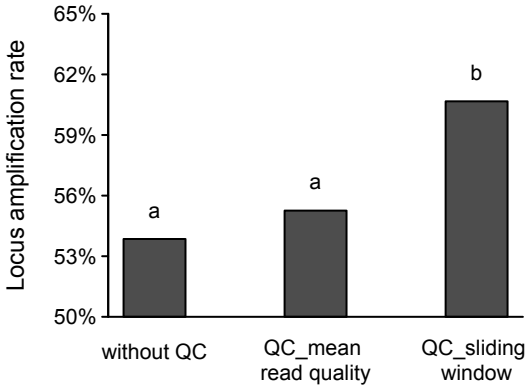

Figure 2.6 The effect of quality control on microsatellite locus amplification. *In silico* microsatellite amplification rate is significantly higher with quality control based on a sliding window approach (QC_sliding window) relative to that without QC, as well as that with QC based on mean read quality score (QC_mean read quality).

Table 2.1 Sequencing capacity and microsatellite throughput of 500-bp genomic shotgun circular consensus sequencing using four SMRT cells per species. CCS, circular consensus sequencing; CCS reads, adapter-free consensus sequences generated by CCS; trimmed CCS reads, CCS reads passing quality control.

	Alchornea costaricensis (Euphorbiaceae)	*Cecropia insignis* (Cecropiaceae)	*Triplaris cumingiana* (Polygonaceae)
Sequence megabases (Mb)	31.4	70.2	65.2
Number of reads			
CCS reads	105 881	198 989	178 122
Trimmed CCS reads	95 265	177 161	157 026
Average read length (bp)			
CCS reads	297	353	366
Trimmed CCS reads	201	225	222
Microsatellite detection			
Non-redundant SSR-containing CCS reads	5433	6212	5793
Non-redundant SSR-containing trimmed CCS reads	3146	3072	3001
SSR loci (≥5 repeats; primer design A + B)	390	512	795

Table 2.2 Cross-platform comparisons of next-generation sequencing (NGS) use in microsatellite development. Platform-specific information of read length, read quantity, and observed microsatellite loci are from this study on PacBio, Elliott *et al.* (2014) on Ion Torrent and 454, and Nowak *et al.* (2014) on Illumina MiSeq. Predicted microsatellite loci are calculated based on read length simulations according to platform-specific mean read length and read quantity, assuming no sequencing errors.

NGS platform	PacBio CCS	Ion Torrent PGM	454 GS-FLX	Illumina MiSeq
Sequencing unit	4 SMRT cells	1 '316' chip	1/8 PTP	1 PE 250 bp
Average read length (bp)	350	150	350	400*
Number of reads	160 000	1 000 000	150 000	6 300 000
Sequencing cost§	~$1000	~$1000	~$2000	~$1400
Predicted SSR loci of ≥5 repeats (primer design A + B)	1769	2213	1658	90 194
Observed SSR loci of ≥5 repeats	1645¶	413	165	81 886
Predicted SSR loci of ≥10 repeats (primer design A + B)	349	319	327	18 269

PTP, PicoTiterPlate; PE, paired-end.
*Predicted mean contig length of MiSeq paired-end 250 bp sequencing (observed contig lengths were not reported in the original data)
§Sequencing cost including library preparation from Glenn (2013)
¶Observed SSR loci of ≥5 repeats (primer design A + B; default parameters used in the QDD program) retrieved from CCS reads without quality control

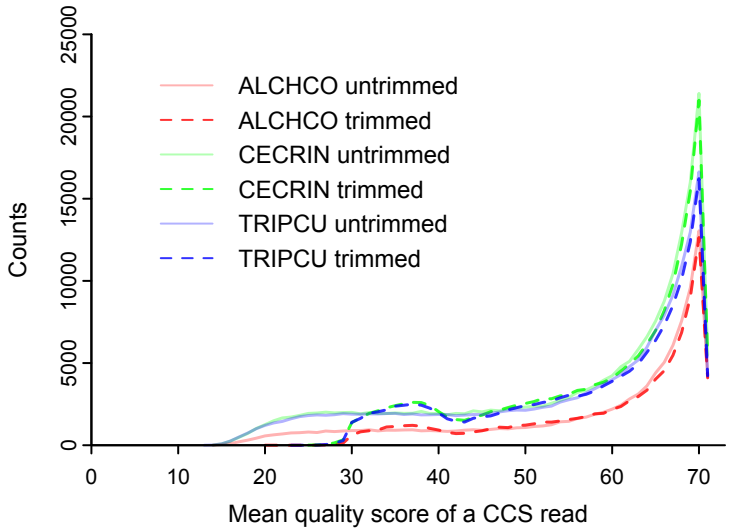

Figure 2S.1 Frequency distribution of mean quality score of individual CCS reads before (untrimmed) and after quality control (trimmed). ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

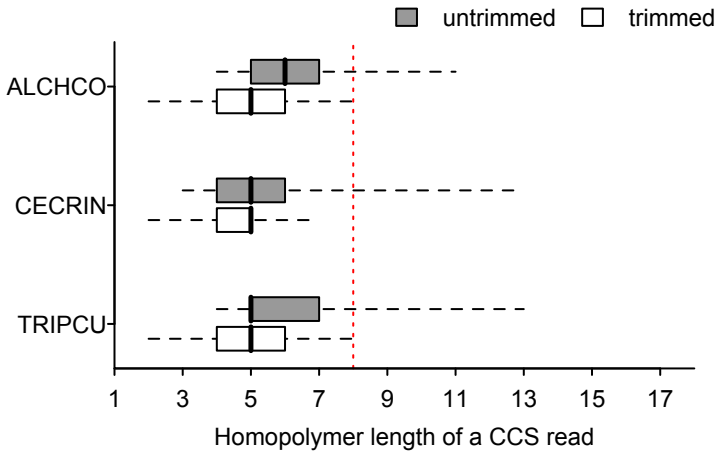

Figure 2S.2 Homopolymer lengths of CCS reads. Dotted whiskers present the 2.5th percentile and 97.5th percentile. Red vertical line indicates a homopolymer length of 8 bases. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

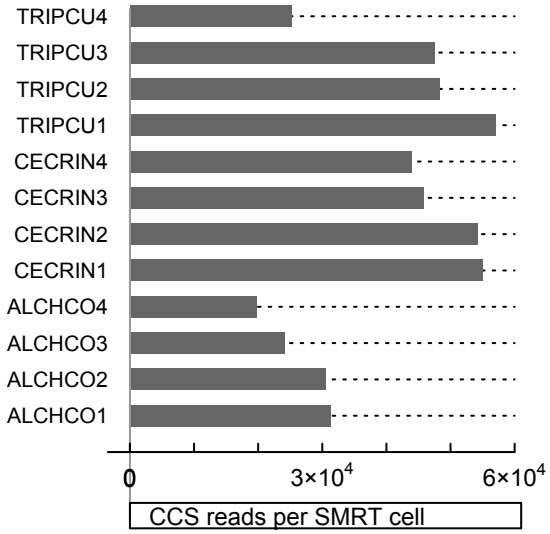

Figure 2S.3 The number of CCS reads generated by a single SMRT cell. Individual species were sequenced using four SMRT cells, designated numerically as 1, 2, 3, and 4. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

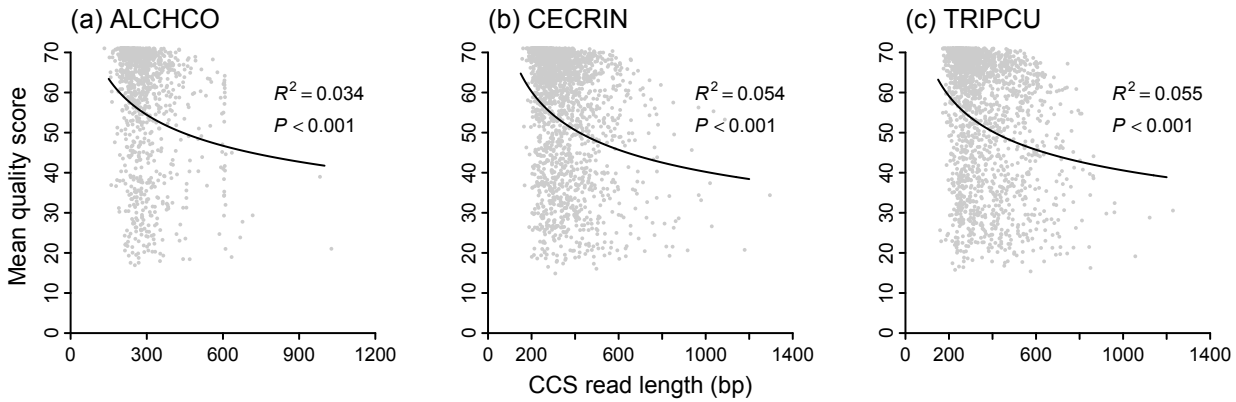

Figure 2S.4 Negative correlation between sequence quality and sequence length in raw CCS reads. The linear regression was fitted between log-transformed mean quality score of individual CCS reads and log-transformed CCS read lengths. One per cent of raw CCS reads are visualized (grey dots). ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

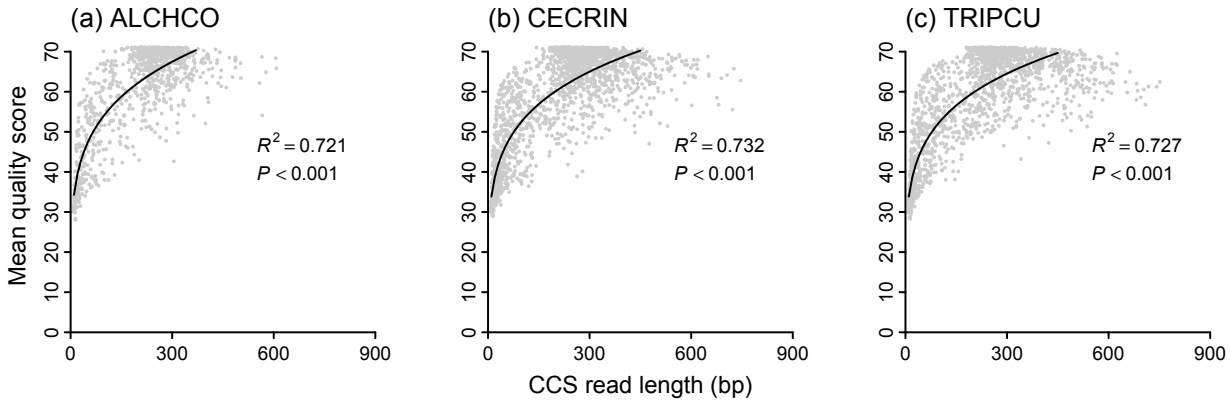

Figure 2S.5 Positive correlation between sequence quality and sequence length in post-QC CCS reads. The linear regression was fitted between log-transformed mean quality score of individual post-QC CCS reads and log-transformed post-QC CCS read lengths. One per cent of post-QC CCS reads are visualized (grey dots). ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

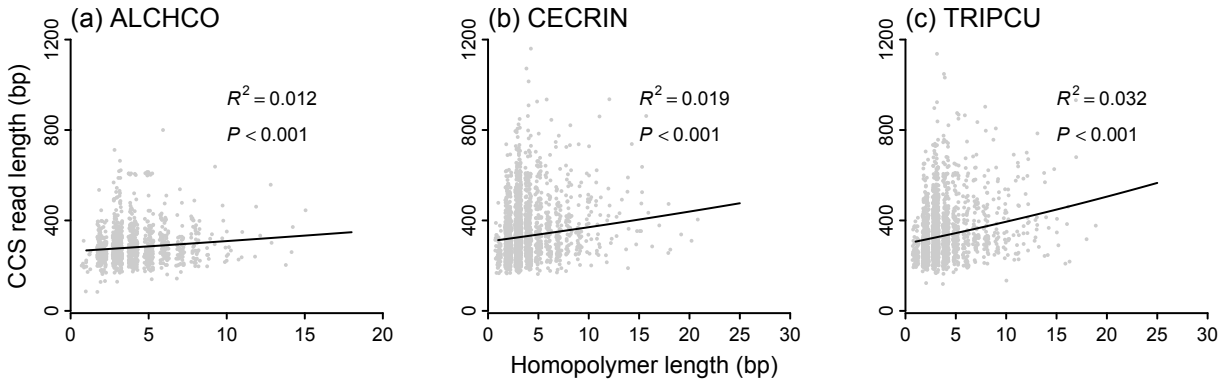

Figure 2S.6 Positive correlation between homopolymer length and raw CCS read length. The linear regression was fitted between homopolymer lengths and square-root transformed CCS read lengths. One per cent of raw CCS reads are visualized (grey dots). ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

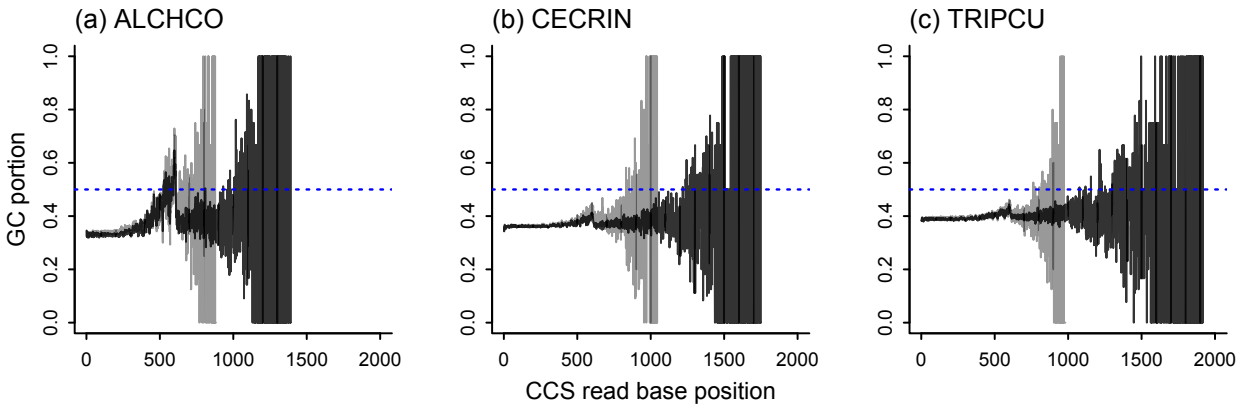

Figure 2S.7 Base position GC content of CCS reads before (black) and after quality control (grey). Blue dotted lines indicate a GC content of 50%. ALCHCO: *Alchornea costaricensis*; CECRIN: *Cecropia insignis*; TRIPCU: *Triplaris cumingiana*.

Table 2S.1 Simulations of microsatellite detection effectiveness in relation to read length when sequencing errors were not introduced. Reads were simulated from the *Populus trichocarpa* genome using 0.1× coverage. Proportion data within parentheses are calculated relative to the number of simulated reads.

Read length (bp)	Number of reads	SSR-containing reads	Redundant grouped sequences	Redundant multihit sequences	Non-redundant SSR-containing reads	Total SSR loci (≥5 repeats; A + B)	Potentially polymorphic SSR loci (≥10 repeats; A + B)
100	394 508	10 373 (2.63%)	15 (0.00%)	0 (0.00%)	9768 (2.48%)	33 (0.01%)	5 (0.00%)
150	263 006	10 368 (3.94%)	352 (0.13%)	1 (0.00%)	9281 (3.53%)	582 (0.22%)	84 (0.03%)
200	197 254	10 154 (5.15%)	609 (0.31%)	27 (0.00%)	8744 (4.43%)	897 (0.45%)	167 (0.08%)
250	157 804	9957 (6.31%)	763 (0.48%)	37 (0.01%)	8348 (5.39%)	1043 (0.66%)	195 (0.12%)
300	131 503	9905 (7.53%)	940 (0.71%)	65 (0.05%)	8104 (6.16%)	1256 (0.96%)	248 (0.19%)
350	112 717	9930 (8.81%)	1058 (0.94%)	134 (0.12%)	7928 (7.03%)	1246 (1.11%)	246 (0.22%)
400	98 627	9972 (10.1%)	1251 (1.27%)	183 (0.19%)	7670 (7.78%)	1412 (1.43%)	286 (0.29%)
500	78 902	9461 (12.0%)	1510 (1.91%)	233 (0.30%)	6980 (8.85%)	1375 (1.74%)	264 (0.33%)
600	65 752	9417 (14.3%)	1633 (2.48%)	439 (0.67%)	6662 (10.1%)	1424 (2.17%)	271 (0.41%)
700	56 359	9355 (16.6%)	1742 (3.09%)	659 (1.17%)	6266 (11.1%)	1477 (2.62%)	297 (0.53%)
800	49 314	9113 (18.5%)	1852 (3.76%)	771 (1.56%)	5807 (11.8%)	1324 (2.68%)	281 (0.57%)
1000	39 451	8786 (22.3%)	1918 (4.86%)	1129 (2.86%)	5158 (13.1%)	1382 (3.50%)	298 (0.76%)
1200	32 876	8526 (25.9%)	2065 (6.28%)	1334 (4.06%)	4565 (13.9%)	1251 (3.81%)	242 (0.74%)

Table 2S.2 Simulations of microsatellite detection effectiveness in relation to read length when PacBio CCS error profiles (Ono *et al.* 2013) were used. Reads were simulated from the *Populus trichocarpa* genome using 0.1× coverage. Proportion data within parentheses are calculated relative to the number of simulated reads.

Read length (bp)	Number of reads	SSR-containing reads	Redundant grouped sequences	Redundant multihit sequences	Non-redundant SSR-containing reads	Total SSR loci (≥5 repeats; A + B)	Potentially polymorphic SSR loci (≥10 repeats; A + B)
100	394 518	9563 (2.42%)	7 (0.00%)	0 (0.00%)	9167 (2.32%)	37 (0.01%)	4 (0.00%)
150	263 015	9519 (3.62%)	578 (0.22%)	2 (0.00%)	8600 (3.27%)	547 (0.21%)	64 (0.02%)
200	197 266	9443 (4.79%)	937 (0.47%)	10 (0.01%)	8173 (4.14%)	815 (0.41%)	117 (0.06%)
250	157 814	9391 (5.95%)	1053 (0.67%)	35 (0.02%)	7987 (5.06%)	1051 (0.67%)	153 (0.10%)
300	131 514	9107 (6.92%)	1130 (0.86%)	41 (0.03%)	7618 (5.79%)	1125 (0.86%)	179 (0.14%)
350	112 727	9214 (8.17%)	1333 (1.18%)	55 (0.05%)	7457 (6.62%)	1248 (1.11%)	198 (0.18%)
400	98 638	9115 (9.24%)	1463 (1.48%)	82 (0.08%)	7192 (7.29%)	1238 (1.26%)	224 (0.23%)
500	78 914	9067 (11.5%)	1590 (2.01%)	234 (0.30%)	6897 (8.74%)	1266 (1.60%)	224 (0.28%)
600	65 763	8688 (13.2%)	1685 (2.56%)	317 (0.48%)	6337 (9.64%)	1245 (1.89%)	243 (0.37%)
700	56 376	8461 (15.0%)	1661 (2.95%)	438 (0.78%)	6034 (10.7%)	1300 (2.31%)	226 (0.40%)
800	49 324	8393 (17.0%)	1839 (3.73%)	531 (1.08%)	5685 (11.5%)	1373 (2.78%)	265 (0.54%)
1000	39 463	8076 (20.5%)	1776 (4.50%)	788 (2.00%)	5175 (13.1%)	1285 (3.26%)	233 (0.59%)
1200	32 887	7889 (24.0%)	1977 (6.01%)	1018 (3.10%)	4562 (13.9%)	1204 (3.66%)	242 (0.74%)

References

Ono Y, Asai K, Hamada M (2013) PBSIM: PacBio reads simulator-toward accurate genome assembly. *Bioinformatics* **29**, 119-121.

APPENDIX A

Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing

Abstract

Polymorphic microsatellite loci were characterized in the dioecious Neotropical rain forest tree *Virola sebifera*. The markers will be used to study ecological and genetic impacts of hunting and landscape change in this vertebrate-dispersed, insect-pollinated tree species. Simple sequence repeats (SSRs) were screened from genomic libraries of South American *V. sebifera* obtained by shotgun 454 pyrosequencing. Primer pairs were tested on Panamanian samples ($N = 42$). Approximately 52% of the 61 tested SSR markers amplified and 16% were polymorphic. Ten selected polymorphic SSR loci contained 7 to 15 alleles per locus, and polymorphic information content (PIC) averaged 0.694. Observed heterozygosity ranged from 0.465 to 0.905, and expected heterozygosity was between 0.477 and 0.876. The ten polymorphic loci will be useful in studying gene flow and genetic structure at local and regional spatial scales in *Virola sebifera*.

Keywords: microsatellite loci; shotgun 454 pyrosequencing; *Virola sebifera*

Introduction

Neotropical nutmeg *Virola sebifera* Aubl. (Myristicaceae) is a wide-ranging canopy tree found in mature tropical forests from Central America to the Amazon Basin and Guiana Shield. Like other species in its genus, *V. sebifera* is dioecious, pollinated by small insects and dispersed by vertebrates (primarily large birds) that consume the nutrient rich red aril covering its seeds (Howe 1981). Given the high mobility and considerable seed loads of large avian dispersers, seed-mediated gene flow in *V. sebifera* may play an important role in maintaining genetic variation within and among populations. However, as increasing anthropogenic activities (e.g., hunting and landscape change) adversely impact the abundance and/or habitat of frugivores (Wright 2003; Vetter *et al.* 2011), it is important to investigate how changing vertebrate densities may impact gene flow and population structure in *V. sebifera* and other tropical forest tree species.

To address these and other questions, we developed a set of polymorphic microsatellite DNA markers for *V. sebifera*, based on genomic DNA libraries obtained from French Guiana samples by shotgun 454 pyrosequencing (Gardner *et al.* 2011b).

Methods and Results

Previously developed genomic libraries of *V. sebifera* (Gardner *et al.* 2011b) were obtained using the combined genomic DNA of six French Guiana individuals, sampled from tagged trees in trails or permanent forest inventory plots in three localities: Sentier la Mirande (4°51'N, 52°20'W; Tag no. S35, S31), Sentier Rorota (4°52'N, 52°15'W; S104, S110), and Iracoubo (5°25' N, 53°5'W; S230, S235). Genomic DNA was isolated from each individual using NucleoSpin Plant II (Macherey Nagal, Düren, Germany), then pooled with

equal concentrations (~0.8 μg/individual) for subsequent 454 pyrosequencing. Standard GS-FLX Titanium library preparation was adopted. After DNA nebulization, small fragments of length <350 bp were removed. Fragmented DNA was then ligated with MID-tagged (MID5, ACGAGTAGACT) adapters. This barcoded *V. sebifera* DNA library was multiplexed with 7 other species in a single run of GS-FLX Titanium, which rendered *V. sebifera* 12.5% of the picotiter plate.

We used the program QDD version 2 (Meglecz *et al.* 2010), set at default parameters, to search for SSR loci with ≥5 uninterrupted motif repeats from 90,164 read sequences (mean read length = 367 bp) (Gardner *et al.* 2011b, a). The SSR marker output was further restricted to A and B primer designs in QDD version 2, so as to exclude loci with complex flanking regions (i.e. containing repeat units). We obtained a total of 526 SSR loci, of which 315 contained di-nucleotide motif, followed by 182 tri-, 21 tetra-, 6 penta-, and 2 hexa-nucleotide motifs. Following the suggestions of Gardner *et al.* (2011b), we first focused on loci containing at least 10 pure repeat units of di-, tetra- and penta-nucleotide SSR motifs, which were expected to be more polymorphic than other motifs. However, because of an unexpected low rate of amplification success and polymorphism, we also included compound motifs, and tri- and hexa-nucleotide microsatellite loci of ≥9 repeats. The final testing array contained 61 candidate SSR markers (57% in di-, 36% in tri-, 3% in tetra-, 2% in penta-, and 2% in hexa-nucleotide motif).

We checked the amplification rate and polymorphism of the 61 SSR primer pairs in 42 *V. sebifera* adult trees (diameter at breast height ≥20 cm; voucher: Pérez 1806 and Pérez 1930, STRI herbarium, Panama), which were randomly collected from the 50-ha Forest Dynamics Plot in the plateau of Barro Colorado Island (9°10'N, 79°51'W), Panama.

Genomic DNA was isolated from silica-dried leaves using the DNeasy Plant Mini Kit (QIAGEN, Valencia, California, USA), quantified using NanoDrop 2000 (Thermo Scientific, Wilmington, Delaware, USA) and diluted to 1.5 ng/μL for subsequent PCR. The 6 μL PCR cocktail contained 1.5 ng DNA template, 0.05 μM of M13 tagged (5′−TGTAAAACGACGGCCAGT−3′) forward primer, 0.4 μM reverse primer, 0.017 μM 6FAM-labeled M13 primer (5′−TGTAAAACGACGGCCAGT−3′), 4 mM $MgCl_2$, and 3 μL GoTaq Colorless Master Mix (Promega, Madison, Wisconsin, USA) with buffer (pH 8.5), 200 μM of each dNTP and 1U *Taq* DNA polymerase. PCRs were carried out in Mastercycler ep thermocycler (Eppendorf, Hamburg, Germany) following an initial denaturation at 94°C for 4 min; 28 cycles of 94°C for 30 s, 55°C for 40 s and 72°C for 60 s; 10 cycles of 94°C for 30 s, 52°C for 40 s and 72°C for 60 s; and a final extension at 72°C for 10 min. PCR product of 1.5 μL was added to 12 μL Hi-Di formamide (Applied Biosystems, Carlsbad, California, USA) and 0.05 μL GeneScan 500 Rox Standard (Applied Biosystems) for subsequent fragment sizing in ABI 3730 DNA Analyzer (Applied Biosystems) by the DNA Sequencing Core Laboratory at the University of Michigan. Alleles were visualized and scored using GeneMarker version 1.7 (Softgenetics, State College, Pennsylvania, USA). Marker polymorphism, including the number of alleles per locus, observed and expected heterozygosity, exclusion probability with one parent known, and Hardy–Weinberg equilibrium (HWE), was estimated in GenAIEx version 6.4 (Peakall & Smouse 2006). Significance levels for multiple tests of HWE (α-level = 0.05) were adjusted by sequential Bonferroni procedure (Rice 1989). In addition, polymorphism information content of each locus was measured using PowerMarker version 3.0 (Liu &

Muse 2005). We tested for the presence of null alleles, allelic dropout, and scoring errors (due to stuttering) using MICRO-CHECKER version 2.2.3 (Van Oosterhout *et al.* 2004).

Our results showed that 17 (49%) di-, 13 (59%) tri-, 1 (50%) tetra-, 0 penta-, and 1 (100%) hexa-nucleotide markers were amplifiable; but 3 (9%) di-, 6 (27%) tri-, 0 tetra-, 0 penta-, and 1 (100%) hexa-nucleotide SSRs were considered as polymorphic (≥6 alleles per locus) in the present study. These ten polymorphic markers (Table A1) had mean allelic richness of 10.3 alleles per locus (Table A2). Observed heterozygosity ranged from 0.465 to 0.905, and expected heterozygosity was between 0.477 and 0.876. PIC per locus averaged 0.694 (Table A2). No allelic dropout or scoring errors were detected, but one locus (VSE02) appeared to contain null alleles. Two (VSE02 and VSE36) of the ten loci showed deviation from Hardy–Weinberg proportions after sequential Bonferroni correction ($P < 0.006$). The overall exclusion probability with one parent known was 0.992.

Conclusion

We found that tri-nucleotide SSR loci exhibited better marker properties, such as higher probability of polymorphism and less stuttering, than the other motifs, particularly di-nucleotide SSRs. Although the 454 genomic libraries were obtained from French Guiana samples, the markers were developed for Panamanian individuals, despite the probable high levels of genomic divergence between populations located east and west of the Andean cordilleras. Genomic divergence may partly explain the unexpected low rate of amplification (52%) and polymorphism (16%) of the markers. Although one marker (VSE02) showed evidence of null alleles, and one other marker showed deviation from HWE, these markers may perform well in the South American populations. The ten

polymorphic loci characterized here will be useful for studies of gene flow and population structure in this widespread, vertebrate dispersed, dioecious tree species.

Acknowledgements

This Appendix was coauthored with Christopher Dick, Andrew Lowe, and Michael Gardner and published in Applications in Plant Sciences in 2013. The authors thank C. Scotti-Saintagne and I. Scotti for contributing *V. sebifera* DNA samples and collection information. The work was supported by a Rackham Graduate Student Research Grant from the University of Michigan.

References

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011a) Data from: Rise of the machines-recommendations for ecologists when using next generation sequencing for microsatellite development. Dryad Digital Repository. doi:10.5061/dryad.f1cb2.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011b) Rise of the machines - recommendations for ecologists when using next generation sequencing for microsatellite development. *Molecular Ecology Resources* **11**, 1093-1101.

Howe HF (1981) Dispersal of a Neotropical nutmeg (*Virola sebifera*) by birds. *The Auk* **98**, 88-98.

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128-2129.

Meglecz E, Costedoat C, Dubut V*, et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404.

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

Rice WR (1989) Analyzing tables of statistical tests. *Evolution* **43**, 223-225.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Vetter D, Hansbauer MM, Vegvari Z, Storch I (2011) Predictors of forest fragmentation sensitivity in Neotropical vertebrates: a quantitative review. *Ecography* **34**, 1-8.

Wright SJ (2003) The myriad consequences of hunting for vertebrates and plants in tropical forests. *Perspectives in Plant Ecology Evolution and Systematics* **6**, 73-86.

Table A1 Characteristics of ten polymorphic SSR markers developed in *Virola sebifera*. T_a = annealing temperature.

Locus	Primer sequence (5′−3′)[§]	Motif	Size range	T_a (°C)	GenBank accession no.
VSE02	F: CGGTAGTCCATTGATTGGCA R: GCTGTCATTGTGCATCTTCCT	$(AG)_{12}$	266−296	55	JX415276
VSE11	F: TATAGATGCCTGCCATTGGA R: TCGTGCGAAATTCCCTTCTA	$(AG)_{10}$	237−267	55	JX415277
VSE30	F: CATGCATGCTGGTCCATA R: TTCAGCATATTCTCATGTTCCA	$(AGT)_{10}$	159−186	55	JX415278
VSE31	F: AACTAGGGCTCTCGCAGCTT R: CCAAAGAAGTGCTCCTCAGC	$(AAT)_{12}$	183−210	55	JX415279
VSE32	F: TGCCCAAGTGGGTTTCTCTA R: CCAGTGTTTCTTCTCTTGCATC	$(AAT)_{15}$	197−221	55	JX415280
VSE36	F: AGACGGATTGAGGAGAAGCC R: CGGAGCACAGGAATGAAATC	$(ACC)_{10}$	222−243	55	JX415281
VSE38	F: CCATTTGCTCTAAGCAATTCATC R: TCACATGCGAATTGTTCACAC	$(ACT)_{14}$	214−253	55	JX415282
VSE42	F: CACCGCTACTGTTTCCTGGT R: GTGGGATGTGCCATAGAAGC	$(AG)_3G(AG)_3G(AG)_{14}$	283−306	55	JX415283
VSE45	F: TGAAATTTGTTCCCTTCTGAGG R: TGATCCATTATTCAGATGAGGC	$(TCA)_5(TCGTCA)_{14}(TCA)_3$	132−163	55	JX415284
VSE55	F: GTTGGAGACTGTCCTCGGTG R: TGCTTAACAGCATGGAATGG	$(AGT)_9$	162−192	55	JX415285

[§] M13 tail (TGTAAAACGACGGCCAGT) added to the 5′ end of each forward primer.

Table A2 Summary statistics of SSR marker polymorphism screened in 42 *V. sebifera* individuals located in the 50-ha Forest Dynamics Plot on Barro Colorado Island, Panama.

Locus	***A***	H_O	H_E	**PE**	**PIC**
VSE02	15	0.487	0.876*	0.604	0.864
VSE11	12	0.810	0.834	0.511	0.816
VSE30	9	0.767	0.695	0.306	0.666
VSE31	10	0.762	0.764	0.384	0.734
VSE32	8	0.756	0.732	0.347	0.703
VSE36	7	0.465	0.477*	0.129	0.456
VSE38	12	0.905	0.827	0.491	0.806
VSE42	11	0.571	0.566	0.197	0.549
VSE45	10	0.561	0.544	0.178	0.524
VSE55	9	0.854	0.844	0.519	0.825
Mean	10.3	0.694	0.716	0.992¶	0.694

Note: A = number of alleles per locus; H_O = observed heterozygosity; H_E = expected heterozygosity; PE = probability of exclusion with one parent known; PIC = polymorphism information content.

* significant deviation from Hardy–Weinberg expectations after sequential Bonferroni correction ($P < 0.006$).

¶ probability of exclusion over all loci.

APPENDIX B

Characterization of twenty-six microsatellite markers for the tropical pioneer tree species *Cecropia insignis* Liebm. (Urticaceae)

Abstract

Cecropia insignis is an ecologically important Neotropical pioneer tree and major vertebrate food source. Although this species is relatively common in faunally intact tropical rainforests, its population dynamics may be negatively impacted by hunting of seed-dispersing animals. To better understand gene flow and regeneration dynamics in *C. insignis*, we characterized twenty-six microsatellite markers in a population sampled from Barro Colorado Island, Panama. Eleven loci of ≥3 alleles were tested on 48 individuals, whereas the remaining fifteen loci of 2 alleles were tested on 12 individuals. Allelic richness ranged from 2 to 9 per locus. Observed and expected heterozygosity averaged 0.478 and 0.440 respectively. Polymorphism information content was between 0.141 and 0.757. Only two loci exhibited deviation from Hardy-Weinberg proportions.

Keywords: *Cecropia insignis*, microsatellite markers, tropical tree, seed dispersal

Introduction

Cecropia insignis is a dioecious, gap-dependent canopy tree species distributed broadly in lowland moist forests of Central and northern South America (Croat 1978). It provides important food resources (e.g., leaves, nectar, fruits) for forest-dwelling animals. Although *Cecropia* trees represent one of the few primarily wind-pollinated taxa, its seed dispersal is mediated by vertebrates including large birds and monkeys. Hunting pressures are increasingly threatening the persistence of such seed-dispersing vertebrates, and have measurably altered tropical forest dynamics (Terborgh *et al.* 2008). We developed twenty-six polymorphic microsatellite markers for *C. insignis* to evaluate the impact of hunting and other anthropogenic changes on gene flow and regeneration in this species.

Methods

Methods used to obtain genomic data using circular consensus sequencing of Pacific Biosciences (PacBio) are described by Wei *et al.* (2014). Briefly, a PacBio 500-bp SMRTbell library was established from the genomic DNA of one *C. insignis* tree, and then sequenced using four SMRT cells with C2 chemistry. In total, 198989 circular consensus reads were generated. A quality-control step (for details, see Wei *et al.* 2014) was performed before searching for microsatellite loci and designing primers in QDD v2.1 (Meglécz *et al.* 2010). In total, 512 microsatellites loci were retrieved. From the pure (non-interrupted) microsatellite loci (n = 404), we synthesized 69 primer pairs (38 di- and 31 tri-nucleotide motifs).

For marker validation, we isolated genomic DNA from 48 reproductive-sized trees of *C. insignis* growing on Barro Colorado Island, Panama. We adjusted the use of DNeasy

Plant Mini Kit (QIAGEN, Valencia, California, USA) for high-throughput DNA isolation by replacing DNA binding columns with E-Z® 96 DNA Plates (Omega Bio-Tek, Norcross, Georgia, USA). After an initial screening of primer amplification on 3 individuals, polymorphic loci were tested on another 9 samples. Then microsatellite loci showing ≥3 alleles based on these12 samples were scored on an additional 36 individuals. PCRs were carried out as follows: 94°C for 4 min; 28 cycles of 94°C for 30 s, 59°C (decreasing 0.2°C per cycle) for 40 s and 72°C for 60 s; 10 cycles of 94°C for 30 s, 53°C for 40 s and 72°C for 60 s; and 72°C for 10 min. Each 8-μL PCR contained 1 μL of 4 ng/μL DNA, 0.05 μL of 1 μM HEX-labeled or 1.5 μM FAM-labeled M13 primer (TGTAAAACGACGGCCAGT), 0.12 μL of 5 μM M13-tagged forward primer, 0.48 μL of 5 μM reverse primer, 0.8 μL of 25 mM $MgCl_2$, 4 μL of GoTaq Colorless Master Mix (Promega, Madison, Wisconsin, USA), and 1.55 μL H_2O. PCR products of two loci labeled by different dyes were sized in a single lane on an ABI 3730 DNA Analyzer (Applied Biosystems, Carlsbad, California, USA). Alleles were then scored using GeneMarker v2.4.1 (SoftGenetics, State College, Pennsylvania, USA). Allelic richness, observed and expected heterozygosity, and Hardy-Weinberg equilibrium (HWE) were estimated using GenAlEx v6.5 (Peakall & Smouse 2012). Polymorphism information content (PIC) was assessed in PowerMarker v3.25 (Liu & Muse 2005).

Results

We described here only the 26 polymorphic microsatellite loci. For the eleven markers screened on 48 individuals (Table B1), allelic richness averaged 5 per locus (range 3–9). H_O ranged from 0.188 to 0.875; H_E varied between 0.205 and 0.785. PIC was

between 0.188 and 0.757 (mean = 0.499). All of these eleven loci conformed to HWE. For the fifteen markers showing two alleles and tested on 12 individuals (Table B1), observed and expected heterozygosity averaged 0.400 and 0.360 respectively. PIC was between 0.141 and 0.375. Two of the 15 loci (CEC_52 and CEC_65) deviated from Hardy-Weinberg expectations.

Acknowledgements

This Appendix was coauthored with Christopher Dick and published in Conservation Genetics Resources in 2014. The work was supported by a CTFS-ForestGEO grant from Smithsonian Tropical Research Institute and Center for Tropical Forest Science.

References

Croat TB (1978) *Flora of Barro Colorado Island* Stanford University Press, Stanford, California, USA.

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128-2129.

Meglécz E, Costedoat C, Dubut V*, et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404.

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537-2539.

Terborgh J, Nun ez-Iturri G, Pitman NCA*, et al.* (2008) Tree recruitment in an empty forest. *Ecology* **89**, 1757-1768.

Wei N, Bemmels JB, Dick CW (2014) The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. *Mol Ecol Resour* **14**, 953-965.

Table B1. Characteristics of 26 microsatellite markers developed in *Cecropia insignis*.

Locus	Primer sequence (5'-3')[§]	Motif	Size range	*A*	H_O	H_E	PIC	Sample size	Accession no.
CEC_08	F: CTGCAATTGACTTGCCACAC R: GGTGTGAAATGAAAGTGACCC	$(AAG)_{11}$	149–206	5	0.771	0.642	0.593	48	KF680367
CEC_10	F: ATTGCTCGTGCAACCAAAG R: TTGTGCCATGTTAATAGCCC	$(AAT)_8$	258–285	5	0.596	0.565	0.523	48	KF680369
CEC_12	F: TTCCAATCCGGAGATAAACG R: AAGCAAGAATCTCAAAGCCG	$(AAG)_{10}$	110–128	4	0.708	0.581	0.524	48	KF680371
CEC_17	F: TTCTTGATCGTGTTTGCTGC R: AAATGTTCAAGGCATTGGTTC	$(AAT)_7$	115–127	4	0.458	0.425	0.364	48	KF680376
CEC_37	F: CAAGAGATGCGTCGAGAGTG R: GGCAATCAATTTGCGTAACC	$(AG)_{16}$	151–157	4	0.479	0.545	0.466	48	KF680388
CEC_43	F: TTCGTGTATGAGGACAACGAG R: AATTCCACGAGGAAGCAGAG	$(AG)_{14}$	293–317	5	0.583	0.688	0.624	48	KF680393
CEC_45	F: TTTACCAAACCCAATTCCC R: ATTCTCAGCAAGTTCCCAGC	$(AG)_{13}$	118–152	9	0.875	0.785	0.757	48	KF680394
CEC_46	F: AGTACAACACCCGGATCGAC R: TCGAATATAACGCCTCTCGC	$(AG)_{13}$	112–136	8	0.604	0.528	0.503	48	KF680395
CEC_56	F: TGGCCTTCTTGAGTTGTTTG R: TCAGCCACTCTCACTCTTCG	$(AC)_{10}$	193–201	3	0.625	0.539	0.447	48	KF680402
CEC_61	F: TCCAAGTAACATCCTCTCCCTC R: TCCCTCAGAAAGCGAAGAAC	$(AG)_{10}$	115–121	3	0.188	0.205	0.188	48	KF680406
CEC_64	F: TTTGTCTTTGGCTTTGGACC R: CAACCTTTGCAAATTGGTCTAC	(AC)9	145–155	4	0.542	0.536	0.497	48	KF680408
CEC_15	F: ACCAGAGCCTTGAACAATCC R: TTCTTTGGACGAGAAATCGG	$(AAG)_7$	119–122	2	0.167	0.278	0.239	12	KF680374
CEC_22	F: CCGCATGGATAATTTCTCTTC R: ACATCGTTGCATGAGCTTTG	$(AAT)_8$	204–207	2	0.333	0.375	0.305	12	KF680381
CEC_31	F: GGGTGTATGCTCTCACACTTG	$(AAT)_7$	129–138	2	0.333	0.278	0.239	12	KF680386

	R: TCCATGATATGGTTTGGGTG								
CEC_34	F: TTAGGACTACTGCCTTCGCAC R: TATTGAGGCATGGAGGCTTG	$(AC)_{19}$	153–163	2	0.417	0.330	0.276	12	KF680387
CEC_38	F: TTACAGAGCATTGTGACCCG R: TGATGGAAGCTCTGAAGCAC	$(AG)_{15}$	159–161	2	0.500	0.486	0.368	12	KF680389
CEC_40	F: TTATGGGCAACTACGGCTTC R: CCATGTTCTAAACAATGTGTCC	$(AG)_{15}$	121–125	2	0.500	0.375	0.305	12	KF680390
CEC_41	F: TGAGCAAGCTGGAAAGGAAG R: TGCAAACCCAGCTATAAATGC	$(AG)_{15}$	156–166	2	0.583	0.413	0.328	12	KF680391
CEC_49	F: GAATTGCACATTGCCCTCTC R: CTCCGGTCTCTTCCTTCCC	$(AG)_{12}$	116–118	2	0.417	0.330	0.276	12	KF680397
CEC_52	F: ACCTTTGACCGTGGGATTC R: TGGTTGTCAAACTGTAAGGCAG	$(AC)_{10}$	126–132	2	1.000*	0.500	0.375	12	KF680398
CEC_53	F: GGCTGAGAGCTTTGGAGATG R: ACACTGTAGCAGAGCGGAGC	$(AG)_{10}$	142–150	2	0.250	0.330	0.276	12	KF680399
CEC_59	F: CCTCGGTGACCTTGAACTTG R: AAGAAACCCTTCAATCTCTGC	$(AG)_{10}$	154–156	2	0.167	0.153	0.141	12	KF680404
CEC_60	F: CTCAGCATAGATCTCGTTGCC R: TCTACTCAACAACCCGACCC	$(AG)_{10}$	184–186	2	0.250	0.413	0.328	12	KF680405
CEC_62	F: GTTTGGTGGGTTCACATGG R: CGATGTGTCACACTTGGGTC	$(AG)_{10}$	115–117	2	0.583	0.413	0.328	12	KF680407
CEC_65	F: TGAGGAATCTCCAAGGGAAG R: TCAGTGATTGGACTTCTGTTCC	$(AC)_{9}$	117–121	2	0.167*	0.444	0.346	12	KF680409
CEC_67	F: CTTGAAACCGGCTCCTGAAC R: TCGGGAATGGAAATAAATATGC	$(AG)_{9}$	157–163	2	0.333	0.278	0.239	12	KF680411

[§]M13 tail (TGTAAAACGACGGCCAGT) attached to the 5′ end of individual forward primers
A = number of alleles per locus; H_O = observed heterozygosity; H_E = expected heterozygosity; PIC = polymorphism information content. Significant deviation from Hardy-Weinberg proportions at $P < 0.05$ (*)

APPENDIX C

Polymorphic microsatellite markers for a wind-dispersed tropical tree species, *Triplaris cumingiana* (Polygonaceae)

Abstract

Novel microsatellite markers were characterized in the wind-dispersed and dioecious Neotropical tree *Triplaris cumingiana*, for use in understanding the ecological processes and genetic impacts of pollen- and seed-mediated gene flow in tropical forests. Sixty-two microsatellite primer pairs were screened, from which 12 markers showing ≥5 alleles per locus (range 5–17) were tested on 47 individuals. Observed and expected heterozygosity averaged 0.692 and 0.731 respectively. Polymorphism information content was between 0.417 and 0.874. Linkage disequilibrium was observed in one of the 66 pairwise comparisons between loci. Two loci showed deviation from Hardy-Weinberg expectations. An additional 14 markers exhibiting lower polymorphism were characterized on a smaller number of individuals. These microsatellite markers have high levels of polymorphism and reproducibility and will be useful in studying gene flow and population structure in *T. cumingiana*.

Keywords: gene flow; microsatellite marker; PacBio sequencing platform; single-molecule real-time sequencing; *Triplaris cumingiana*; wind dispersal

Introduction

Triplaris cumingiana Fisch. & C.A. Mey. ex C.A. Mey. (Polygonaceae) is a wind-dispersed, dioecious tree species found in humid forests of lower Central and western South America (Croat 1978). It forms an obligate mutualistic relationship with stinging ant *Pseudomyrmex triplarinus* (Croat 1978) as an anti-herbivore defense. Unlike most dioecious tree species that have inconspicuous unisexual flowers (Bawa & Opler 1975), flower sexual dimorphism is pronounced in *T. cumingiana*, which produces bright red bracts signaling flowers on female trees during the dry season of Panama. The dioecious mating system, which permits identification of the maternal and paternal contribution to seedlings, combined with the ease of sexual determination, make *T. cumingiana* of particular interest for studies of pollen- and seed-mediated gene flow in tropical tree species. To investigate the ecological and genetic impacts of pollen and seed dispersal in *T. cumingiana* as compared to tropical trees of alternative pollination and dispersal syndromes, we developed polymorphic microsatellite markers in *T. cumingiana.* We used single-molecule real-time sequencing (SMRT) implemented in the PacBio RS platform (Pacific Biosciences, Menlo Park, California, USA) because it is capable of generating long reads (Wei *et al.* 2014).

Methods and Results

Genome shotgun sequences were obtained using PacBio's high accuracy mode of circular consensus sequencing (Wei *et al.* 2014). In brief, genomic DNA from one reproductive-sized tree of *T. cumingiana* that grows in the 50-ha Forest Dynamics Plot (FDP) on Barro Colorado Island, Panama (9°10'N, 79°51'W; tag no. 199017), was used for

PacBio 500-bp SMRTbell library preparation. Following sonication, DNA fragments averaging 500 bp were ligated with two 55-nt hairpin adapters, and then sequenced on PacBio RS platform using C2 chemistry. Four SMRT cells generated a total of 178,122 circular consensus reads. Quality control was performed to remove homopolymer-rich sequences and poor-quality portions of individual reads (for details, see Wei *et al.* 2014). The resulting high-quality sequences were used for microsatellite searching and primer design in QDD v2.1 (Meglécz *et al.* 2010). In total, 795 microsatellite loci were obtained, in which 686 loci contained pure repeat motifs (524 di-, 143 tri-, 15 tetra-, 3 penta-, and 1 hexanucleotide repeat motifs). From these pure microsatellites, loci of ≥9 repeats with dinucleotide motifs and loci of ≥7 repeats with other repeat motifs were retained for marker validation. This test array was comprised of 62 microsatellite loci (32 di-, 25 tri-, 3 tetra-, 1 penta- and 1 hexanucleotide motifs).

To test these primers, we isolated genomic DNA using a modified CTAB method (Doyle & Doyle 1987) from lyophilized leaves of 47 *T. cumingiana* adult trees growing in the 50-ha FDP (voucher no. Pérez 1862; STRI herbarium, Panama). After the initial check of primer amplification on three individuals, we found 39 primer pairs generated easily interpretable allelic patterns, in which 11 loci were monomorphic. Then we tested the remaining 28 polymorphic markers on another 9 individuals, two loci of which showing apparent null alleles were excluded from further analyses. Microsatellite loci of ≥5 alleles based on these 12 samples were screened on an additional 35 individuals. The 8-μL PCR reactions contained 4 ng DNA, 6.25 nM HEX- or 9.40 nM FAM-labeled M13 primer (TGTAAAACGACGGCCAGT), 0.075 μM M13-tagged forward primer, 0.3 μM reverse primer, 4 mM $MgCl_2$, 4 μL GoTaq Colorless Master Mix (Promega Corporation, Madison,

Wisconsin, USA) including 200 μM of each dNTP and 1U *Taq* DNA polymerase, and H_2O. PCRs were carried out using two different thermocycling conditions. For most of the tested primers, we used a touchdown protocol (a; Table C1): 94°C for 4 min; 28 cycles of 94°C for 30 s, 59°C (decreasing 0.2°C per cycle) for 40 s, and 72°C for 60 s; 10 cycles of 94°C for 30 s, 53°C for 40 s, and 72°C for 60 s; and a final extension at 72°C for 10 min. When the above protocol produced weak PCR amplicons, we followed a non-touchdown protocol (b; Table C1): 94°C for 4 min; 28 cycles of 94°C for 30 s, 54.5°C for 40 s, and 72°C for 60 s; 10 cycles of 94°C for 30 s, 51.5°C for 40 s, and 72°C for 60 s; 72°C for 10 min. PCR amplicons were multiplexed by combining one HEX-labeled locus of 1.6 μL and one FAM-labeled locus of 1.4 μL, with 11.5 μL Hi-Di formamide (Life Technologies, Carlsbad, California, USA) and 0.05 μL GeneScan 500 Rox (Life Technologies), before loading to a single lane on an ABI 3730 DNA Analyzer (Life Technologies). Alleles were called using GeneMarker v2.4.1 (SoftGenetics, State College, Pennsylvania, USA).

We examined marker characteristics including allelic richness, observed and expected heterozygosity, and probability of exclusion (PE_2, when one parent known; PE_3, of a parent pair) using GenAlEx v6.5 (Peakall & Smouse 2012). Polymorphism information content (PIC) was estimated using PowerMarker v3.25 (Liu & Muse 2005). Exact tests of Hardy-Weinberg equilibrium (HWE) and locus pairwise linkage disequilibrium (LD) were conducted in GENEPOP v4.2.2 (Rousset 2008). The *P*-values of HWE and LD tests were adjusted for multiple comparisons using Holm's correction (Holm 1979).

We first focused on 12 markers with an average allelic richness of 9 (range 5–17) (Table C2). Observed heterozygosity ranged from 0.426 to 0.936 (mean = 0.692), and expected heterozygosity was between 0.475 and 0.885 (mean = 0.731). Locus PIC averaged

0.705, and overall exclusion probability was around 0.998 (with one parent known) and 1.000 (of a parent pair). We observed linkage disequilibrium between the loci TRI_27 and TRI_31 (Holm's adjusted $P = 0.038$). Two loci (TRI_26 and TRI_38) showed HWE deviation (Holm's adjusted $P < 0.007$). These 12 microsatellite markers should provide resolution for studying gene flow and genetic structure in *T. cumingiana*. In addition, we provide information on 14 less polymorphic loci (2–4 alleles per locus tested on 12 individuals) (Table C3), which can be potential candidate markers if more genetic information is required.

Conclusion

We characterized novel microsatellite markers in the dioecious insect-pollinated, wind-dispersed tropical tree *T. cumingiana* for understanding the processes of pollen- and seed-mediated gene flow in tropical forests. This will be done in parallel with studies of tree species with alternative pollination and dispersal syndromes. These markers can also be useful for studying the ecological responses of *T. cumingiana* (e.g. dispersal, recruitment), to rapid changes in temperature and rainfall patterns, as the distribution of this species is associated with high soil phosphorous and high dry-season intensity (Condit *et al.* 2013).

Acknowledgements

This Appendix was coauthored with Christopher Dick and published in Applications in Plant Sciences in 2014. The work was supported by a CTFS-ForestGEO grant from Smithsonian Tropical Research Institute and Center for Tropical Forest Science.

References

Bawa KS, Opler PA (1975) Dioecism in tropical forest trees. *Evolution* **29**, 167-179.

Condit R, Engelbrecht BMJ, Pino D, Perez R, Turner BL (2013) Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5064-5068.

Croat TB (1978) *Flora of Barro Colorado Island* Stanford University Press, Stanford, California, USA.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11-15.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70.

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**, 2128-2129.

Meglécz E, Costedoat C, Dubut V*, et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404.

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537-2539.

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Wei N, Bemmels JB, Dick CW (2014) The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. *Molecular Ecology Resources* **14**, 953-965.

Table C1. Characteristics of twelve polymorphic microsatellite markers developed in *Triplaris cumingiana*.

Locus	Primer sequences (5'–3')[§]	Repeat motif	Size (bp)	T_a (ºC)[¶]	GenBank accession no.
TRI_01	F: GGCTTTAATTCACCATTTAGCC R: TTGCATCCACACCTAGCAAC	$(AAT)_8$	337–418	b	KF680412
TRI_07	F: GCCTGACATGATCAAATCCTC R: TTTCAATTGTTGACGGGATG	$(ACAT)_8$	220–364	b	KF680415
TRI_09	F: GAAGTTGGCAGTCGAGGTTC R: CAAGCTCCAAACTCCCTCAG	$(AAAG)_8$	194–242	a	KF680417
TRI_20	F: ATTTGCCATCCGCTACTTG R: CTCATCATACGATGGCGTTC	$(AAG)_9$	196–217	a	KF680422
TRI_26	F: ATAGCCTCTAGCCCGACCTG R: GGGCTCTTCTGCTAGGGTTC	$(ACATAT)_7$	196–238	a	KF680426
TRI_27	F: TCCCTCAGACTGTCCAAAGC R: AGCCAATTGATTGGTTTCAAG	$(AAG)_{17}$	154–238	a	KF680427
TRI_31	F: GCAAATCATAATTGGGCTTACC R: CTGCCCTAAACGATCTCACC	$(AT)_9$	200–224	b	KF680430
TRI_38	F: TGGCTTGACTTGTCGATGTG R: CCACAATTTACAAACCACAAAG	$(AT)_{12}$	109–127	b	KF680432
TRI_40	F: TACACGGGAGCTTGATTTCC R: ATAAACCTAGGCACGGAGGC	$(AG)_{10}$	232–254	a	KF680433
TRI_45	F: TCATGAGGGAAGATGAGTTCG R: AAATAAATTGGGCACGATAGC	$(AG)_{26}$	106–122	a	KF680437
TRI_49	F: GTCGGCCTGCTTCTTTCTC R: TGCGACTTGTAACTGCAACG	$(AG)_{19}$	123–149	a	KF680440
TRI_55	F: AACCCTTGACGAGTCATTGC R: CAATTTGAAGCAAGCTGAGTG	$(AG)_{17}$	288–304	a	KF680444

Note: T_a = annealing temperature.

[§]M13 tail (TGTAAAACGACGGCCAGT) added to the 5′ end of each forward primer.

[¶]a, 59ºC (decreasing 0.2ºC/cycle) in a touchdown PCR; b, 54.5ºC in a non-touchdown PCR (see text for details).

Table C2. Summary statistics of microsatellite marker polymorphism tested on 47 reproductively mature trees of *Triplaris cumingiana*, growing in the 50-ha Forest Dynamics Plot on Barro Colorado Island, Panama.

Locus	***A***	H_O	H_E	PE_2	PE_3	**PIC**
TRI_01	13	0.915	0.877	0.605	0.909	0.866
TRI_07	17	0.800	0.847	0.554	0.890	0.835
TRI_09	8	0.766	0.740	0.366	0.759	0.717
TRI_20	7	0.830	0.775	0.398	0.769	0.747
TRI_26	8	0.511*	0.691	0.303	0.700	0.664
TRI_27	15	0.936	0.885	0.626	0.920	0.874
TRI_31	7	0.652	0.790	0.419	0.786	0.762
TRI_38	9	0.511*	0.825	0.484	0.834	0.804
TRI_40	8	0.660	0.655	0.269	0.670	0.631
TRI_45	6	0.489	0.481	0.127	0.466	0.454
TRI_49	5	0.809	0.736	0.317	0.666	0.688
TRI_55	5	0.426	0.475	0.116	0.384	0.417
Mean	9.0	0.692	0.731	0.998§	1.000§	0.705

Note: A = number of alleles per locus; H_O = observed heterozygosity; H_E = expected heterozygosity; PE_2 = probability of exclusion with one parent known; PE_3 = probability of exclusion of a parent pair; PIC = polymorphism information content.

* Significant deviation from Hardy-Weinberg proportions after Holm's correction (adjusted $P < 0.007$).

§ Cumulative probability of exclusion over multiallelic loci.

Table C3. Additional 14 polymorphic microsatellite markers of *Triplaris cumingiana* screened on 12 individuals sampled from the 50-ha Forest Dynamics Plot on Barro Colorado Island, Panama. PCRs follow a touchdown protocol (see text for details).

Locus	**Primer sequences (5'–3')**	**Repeat motif**	**Size (bp)**	***A***	H_O	H_E	**PIC**	**GenBank accession no.**
TRI_06	F: CCTTTCCAAACAAGGCTTACC R: GGTCTTGGATCAGCTGAAGG	$(ACT)_7$	272–281	2	0.083	0.080	0.077	KF680414
TRI_13	F: TGTGTATACCACAAAGCCGAAG R: TCTTCAATCGTTCTGCCTCC	$(AGG)_{15}$	115–118	2	0.667	0.444	0.346	KF680419
TRI_28	F: TCAAACGATACATTCCATTCTG R: TTGGAATGTTAGGATTGGCG	$(AAT)_{14}$	108–114	2	0.167	0.444	0.346	KF680428
TRI_30	F: AAAGGGAGGAGAAGAATGGTG R: TCTGCATGGTTGTCTCATAAAC	$(AAT)_{11}$	125–212	4	0.250	0.358	0.338	KF680429
TRI_36	F: GGAGTTGACTTGCATTTGGG R: TCATACCCAGTTAACCCATGC	$(AG)_9$	163–177	2	0.222	0.444	0.555	KF680431
TRI_44	F: TTTAGCCACAATTGCTCAAGAC R: AAAGATCGTCGTTCTCCCAC	$(AT)_{28}$	148–166	4	0.250	0.705	0.651	KF680436
TRI_51	F: CATGTACCAAACTGAACCTGTC R: CTCTTGACCGACCGACGAG	$(AC)_{20}$	147–153	3	0.455	0.368	0.425	KF680441
TRI_52	F: TTTCTTGGGTAATTAGTGAGGG R: TAATCCCTGTAGCGTAATCCC	$(AG)_{23}$	115–121	4	0.182	0.380	0.446	KF680442
TRI_54	F: GTTTGACCAAGGTTGACCAG R: GGGAAAGAACAAGAAGGAAGG	$(AG)_{26}$	121–123	2	0.083	0.080	0.077	KF680443
TRI_56	F: CTAATCGATTGAGGTTCGTGG R: TTGGCAGCAATCTAAGTCCC	$(AG)_{15}$	138–142	3	0.818	0.661	0.652	KF680445
TRI_57	F: CAGCTGCTATTGCTCTCAGC R: TATTTCCAACCAATCTCCCG	$(AG)_{14}$	147–155	3	0.833	0.517	0.420	KF680446
TRI_58	F: GAACATCCCAACAACATCCC R: TAGTGGTCGGCAAGCTAGTG	$(AG)_{14}$	266–268	2	0.636	0.483	0.471	KF680447
TRI_59	F: GGTGGATGTGGCAGTGTTAG R: GATCCGAAATTTGCCGTTAC	$(AG)_{13}$	190–192	2	0.667	0.444	0.346	KF680448

TRI_62	F: TAGCGACGGATAAGCTAGGG R: TTATTCTGCCATCACCGCTC	$(AG)_{11}$	175–187	3	0.091	0.169	0.281	KF680450

Chapter III

Seed dispersal drives spatial genetic patterns in tropical trees

Abstract

Seed dispersal is broadly recognized for ecological significance in species-rich tropical forests, but its genetic importance relative to pollen dispersal is less understood in many tropical trees. Efforts to assess the extent to which contemporary seed dispersal impacts spatial genetic structure (SGS) are constrained by challenges in retrieving seed and pollen dispersal processes from resultant SGS. Here we use approximate Bayesian computation (ABC) to evaluate the respective contributions of seed and pollen dispersal to the formation of SGS. Our study is focused on seedling banks of four tropical tree species with contrasting pollination and seed dispersal syndromes. Following prior expectations, variation in seedling SGS among species reflected the underlying differences in seed dispersal distance based on ABC inferences. Genetic affinity declined with logarithm-transformed distance three to four times faster in wind or mammal-dispersed trees than in avian-dispersed species, and the corresponding ABC-derived estimate of median seed dispersal distance was four to five times shorter. Pollen dispersal, however, could not be precisely inferred from seedling SGS, as the inference accuracy depends upon how large of an effect pollen dispersal has on SGS. Using simulations, we show that once pollination occurs beyond near neighbors, it has limited influence on seedling SGS, as a result of which its ABC inference suffers from low accuracy. Our study indicates that seed dispersal

is the primary force driving spatial genetic patterns in these tropical trees. This result has management implications for predicting how tropical trees will respond to the anthropogenic loss of seed dispersing animals.

Keywords: seed dispersal, pollen dispersal, tropical trees, spatial genetic structure, genetically marked point process, approximate Bayesian computation

Introduction

The ecological importance of seed dispersal has been broadly recognized in species-rich tropical rain forests as a key factor governing local and regional species diversity patterns (Webb & Peart 2001; Condit *et al.* 2002). However, the population genetic impacts of seed dispersal have received less attention than gene dispersal via pollen (Hamilton 1999; Hamilton & Miller 2002). Accumulated empirical evidence, often from wind-pollinated temperate forest trees, suggests that pollen dispersal has a greater impact on genetic connectivity at broad spatial scales (Ennos 1994; Petit *et al.* 2005). Yet, a significant deficiency exists in our understanding of the respective roles of seed and pollen dispersal in structuring genetic variation in tropical trees, which are primarily animal pollinated (Bawa *et al.* 1985) and dispersed (Howe & Smallwood 1982). In particular, in tropical regions growing intensities of anthropogenic disturbance (e.g. overhunting, habitat fragmentaion; Wright 2005) are raising alarm about potential cascading effects on the dispersal, recruitment and community composition of tropical trees due to the functional loss of vertebrate seed dispersers (Terborgh *et al.* 2008; Harrison *et al.* 2013). Determining the genetic consequences of this disturbance entails first quantifying the extent to which contemporary gene dispersal via seeds, relative to pollen, affects genetic structure in tropical tree species. Progress in this area is hindered, however, by the difficulties inherent in retrieving and separating seed and pollen dispersal processes from resulting spatial genetic patterns.

Seed dispersal gives rise to the spatial distribution of plants and their associated genotypes, half of which come from pollen gametes. Therefore, seed dispersal interacts with pollen movement to exert a composite effect on spatial genetic structure (SGS).

Distinguishing the respective contribution to SGS requires isolation of seed and pollen component from overall gene dispersal into seedling cohorts. Simultaneous inferences of seed and pollen dispersal can be realized, for instance, by assigning dispersed seedlings to respective maternal and paternal trees using genetic data (e.g. Hardesty *et al.* 2006). This approach holds promise for estimating spatially explicit movement of seeds and pollen, especially for primarily vertebrate-dispersed tropical trees. However, parental assignment does not provide information relating to how seed and pollen dispersal generate SGS. As an alternative to precise parentage identification, it is possible to infer the effective numbers of seed and pollen donors of dispersed seedlings (Grivet *et al.* 2009); yet this parental correlation-based method has not been explored to recover the underlying processes of seed and pollen dispersal.

The most pertinent approach to linking seed and pollen dispersal to their population genetic impacts is to extract the signature of these two ecological processes coded in resultant SGS. Spatial genetic information to date has been almost exclusively employed for historical gene flow inference from adult populations, at a spatial scale to which expectations of the infinite island model (Wright 1965) or isolation by distance (Slatkin 1991; Rousset 1997, 2000) at evolutionary equilibrium are assumed to hold. Separation of seed and pollen dispersal from genetic structure at broad geographic scales has been facilitated by uniparentally inherited genetic markers (Ennos 1994; Oddou-Muratorio *et al.* 2001; Petit *et al.* 2005). However, plastid markers often do not provide sufficient variation at finer scales (e.g. a forest stand) particularly in angiosperms. The potential to simultaneously reconstruct contemporary seed and pollen dispersal processes from SGS, in which theoretical population genetic models are unsupported, has not been explored (for

historical gene dispersal, see Heuertz *et al.* 2003), despite potentially strong signals of gene dispersal retained in seedling banks.

Here we develop a new analytical framework to infer and separate the impacts of seed and pollen dispersal on the SGS of seedling banks, using an approximate Bayesian computation (ABC) method. In this ABC-enabled framework, our spatially explicit individual-based simulation model follows population demographic characteristics observed in the field, but alters pollination environment and seed movement to examine how seedling spatial genetic patterns behave as a function of these two interacting processes. The best-fitted scenarios of seed and pollen dispersal provide information about their respective importance in determining SGS. We apply this new method to genetic data obtained for four dioecious tree species growing in a tropical moist forest in Panama. These species exhibit distinct strategies of life history (shade-tolerant or gap-dependent), pollination (wind or insect-pollinated) and seed dispersal syndrome (avian, mammal, or wind-dispersed). We hypothesized that species dispersed by highly mobile animals (e.g. large-sized birds and bats) should exhibit weaker seedling SGS than species dispersed primarily by mammals or by abiotic agents (e.g. wind). We first characterize seedling SGS using genetically marked point process, which enables the robustness of SGS characterization to arbitrarily defined distance intervals. Secondly, we assess empirically whether seed-mediated gene flow is the primary mechanism driving seedling SGS, by comparing spatial genetic affinity of seedlings to female trees and to male trees. Thirdly, we infer seed and pollen dispersal distance from seedling SGS using the ABC method. We then discuss the potential ecological and evolutionary implications.

Methods

Species and sampling

Four dioecious Neotropical tree species considered here are insect-pollinated and vertebrate-dispersed *Virola sebifera* (Myristicaceae) and *Tetragastris panamensis* (Burseraceae), wind-pollinated and vertebrate-dispersed *Cecropia insignis* (Urticaceae), and insect-pollinated and wind-dispersed *Triplaris cumingiana* (Polygonaceae). *Virola sebifera* is a shade-tolerant canopy tree, distributed broadly in mature tropical forests from Central America to the Amazon Basin and Guiana Shield (Croat 1978). The dominant agents dispersing nutrient-rich fruits of *V. sebifera* are big birds, such as toucans on Barro Colorado Island (BCI), Panama (Howe 1981). *Tetragastris panamensis* has a similar life form and geographic distribution as *V. sebifera*, but is more abundant locally and regionally. This species produces low-nutrient fruits that are primarily dispersed by monkeys (Howe 1980). *Cecropia insignis* is a wide-ranging pioneer canopy tree found in lowland moist forests of Central and northern South America (Croat 1978), the dominant dispersers of which are bats and birds (Brokaw 1986). Distinctively, *Triplaris cumingiana* is a midstory tree and its fruits are attached to bright red bracts that facilitate wind dispersal (Croat 1978). Taxa are henceforth referred to by the genus name.

Our study was carried out in the 50-ha (1000 × 500 m) Forest Dynamics Plot (FDP) on BCI (Condit 1998; Hubbell *et al.* 1999). A complete census of seedlings (height ≤ 10 cm) has not been incorporated into the plot protocol. Hence, we conducted the seedling census in a central 18-ha (600 × 300 m) subplot within FDP for *Virola*, *Cecropia* and *Triplaris* between 2012 and 2013. An exhaustive sampling was performed for *Virola* (n = 377) and *Cecropia* (n = 503) seedlings: we mapped individual seedlings and harvested a

small portion (1 cm^2) of a single leaf from each seedling. With *Triplaris*, we mapped and collected up to 30 seedlings from each 5 × 5 m quadrat of the 18-ha subplot (7200 quadrats being visited), totaling 369 seedlings in our sampling. *Tetragastris* seedlings were surveyed and collected in 2010 within a smaller central subplot of 2 ha (200 × 100 m), because of a markedly abundant seedling bank in this species. With *Tetragastris*, we first recorded the number of seedlings every 5 × 5 m quadrat in the 2-ha subplot (800 quadrats being visited), and then sampled *ca*. 15% (n = 269) of the seedlings based on an *in silico* randomization. Reproductive-sized trees (with dbh as low as 7.5 to 15 cm depending upon species) were sexed and collected for leaf tissues from 2010 to 2013 from the entire FDP. Sex information of *Cecropia*, *Triplaris* and partial *Virola* adult trees was kindly provided by M. Bruijning (Radboud University), and the remaining *Virola* and *Tetragastris* adult trees were surveyed by the authors. Reproductively mature trees were sexed on the basis of flower forms, or the evidence of fruits on the tree and/or seedling carpets under tree crowns during different flowering and fruiting seasons. Leaf tissues were freeze-dried prior to DNA isolation.

Microsatellite genotyping

Microsatellite markers have been developed for the four tree species using traditional (Kenfack & Dick 2009) and next-generation sequencing approaches (Wei *et al.* 2013; Wei *et al.* 2014). We extracted genomic DNA from 1518 seedlings and 789 adult trees of these species. A modified CTAB method (Doyle & Doyle 1987) was applied to *Tetragastris* and *Triplaris*. For *Virola* and *Cecropia*, we adjusted the use of DNeasy Plant Mini Kit (QIAGEN, Valencia, CA, USA) by replacing binding columns with E-Z® 96

DNA Plates (Omega Bio-Tek, Norcross, GA, USA). After quality and quantity check using NanoDrop 2000 (Thermo Scientific, Wilmington, DE, USA), DNA was standardized to 4 ng/μL and genotyped at a subset of polymorphic loci (n = 9–11 per species; Table 3S.1) from previous publications. In addition, we included three newly developed microsatellite loci for *Virola* (Table 3S.1). PCR reactions followed previous protocols (Kenfack & Dick 2009; Wei *et al.* 2013; Wei & Dick 2014a, b). Amplicons were sized in ABI 3730 DNA Analyzer (Life Technologies, Carlsbad, CA, USA). Alleles were then scored using GeneMarker v2.4.1 (SoftGenetics, State College, PA, USA). Approximately 10% of the total samples were duplicated to ensure the consistency of genotyping, and no mismatches were detected. Some *Virola* seedlings (n = 14) were excluded from our analyses due to poor DNA quality and the resulting PCR failures.

Genetically marked point process

Spatial genetic structure is often characterized using autocorrelation methods that test for the dependence of genetic relatedness on spatial patterns (Smouse & Peakall 1999; Vekemans & Hardy 2004; Smouse *et al.* 2008). In the context of autocorrelation, pairs of individuals are allocated into discrete distance intervals according to their physical distances, and mean pairwise genetic relatedness at a distance interval $(r, r + \Delta r]$, $F(r)$, is

$$F(r) = \frac{\sum_{p=1}^{n} I_p(r) g_p}{\sum_{p=1}^{n} I_p(r)}, \qquad (1)$$

where g_p is the genetic relatedness between a pair of individuals p, n is the total number of pairs and $I_p(r)$ is an indicator variable equal to 1 if the pair p resides in $(r, r + \Delta r]$ or 0 if

not. As the size of Δr influences $F(r)$ estimates, SGS characterization is sensitive to pre-defined distance intervals. Here we used genetically marked point process (see also Shimatani 2002; Shimatani & Takahashi 2003) to ameliorate this problem by fitting the underlying spatial distribution using smoothing kernels with an optimized bandwidth.

Spatial genetic structure can be viewed as a marked point process, in which individuals are the ground points and the genotypes that individuals possess are the marks (Shimatani 2002; Shimatani & Takahashi 2003). Spatial distribution of individuals, arising from underlying processes such as seed dispersal, is described by the probability of finding an individual at distance r from a focal individual, $f(r)$, as

$$f(r) = \sum_{p=1}^{n} w_p K_h(r - r_p), \tag{2}$$

where $K_h(u) = (1/h)K(u/h)$ is a smoothing kernel (e.g. a Gaussian kernel used here) with a bandwidth of h, and r_p is the distance between a pair of individual p and w_p is the edge-correction weight of r_p. We set h to $1.06\sigma n^{-1/5}$ (Silverman 1986), where σ is the standard deviation of the observed pairwise distances. In this context, $F(r)$ is the mean of the conditional probability distribution of the genetic marks, resulting from processes such as seed and pollen dispersal, given distance r:

$$F(r) = \frac{\sum_{p=1}^{n} w_p K_h\left(r - r_p\right) g_p}{\sum_{p=1}^{n} w_p K_h\left(r - r_p\right)}. \tag{3}$$

Equation (3) replaces the indicator variable $I_p(r)$ in equation (1) with a smoothing kernel (and an optimized bandwidth), which thereby considers and weights all pairs of individuals according to their distances to r. As w_p appears in both the numerator and denominator of

equation (3), we used $w_p = 1$ to estimate the (un-weighted) pairwise distances, for ease of comparison with previous literature.

Dependence of genetic marks on the spatial point process indicates the presence of non-random SGS. With $F(r)$ asymptotically normally distributed, confidence intervals of the null hypothesis of randomness in SGS can be analytically approximated:

$$\left[\mu_g - z_{1-\alpha/2}\sqrt{\frac{c_K\sigma_g^2}{nh\,f(r)}}, \mu_g + z_{1-\alpha/2}\sqrt{\frac{c_K\sigma_g^2}{nh\,f(r)}}\right], \quad (4)$$

where $c_K = \int K(u)^2 du$, μ_g and σ_g are the mean and standard deviation of pairwise genetic relatedness, and $z_{1-\alpha/2}$ is upper critical value of the standard normal distribution. This analytic CI agrees well with that derived from permutations (Fig. 3S.1), which are commonly used in autocorrelation methods. R scripts of SGS characterization using genetically marked point process are available in Appendix 3S.1 (Supporting Information).

Pairwise genetic relatedness g_p was estimated using kinship coefficient (Loiselle *et al.* 1995; Appendix 3S.2), which measures the genetic relatedness between a pair of individuals relative to the mean relatedness between individuals drawn at random from the sampled population (Hardy *et al.* 2006). Quantifying SGS involves two forms of summary statistics: SGS intensity and spatial extent. SGS intensity measures how fast genetic affinity changes with distance. We estimated SGS intensity statistic, b, using the slope of linear regression between conditional mean $F(r)$ in equation (3) and the midpoints of distance intervals after natural logarithmic transformation. The sign of b (positive or negative) indicates genetic affinity either ascending or decaying with distance; the absolute value of b reflects SGS intensity. In addition, the commonly used *Sp* statistic (Vekemans & Hardy

2004) was included for estimating SGS intensity, defined as $Sp = b_F/(F_1 - 1)$, where b_F is the slope of linear regression between pairwise genetic relatedness g_p and the natural logarithm of distance r_p, and F_1 is $F(r)$ at the first distance interval. SGS spatial extent or genetic aggregation scale (S_g) measures the distance until which genetic affinity is no longer stronger than expected under the null hypothesis; that is, the distance at which $F(r)$ first intersects the 95% CI in equation (4). The S_g statistic, however, needs to be used and interpreted with caution due to its strong dependence on sampling area (Vekemans & Hardy 2004), despite that its sensitivity to pre-defined distance intervals is amended using genetically marked point process (Appendix 3S.3).

Patterns of SGS were evaluated for seedlings and adult trees, where g_p was estimated against the reference allele frequencies in their respective population. Once significant genetic aggregation was discerned in seedlings, we assessed whether the non-random SGS results primarily from seed-mediated gene flow, rather than pollen-mediated gene flow (from pollen donors to seedlings). To do so, we used analysis of covariance (ANCOVA) to compare intensity statistic b between female tree–seedling and male tree–seedling SGS pattern, in which between-cohort g_p was calculated against the reference allele frequencies in the pooled population of seedlings and adult trees. A significantly higher estimate of b (in absolute value) in female tree–seedling SGS than in male tree–seedling SGS is regarded as the evidence of seed flow-mediated genetic aggregation in seedlings. Furthermore, the SGS pattern between female trees and seedlings was employed to distinguish the effect on SGS of localized seed dispersal from of long-distance clumped seed dispersal (sensu Russo & Augspurger 2004). In the latter case, seedlings still form spatial aggregation; however, disassociation occurs between spatial and genetic affinity of

seedlings to their mother trees, therefore unlikely resulting in strong SGS between female trees and seedlings. All the analyses were conducted in R v3.0.2 (R Development Core Team 2013).

Inferring seed and pollen dispersal using approximate Bayesian computation

Inferring seed and pollen dispersal from spatial genetic patterns using ABC involves: 1) building a mechanistic model that simulates how the two ecological processes generate seedling SGS; 2) sampling broadly from model parameter space, in our model including mating neighborhood size (see below) and seed dispersal distance; 3) retrieving summary statistics of simulated seedling SGS from each independent combination of parameter priors; 4) approximating posterior distributions of parameters based on the simulations in which the summary statistics of seedling SGS closely approximate the observed values.

We consider the simulation algorithm in an individual-based spatially explicit setting. Reproductive-sized trees are distributed in a continuous landscape of size L = 1000 × 500 m. The initial population configuration follows the observed population in the field, including the location, gender and genetic information of individual adult trees (in each species). Pollination events occur at random within the mating neighborhood, which is constituted by the nearest N_m potential mates, of each female tree. A female tree k produces N_{sk} offspring, following a Poisson probability function:

$$f(N_{sk}) = \frac{\lambda_k^{N_{sk}}}{N_{sk}!} e^{-\lambda_k}, \quad (5)$$

where intensity λ_k is proportional to the dbh of female tree k, $\lambda_k \propto (\mathrm{dbh}_k)^{\theta}$ (Appendix 3S.4),

and θ is the power parameter describing the relationship between female tree fecundity and dbh (Greene *et al.* 2004; Uriarte *et al.* 2005). An offspring is then being displaced *x* meters away in a random direction from the maternal tree, following a two-parameter gamma distribution (shape α and rate β). The maternal and paternal tree of each offspring in the simulation were registered for estimating pollen dispersal distance. For simplicity, this model simulates a closed environment, allowing no immigrating seed and pollen flow. Although alternative models were also tested, including seed and pollen dispersal following a lognormal and exponential kernel respectively with female and male fecundity both allometrically or non-allometrically modeled (Appendix 3S.5), we focused on this simpler case as it was favored over the others by ABC model selection (Appendix 3S.5).

We chose uniform prior distributions for parameters in the simulation model: mating neighborhood size, $1 \leq N_m \leq 60$, except in *Triplaris* ($1 \leq N_m \leq 20$) which has a small population size of male trees (<40) in FDP; power parameter (θ) relating female reproductive success to dbh, $0 \leq \theta \leq 10$; seed dispersal kernel parameters ($1 \leq \alpha \leq 5$ and $0.01 \leq \beta \leq 0.2$), with mean seed dispersal distance ($= \alpha/\beta$) ranging from 5 to 500 m and median distance (the 50th percentile of the gamma distribution) between 3 and 467 m. To sample the simulated seedlings, we applied the same sampling strategy as in the field collection (Appendix 3S.6). Spatial genetic structure of simulated seedlings was characterized using genetically marked point process based on four summary statistics: spatial aggregation statistic pcd_1 (i.e. pair correlation density or relative neighborhood density at the first distance interval, Condit *et al.* 2000; see Appendix 3S.7 for details), describing the ground spatial distribution of seedlings; SGS intensity statistic *b*, genetic

aggregation scale S_g and genetic relatedness at the first distance interval F_1, describing the distribution of genetic marks in seedlings. The use of S_g is justified in the ABC framework, as the sampling scheme was the same for the observed and simulated seedlings. We also tested alternative sets of summary statistics, but no significant differences in parameter estimation were observed (data not shown).

In each species (except *Tetragastris*), 400 000 independent simulations were run for the model using the ABC method implemented in R package 'EasyABC' (Jabot *et al.* 2013). We kept 400 simulations according to standard ABC rejection algorithm, in which summary statistics of the simulated seedlings were close to those of the observed seedlings within a given tolerance threshold (e.g. 0.1% here) using R package 'abc' (Csillery *et al.* 2012). Posterior estimates of median seed dispersal distance, mating neighborhood size, as well as median pollen dispersal distance that occurred in the simulated population, were derived from these 400 simulations. For *Tetragastris*, we ran 200 000 simulations for the purpose of reducing computation time due to a large number of seedlings generated in each simulation, and retained 400 simulations as above for parameter estimation. To assess whether seed and pollen dispersal could be reasonably inferred using this ABC method, we simulated seedling banks, each of 1000 replications, with different model parameters known *a priori*, and then checked whether posterior estimates can approximate the actual parameter values with low bias and low root mean square error (RMSE).

Aside from the attempt to estimate seed and pollen dispersal distance using ABC, we evaluated how seedling SGS responds to changes in seed and pollen dispersal. All else being equal, we altered gene dispersal processes: median seed dispersal distance from 12 to 197 m, based on a gamma kernel (fixed rate $\beta = 0.1$) with shape α increasing every 0.5

from 1.5 to 20; mating neighborhood size N_m from 2 to 80 (i.e. 2, 5, 7, 10,…,80). Collectively, for each of the 1280 combinations of seed and pollen dispersal parameters, 300 simulations were conducted to obtain the mean estimates of SGS summary statistics.

Results

SGS measures and empirical inference of the role of seed dispersal

Consistent with our expectation, strong genetic aggregation in seedling banks was observed in wind-dispersed *Triplaris* [b = -0.0273 ± 0.0026 (standard error by jackknifing loci); Sp = 0.0292 ± 0.0035] and primarily monkey-dispersed *Tetragastris* (b = -0.0209 ± 0.0059; Sp = 0.0230 ± 0.0067; Fig. 3.1a). SGS intensity (b statistic) was not different between seedlings of these two species (ANCOVA: $F_{1,36} = 0.776$, $P = 0.384$), but seedling spatial aggregation (pcd_1) in *Triplaris* was 3.09 times as high as in *Tetragastris* (Fig. 3S.2a). In contrast, avian-dispersed *Virola* and *Cecropia* exhibited three to four times lower magnitude of seedling genetic aggregation (*Virola*: b = -0.0079 ± 0.0005, Sp = 0.0086 ± 0.0006; *Cecropia*: b = -0.0076 ± 0.0008, Sp = 0.0105 ± 0.0011; Fig. 3.1a). *Cecropia* seedlings were 4.74 times more aggregated than *Virola* with respect to pcd_1 (Fig. 3S.2a), despite a comparable level of SGS intensity (b statistic, ANCOVA: $F_{1,56} = 0.057$, P = 0.813).

SGS intensity was significantly lowered in adult trees in *Virola* (b = -0.0028 ± 0.0012; relative to seedlings, ANCOVA: $F_{1,51} = 33.72$, $P < 0.001$), *Tetragastris* (b = -0.0012 ± 0.0009; ANCOVA: $F_{1,31} = 29.95$, $P < 0.001$) and *Cecropia* (b = -0.0046 ± 0.0010; ANCOVA: $F_{1,51} = 16.19$, $P < 0.001$; Fig. 3.1). But this decrease in genetic aggregation from seedlings to adult trees was not yet significant in *Triplaris* (b = -0.0244 ± 0.0041;

ANCOVA: $F_{1,51} = 0.54$, $P = 0.466$). Interspecific variation in SGS pattern of adult trees was inconsistent with that observed in seedling banks (Fig. 3.1).

Empirically inferring the contribution of seed dispersal to seedling SGS entails teasing apart the confounding effect of pollen-mediated gene flow from male trees to seedlings. Intensity statistic b in female tree–seedling SGS (denoted as b_{fs}) was significantly higher (in absolute value) than in male tree–seedling SGS (b_{ms}) in *Virola* (b_{fs} = -0.0078 ± 0.0016, b_{ms} = -0.0012 ± 0.0012; ANCOVA: $F_{1,56} = 267.7$, $P < 0.001$), *Tetragastris* (b_{fs} = -0.0190 ± 0.0028, b_{ms} = 0.0028 ± 0.0036; $F_{1,56} = 121.0$, $P < 0.001$) and *Cecropia* (b_{fs} = -0.0121 ± 0.0018, b_{ms} = -0.0053 ± 0.0017; $F_{1,56} = 57.9$, $P < 0.001$; Fig. 3.2). With *Triplaris*, b was not different between female tree–seedling (b_{fs} = -0.0249 ± 0.0024) and male tree–seedling comparison (b_{ms} = -0.0223 ± 0.0048; ANCOVA: $F_{1,56} = 1.54$, P = 0.220; Fig. 3.2). A similar pattern was detected when considering the spatial genetic affinity of seedlings to the subset of female and male trees residing in the same sampling area as seedlings (Fig. 3S.3). By comparing female tree–seedling and male tree–seedling SGS, we identified the predominant role of seed dispersal in governing seedling SGS, with an exception of *Triplaris*.

Additionally, genetic affinity of seedlings to female trees declined significantly with their natural logarithm-transformed distances (linear regression: *Virola*, $F_{1,28} = 499.2$, R^2 = 0.945, $P < 0.001$; *Tetragastris*, $F_{1,28} = 103.5$, $R^2 = 0.779$, $P < 0.001$; *Cecropia*, $F_{1,28}$ = 250.7, $R^2 = 0.896$, $P < 0.001$; *Triplaris*, $F_{1,28} = 472$, $R^2 = 0.942$, $P < 0.001$), suggesting long-distance clumped seed dispersal is unlikely the primary mechanism driving non-random SGS in seedling banks in our study system.

ABC-enabled inference of seed and pollen dispersal

Heterospecific variation in seedling SGS (Fig. 3.1a) reflected the underlying differences in seed dispersal distance among species (Fig. 3.3). Longer median seed dispersal distance was detected in *Virola* (posterior median = 62.2 m, 2.5–97.5% quantile range 53.4–78.8 m) and *Cecropia* (61.2 m, 41.0–114.2 m). Although the posterior probability in *Cecropia* of median seed dispersal distance was less concentrated than the other species (Fig. 3.3), it deviated substantially from the prior values towards longer distances. Median seed dispersal distance was approximately four to five times shorter in *Tetragastris* (13.1 m, 9.9–16.0 m) and *Triplaris* (12.3 m, 8.5–14.8 m). Mating neighborhood size, however, could not be precisely estimated from seedling SGS in *Tetragastris* and *Virola*, given that the posteriors hardly or only slightly deviated from the uniform priors (Fig. 3S.4); with *Cecropia* and *Triplaris*, in which male tree–seedling SGS was significant (Fig. 3.2), N_m approximated a median of 25 (quantile range 4–56) and of 7 (quantile range 1–17) respectively. A similar pattern of posterior-to-prior probability distribution was observed in median pollen dispersal distance (Fig. 3.4), reflecting the inherent correlation between these two parameters in the simulation model (Appendix 3S.8), by assuming random mating within the neighborhood of individual female trees. Inferred median pollen dispersal distance was above 100 m in *Virola* (168.0 m, quantile range 86.1–288.4 m), *Cecropia* (130.7 m, quantile range 45.2–207.6 m) and *Triplaris* (110.1 m, quantile range 62.5–174.8 m), and was 89.0 m in *Tetragastris* (quantile range 40.2–141.0 m). However, the broad posterior distribution (or the wide quantile range) of median pollen dispersal distance in these species, unlike that of median seed dispersal distance (Fig. 3.3), made pollen dispersal inference less accurate.

Our lack of confidence in inferring pollination neighborhood and pollen dispersal distance from seedling SGS could arise from two possible sources: low ABC approximation efficacy or limited impacts of pollen dispersal on determining SGS in a forest stand. The possibility of low estimation efficiency was ruled out because model parameters were reasonably recovered by the ABC method using simulated data (Table 3S.2). The estimate of median seed dispersal distance had high accuracy (i.e. small bias and low RMSE, Table 3S.2; Fig. 3S.5), despite a relatively larger positive bias and RMSE observed in the shape (α) and rate parameter (β) of seed dispersal kernel. An accurate estimate of mating neighborhood N_m was only achieved when N_m was small (Fig. 3S.6 and Table 3S.2); when N_m was larger than 10–15, by which it may have an negligible impact on seedling SGS, N_m cannot be accurately inferred (Fig. 3S.6). Indeed, we found that seedling SGS exhibited distinct responses to changes in seed vs. pollen dispersal distance (Figs. 3.5 and 3S.7), being nearly invariant to pollen dispersal, except when pollination was restricted (e.g. median pollen dispersal distance < 100 m or N_m < 10), in which it affected SGS in seedling banks due to the excess of full siblings.

Discussion

By quantifying spatial genetic structure using genetically marked point process and inferring the underlying processes, that is, seed and pollen dispersal, from resultant SGS patterns, we found consistent lines of evidence for a dominant role of contemporary seed dispersal in governing SGS in these tropical trees. The evidence is threefold: first, female tree–seedling SGS is, in general, substantially stronger than male tree–seedling SGS;

second, seed dispersal can be accurately inferred from seedling SGS, unlike pollen dispersal in these species; third, simulations show that seedling SGS responds strongly to seed dispersal, but only to pollen dispersal when it is restricted to a few near neighbors. This asymmetric role of seed and pollen dispersal limitation on SGS provides insights into predicting how spatial genetic variation responds to potential alterations in gene dispersal pathways, as a result of increasing anthropogenic disturbance in tropical rain forests.

Spatial genetic structure and seed dispersal syndrome and distance

Variation in ability to disperse seeds (Muller-Landau *et al.* 2008) is a potentially important niche dimension that may promote the ecological coexistence of trees in species-rich tropical forests (Nathan & Muller-Landau 2000). Heterospecific differences in seed dispersal distance translate into their varying levels of spatial structuring (Seidler & Plotkin 2006), with weaker spatial aggregation in animal-dispersed trees relative to species that are dispersed by abiotic means, especially explosive dispersal and gyration. Such an effect of seed dispersal is also postulated on SGS patterns (Vekemans & Hardy 2004). A comparison of SGS intensity among 47 plant species (Vekemans & Hardy 2004), involving both temperate and tropical herbs and trees, revealed a trend of reduced SGS intensity in animal-dispersed taxa than either wind or gravity-dispersed species, albeit the contrast was not statistically significant. Apart from large heterogeneities among taxa in life form and population demography, inspecting SGS at later life stages, such as adult trees, can add another layer of complexity. During seedling establishment into adulthood, thinning processes, often acting in a density-dependent manner (Terborgh 2012), may preferentially target aggregated individuals with higher than average genetic affinity, and thereby

gradually erase the genetic signature of dispersal processes. One implication is that the lack of differentiation among species in adult genetic aggregation does not necessarily imply comparable levels among them of gene dispersal via seeds and pollen. For instance, despite the similar levels of SGS intensity in adult trees of *Virola*, *Tetragastris* and *Cecropia*, we found that seed dispersal distance is substantially shorter in *Tetragastris* relative to *Virola* and *Cecropia* inferred from SGS in seedling banks. The SGS of adult trees is the combined outcome of seed and pollen dispersal and various post-dispersal processes (e.g. habitat filtering, competition, predation). Assessing deviation from genetic randomness in seedlings is anticipated to provide a more direct account of the effect of seed and pollen dispersal, by minimizing the effect of other confounding ecological mechanisms.

In this study, we found tree species that disperse seeds or seedlings at longer distances via avian agents (*Virola* and *Cecropia*; Fig. 3.3) have significantly lower magnitude of genetic aggregation in seedlings (Fig. 3.1a) than do species that have shorter seed dispersal distances (mammal-dispersed *Tetragastris* and wind-dispersed *Triplaris*). In *Cecropia*, the genetic pattern holds despite the high level of spatial aggregation in seedlings (Fig. 3S.2a), which is caused by its life strategy as a gap specialist rather than seed dispersal limitation. We recognize, however, that predicting SGS on the basis of seed dispersal syndrome may prove impractical in light of substantial variation in seed dispersal distance for species possessing the same dispersal mode (Clark *et al.* 2005). Nevertheless, our results suggest that seed dispersal distance itself bears a close relation with the degree of genetic aggregation in tropical trees, at the least during early life-history stages.

Inferring seed and pollen dispersal using spatial genetic information

Inferring seed and pollen dispersal from resulting SGS has long been exploited for investigating historical gene flow, but such inference is contingent upon specific population genetic models and the availability of genetic makers with uniparental inheritance (e.g. Ennos 1994; Petit *et al.* 2005). In the absence of uniparental markers, Heuertz *et al.* (2003) used a dynamic lattice model to simulate SGS of adult populations in the temperate ash tree *Fraxinus excelsior* for estimating seed and pollen dispersal. The study was unable to derive precise estimates of seed and pollen dispersal distance from SGS, partly because of the limited number of prior combinations (n = 56) of seed and pollen dispersal distance used in the simulation. In contrast, our model included several hundred thousand parameter combinations to gauge seed and pollen dispersal from the SGS in seedling banks.

Our approach can adequately infer the processes that impact SGS, such as seed dispersal in our study system. As simulation-enabled inferences are nevertheless constrained by the underlying assumptions, our method models a closed environment and thus cannot estimate long-distance seed and pollen dispersal events, which are hypothesized to be disproportionately important for tree populations at landscape scales (Kremer *et al.* 2012). Immigrating seed and pollen flow into the 50-ha FDP could lower the average genetic relatedness in seedling banks and elevate SGS intensity compared to a closed setting, because genetic relatedness F_{ij} used here is a relative estimate to the mean between individuals drawn at random. Therefore, we may underestimate local seed and pollen dispersal using a closed-system model. Although the death of parental trees within the plot could introduce another source of biases, this effect is likely to be minor because of the long life spans of trees and an inclusion of young seedlings. Despite a conservative estimate of local seed dispersal, our model achieves a relatively unbiased inference of species

variation in seed dispersal from SGS. The extent to which species differ in seed dispersal ability in our study is consistent with that detected using inverse model fitting of seed rain data by Muller-Landau *et al.* (2008), although the estimate of effective seed dispersal based on seedlings here is two to three-fold higher than primary seed shadow.

This ABC-based approach cannot accurately infer processes that have negligible effects on seedling SGS, such as pollen dispersal events that exceed near neighbors. In our study system especially in *Virola* and *Tetragastris*, pollen is likely disseminated in a broad manner, resulting in a large mating neighborhood and thus the low estimation accuracy of N_m and pollen dispersal distance from seedling SGS. But in *Triplaris*, strong spatial aggregation in adult trees (Fig. 3S.2b) may lead to localized pollination and a smaller mating neighborhood, likely as a result of density-dependent foraging behaviors of insect pollinators (Ghazoul 2005). Although our model considered only dioecious species here, it can be easily modified to accommodate monoecious and hermaphroditic mating systems by incorporating a selfing parameter. In the case of selfing or inbreeding by which pollination contributes to the formation of offspring SGS, our method is expected to be able to approximate its magnitude.

Using simulations, we demonstrate the disparate responses of seedling SGS to changes in seed vs. pollen dispersal distance. Seed dispersal exerts a strong and continuous effect on seedling SGS, whereas the effect of pollen dispersal is pronounced mainly when it is restricted. Beyond the local mating neighborhood, the extent that increased pollen dispersal distance lessens SGS intensity in seedling banks is too weak to be detectable at local scales; this underlines the asymmetric role of pollen and seed dispersal limitation.

Contemporary seed dispersal driving spatial genetic structure and its implications

Our results indicate that contemporary seed dispersal is the primary mechanism determining seedling genetic patterns at fine spatial scales in these species. Particularly, we found that even extensive pollen dispersal at a forest stand may not offset SGS in seedling banks established by seed dispersal. Our study system, involving diverse pollination and seed dispersal syndromes and life-history traits, may permit us to anticipate a prevalent effect of seed dispersal on governing SGS in tropical trees.

Understanding how contemporary seed and pollen dispersal affect spatial genetic variation in currently intact tropical rain forests, as in Barro Colorado Island, is critical for predicting potential genetic changes in trees residing in faunally depauperate forests. Compelling evidence exists demonstrating that interrupted seed dispersal processes accompany the decline in vertebrate dispersers due to hunting pressures (Wright *et al.* 2000; Harrison *et al.* 2013) and habitat fragmentation (Wright & Duber 2001) in tropical forests. Therefore, curtailed seed deposition, due to the functional loss of animal dispersers, could substantially magnify genetic aggregation. As a consequence, denser individuals with greater genetic relatedness than random are expected at the scale where many ecological processes (e.g. competition, predation) act. To the extent that genetic variation confers functional diversity among conspecifics to moderate niche overlapping in resource competition and/or susceptibility to herbivore predation and disease infection (reviewed in Vellend & Geber 2005; Hughes *et al.* 2008), elevated neighboring genetic similarities would predict a further reduction in individual fitness to lower population recruitment, than expected by assuming individual equivalency within species. Reduced seed dispersal could also influence pollen dispersal in the long term. If mature trees become more aggregated

over longer temporal scales with shortened seed dispersal distance, pollination, mediated by biotic agents, will tend to increase in short-distance fractions (Ghazoul 2005). This positive feedback between seed and pollen dispersal (Hardy *et al.* 2006) would invoke steady erosion in population genetic variation and the ability to respond to changing environments for tree species, whose effective seed dispersers are or will be deprived by human disturbance in tropical forests. Therefore, ensuring seed dispersal processes has immediate and long-term relevance for sustaining tree population demographic and evolutionary dynamics in tropical forests.

Acknowledgements

This chapter was coauthored with Matteo Detto and Christopher Dick, and is currently in revision after being submitted to journals. We thank Jordan Bemmels and Ashley Thomson for helpful discussions on the manuscript. We also thank the Smithsonian Tropical Research Institute and Center for Tropical Forest Science for facilitating fieldwork on Barro Colorado Island and providing a CTFS-ForestGEO grant to N.W. and C.W.D for the molecular laboratory work. N.W. was supported by a Barbour Scholarship from the University of Michigan.

References

Bawa KS, Bullock SH, Perry DR, Coville RE, Grayum MH (1985) Reproductive biology of tropical lowland rain forest trees. II. Pollination systems. *American Journal of Botany* **72**, 346-356.

Brokaw NL (1986) Seed dispersal, gap colonization, and the case of *Cecropia insignis*. In: *Frugivores and seed dispersal* (eds. Estrada A, Fleming T), pp. 323-331. Kluwer Academic Publishers, Dordrecht, Netherlands.

Clark CJ, Poulsen JR, Bolker BM, Connor EF, Parker VT (2005) Comparative seed shadows of bird-, monkey-, and wind-dispersed trees. *Ecology* **86**, 2684-2694.

Condit R (1998) *Tropical forest census plots* Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas, USA.

Condit R, Ashton PS, Baker P*, et al.* (2000) Spatial patterns in the distribution of tropical tree species. *Science* **288**, 1414-1418.

Condit R, Pitman N, Leigh EG*, et al.* (2002) Beta-diversity in tropical forest trees. *Science* **295**, 666-669.

Croat TB (1978) *Flora of Barro Colorado Island* Stanford University Press, Stanford, California, USA.

Csillery K, Francois O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* **3**, 475-479.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11-15.

Ennos RA (1994) Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72**, 250-259.

Ghazoul J (2005) Pollen and seed dispersal among dispersed plants. *Biological Reviews* **80**, 413-443.

Greene DF, Canham CD, Coates KD, Lepage PT (2004) An evaluation of alternative dispersal functions for trees. *Journal of Ecology* **92**, 758-766.

Grivet D, Robledo-Arnuncio JJ, Smouse PE, Sork VL (2009) Relative contribution of contemporary pollen and seed dispersal to the effective parental size of seedling population of California valley oak (*Quercus lobata*, Nee). *Molecular Ecology* **18**, 3967-3979.

Hamilton MB (1999) Tropical tree gene flow and seed dispersal. *Nature* **401**, 129-130.

Hamilton MB, Miller JR (2002) Comparing relative rates of pollen and seed gene flow in the island model using nuclear and organelle measures of population structure. *Genetics* **162**, 1897-1909.

Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters* **9**, 516-525.

Hardy OJ, Maggia L, Bandou E*, et al.* (2006) Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. *Molecular Ecology* **15**, 559-571.

Harrison RD, Tan S, Plotkin JB*, et al.* (2013) Consequences of defaunation for a tropical tree community. *Ecology Letters* **16**, 687-694.

Heuertz M, Vekemans X, Hausman J, Palada M, Hardy O (2003) Estimating seed vs. pollen dispersal from spatial genetic structure in the common ash. *Molecular Ecology* **12**, 2483-2495.

Howe HF (1980) Monkey dispersal and waste of a neotropical fruit. *Ecology* **61**, 944-959.
Howe HF (1981) Dispersal of a Neotropical nutmeg (*Virola sebifera*) by birds. *The Auk* **98**, 88-98.
Howe HF, Smallwood J (1982) Ecology of seed dispersal. *Annual Review of Ecology and Systematics* **13**, 201-228.
Hubbell SP, Foster RB, O'Brien ST*, et al.* (1999) Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* **283**, 554-557.
Hughes AR, Inouye BD, Johnson MTJ, Underwood N, Vellend M (2008) Ecological consequences of genetic diversity. *Ecology Letters* **11**, 609-623.
Jabot F, Faure T, Dumoulin N (2013) EasyABC: performing efficient approximate Bayesian computation sampling schemes using R. *Methods in Ecology and Evolution* **4**, 684-687.
Kenfack D, Dick CW (2009) Isolation and characterization of 15 polymorphic microsatellite loci in *Tetragastris panamensis* (Burseraceae), a widespread Neotropical forest tree. *Conservation Genetics Resources* **1**, 385-387.
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.
Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **82**, 1420-1425.
Muller-Landau HC, Wright SJ, Calderon O, Condit R, Hubbell SP (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology* **96**, 653-667.
Nathan R, Muller-Landau HC (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology & Evolution* **15**, 278-285.
Oddou-Muratorio S, Petit RJ, Le Guerroue B, Guesnet D, Demesure B (2001) Pollen- versus seed-mediated gene flow in a scattered forest tree species. *Evolution* **55**, 1123-1135.
Petit RJ, Duminil J, Fineschi S*, et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* **14**, 689-701.
R Development Core Team (2013) *R: A language and environment for statistical computing, Version 3.0.2* R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* **145**, 1219-1228.
Rousset F (2000) Genetic differentiation between individuals. *Journal of evolutionary biology* **13**, 58-62.
Russo SE, Augspurger CK (2004) Aggregated seed dispersal by spider monkeys limits recruitment to clumped patterns in *Virola calophylla*. *Ecology Letters* **7**, 1058-1067.
Seidler TG, Plotkin JB (2006) Seed dispersal and spatial pattern in tropical trees. *Plos Biology* **4**, 2132-2137.
Shimatani K (2002) Point processes for fine-scale spatial genetics and molecular ecology. *Biometrical Journal* **44**, 325-352.

Shimatani K, Takahashi M (2003) On methods of spatial analysis for genotyped individuals. *Heredity* **91**, 173-180.
Silverman BW (1986) *Density estimation* Chapman and Hall, London, UK.
Slatkin M (1991) F_{ST} in a hierarchical island model. *Genetics* **127**, 627.
Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**, 561-573.
Smouse PE, Peakall R, Gonzales E (2008) A heterogeneity test for fine-scale genetic structure. *Molecular Ecology* **17**, 3389-3400.
Terborgh J (2012) Enemies maintain hyperdiverse tropical forests. *American Naturalist* **179**, 303-314.
Terborgh J, Nunez-Iturri G, Pitman NC*, et al.* (2008) Tree recruitment in an empty forest. *Ecology* **89**, 1757-1768.
Uriarte M, Canham CD, Thompson J, Zimmerman JK, Brokaw N (2005) Seedling recruitment in a hurricane-driven tropical forest: light limitation, density-dependence and the spatial distribution of parent trees. *Journal of Ecology* **93**, 291-304.
Vekemans X, Hardy OJ (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**, 921-935.
Vellend M, Geber MA (2005) Connections between species diversity and genetic diversity. *Ecology Letters* **8**, 767-781.
Webb CO, Peart DR (2001) High seed dispersal rates in faunally intact tropical rain forest: theoretical and conservation implications. *Ecology Letters* **4**, 491-499.
Wei N, Bemmels JB, Dick CW (2014) The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio. *Molecular Ecology Resources* **14**, 953-965.
Wei N, Dick CW (2014a) Characterization of twenty-six microsatellite markers for the tropical pioneer tree species *Cecropia insignis* Liebm (Urticaceae). *Conservation Genetics Resources* **6**, 987-989.
Wei N, Dick CW (2014b) Polymorphic microsatellite markers for a wind-dispersed tropical tree species, *Triplaris cumingiana* (Polygonaceae). *Applications in Plant Sciences* **2**, 1400051.
Wei N, Dick CW, Lowe AJ, Gardner MG (2013) Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing. *Applications in Plant Sciences* **1**, 1200295.
Wright S (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* **19**, 395-420.
Wright SJ (2005) Tropical forests in a changing environment. *Trends in Ecology & Evolution* **20**, 553-560.
Wright SJ, Duber HC (2001) Poachers and forest fragmentation alter seed dispersal, seed survival, and seedling recruitment in the palm *Attalea butyraceae*, with implications for tropical tree diversity. *Biotropica* **33**, 583-595.
Wright SJ, Zeballos H, Dominguez I*, et al.* (2000) Poachers alter mammal abundance, seed dispersal, and seed predation in a neotropical forest. *Conservation Biology* **14**, 227-239.

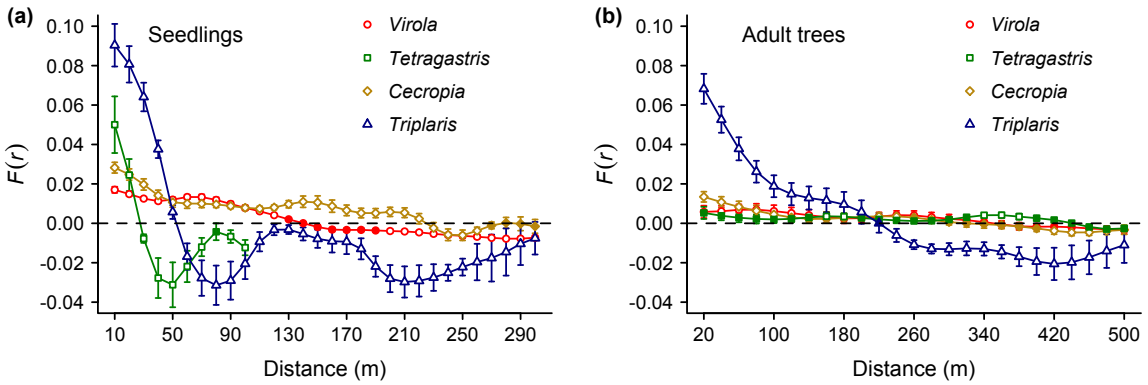

Figure 3.1 Spatial genetic structure in seedling banks (a) and adults trees (b) of the four tropical tree species. SGS is represented by the condition mean of genetic relatedness $F(r)$ over distance. Empty symbols indicate significant values falling outside the 95% confidence limits defined by the null expectation of randomness in SGS. Error bars (sometimes too small to be seen) represent one standard error obtained by jackknifing over loci.

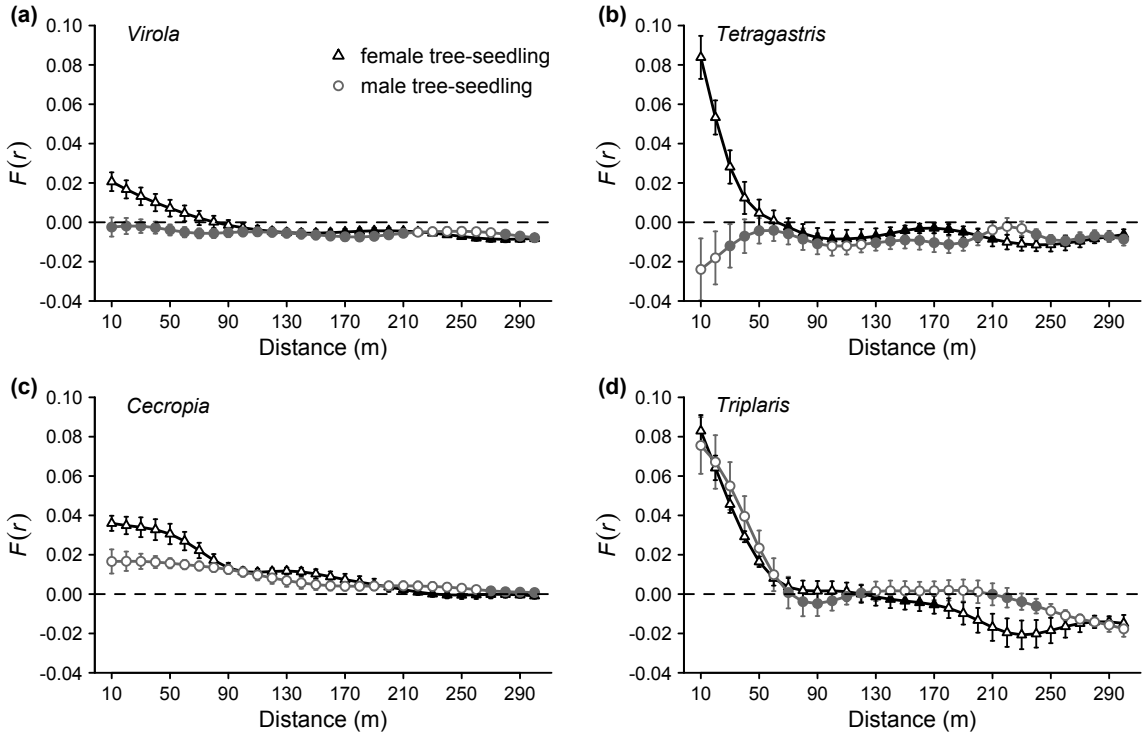

Figure 3.2 Spatial genetic structure of seedlings to female trees and to male trees. Empty symbols indicate significant values falling outside of the 95% confidence limits defined by the null expectation of randomness in between-cohort SGS. Error bars represent one standard error obtained by jackknifing over loci.

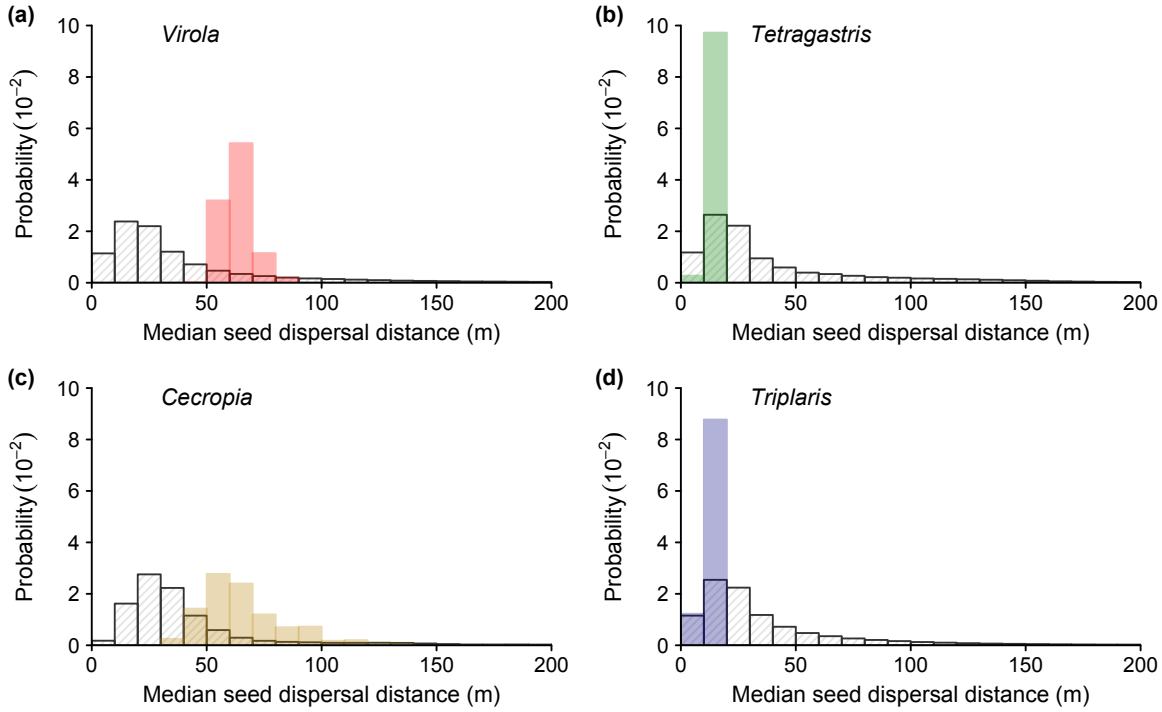

Figure 3.3 Posterior distribution of median seed dispersal distance inferred from the spatial genetic structure of seedlings using the approximate Bayesian computation method. Hatched bars represent the prior distributions of median seed dispersal distance. Solid bars represent the posterior median distance based on the retained simulations (n = 400) that closely approximated the observed seedling SGS in individual species.

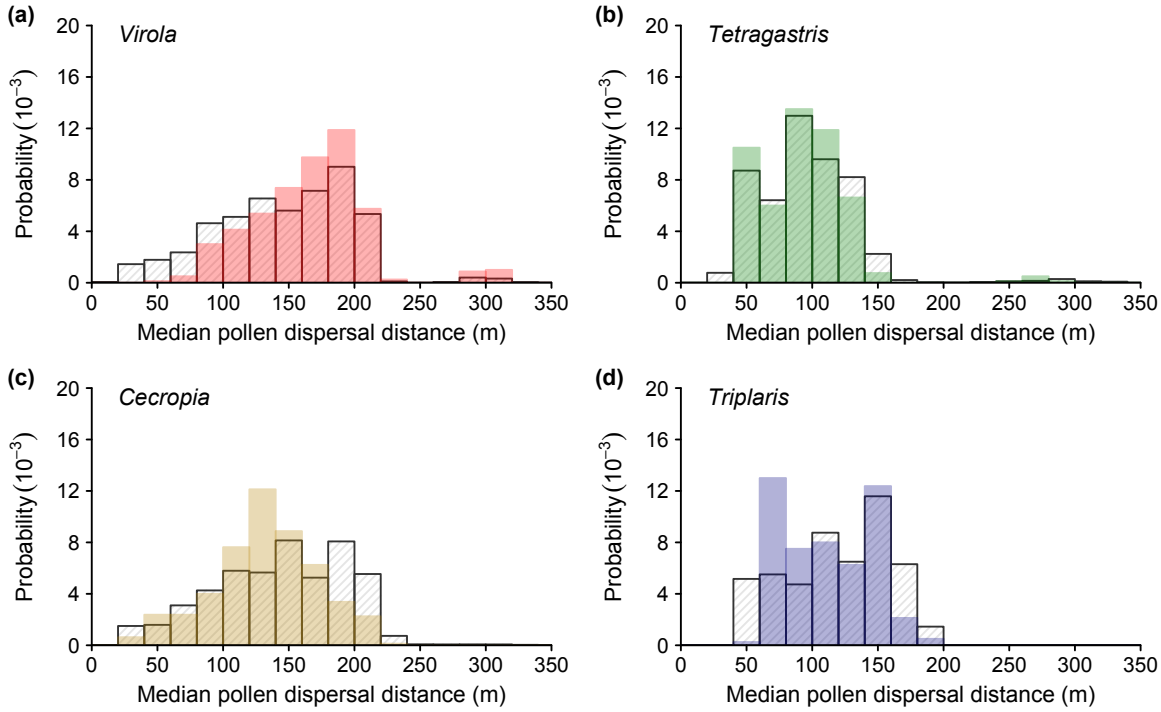

Figure 3.4 Posterior distribution of median pollen dispersal distance inferred from the spatial genetic structure of seedlings using the approximate Bayesian computation method. Hatched bars represent pollen dispersal (median distance) that occurred in the original ABC simulations (n = 400 000 or 200 000; see the main text). Solid bars represent the posterior median distance based on the retained simulations (n = 400) that closely approximated the observed seedling SGS in individual species.

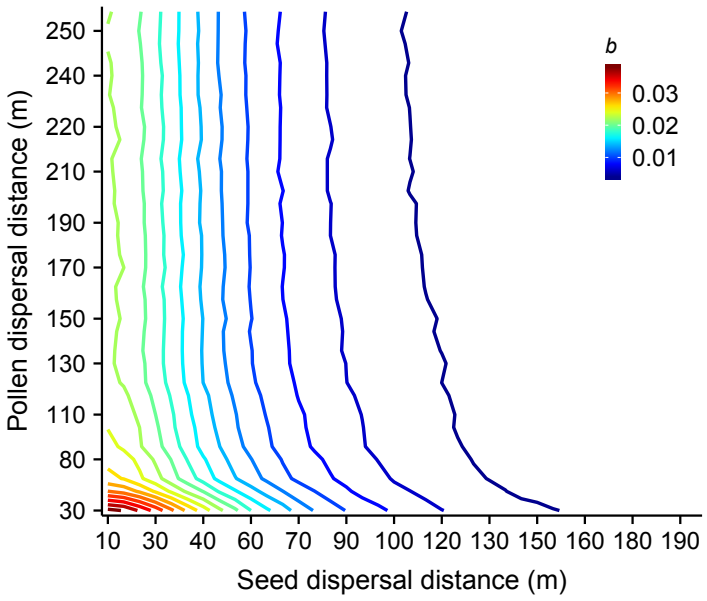

Figure 3.5 Spatial genetic structure of seedlings as a function of median seed and pollen dispersal distance. In this contour plot, isolines are drawn at a distance of 0.002, with respect to SGS intensity statistic *b*, from the bottom left (0.038) to the right (0.004). Seed dispersal was modeled using a gamma kernel (fixed rate parameter $\beta = 0.1$), with shape parameter (α) varying from 1.5 to 20. Female tree fecundity related parameter θ was set to 2. Mating neighborhood size N_m changed from 2 to 80. Initial model configuration used the population information from *Virola*, as an illustrative example.

Appendix 3S.1 R scripts of genetically marked point process.

```
#rp:    pairwise distances
#gp:    pairwise genetic relatedness
#wp:    edge-correction weights for rp
#r:     distance intervals, e.g. c(10, 20, 30,...)

FR<-function(rp, gp, wp=1, r)
{
        n=length(rp)            #total number of pairwise comparisons
        if (is.null(r))
        {
                dr=(max(rp)-min(rp))/99
                r=seq(min(rp), max(rp), dr)
        }

        #a global optimal bandwidth h for a Gaussian smoothing kernel
        sig1 = 1.06*sd(rp)
        if (sig1 <= 0) sig1 = max(rp) -min(rp)
        if (sig1 > 0)     h = sig1*(1/n)^(1/5) else h = 1

        #normalizing edge-correction weights for pairwise distances
        if (length(wp) == 1) wp=rep(1/n, n)
        if (length(wp) > 1) wp=wp/sum(wp)

        I=length(r)                     #number of distance intervals
        fx=rep(0,I)
        Fr=rep(0,I)                     #mean pairwise relatedness at each distance interval

        for (i in 1:I)
        {
                Kx = wp*exp(-1/2*((rp-r[i])/h)^2)
                fx[i] =  sum(Kx)
                Fr[i] = (Kx %*% gp)/fx[i]
        }

        fx=fx/sqrt(2*pi)/h
        ck=1/2/sqrt(pi); se2=var(gp); m4=mean((gp-mean(gp))^4)
        Fr.ci.low=mean(gp)-qnorm(.975)*sqrt(ck*se2/fx/n/h) #lower bound of 95% CI
        Fr.ci.up=mean(gp)+qnorm(.975)*sqrt(ck*se2/fx/n/h) #upper bound of 95% CI

        return(data.frame(r, Fr, Fr.ci.low, Fr.ci.up))

}
```

```
#using FR function
Fr=FR(rp, gp, r)
#Visualizing spatial autocorrelation of genetic relatedness
with(Fr, plot(r, Fr, type='b')); with(Fr, lines(r, Fr.ci.low)); with(Fr, lines(r, Fr.ci.up))
```

Appendix 3S.2 Genetic relatedness F_{ij}.

Pairwise genetic relatedness g_p was estimated using kinship coefficient F_{ij} (Loiselle *et al.* 1995):

$$F_{ij} = \frac{\sum_l \sum_a [(p_{ila} - p_{la})(p_{jla} - p_{la}) + p_{la}(1 - p_{la}) / (2n_l - 1)]}{\sum_l \sum_a p_{la}(1 - p_{la})},$$

where p_{la} is the average frequency of allele a at locus l in the reference population, p_{ila} and p_{jla} are the frequency of allele a at locus l in (a focal pair of) individual i and j respectively, and n_l is the number of individuals being genotyped that contained non-missing alleles at locus l.

Appendix 3S.3 Performance evaluation of genetically marked point process.

The sensitivity of SGS spatial extent (S_g) to an arbitrarily defined distance interval size was substantially lowered using genetically marked point process, compared to the conventional autocorrelation method using equation (1) (Fig. AS1). Although S_g inflated with an enlarged interval size using both methods, the rate of increase was significantly higher for the traditional means (ANCOVA: $F_{1,8} = 25.78$, $P < 0.001$; Fig. AS1). When applying the genetically marked point process, changes in S_g only became apparent when interval size was expanded to 30 m, at which few data points were available for estimating S_g (Fig. S1).

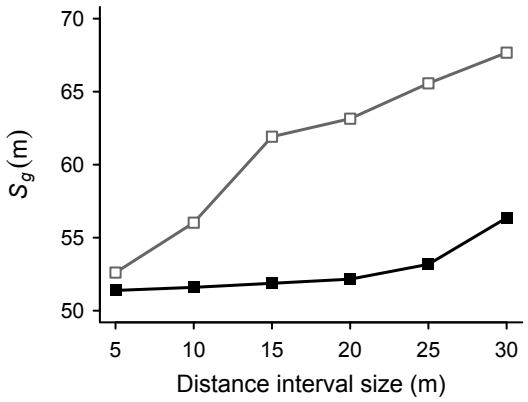

Figure AS1 SGS spatial extent (S_g) in response to distance interval size. Summary statistic S_g derived from genetically marked point process is represented by filled black symbols. Empty symbols are the same estimate based on the conventional means of autocorrelation using equation (1). We used *Triplaris* seedlings here as an illustrative example. With an interval size of Δr, the distance intervals used in SGS quantification, in our case, were from 10 to 300 m being (10, 10 + Δr], (10+ Δr, 10 + 2Δr], etc.

Appendix 3S.4 Modeling female fecundity as a function of dbh.

In our model, the number of offspring (N_{sk}) produced by female tree k follows a Poisson distribution with parameter λ_k. We related λ_k to the dbh of female tree k, $\lambda_k = \gamma(\mathrm{dbh}_k)^{\theta}$, where γ is a scaling factor and θ is the power parameter describing the relationship between female fecundity and dbh. The scaling factor γ is the weighted average fecundity:

$$\gamma = \frac{N_s}{\sum_{k=1}^{nf} \left(dbh_k\right)^{\theta}},$$

where nf is the total number of female trees, and N_s is the number of offspring (e.g. seedlings here) produced by all the female trees.

To calculate N_s, we first obtained the number of seedlings ($N_{s\text{-sub}}$) found in the subplot (A_{sub}) within the Forest Dynamics Plot (A = 50 ha). With *Virola* and *Cecropia*, the number of seedlings in the 18-ha subplot (A_{sub} = 18 ha) is the number of collected seedlings ($N_{s\text{-sampled}}$), because of an exhaustive sampling. For *Triplaris*, we assumed 90% seedlings were collected, and thus $N_{s\text{-sub}} = N_{s\text{-sampled}}/0.9$. For *Tetragastris*, *ca*. 15% seedlings were collected from the 2-ha subplot (A_{sub} = 2 ha), and thus $N_{s\text{-sub}} = N_{s\text{-sampled}}/0.15$. Then, assuming constant seedling density between on and off the subplot, N_s can be approximated as $N_s = (N_{s\text{-sub}}/A_{\text{sub}}) \times A$. But in *Triplaris*, population density differs between on and off the 18-ha subplot and population sex ratio is female biased, so we estimated $N_s = N_{s\text{-sub}}/nf_{\text{sub}} \times nf$, where nf_{sub} is the number of female trees in the subplot.

Appendix 3S.5 Alternative simulation models and model selection.

	Simulation model 1 (used in the main text)	**Simulation model 2**	**Simulation model 3**
Seed dispersal	Gamma kernel	Lognormal kernel	Lognormal kernel
Pollen dispersal	Mating neighborhood	Exponential kernel	Exponential kernel
Female fecundity	Allometric	Allometric	Non-allometric
Male fecundity	Random	Allometric	Non-allometric

Simulation model 2. Reproductive-sized trees are distributed in a continuous landscape of size L = 1000 × 500 m. The initial population configuration follows the observed population in the field, including the location, gender and genetic information of individual adult trees.

For a focal female tree k, the mating probability of a male tree j, P_{jk}, is determined by its distance to k, d_{jk}, according to an exponential kernel, $p(d_{jk})$, and its fecundity, $\lambda_j \propto (\mathrm{dbh}_j)^v$:

$$P_{jk} = \frac{p(d_{jk})\lambda_j}{\sum_j p(d_{jk})\lambda_j}.$$

Female tree k produces N_{sk} offspring, following a Poisson probability function, with intensity λ_k being proportional to the dbh of female tree k, $\lambda_k \propto (\mathrm{dbh}_k)^\theta$ (Appendix 3S.4).

An offspring is then being displaced x meters away in a random direction from the maternal tree, following a lognormal distribution (meanlog μ and sdlog σ).

Simulation model 3. Reproductive-sized trees are distributed in a continuous landscape of size L = 1000 × 500 m. The initial population configuration follows the observed

population in the field, including the location, gender and genetic information of individual adult trees.

For a focal female tree k, the mating probability of a male tree j, P_{jk}, is determined by its distance to k, d_{jk}, according to an exponential kernel, $p(d_{jk})$, and its fecundity, λ_j:

$$P_{jk} = \frac{p(d_{jk})\lambda_j}{\sum_j p(d_{jk})\lambda_j}.$$

Male tree fecundity λ_j follows a negative binomial probability function, with parameter *mu* being the average seedlings per male tree (= N_s/nm, where N_s is the total number of seedlings as in Appendix 4 and *nm* is the number of male trees) and the over-dispersion parameter *sm*.

Female tree k produces N_{sk} offspring, following a negative binomial probability function, with parameter *mu* being the average seedlings per female tree (= N_s/nf, where *nf* is the number of female trees) and the over-dispersion parameter *sf*.

An offspring is then being displaced x meters away in a random direction from the maternal tree, following a lognormal distribution (meanlog μ and sdlog σ).

Simulation implementation. In addition to simulation model 1, we ran model 2 and 3 using *Virola* and *Triplaris* as examples. Uniform parameter priors in model 2 included: power parameter θ relating female fecundity to dbh, $0 \le \theta \le 10$; power parameter ν relating male fecundity to dbh, $0 \le \nu \le 10$; lognormal seed dispersal kernel parameters ($\log(10) \le \mu \le \log(250)$ and $0.5 \le \sigma \le 2$); exponential pollen dispersal kernel parameter (rate *rp*, $0.001 \le rp \le 0.05$). In model 3, uniform priors included: over-dispersion parameter ($0.1 \le sf \le 5$

and $0.1 \leq sm \leq 5$) in negative binomial distribution describing female and male fecundity; lognormal seed dispersal kernel parameters ($\log(10) \leq \mu \leq \log(250)$ and $0.5 \leq \sigma \leq 2$); exponential pollen dispersal kernel parameter (rate rp, $0.001 \leq rp \leq 0.05$). We ran 400 000 independent simulations of each model for each species. Summary statistics (see the main text) were calculated from simulated seedlings and from observed seedlings.

Model selection. The posterior probability of each model was estimated based on the rejection method in 'abc' package (Csillery *et al.* 2012), which is approximated by the proportion of accepted simulations from each model given the threshold distance between the simulated and observed seedlings. We set the threshold to 0.1%. The Bayes factor, defined as the ratio of the posterior probability between two models was used for model selection.

	Triplaris	*Virola*
Proportion of accepted simulations		
model 1	0.818	0.639
model 2	0.174	0.059
model 3	0.008	0.302
Bayes factor		
model 1 vs. model 2	4.7	10.8
model 1 vs. model 3	109.1	2.1
model 2 vs. model 3	23.2	0.2

Appendix 3S.6 Sampling simulated seedling banks.

To sample the simulated seedlings, we applied the same sampling strategy as in the field collection; that is, a central 18-ha subplot was used in *Virola* and *Triplaris*, forest gaps in the 18-ha subplot for *Cecropia* seedlings and 2-ha subplot for *Tetragastris*. Forest gaps, from where *Cecropia* seedlings were collected, were reconstructed using R package 'spatstat' (Fig. AS2). Only offspring being dispersed to gap areas are assumed to survive in the simulation. For all the species except *Tetragastris*, an exhaustive seedling collection was conducted, if there were fewer simulated than observed seedlings; otherwise, we sampled maximally the observed number of seedlings at random for calculating the summary statistics. With *Tetragastris*, we randomly selected 269 simulated seedlings from the central 2-ha subplot.

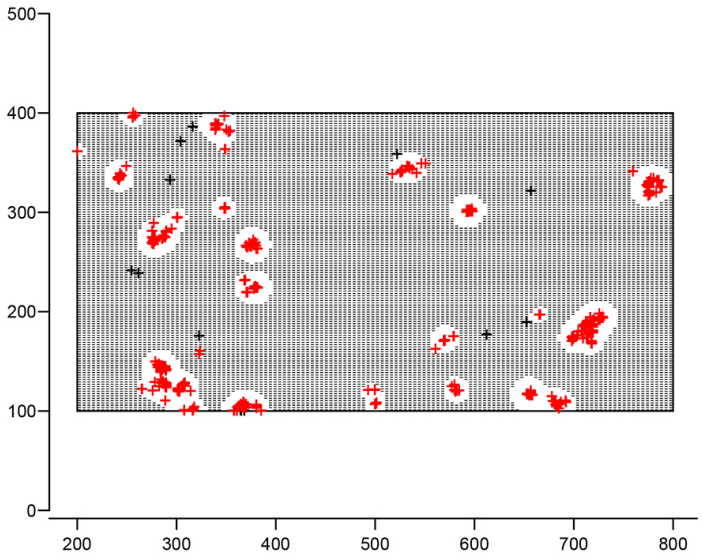

Figure AS2 *Cecropia* seedlings and the reconstructed forest gaps in a central 18-ha subplot of the 50-ha Forest Dynamics Plot on Barro Colorado Island, Panama. White regions are the reconstructed forest gaps. Seedlings are denoted by '+'. Seedlings (n = 485) in red are those within the built gaps, and the ones in black (n = 18) are not captured by our reconstructed gaps.

Appendix 3S.7 Spatial analysis using pair correlation density.

Spatial distribution in genetically marked point process can be quantified using pair correlation density pcd(r), which estimates relative neighborhood density as a function of distance r (Wiegand & Moloney 2004). In relation to equation (2), pcd(r) = f(r)/D^2, where D is the population density. Complete spatial randomness at r is expected when pcd(r) = 1, whereas clumping is indicated by pcd(r) > 1 and underdispersion by pcd(r) < 1. A homogeneous Poisson point process was simulated 199 times, and the fifth highest and lowest pcd(r) were used to approximate the 95% confidence limits of complete spatial randomness. Pair correlation density was estimated with a Gaussian smoothing kernel using package 'spatstat' in R v3.0.2 (R Development Core Team 2013). Pair correlation density or relative neighborhood density at the first distance interval (pcd_1) was used as a simple measure of spatial aggregation intensity, as pcd(r) at short distances (e.g. the first few distance intervals) are highly correlated (Condit *et al.* 2000). The first distance interval was set to (0, 10 m] in this study for measuring spatial aggregation in both observed and simulated seedlings.

Appendix 3S.8 Relationship between median pollen dispersal distance and N_m.

In our simulation model, pollen dispersal is assumed to occur at random within a mating neighborhood, which constitutes N_m potential male trees, for each female tree. Under the ABC framework, in each simulation N_m is chosen randomly between 1 and 60, except in *Triplaris* (between 1 and 20); median pollen dispersal distance that occurred in the simulation was recorded. Based on 400 000 independent simulations (or 200 000 simulations in *Tetragastris*), mating neighborhood size N_m is positively correlated with median pollen dispersal distance due to the assumption of random mating (Fig. AS3).

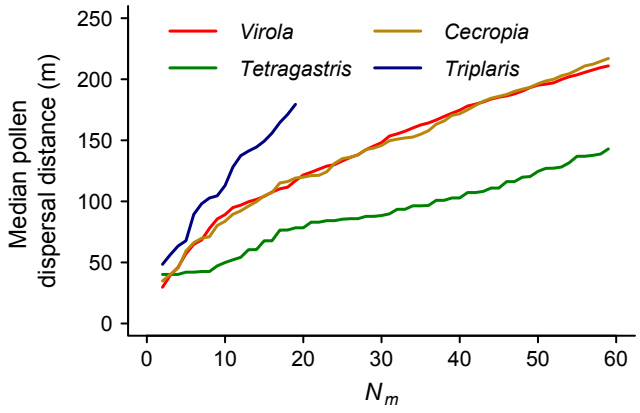

Figure AS3 Positive correlation between potential mating neighborhood size (N_m) and median pollen dispersal distance in the ABC framework.

Table 3S.1 Allelic richness (*A*) of microsatellite markers used in each species.

Virola[1]		*Tetragastris*[2]		*Cecropia*[3]		*Triplaris*[4]	
Locus	*A*	Locus	*A*	Locus	*A*	Locus	*A*
VSE11	16	Tpan014	6	CEC_08	6	TRI_01	17
VSE30	12	Tpan015	14	CEC_10	6	TRI_09	12
VSE32	12	Tpan152	8	CEC_12	8	TRI_20	8
VSE38	13	Tpan241	7	CEC_17	5	TRI_27	20
VSE45	12	Tpan301	5	CEC_37	7	TRI_31	11
VSE55	13	Tpan321	5	CEC_43	10	TRI_40	8
VSE59[¶]	11	Tpan441	10	CEC_45	12	TRI_45	8
VSE68[¶]	11	Tpan681	10	CEC_46	11	TRI_49	9
VSE76[¶]	7	Tpan882	2	CEC_56	5	TRI_55	6
		Tpan893	6	CEC_61	3		
				CEC_64	5		

[1]Wei *et al.* (2013); [2]Kenfack and Dick (2009); [3]Wei and Dick (2014a); [4]Wei and Dick (2014b)

[¶]Newly developed microsatellite markers for *Virola*:
VSE59 (F: GGGAACTTGAGAATAACCCACA; R: ACGTGGAAAGAAAGTGCGAA),
VSE68 (F: GGAAGCTGCAAGAAAGATGC; R: GCAGACCCTGTGATCCATGT),
VSE76 (F: TGGTTTGGTCATCTGCAACA; R: TCACCATCATGCATCTTTGC).

Table 3S.2 Relative bias (rBias) and relative root mean square error (rRMSE) of ABC parameter estimates. Errors were calculated based on 1000 replicates of the simulation model using *Triplaris* and *Virola* as examples with the following parameters known *a priori*: when considering the effect of variation in female fecundity (i.e. varying θ) on parameter estimation, we set shape α = 2, rate β = 0.1 and N_m = 10 (*Triplaris*) and 5 (*Virola*); when considering the effect of mating neighborhood (i.e. varying N_m), we set θ = 4 (*Triplaris*) or 2 (*Virola*), shape α = 2 and rate β = 0.1. D_S and D_P denote median seed and pollen dispersal distance respectively.

			$\widehat{D_S}$		$\widehat{D_P}$		$\hat{\theta}$		$\hat{\alpha}$		$\hat{\beta}$		$\widehat{N_m}$	
			rBias	rRMSE	rBias	rRMSE	rBias	rRMSE	rBias	rRMSE	rBias	rRMSE	rBias	rRMSE
Triplaris	θ	0	0.026	0.097	0.000	0.162	–	–	0.377	0.389	0.441	0.444	-0.029	0.184
		2	0.434	0.455	0.053	0.239	1.403	1.543	0.588	0.602	0.249	0.281	-0.017	0.343
		4	0.111	0.160	0.234	0.251	0.261	0.386	0.408	0.442	0.430	0.443	0.285	0.344
		6	0.020	0.108	0.155	0.194	0.048	0.225	0.312	0.367	0.450	0.463	0.137	0.222
		8	-0.011	0.084	0.092	0.153	-0.039	0.136	0.296	0.339	0.430	0.442	0.035	0.175
	N_m	5	0.114	0.164	0.418	0.574	0.330	0.446	0.179	0.256	0.331	0.354	0.540	0.880
		10	0.107	0.157	0.234	0.253	0.246	0.371	0.391	0.423	0.424	0.438	0.288	0.349
		15	0.044	0.114	-0.016	0.057	0.071	0.224	0.333	0.361	0.439	0.449	-0.044	0.128
		20	0.030	0.095	-0.147	0.152	-0.003	0.139	0.260	0.278	0.369	0.382	-0.186	0.200
Virola	θ	0	0.107	0.135	-0.007	0.158	–	–	0.395	0.405	0.368	0.374	0.023	0.337
		2	0.064	0.109	0.106	0.265	0.069	0.271	0.334	0.348	0.345	0.360	0.264	0.678
		4	0.057	0.111	0.242	0.471	0.039	0.147	0.299	0.336	0.297	0.336	0.642	1.327
		6	0.061	0.113	0.497	0.772	0.043	0.119	0.324	0.365	0.313	0.356	1.285	2.146
		8	0.024	0.082	0.933	1.174	0.012	0.072	0.305	0.330	0.333	0.359	2.043	2.808
	N_m	5	0.063	0.112	0.115	0.272	0.074	0.282	0.326	0.342	0.342	0.359	0.285	0.682
		10	0.027	0.087	0.368	0.486	0.048	0.287	0.301	0.339	0.331	0.370	1.091	1.429
		15	0.027	0.089	0.364	0.428	-0.002	0.283	0.284	0.332	0.317	0.368	0.886	1.044
		20	0.053	0.102	0.276	0.323	0.024	0.280	0.283	0.344	0.277	0.347	0.598	0.696
		25	0.041	0.097	0.203	0.235	-0.046	0.284	0.279	0.339	0.286	0.357	0.372	0.439

Note: when a parameter known *a priori* is set to zero (e.g. θ = 0), the relative bias and relative RMSE cannot be estimated (as the denominator is zero).

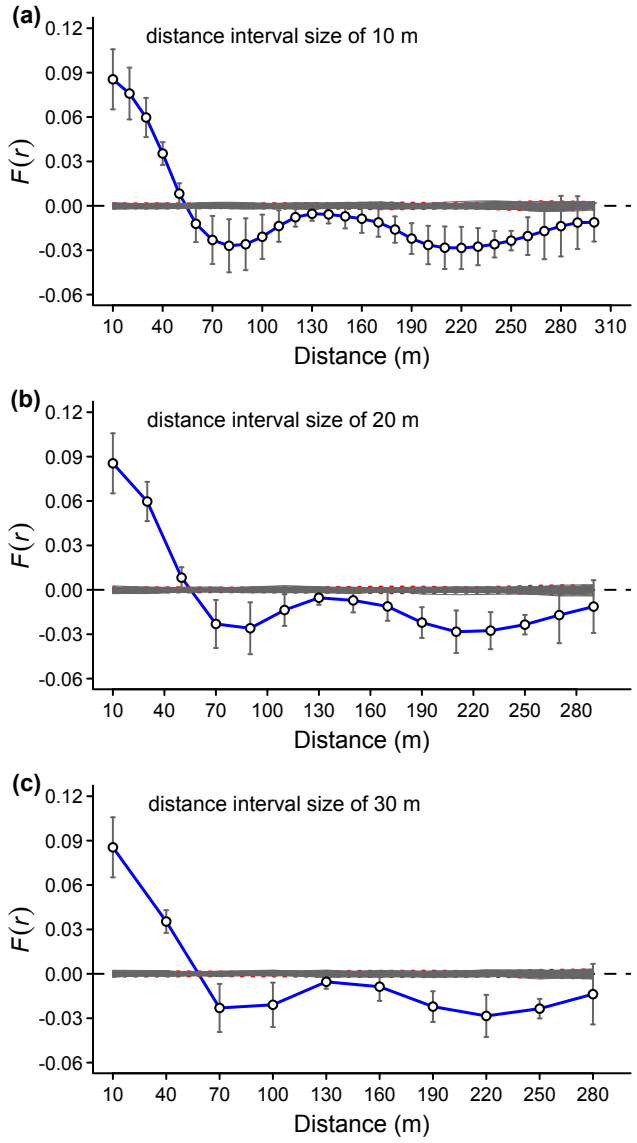

Figure 3S.1 Correlogram of spatial genetic structure based on genetically marked point process, using *Triplaris* seedlings as an illustrative example. Red dotted lines are the analytical 95% confidence limits of randomness in SGS; grey lines represent 100 permutations. Two standard errors by jackknifing over loci are indicated. Three different distance interval sizes were used, as (a) 10 m, (b) 20 m and (c) 30 m.

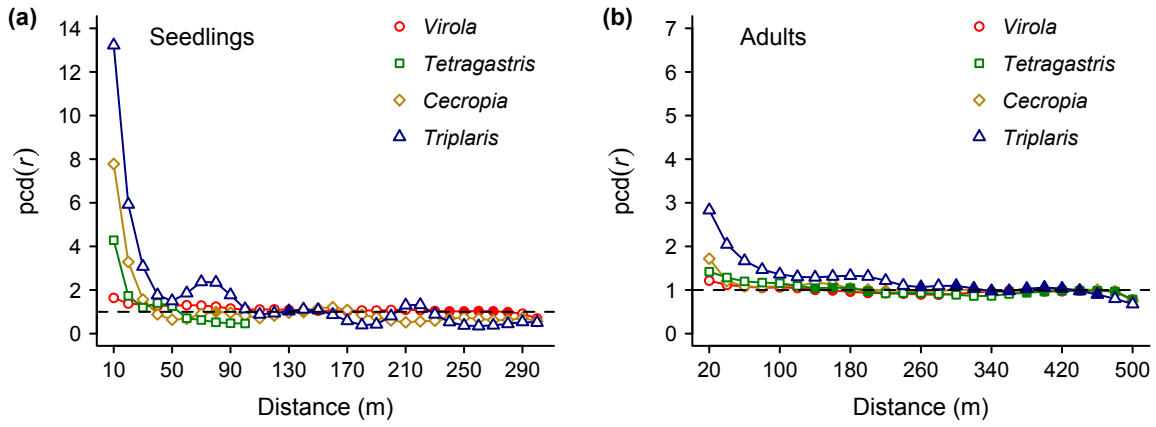

Figure 3S.2 Spatial structure of seedlings and adult trees measured by pair correlation density, pcd(*r*). Solid symbols indicate non-significant values within the 95% confidence limits of the null expectation of complete spatial randomness (see Appendix 3S.7).

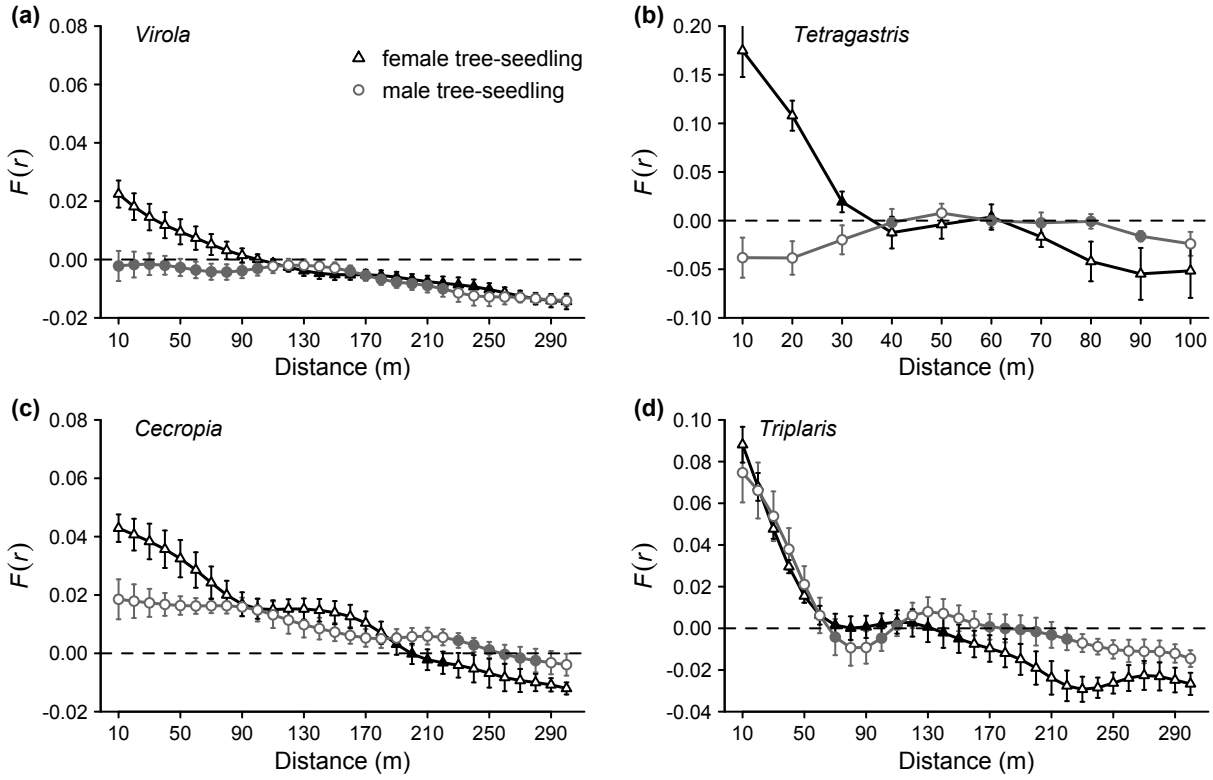

Figure 3S.3 Spatial genetic structure of seedlings to female trees (triangles) and of seedlings to male trees (circles). The female and male trees considered here were from the same subplot as the seedlings. Empty symbols indicate significant values outside the 95% confidence limits of the null expectation of randomness in between-cohort SGS. Error bars represent one standard error obtained by jackknifing over loci.

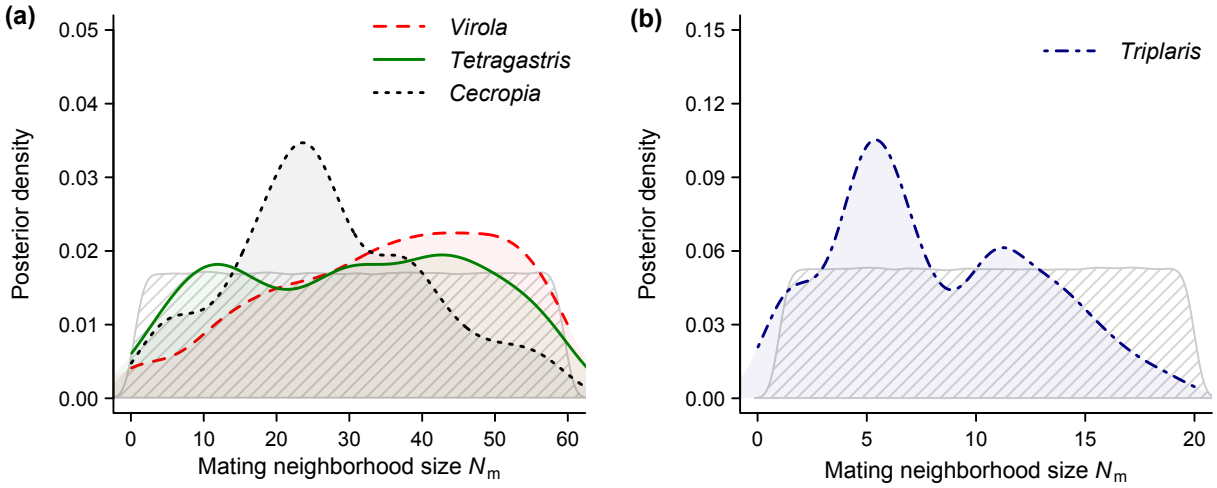

Figure 3S.4 Posterior distribution of mating neighborhood size (N_m) inferred from seedling spatial genetic structure using the approximate Bayesian computation method. Hatched regions represent prior uniform distributions.

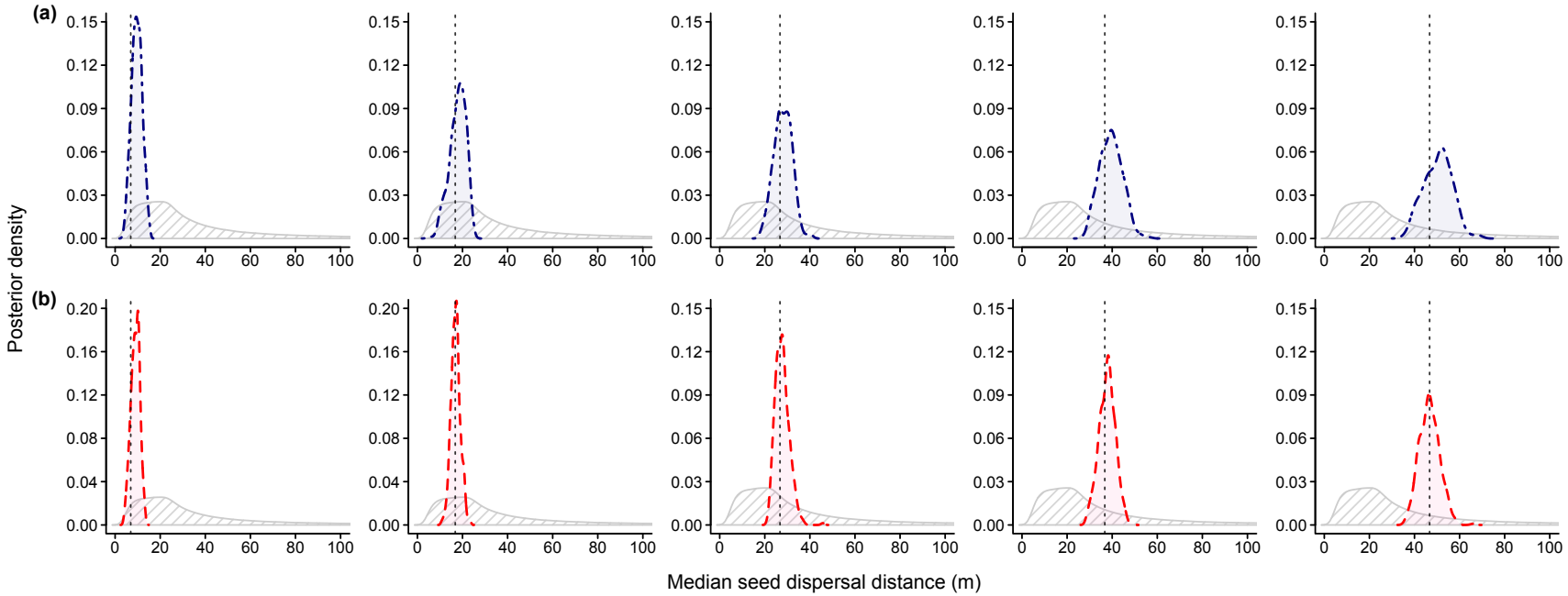

Figure 3S.5 Validating ABC-based inference of median seed dispersal distance using simulated data. *Triplaris* (panel a) and *Virola* population (panel b) were used as illustrative examples. Hatched regions represent parameter prior distributions. Dotted lines indicate the actual values of median seed dispersal distance used to simulate seedling banks, with the other model parameters fixed following $\theta = 4$ (*Triplaris*) or 2 (*Virola*) and $N_m = 10$ in both cases.

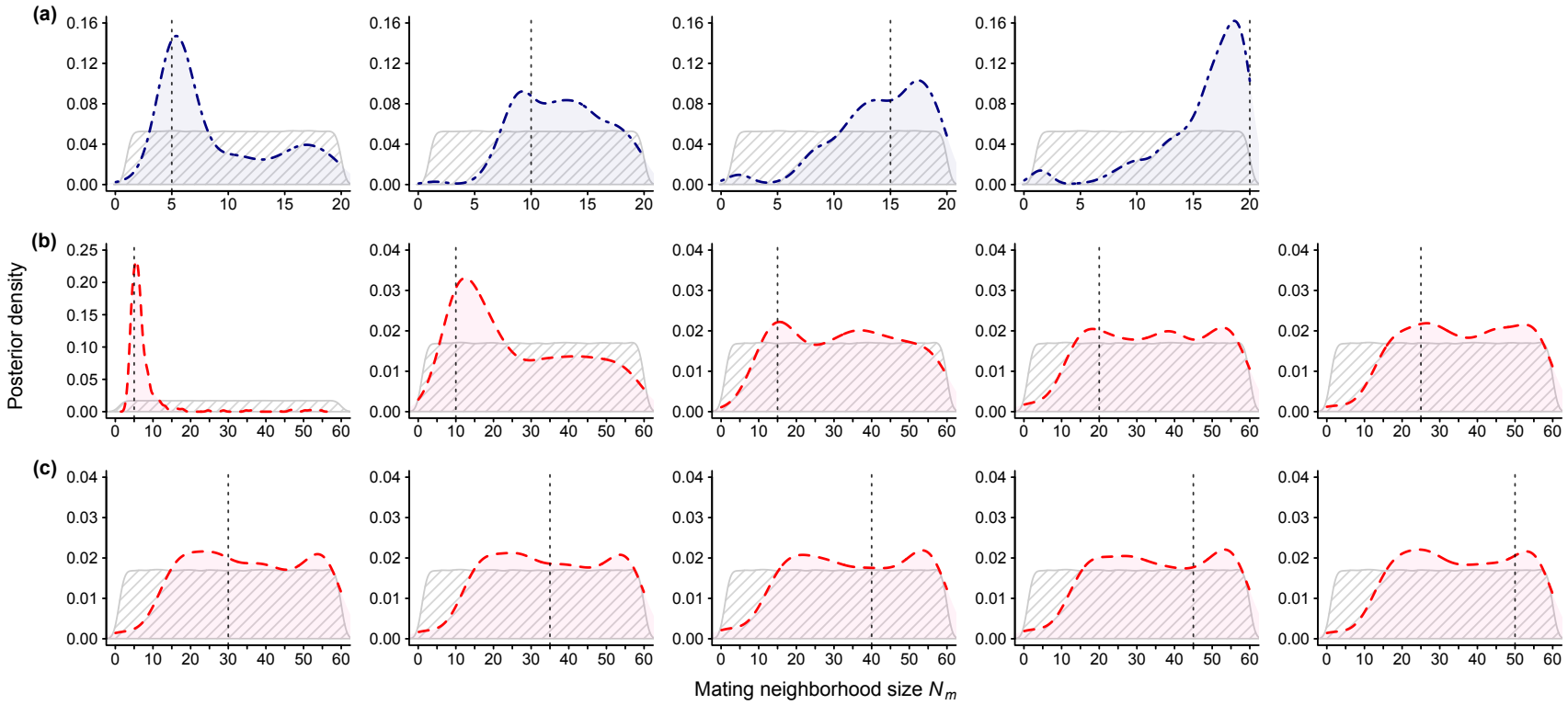

Figure 3S.6 Validating ABC-based inference of mating neighborhood size using simulated data. *Triplaris* (panel a) and *Virola* population (panel b and c) were used as illustrative examples. Hatched regions represent parameter prior distributions. Dotted lines indicate the actual values of N_m used to simulate seedling banks, with the other model parameters fixed following $\theta = 4$ (*Triplaris*) or 2 (*Virola*), shape $\alpha = 2$ and rate $\beta = 0.1$ (see also Table 3S.2).

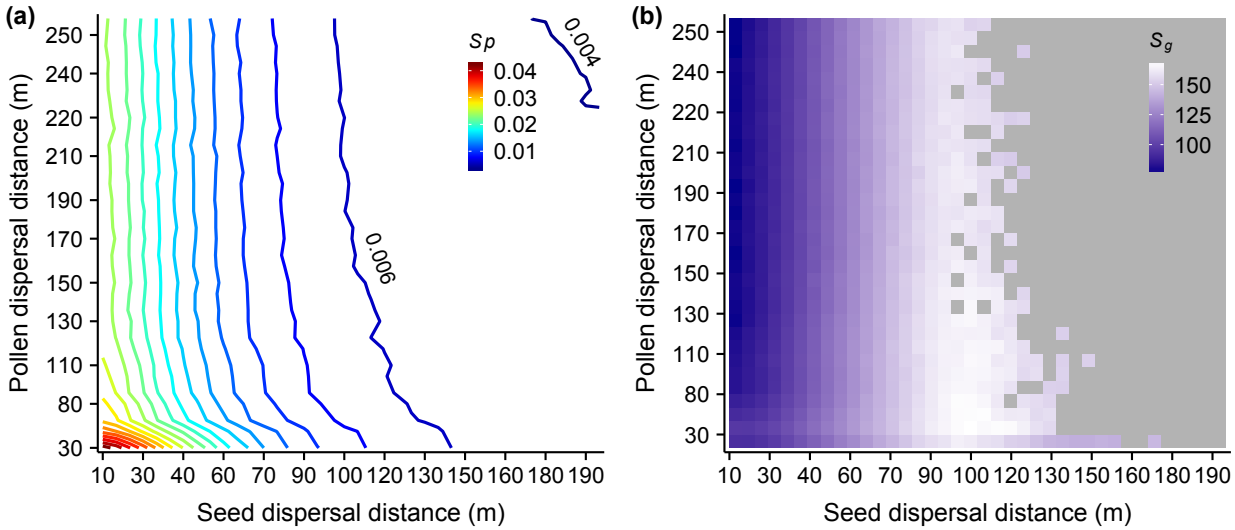

Figure 3S.7 Spatial genetic structure in seedlings as a function of median seed and pollen dispersal distance. In (a), contour isolines are drawn at a distance of 0.002, with respect to SGS intensity statistic *Sp*, from the bottom left (0.042) to the upper right (0.004). In (b), grey regions indicate that genetic aggregation scale (S_g) cannot be estimated due to the randomness of SGS, that is, $F(r)$ resided within the 95% confidence limits. Seed dispersal was modeled using a Gamma kernel (fixed rate parameter $\beta = 0.1$), with shape parameter α varying from 1.5 to 20. Female tree fecundity related parameter θ was set to 2. Mating neighborhood size N_m changed from 2 to 80. Initial model configuration used the population information

References

Condit R, Ashton PS, Baker P*, et al.* (2000) Spatial patterns in the distribution of tropical tree species. *Science* **288**, 1414-1418.

Csillery K, Francois O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* **3**, 475-479.

Kenfack D, Dick CW (2009) Isolation and characterization of 15 polymorphic microsatellite loci in *Tetragastris panamensis* (Burseraceae), a widespread Neotropical forest tree. *Conservation Genetics Resources* **1**, 385-387.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **82**, 1420-1425.

R Development Core Team (2013) *R: A language and environment for statistical computing, Version 3.0.2* R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Wei N, Dick CW (2014a) Characterization of twenty-six microsatellite markers for the tropical pioneer tree species *Cecropia insignis* Liebm (Urticaceae). *Conservation Genetics Resources* **6**, 987-989.

Wei N, Dick CW (2014b) Polymorphic microsatellite markers for a wind-dispersed tropical tree species, *Triplaris cumingiana* (Polygonaceae). *Applications in Plant Sciences* **2**, 1400051.

Wei N, Dick CW, Lowe AJ, Gardner MG (2013) Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing. *Applications in Plant Sciences* **1**, 1200295.

Wiegand T, Moloney KA (2004) Rings, circles, and null-models for point pattern analysis in ecology. *Oikos* **104**, 209-229.

CHAPTER IV

Frequent long-distance seed and pollen dispersal and their genetic impacts in tropical trees

Abstract

Long-distance gene dispersal by seeds and pollen has profound influences on the ecological and evolutionary dynamics of tree populations. Relative to the total diversity of tropical trees, we know in disproportionately few taxa how far seeds and pollen can move. Using inverse modeling integrated with unambiguous maternal and paternal inferences of established seedlings, we quantified seed and pollen dispersal in four Neotropical tree species that differ in dispersal and pollination syndromes. We found that seedlings were frequently established over 100 m from source trees, especially in avian-dispersed *Virola sebifera* and *Cecropia insignis*. Model-inferred median seed dispersal distance varied substantially from 25.4 m in wind-dispersed *Triplaris cumingiana* to 274.4 m in *V. sebifera*, whereas model-inferred pollen dispersal distances (median = 306–412 m) were similar among wind and insect pollinated species. Although our findings support the broad consensus of pollen-dominated gene dispersal in trees, we saw increased significance of gene dispersal by seeds in vertebrate-dispersed tropical trees. The best-fitted seed dispersal kernels, principally the lognormal, predicted the fraction of dispersal events >1 km as high as 17.7% in *V. sebifera* and *ca*. 1–2% in the others. This fraction of pollen dispersal >1 km was approximately 10–20% in these species. Our spatially explicit simulations examining

the genetic impacts of near vs. far tail of gene dispersal suggest that seed and pollen dispersal limitation would exacerbate inbreeding and genetic diversity loss due to drift, potentially constraining adaptive responses to changing environments.

Keywords: long-distance gene dispersal, seed dispersal, pollen dispersal, parentage inference, inverse modeling, dispersal kernel, genetic diversity

Introduction

Gene dispersal mediated by seeds and pollen influences the ecological and evolutionary dynamics of plants (Levin *et al.* 2003; Kremer *et al.* 2012). The spatial scale and frequency with which seeds arrive and establish factor profoundly on species abundance and forest community assembly (Hubbell 2001). This ability to disperse progenies, especially the long-distance events (Clark *et al.* 1999; Cain *et al.* 2000; Nathan *et al.* 2008), can help plants track optimal climatic niches during historical (Clark *et al.* 1998) and current climate change (Nathan *et al.* 2011; Corlett & Westcott 2013). Apart from migration responses, genetic variation maintained by seed and pollen-mediated gene dispersal may enable adaptation to changing abiotic and biotic environments *in situ*, as well as *ex situ* at the leading front of colonizing new habitats (Hamrick 2004; Aitken *et al.* 2008; Kremer *et al.* 2012). It is thus essential to know how far seeds and pollen can move for broad categories of plants. Our empirical knowledge is, nevertheless, lagging behind the need of this relevant information for an improved understanding of biodiversity maintenance at gene and species levels, especially in the context of human-induced environmental changes.

Seed-mediated gene dispersal is realized through the displacement of progenies from seed trees; pollen-mediated gene dispersal occurs through two stages–pollen dispersal from pollen donors to maternal trees and then through seed dispersal to progenies. Mortality at establishment and other stages of progeny development highlights the distinction between primary and effective seed and pollen dispersal, based respectively upon seeds and established seedlings (or even later life stages). Various approaches to measuring primary seed and pollen dispersal have been applied to diverse plant taxa. These disparate yet often complementary approaches fall into three broad categories: (1)

prospective monitoring (Williams 2010; Jansen *et al.* 2012), (2) mechanistic models (Nathan *et al.* 2002; Klein *et al.* 2003; Cortes & Uriarte 2013), and (3) retrospective reconstruction involving non-genetic (Clark *et al.* 1999; Muller-Landau *et al.* 2008) and genetic approaches (Austerlitz *et al.* 2004; Grivet *et al.* 2005). For example, prospective tracking of airborne pollen clouds permits detection of pollen movements over kilometers (reviewed in Kremer *et al.* 2012), although the likelihood of contributing to actual seed and seedling production remains largely uncertain. Mechanistic models of animal-mediated seed dispersal take into account foraging behaviors of the primary disperser, observed in the field using prospective monitoring methods, for predicting the probability of seed deposition at certain distances (Westcott *et al.* 2005; Russo *et al.* 2006). This type of methods fits satisfactorily plant species with a dominant seed disperser or a small disperser assemblage. Retrospective reconstruction examines the outcomes for causal processes, including non-genetic methods such as inverse models (Muller-Landau *et al.* 2008), and genetic methods such as maternity and paternity inferences (reviewed in Ashley 2010).

These approaches have shown varying degrees of success in quantifying effective seed and pollen dispersal. Mechanistic models of seed dispersal, for example, requires information on not only initial seed position but also site-specific survival, which may involve complex processes, including distance and density-dependent recruitment (Janzen 1970; Connell 1971). Inverse models employed in seedlings (LePage *et al.* 2000; Greene *et al.* 2004; Uriarte *et al.* 2005) estimate primary seed dispersal and ecological correlates of post-dispersal establishment for elucidating recruitment patterns (i.e. effective seed dispersal). The genetic retrospective methods (e.g. parentage inference) can be directly used to locate parental trees of recruited seedlings as the net outcome of primary gene dispersal

and post-dispersal processes (Hardesty *et al.* 2006; Chybicki & Burczyk 2010; Moran & Clark 2011).

Despite the unifying nature of seed and pollen dispersal from the plant perspective, investigations of seed and pollen dispersal have often been undertaken in isolation by separate fields. As a consequence, there are very few species in which both seed and pollen dispersal have been quantified, as compared to species in which either seed or pollen dispersal has been studied (Hardesty *et al.* 2006; Ashley 2010). This asymmetric emphasis on only one of the gene dispersal processes hinders a complete understanding of responses to environmental changes in many forest trees. In addition, empirical focus has been placed on primary rather than effective seed and pollen dispersal, albeit the latter are of higher relevance to ecological and evolutionary dynamics. This is likely due to the added challenge of losing both maternal and paternal cues after progenies being dispersed and established. For example, in the case of paternity inference, it is more feasible to characterize pollen dispersal based on seed arrays collected from maternal trees than based on established seedlings. Furthermore, we know better the processes of gene dispersal mediated by abiotic means, such as wind, than more complex processes driven by idiosyncratic behaviors of mutualistic partners (e.g. wind vs. animal-mediated seed dispersal, Clark *et al.* 1999). This may in part explain our limited empirical knowledge of primarily animal-mediated seed and pollen dispersal in tropical forests relative to temperate forests.

In low-diversity temperate forests, rare events of long-distance seed dispersal have been invoked and modeled to explain rapid postglacial re-colonization of temperate zone trees, despite the observation that most seeds are deposited very locally (i.e. Reid's Paradox,

Clark *et al.* 1998); relative to seed dispersal, more frequent long-distance pollen dispersal by wind has been inferred using genetic methods, explaining low genetic differentiation in many temperate trees (Ouborg *et al.* 1999; Hamrick 2004). In species-rich tropical forests, compared to the total diversity of tree species, disproportionally few taxa have been investigated, especially those dispersed and pollinated by animals. Insights gained from existing empirical studies emphasize the potential of long-distance seed dispersal by vertebrate dispersers in tropical trees (Hardesty *et al.* 2006; Russo *et al.* 2006; Hanson *et al.* 2007; Sezen *et al.* 2009). Some such studies (Sezen *et al.* 2005; Hardesty *et al.* 2006) have highlighted the possibility of comparable or even longer gene dispersal by seeds than by pollen, in marked contrast to the broad consensus of pollen-dominated gene dispersal in temperate taxa (Ouborg *et al.* 1999; Petit & Hampe 2006; Kremer *et al.* 2012). Yet, this increased importance of seed-mediated gene dispersal from temperate to tropical forests awaits additional empirical support from other frugivore-dispersed tropical trees.

In this study, we quantified effective seed and pollen dispersal (referred to as seed and pollen dispersal for the sake of simplicity) in four sympatric tropical dioecious tree species that differ in seed dispersal and pollination syndromes. By integrating parentage inference with inverse modeling (Jones & Muller-Landau 2008), we aimed to evaluate (i) the frequency of long-distance seed and pollen dispersal in these tropical trees, (ii) the magnitude of seed vs. pollen dispersal and of seed vs. pollen-mediated gene dispersal and (iii) the best-fitting seed and pollen dispersal kernel, whether light or heavy tailed, the latter indicating the potential of long-distance events. Then using spatially explicit dynamic simulations, in which seed and pollen dispersal are constrained to the near or far tail of the

fitted kernels, we examined (iv) the genetic impacts of short vs. long-distance seed and pollen dispersal at an evolutionary timescale.

Materials and Methods

Study species

Our study focused on four dioecious tree species, in a seasonal tropical lowland forest on Barro Colorado Island (BCI, 9°10' N, 79°51' W), Panama. These four species, *Virola sebifera* (Myristicaceae), *Tetragastris panamensis* (Burseraceae), *Cecropia insignis* (Urticaceae) and *Triplaris cumingiana* (Polygonaceae), are differentiated across a variety of organismal and ecological attributes. *Virola sebifera*, *T. panamensis* and *C. insignis* are common canopy trees, attaining heights of 30–40 m, and the diameter at breast height (dbh) can reach respectively 30, 60 and 70 cm (Croat 1978). In contrast, *T. cumingiana* is a less-abundant midstory species, 10–20 m tall and 12–30 cm in dbh at maturity, and is spatially aggregated (Croat 1978). These species are positioned disparately along the spectrum of shade tolerance, with *V. sebifera* and *T. panamensis* being tolerant to low light, *C. insignis* light demanding and *T. cumingiana* intermediate in this respect (Comita *et al.* 2007).

Consistent with the characteristic floral morphology of dioecious trees (Bawa & Opler 1975), these species display dull-colored, small-sized flowers, expect for the sexually dimorphic species *T. cumingiana*, in which female flowers are bright pink and considerably larger (*ca*. 3 cm long). Generalist insect pollinators (e.g. small bees, beetles, wasps) are the dominant pollen vectors of dioecious tree species (Bawa & Opler 1975; Bawa *et al.* 1985). *Cecropia* is, however, among the few plant taxa that are wind pollinated in tropical rain forests (Bawa & Opler 1975; Croat 1978). Flowering concentrates either in dry season

between January and April (*C. insignis* and *T. cumingiana*) or in early rainy season in June and July (*T. panamensis*), whereas both the two flowering peaks occur in *V. sebifera* (Croat 1978).

Fruits of *V. sebifera*, *T. panamensis* and *C. insignis* are consumed by diverse vertebrate frugivores on BCI. The disperser assemblage of *V. sebifera*, comprised of six bird species but primarily toucans (Howe 1981), is smaller than that of *T. panamensis*, which includes eight mammals, howler monkeys in particular, and nine bird species (Howe 1980). *Cecropia insignis* is primarily dispersed by large avian frugivores and bats, as well as mammals (Brokaw 1986). The large calyx of female flowers in *T. cumingiana* facilitates seed dispersal by wind (Croat 1978). For the sake of brevity, we refer to the study species by genus name hereafter.

Sampling and genotyping

Field studies were carried out in the 50-ha (1000 × 500 m) Forest Dynamics Plot (FDP) on BCI, in which each freestanding woody stem of dbh ≥1 cm has been permanently tagged, mapped and taxonomically identified to species (Condit 1998; Hubbell *et al.* 1999). A census of stem growth, recruitment and mortality is conducted every five years. We based our collection of reproductive-sized trees on 2005 and 2010 FDP census data. Species-specific reproductive size threshold, defined as the dbh above which trees become fully fertile (R.B. Foster, unpubl. data) on BCI, provides an upper limit of the minimal dbh for consideration as adult trees (20 cm in *Virola* and *Triplaris*; 30 cm in *Tetragastris* and *Cecropia*). To minimize the odds of excluding some smaller yet fertile trees, we evaluated the reproductive status of individual trees in FDP using a lowered dbh threshold, as, 7 cm

in *Cecropia* and *Triplaris*, 10 cm in *Virola* and 15 cm in *Tetragastris*. Tree gender is designated according to flower forms, or the presence of fruits on the tree and/or seedling carpets under tree crowns. From 2010 to 2013, we collected leaf tissues from 789 reproductive-sized trees: 214 (76 female trees; 94 male trees; 44 sex-unknown trees) in *Virola*, 263 (104, 107, 52) in *Tetragastris*, 230 (100, 111, 19) in *Cecropia* and 82 (45, 31, 6) in *Triplaris*. Trees for which we could not ascertain gender (i.e. sex-unknown) were often smaller in size or infested by lianas. As female trees were identified on the basis of fruit and/or female flower production, sex-unknown trees were subsumed as potential male parent candidates for parentage inferences.

Seedlings (height ≤ 10 cm) were exhaustively surveyed, mapped and collected between 2012 and 2013 from a central subplot of 18 ha (600 × 300 m) within FDP for *Virola*, *Cecropia* and *Triplaris*, and from a smaller central subplot of 2 ha (200 × 100 m) for *Tetragastris* in 2010. We nondestructively sampled all the 377 *Virola* and 503 *Cecropia* seedlings, most *Triplaris* seedlings (n = 369) and a representative subsample (n = 269) of *Tetragastris* seedlings (Wei *et al.* in review). Details of seedling sampling strategy and microsatellite genotyping (at averagely 10 loci) of the 1518 seedlings and 789 adult trees of the four species have been described in our previous study (Wei *et al.* in review).

Parentage inference

Simultaneous inferences of maternity and paternity were realized using a pedigree-based likelihood approach in COLONY v2.0 (Wang & Santure 2009; Jones & Wang 2010), which identifies the most likely configurations of family groups. We parameterized COLONY to accommodate the polygamous dioecious trees, allowing the possibility of

inbreeding. All female trees were considered as candidate maternal parents, and the male and sex-unknown trees as candidate paternal parents. We assumed that approximately eighty percent of the actual maternal and paternal trees were included in our sampling. Although the misspecification of this sampling fraction could result in erroneous parentage assignments, this effect is likely minor using COLONY (Wang & Santure 2009). To improve the accuracy of parentage inference, we ran a parallel Linux version of COLONY under the maximum allowed searching length and the highest likelihood-computing precision, and allowed for modestly high genotyping error rates (1–6% averaged over loci; Table 4S.1).

Seedlings with uniquely identified maternal trees within FDP based upon a confidence level of ≥80% were first retained. In the cases of low-confidence (<80%) maternity assignments where two or more likely mother candidates appeared, we assumed conservatively the nearest female adult as the mother tree, conditional on no genotypic mismatches being detected. Seed dispersal distances were then gauged based upon confident and conservative mother-seedling relationships; the remaining seedlings were regarded as originating from off-plot seed dispersal. Confident paternity assignments (≥80%) were used to measure pollen-mediated gene dispersal from father trees to seedlings. For pollen dispersal that occurred between female and male trees, we considered only the situations where maternity and paternity of individual seedlings were both detected with confidence (≥80%) within FDP. Although we selected a relaxed confidence level of 80% for parentage inference, confident maternity and paternity assignments were obtained at an average level of 99%.

Inverse modeling integrated with genetic data

Parentage inference alone provides important yet limited insights into seed and pollen dispersal, because distances of immigrant seeds and pollen into FDP cannot be estimated. To incorporate the off-plot events, we used modeling approaches that integrate inverse models, parentage inference and off-plot integration (Jones & Muller-Landau 2008). To estimate seed dispersal distance and kernel, we considered primarily the gene shadow model accounting for immigrants (GSMi, Jones & Muller-Landau 2008). For comparisons, we also included gene shadow model without immigrants (GSM, Jones & Muller-Landau 2008), seed shadow model (SSM, Ribbens *et al.* 1994) and SSM with immigrants (SSMi, Muller-Landau *et al.* 2008). To estimate pollen dispersal distance and kernel, we treated each female tree as a pollen trap with on- and off-plot male trees competing to fertilize each source tree (or to reach each trap), namely, competing source model taking into account immigrant pollen (CSMi, Jones & Muller-Landau 2008), which although was originally developed for measuring seed dispersal (Robledo-Arnuncio & Garcia 2007). These models were described in Appendix 4S.1 (Supporting Information).

Model fitting was conducted using maximum likelihood in R v3.0.3 (R Development Core Team 2013). The two-dimensional kernels of seed and pollen dispersal, describing deposition probability per unit area at a certain distance, included two-parameter lognormal, gamma and Weibull distribution and one-parameter exponential distribution (Appendix 4S.1). Although we also tested the two-dimensional *t* distribution ('2Dt', Clark *et al.* 1999), it was not reported here because of frequent convergence failures for certain models and species. For each species, the best-fitted seed and pollen dispersal kernel were evaluated under each model using Akaike information criterion (AIC).

Simulating genetic impacts of near vs. far-tail seed and pollen dispersal

Using spatial explicit individual-based simulations, we assessed the effects of short-distance (near-tail) vs. long-distance (far-tail) seed and pollen dispersal on population genetic diversity, inbreeding and spatial genetic structure (SGS) at an evolutionary timescale. Short-distance seed dispersal was defined as the events within 100 m (Cain *et al.* 2000). Because pollen dispersal depends strongly upon population geometry (i.e. inter-mate spatial distribution), we defined short-distance pollen dispersal as the events between near neighbors of opposite sexes. A cut-off of 10 nearest female neighbors for individual male trees was chosen because median inter-mate distances above this mating neighborhood typically exceed 100 m in these species.

The genetic effects of seed and pollen dispersal were examined separately. When pollen dispersal under consideration, the best-fitted seed dispersal kernel inferred from GSMi was applied for the three scenarios: standard pollen dispersal scenario, as a control case (PC), where pollination occurs between on-plot mates following the best-fitted pollen dispersal kernel, with the probability of immigrant pollen being m_p; long-distance scenario (PL), which is similar to PC yet differs in on-plot pollination being beyond near neighbors; short-distance scenario (PS), in which pollination only occurs between near neighbors inside the plot. Likewise, we used the standard pollen dispersal (PC) for the following three scenarios of seed dispersal: standard seed dispersal scenario (SC), in which seed dispersal distances, d_s, are drawn from the best-fitted seed dispersal kernel, $p(d_s)$; long-distance scenario (SL) with d_s of >100 m drawn from $p(d_s)$; short-distance scenario (SS) with d_s of ≤100 m.

The initial simulation setup follows the observed adult population, including tree locations, sexes and genotypes, in a continuous landscape of 1000 × 500 m. A constant population size, N, is maintained through the birth-death equilibrium. Specifically, a female tree i, chosen at random, produces a progeny k that survives into adulthood, accompanying the death of a random adult tree in the population. When the female tree is from outside the plot (with a probability of m_s), offspring k is dispersed and established at a random location within the plot, carrying a background genotype based on population allele frequencies. Alternatively, when the female tree i is located inside the plot, offspring k is dispersed from the source i at a random direction over a distance of d_s, according to specific seed dispersal scenario (SC, SL or SS). In this case, the maternal haplotype of offspring k is from the female tree i according to Mendelian inheritance. The paternal haplotype could be from background if the father tree is outside the plot (with a probability of m_p), or from an on-plot father tree j. The probability of an on-plot male tree being the pollen donor, relative to all the other male trees, is determined by the distance to the female tree i, according to specific pollen dispersal scenario (PC, PL or PS); thus, the one with the highest probability is chosen as the pollen donor. Offspring k will replace the membership of a dead adult tree, chosen at random, yet retaining the dead tree's sex status. To minimize plot edge effects, we focused on the reflection method (Pretzsch 2009) here. The simulation algorithm is described in Appendix 4S.2, in conjunction with the algorithms using alternative torus edge-correction method and alternative genotype determination of immigrant offspring k.

After 1000 generations (i.e. $N \times 1000$ birth-death events), genetic diversity (observed heterozygosity, H_O), population inbreeding (inbreeding coefficient, F_{IS}) and spatial genetic pattern (SGS intensity statistic b; Wei *et al.* in review) were examined for

the simulated population. SGS intensity statistic *b* quantifies the rate of change of genetic relatedness with (log-transformed) distance; thereby, the sign of *b* (positive or negative) indicates genetic relatedness ascending or decaying with distance, and the absolute value of *b* represents the magnitude of SGS. With each seed or pollen dispersal scenario, we simulated 200 population data sets to obtain the mean estimates of the three genetic parameters in R v3.0.3. We parameterized simulation models using the observed data, such as adult tree population and the best-fitted seed and pollen dispersal kernel, from *Virola*, *Tetragastris* and *Triplaris* that vary in gene dispersal ability, to assess the generality of model results and inferences. Information from *Cecropia* population was not used because the effect of seed dispersal would be confounded by stochastic events (i.e. gap formation). We set the rate of immigrant seeds $m_s = 0.05$ and of pollen $m_p = 0.05$ (low) and 0.4 (high).

Results

Maternity inference and seed dispersal distance

Parentage inference revealed that most seedlings were produced by seed trees within the 50-ha FDP; that is, 64% ($n = 234$) *Virola*, 99% (267) *Tetragastris*, 99% (497) *Cecropia* and 94% (348) *Triplaris* seedlings were assigned with maternal trees either confidently (i.e. at a confidence of ≥80%; $n = 176$, *Virola*; 158, *Tetragastris*; 404, *Cecropia*; 322, *Triplaris*) or conservatively by assuming the nearest female tree, among equally likely maternal candidates, as the mother tree (58, *Virola*; 109, *Tetragastris*; 93, *Cecropia*; 26, *Triplaris*).

Considering confident maternity assignments, approximately 40–50% of female trees (42% in *Virola*, 55% in *Cecropia*, 42% in *Triplaris*) within FDP contributed

reproductively to the seedling banks in our sampling, with an exception of 19% in *Tetragastris*, likely resulting from a much smaller seedling subplot (2 ha) used for this species. Seed dispersal distances estimated from confident maternity inferences varied among dispersal syndromes. With primarily avian-dispersed *Virola* and *Cecropia*, seedlings were established from sources trees at a median distance of 102.4 m (SD = 134.1 m, mean = 150.5 m, observed range 4.7–724.2 m) and of 121.4 m (SD = 138.2 m, mean = 160.9 m, range 8.4–750.5 m) respectively. The same estimate decreased by twofold to 52.8 m (SD = 137.0 m, mean = 113.8 m, range 1.2–596.3 m) in monkey-dispersed *Tetragastris*. The shortest seed dispersal distance was observed in wind-dispersed *Triplaris* of 16.1 m (SD = 36.7 m, mean = 27.4 m, range 0.9–309.8 m). Of these assigned seedlings, 53% were dispersed over 100 m in *Virola* and 56% in *Cecropia* followed by 40% in *Tetragastris*, but only 6% in *Triplaris*.

Overall, combining confident and conservative maternity inferences, seed dispersal distances were comparable to that based on confident maternity inferences alone, with median distance being over 100 m in *Virola* (129.5 m, SD = 149.3 m, mean = 175.0 m; Table 4.1 and Fig. 4.1) and *Cecropia* (108.4 m, SD = 135.2 m, mean = 152.3 m), more than twice as long as in *Tetragastris* (40.2 m, SD = 124.1 m, mean = 98.2 m), and approximately tenfold higher than in *Triplaris* (15.8 m, SD = 58.1 m, mean = 31.4 m). The proportion of these seedlings (not including immigrants) being dispersed and established at distances longer than 100 m was 60% and 53% in *Virola* and *Cecropia* respectively, 35% in *Tetragastris* and 7% in *Triplaris*. Due to long-distance seed dispersal in avian-dispersed *Virola* and *Cecropia*, respectively 91% and 84% seedlings were established closer to other

female trees than their source trees. This fraction dropped to 60% in *Tetragastris* and 40% in *Triplaris*.

Paternity inference and pollen dispersal distance

Two-thirds of the sampled seedlings were assigned with father trees unambiguously (at a confidence of ≥80%): 56% (n = 202) *Virola*, 59% (158) *Tetragastris*, 52% (264) *Cecropia* and 65% (239) *Triplaris* seedlings. We did not consider conservative paternity assignments because of the ambiguity in resolving triad (mother–offspring–father) relationships. Parentage inference identified 45 unique pollen donors in *Virola*, 27 in *Tetragastris*, 43 in *Cecropia* and 25 in *Triplaris*. The median distance of pollen-mediated gene dispersal (i.e. distances from father trees to dispersed seedlings) was 208.4 m in *Virola* (SD = 151.8 m, mean = 237.7 m, range 13.9–703.1 m; Fig. 4S.1) and 207.6 m in *Cecropia* (SD = 157.0 m, mean = 231.3 m, range 16.5–815.6 m), twofold longer than that of 107.2 m in *Tetragastris* (SD = 131.1 m, mean = 157.1 m, range 6.4–538.6 m) and that of 116.4 m in *Triplaris* (SD = 121.3 m, mean = 136.2 m, range 3.9–670.9 m). Compared to seed dispersal, the proportion of pollen-mediated gene dispersal exceeding 100 m was consistently higher, as 80% in *Virola*, 75% in *Cecropia*, 52% in *Tetragastris* and 54% in *Triplaris*.

Pollen dispersal distances between paternal and maternal trees were estimated from 30–60% of the sampled seedlings (31% in *Virola*, 33% in *Tetragastris*, 45% in *Cecropia*, 57% in *Triplaris*), in which both confident paternity and maternity were inferred. Median pollen dispersal distances, between 100 and 200 m (Table 4.1), were relatively comparable among pollination syndromes. In insect-pollinated *Virola*, *Tetragastris* and *Triplaris* (Fig.

4.2), the median distance was 176.9 m (SD = 165.4 m, mean = 213.3 m, range 10.0–733.0 m), 129.9 m (SD = 156.2 m, mean = 193.0 m, range 34.5–573.2 m) and 118.7 m (SD = 109.4 m, mean = 132.6 m, range 4.6–680.2 m) respectively. In wind-pollinated *Cecropia*, pollen dispersal occurred at a median distance of 145.9 m (SD = 193.7 m, mean = 205.6 m, range 3.0–914.1 m; Fig. 4.2). To evaluate whether pollen dispersal was primarily between near neighbors, we compared distances from father trees to mother trees with distances to the tenth nearest female neighbors (i.e. local mating neighborhood; Fig. 4S.2). Although approximately half of pollination events (from 41% in *Cecropia* to 60% in *Triplaris*) occurred within the defined mating neighborhood, paired *t*-tests showed that pollen dispersal significantly exceeded distances between near neighbors (Fig. 4S.2), except in *Triplaris* ($t = 0.76$, df = 208, $P = 0.45$).

In line with the broad consensus of pollen-dominated gene dispersal in temperate trees, distances of on-plot pollen dispersal (Fig. 4.3) and pollen-mediated gene flow (Fig. 4S.3) were significantly longer than distances of on-plot seed dispersal in these tropical trees, revealed by paired *t*-tests.

Model fitting of seed and pollen dispersal

The lognormal kernel outperformed other tested kernels in modeling seed dispersal in GSMi, except for *Cecropia*, in which the lowest AIC was associated with the gamma kernel, followed by the lognormal ($\Delta AIC = 5$). Due to the integration of immigrant seeds in GSMi, median distance of seed dispersal estimated with respective best-fitted kernel was longer than that using parentage inference alone (Table 4.1). Especially in *Virola*, in which

36% seedling were not assigned with maternal trees within FDP, GSMi reported a median dispersal distance of 274.4 m (CI 237.3–317.6 m), 2.12 times longer than did parentage inference. But in *Tetragastris*, in which only 1% seedlings were immigrants based on parentage inference, confidence intervals of median seed dispersal distance overlapped entirely between GSMi and parentage inference (Table 4.1). In *Cecropia*, seed dispersal distance estimated with the best-fitted gamma kernel (median = 144.1 m, CI 118.8–174.3 m; Table 4.1) had a wider confidence interval than with the lognormal kernel (median = 123.7 m, CI 113.8–134.2 m), both of which partly overlapped with the estimate based on parentage inference. With respect to the fraction of immigrants, estimates in GSMi (*Virola*, 37.2%; *Tetragastris*, 5.3%; *Cecropia*, 4.9%; *Triplaris*, 1.4%) approximated those inferred from parentage assignments.

Similar to GSMi, other models–GSM, SSSi and SSS–also found that the lognormal kernel provided the best fit to seed dispersal, regardless of species identity. However, estimated median distance was in general discordant among models (Table 4.1), except in wind-dispersed *Triplaris*. In *Triplaris*, GSM, considering only on-plot seed dispersal, reported a median distance consistent with parentage inference. But dispersal distance was biased upwards using GSM in the other species, in comparison with parentage inference. In the absence of genetic data, SSM and SSMi obtained a median distance in close agreement with on-plot seed dispersal from GSM and parentage inference in wind-dispersed *Triplaris*, albeit shorter than the estimate in GSMi. But SSM and SSMi significantly underestimated dispersal distances in *Tetragastris*, according to non-overlapping confidence intervals between these two models and parentage inference or GSMi, whereas overestimated dispersal distances in gap-specialist *Cecropia*, in which initial seed shadow was

dramatically modified by gap dynamics. In *Virola*, median dispersal distance from SSM and SSMi was only half of that estimated from GSMi (Table 4.1), albeit concordant with that of on-plot seed dispersal in parentage inference. Parameter estimates of GSMi, GSM, SSSi and SSS were given in Table 4S.2.

Unlike seed dispersal, the best-fitted pollen dispersal kernel varied among species, with the lognormal kernel in *Tetragastris* and *Triplaris*, and Weibull or exponential kernel in *Virola* and *Cecropia*. But irrespective of the differences in the best-fitted kernels, median pollen dispersal distance under CSMi broadly overlapped in confidence intervals among these four species, and was 1.7–3.2 times as long as on-plot pollen dispersal based on parentage inference. In *Virola* and *Cecropia*, median distance was 305.6 m (CI 255.5–355.3 m) and 386.6 m (CI 338.8–434.9 m) respectively with the Weibull kernel (Table 4.1), and 297.7 m (CI 252.6–351.0 m) and 391.1 m (CI 346.9–440.8 m) with the exponential kernel. Parameter estimates of CSMi were given in Table 4S.3.

Genetic impacts of short vs. long-distance seed and pollen dispersal

Seed dispersal strongly influences population genetic diversity, inbreeding and spatial genetic structure (Tables 4.2 and 4S.4). Specifically, reduced magnitude of seed dispersal, from long-distance (SL) to short-distance scenario (SS), resulted in a significant loss in observed heterozygosity H_O and an increase in inbreeding coefficient F_{IS} and genetic aggregation (intensity statistic b) for both good dispersers (e.g. *Virola*; Table 4.2) and poor dispersers (e.g. *Triplaris*; Table 4S.4).

In the case of good dispersers (Table 4.2), pronounced population genetic changes were observed when seed dispersal was constrained to the near tail within 100 m from SC

to SS. As median seed dispersal distance was reduced by four-fold from 196.9 m in SC to 52.4 m in SS, individuals became more aggregated spatially, with relative neighborhood density (Condit *et al.* 2000) at the first distance interval (20 m) twofold higher in SS. Spatial aggregation influenced mating environment, as median on-plot pollen dispersal distance was 105.2 m in SS relative to 147.1 m in SC. The changes in seed dispersal and subsequent pollination intensified spatial genetic aggregation (b = -0.019, SS; -0.005, SC) and inbreeding (F_{IS} = 0.023, SS; 0.001, SC) owing to assortative mating with relatives; these phenomena were particularly salient in the case of limited gene dispersal influx via pollen (i.e. m_p = 0.05).

For poor dispersers (e.g. *Triplaris*; Table 4S.4), the majority of seed dispersal events occurred within 100 m in SC, thereby population genetic changes were less prominent between SC and short-distance scenario (SS) than between long-distance scenario (SL) and SC or SS. For instance, in simulations parameterized with the observed data of *Triplaris*, median seed dispersal was ten times longer in SL (182.5 m) than in SS (18.1 m). As a result, a significant decay in H_O was observed (t = 12.6, df = 384.9, P < 0.001 when m_p = 0.05; t = 7.0, df = 397.8, P < 0.001 when m_p = 0.4), as well as significantly augmented population inbreeding and spatial genetic aggregation (Table 4S.4).

Pollen dispersal impacts genetic diversity and population inbreeding through non-random mating (e.g. mating with near neighbors in short-distance pollen dispersal scenario, PS) (Tables 4.2 and 4S.4). For instance, in simulations parameterized with the observed data of *Virola* (Table 4.2), H_O was significantly reduced in PS relative to PC (t = 15.7, df = 323.5, P < 0.001) and F_{IS} was elevated but not yet significant (t = 1.65, df = 388.3, P =

0.1), likely due to the relatively low population genetic aggregation (i.e. near neighbors are not more related than expected on average) in good dispersers. But near-neighbor pollination resulted in a significant increase in F_{IS} estimate in poor dispersers (Table 4S.4). The population genetic function of pollen dispersal in maintaining genetic diversity and reducing inbreeding could also be found in seed dispersal simulations (SL, SC and SS) between $m_p = 0.4$ and $m_p = 0.05$ (Tables 4.2 and 4S.4). On the other hand, pollen dispersal had a minor influence on SGS relative to seed dispersal (Tables 4.2 and 4S.4).

The genetic impacts of seed and pollen dispersal were in general robust to different edge-correction methods (Tables 4S.5 and 4S.6). However, systematic biases towards unexpected high genetic diversity and low inbreeding and SGS intensity arose in short-distance seed dispersal scenario (SS) using the torus (Table 4S.5) and particularly reflective edge correction (Table 4S.6) with background immigrants supplementing individuals being dispersed outside of the plot (Appendix 4S.2). These biases were caused by spatial clumping of individuals, resulting from short-distance seed dispersal adjacent to the edges of plots, which artificially increased the likelihood of individuals being dispersed outside the plot and being replaced by background immigrants. As a result of this frequent gene flow influx in the short-distance seed dispersal scenario, we observed comparable levels of H_O, F_{IS} and SGS intensity statistic b between SS and SC (Tables 4S.5 and 4S.6), particularly in poor dispersers.

Discussion

Our study presents one of the few empirical evaluations of effective seed and pollen dispersal in tropical trees (e.g. Hardesty *et al.* 2006; Gaino *et al.* 2010), through

unambiguous maternal and paternal inferences of established seedlings. In line with the growing appreciation of potential long-distance seed dispersal in tropical trees mediated by vertebrate frugivores, >50% seedlings were established over 100 m from source trees in primarily avian-dispersed *Virola* and *Cecropia* and over a third of seedlings in primarily mammal-dispersed *Tetragastris*. The fraction of >1 km, predicted by the best-fitted seed dispersal kernels (principally the lognormal), is as high as 17.7% in *Virola*, and 1.2% and 2.4% in *Cecropia* and *Tetragastris* respectively. Pollination by small insects or wind occurred at comparable scales and magnitude, significantly beyond near neighbors. Approximately 10–20% of pollination events could potentially exceed 1 km in all focal species, according to the best-fitted pollen dispersal kernels. Although our results showed no evidence for the suggested seed-dominated gene dispersal in vertebrate-dispersed tropical trees, the ratio of pollen to seed-mediated gene dispersal (i.e. father–seedling vs. mother–seedling distance; Fig. 4S.3) and of pollen to seed dispersal (i.e. father–mother vs. mother–seedling distance; Table 4.1 and Fig. 4.3) in *Virola* and *Cecropia* are considerably lower than in wind-dispersed *Triplaris* and many temperate zone trees of wind or animal dispersal syndromes (Ouborg *et al.* 1999; Petit *et al.* 2005).

Effective seed dispersal in tropical trees

Long-distance seed arrival and establishment can be selectively advantageous (Howe & Smallwood 1982), as it may confer the chance to escape from high mortality near source trees (escape hypothesis; see also Janzen 1970; Connell 1971), to colonize new areas (colonization hypothesis; Clark *et al.* 1998) and to locate critical habitats for recruitment (directed dispersal hypothesis; Wenny & Levey 1998). Our results of lower recruitment

closer to the bole of mother trees in *Virola* (Fig. 4.1A) may suggest the action of the escape or Janzen-Connell (J-C) effect, given that primary seed deposition peaks under or near source trees (Howe 1989). Although seedlings of *Tetragastris* and *Triplaris* recruited most frequently near maternal trees (Fig. 4.1B, D), the median distance shifted from 10.0 and 6.9 m of primary seed dispersal (Muller-Landau *et al.* 2008) to 40.2 and 15.8 m respectively of effective seed dispersal based on parentage inference, suggesting density and/or distance-dependent mortality at short distances. Despite the resemblance of *Cecropia* (Fig. 4.1C) to *Virola* (Fig. 4.1A) in exhibiting a J-C recruitment pattern, the underlying mechanism in *Cecropia* can be disparate: asymmetric influence from source trees by canopy shading and sporadic emergence of new forest gaps.

Local processes of density and distance dependence do not provide a full account of effective seed dispersal in these species, especially at larger spatial scales. Vertebrate dispersers clearly play a central role in mediating long-distance seed deposition. Seedlings of *Virola* and *Cecropia* were frequently found over 100 m from mother trees, some of which reached *ca*. 700 m within FDP. The heavy investment of *Virola* in nutrient-rich arils ensures the removal and transportation of seeds by obligate bird dispersers such as toucans (Howe 1981, 1993) that are capable of long-distance flights (Kays *et al.* 2011). On the other hand, numerous small-seeded *Cecropia* fruits are consumed and moved by diverse opportunistic dispersers of birds, bats and mammals (Brokaw 1986; Howe 1993), which collectively could promote long-distance events adapted for irregular formation of critical habitats. Effective seed dispersal kernels of *Virola* and *Cecropia* were heavy tailed, as lognormal and gamma (shape > 1) distribution decline more slowly than an exponential distribution. The same heavy-tailed lognormal kernel was also found in *Tetragastris*, whose

seeds are dispersed by various generalist vertebrate frugivores. Although the majority of *Tetragastris* seeds were dropped under source trees by the primary disperser, mantled howler monkey (*Alouatta palliata*) (Howe 1980), 35% of the seedlings were established at distances greater than 100 m and some reached nearly 600 m.

Wind dispersal is not as common as animal dispersal in tropical moist forests (Howe & Smallwood 1982). Relative to the animal-dispersed species, effective seed dispersal was considerably shorter in wind-dispersed *Triplaris*, in which only 7.2% seedlings were found over 100 m. It is likely due to the short stature of *Triplaris* being a midstory tree, as tree height and horizontal wind speed affect wind dispersal distance (Nathan *et al.* 2002). Nevertheless, median distance of effective seed dispersal in *Triplaris* (parentage inference: 15.8 m; GSMi: 25.4 m) is comparable to that of primary seed dispersal in a more typical wind-dispersed tropical tree on BCI, *Jacaranda copaia* (parentage inference: 27.0 m, Jones *et al.* 2005; GSM: 17.9 m, Jones & Muller-Landau 2008) that is much taller and larger. But unlike *Jacaranda* in which long-distance dispersal realized by potential wind uplift is essential for regeneration in large forest gaps (Jones *et al.* 2005), this selection differential (i.e. fitness in relation to dispersal) might be of less or minimal significance in *Triplaris*, because habitat filtering of non-random soil phosphorus limitation on BCI (Condit *et al.* 2013) may confer a good chance of establishment close to adult trees in this species.

Seed dispersal kernel is of particular interest and importance in modeling efforts pertaining to population (e.g. migration, Clark *et al.* 1999; Levin *et al.* 2003) and community processes (e.g. assembly, Levin *et al.* 2003). A compound normal distribution with the inverse of the variance following a gamma distribution–'2Dt' distribution (Clark *et*

al. 1999)–has often been used to model primary seed dispersal (Muller-Landau *et al.* 2008) that is presumably uniform near source trees and is fat in the far tail. However, 2Dt did not fit effective seed dispersal well in our study. This is likely due to two factors: first, 2Dt was unstable given an unconstrained shape parameter (Clark *et al.* 1999; Greene *et al.* 2004), as it often failed to converge; second, a mode at zero does not reflect what we have observed of the recruitment pattern at short distances, especially in *Virola* and *Cecropia*. But once converged, 2Dt fitted wind-dispersed *Triplaris* better than did other kernels (including the lognormal), but not in the other species (data not shown). In general, lognormal distributions provide a better fit to effective seed dispersal in these animal-dispersed tropical trees, as they could account for density and distance-dependent mortality near source trees.

The incorporation of genetic information into inverse modeling (Jones & Muller-Landau 2008) is important for an accurate characterization of seed dispersal by vertebrate frugivores. Our study and others have shown that proximity is a poor predictor of maternity in animal-dispersed tropical taxa (Hardesty *et al.* 2006; Sezen *et al.* 2009), as more than 80% seedlings were established closer to other conspecific female adults rather than their mother trees in *Virola* and *Cecropia* and about 60% in *Tetragastris*. Although inverse models, SSMi and SSM, do not assume the nearest seed tree as the mother tree (Muller-Landau *et al.* 2008), genetic inverse models such as GSMi made better predictions of animal-mediated dispersal distances that approximated what were inferred from parentage assignments (Table 4.1). But in wind-dispersed *Triplaris*, inverse models performed equally well in estimating median distances without genetic data (Table 4.1), as shown previously in wind-dispersed *Jacaranda* (Jones & Muller-Landau 2008). This validates the

accuracy and continued popularity of inverse modeling in wind-dispersed trees. Furthermore, immigration needs to be taken into account in non-genetic and genetic inverse modeling. Even when the occurrence of immigrant progenies was as low as 1% in *Cecropia* and *Tetragastris*, genetic inverse models without immigrant integration, GSM, significantly overestimated effective dispersal distance (see also Jones & Muller-Landau 2008).

Effective pollen dispersal in tropical trees

Genetic findings of long-distance contemporary pollen dispersal reconcile Slatkin's Paradox in trees, namely, the incongruence between anticipated gene dispersal limitation from local observations and inferred high gene flow from weak genetic differentiation among populations (Slatkin 1987; Mallet 2001; Ashley 2010; Jones 2010). In tropical trees, insect-mediated pollen dispersal exceeding several hundred or thousand meters is not uncommon based upon paternity inference; many such studies have focused on fragmented habitats (e.g. Chase *et al.* 1996; White *et al.* 2002; Dick *et al.* 2003). Despite the potential of long-distance flights, density-dependent foraging of these pollinators (reviewed in Ghazoul 2005) would predict less extreme long-distance pollination in continuous tropical forests. Median on-plot pollen dispersal distance inferred from seedlings in our study was comparable among insect-pollinated species from 118.7 m in *Triplaris* to 176.9 m in *Virola*. Similar extents of pollen dispersal were also observed in other insect-pollinated trees on BCI (Stacy *et al.* 1996; Hufford *et al.* 2009). In a more comparable case study of insect-pollinated dioecious canopy tree *Simarouba amara* on BCI (Hardesty *et al.* 2006), pollen dispersal distance (mean = 334.4 m) unambiguously inferred from 33 seedlings is greater than the average distance reported here, as 132.6 m from 209 *Triplaris* seedlings to

213.3 m from 113 *Virola* seedlings. It is likely due to larger sampling areas than ours of *Simarouba* seedlings (~40 ha) and adult trees (84 ha), allowing the detection of even longer-distance pollination events (maximum = 1063 m vs. 733 m in *Virola*). Moreover, although distance between mates is an important correlate of pollination success (Fig. 4.2), insect-mediated pollen dispersal is significantly beyond near neighbors (Fig. 4S.2). But in *Triplaris* where adult trees are spatially more aggregated than are in the other species (Wei *et al.* in review), on-plot pollen dispersal was primarily constrained between nearest ten neighbors (Fig. 4S.2; see also Wei *et al.* in review), probably reflecting density-dependent pollination. Considering immigrant pollen, approximately half came from outside the FDP (but *ca*. 35% in *Triplaris*). Median pollen dispersal distance, estimated by CSMi that integrates immigrants, ranged between 305.6 m in *Virola* and 411.7 m in *Tetragastris*.

Wind pollination is rare in tropical trees (Bawa *et al.* 1985), in striking contrast to temperate trees (Regal 1982). As the efficacy of wind pollination is strongly density and distance dependent, anemophily is ineffective and disfavored in most trees of low density in species-rich tropical forests (Regal 1982; Ghazoul 2005). The tallness of *Cecropia*, coupled with flowering in dry windy season on BCI renders selective advantages of wind pollination in this species. We found wind pollination is as effective as insect pollination, if not more, in the study species, with the maximum on-plot distance exceeding 900 m.

Pollen dispersal kernel is fundamental for predicting changes in genetic connectivity and the spread of adaptive variation in response to environmental changes. Various kernel functions have been evaluated in fitting pollen dispersal of both wind and animal-pollinated trees, in which Weibull and exponential power distributions often rank better than others (Tufto *et al.* 1997; Austerlitz *et al.* 2004). Consistently, the Weibull kernel provided a

better fit to pollen dispersal inferred from seedlings in *Virola* and *Cecropia*, along with the exponential kernel. As the estimated shape parameter (*a*) was close to one (Table 4S.3), the Weibull kernel resembled an exponential distribution, explaining the similar performance of these two kernel functions in *Virola* and *Cecropia*. It also suggests that pollen dispersal curves of these two species are slightly light tailed or exponential. Even lighter-tailed Weibull distributions, in which the shape parameter was greater than ours, were found in insect-pollinated tropical tree *Dinizia excelsa* and wind-pollinated temperate tree *Quercus lobata* (Austerlitz *et al.* 2004). But the best-fitted pollen dispersal kernel in *Tetragastris* and *Triplaris* was the lognormal, suggesting nearest mates might not have the highest mating success. In *Tetragastris*, individual trees may flower every two years (Croat 1978); this temporal mismatch could potentially cause lower mating probabilities between nearby individuals. With *Triplaris*, we suspect that some fecund pollen trees dominating the reproduction of seedlings (coefficient of variation in reproductive success, CV = 123%) could lead to higher mating probabilities certain distances away from female trees. Another possible explanation is that inbreeding depression may operate to reduce mating success between closely related nearby trees, as there is strong spatial genetic structure in adult trees of *Triplaris* (Wei *et al.* in review). On the other hand, the 2Dt kernel was unstable in fitting pollen dispersal in these species; even when convergence was achieved, it was still inferior (data not shown).

Gene dispersal by seeds vs. pollen in tropical trees and the implications

Trees in temperate forests are expected to have pollen-dominated gene dispersal, because airborne pollen travels substantially longer distances than do seeds by wind or

animals (Ouborg *et al.* 1999; Hamrick 2004; Petit *et al.* 2005; but see Bacles *et al.* 2006). However, wind pollination declines in frequency from temperate to tropical forests (Regal 1982); meanwhile, the opposite latitudinal gradient is found in seed dispersal by animals (Jordano 2000; Moles *et al.* 2007). In light of the sparse yet growing evidence of long-distance seed dispersal by vertebrate frugivores in the tropics (e.g. Sezen *et al.* 2005; Hardesty *et al.* 2006; Russo *et al.* 2006), one would anticipate an increased importance of seed-mediated gene dispersal in animal-dispersed tropical trees. Indeed, we found that, although pollen remains the primary means of gene dispersal (Fig. 4S.3), the ratio of pollen to seed-mediated gene dispersal (father-seedling vs. mother-seedling median distance) was less than 2 in *Virola* and *Cecropia* and 2.7 in *Tetragastris*, in contrast to the typical one order of magnitude in many temperate trees (Ouborg *et al.* 1999; Petit *et al.* 2005). One such study like ours by Hardesty *et al.* (2006) showed a comparable level (or a ratio of ~1) of on-plot effective seed vs. pollen-mediated gene dispersal in vertebrate-dispersed *Simarouba*. On the other hand, in wind-dispersed *Triplaris*, gene dispersal was mainly realized by pollen movement with the corresponding ratio being 7.4, similar to many temperate trees. The above inferences, nevertheless, did not consider immigrant gene dispersal. Although immigrant pollen is integrated in CSMi, it does not estimate pollen-mediated gene dispersal from father trees to seedlings, but only pollen dispersal from father trees to mother trees. The ratio of pollen to seed dispersal (father-mother vs. mother-seedling median distance), including immigrants, was 1.1 in *Virola* and 2.7 in *Cecropia* but 9.7 in *Tetragastris*. Given that diploid seeds carry twice as much as genetic material of pollen, we still expect the significance of seed-mediated gene dispersal in *Virola* and *Cecropia*.

With simulations realistically parameterized, we show that seed dispersal affects not only spatial patterns but also spatial genetic structure, population inbreeding and genetic diversity. In spite of substantial pollen flow (e.g. $m_p = 0.4$), constrained seed dispersal to the near tail in good dispersers (e.g. *Virola*) can lead to spatial clumping of relatives, which increases the odds of assortative mating and thus population inbreeding levels. In this situation, a non-negligible loss of genetic diversity is seen in the face of long-distance gene dispersal by pollen (Table 4.2). Yet, our simulations could underestimate the negative effects of shortened seed dispersal (SS vs. SC), because immigrant seed dispersal was unrealistically small ($m_s = 0.05$) in control conditions (SC) given the potential of long-distance seed dispersal in animal-dispersed tropical trees. Such scenario would predict the ecological and genetic consequences of ongoing and future decline in vertebrate dispersers due to factors such as hunting and habitat fragmentation. Again, our estimates could be an underestimate if impacted tree species are adapted to dispersal by a small assemblage of frugivores, the loss of which can potentially limit seed dispersal more than assumed (<100 m) in our short-distance scenario. On the other hand, poor dispersers (e.g. *Triplaris*) were scarcely affected by restraining seed dispersal to the near tail. Differential responses among species to changes in seed dispersal would generate community consequences (e.g. changes in composition and structure). Pollen dispersal, on the other hand, has a weaker effect on fine-scale spatial genetic patterns than does seed dispersal, as previously found at an ecological timescale (Wei *et al.* in review). The extent to which pollen dispersal affects population inbreeding is contingent upon spatial genetic aggregation set up by seed dispersal. For instance, in the cases of PS where pollen dispersal is confined between near neighbors, elevated population inbreeding was only pronounced in poor dispersers (e.g.

Triplaris) in which localized seed dispersal is common. Long-distance pollen dispersal is essential to prevent the loss of genetic diversity due to drift, as scenarios of PC maintained higher H_O than do scenarios of PS where $m_p = 0$.

To conclude, our results suggest that long-distance effective seed dispersal can be common in tropical trees. Despite the broad consensus of pollen-dominated gene dispersal in trees, there is an increased importance of gene dispersal by seeds in vertebrate-dispersed tropical trees. The near and far tail of seed and pollen dispersal kernel exhibit differential impacts on population genetic variation. Yet, these effects are examined on a fine spatial scale here. Future studies, extending to landscape scales with seed and pollen dispersal kernels realistically represented like the ones we quantified here, would potentially provide important perspectives into range-wide responses of tropical trees to environmental changes.

Acknowledgements

This chapter was coauthored with Marjolein Bruijning and Christopher Dick, and is in preparation for submission to journals. We thank the Smithsonian Tropical Research Institute and Center for Tropical Forest Science for facilitating fieldwork on Barro Colorado Island and providing a CTFS-ForestGEO grant to N.W. and C.W.D for the molecular laboratory work.

References

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.

Ashley MV (2010) Plant parentage, pollination, and dispersal: how DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences* **29**, 148-161.

Austerlitz F, Dick CW, Dutech C*, et al.* (2004) Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology* **13**, 937-954.

Bacles CFE, Lowe AJ, Ennos RA (2006) Effective seed dispersal across a fragmented landscape. *Science* **311**, 628-628.

Bawa KS, Bullock SH, Perry DR, Coville RE, Grayum MH (1985) Reproductive biology of tropical lowland rain forest trees. II. Pollination systems. *American Journal of Botany* **72**, 346-356.

Bawa KS, Opler PA (1975) Dioecism in tropical forest trees. *Evolution* **29**, 167-179.

Brokaw NL (1986) Seed dispersal, gap colonization, and the case of *Cecropia insignis*. In: *Frugivores and seed dispersal* (eds. Estrada A, Fleming T), pp. 323-331. Kluwer Academic Publishers, Dordrecht, Netherlands.

Cain ML, Milligan BG, Strand AE (2000) Long-distance seed dispersal in plant populations. *American Journal of Botany* **87**, 1217-1227.

Chase MR, Moller C, Kesseli R, Bawa KS (1996) Distant gene flow in tropical trees. *Nature* **383**, 398-399.

Chybicki IJ, Burczyk J (2010) Realized gene flow within mixed stands of *Quercus robur* L. and *Q. petraea* (Matt.) L. revealed at the stage of naturally established seedling. *Molecular Ecology* **19**, 2137-2151.

Clark J, Silman M, Kern R, Macklin E, HilleRisLambers J (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology* **80**, 1475-1494.

Clark JS, Fastie C, Hurtt G*, et al.* (1998) Reid's paradox of rapid plant migration. *BioScience* **48**, 13-24.

Comita LS, Aguilar S, Perez R, Lao S, Hubbell SP (2007) Patterns of woody plant species abundance and diversity in the seedling layer of a tropical forest. *Journal of Vegetation Science* **18**, 163-174.

Condit R (1998) *Tropical forest census plots* Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas, USA.

Condit R, Ashton PS, Baker P*, et al.* (2000) Spatial patterns in the distribution of tropical tree species. *Science* **288**, 1414-1418.

Condit R, Engelbrecht BMJ, Pino D, Perez R, Turner BL (2013) Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 5064-5068.

Connell JH (1971) On the role of natural enemies in preventing competitive exclusion in some marine animals and rain forest trees. In: *Dynamics of Populations* (eds. Boer PJD, Gradwell GR), pp. 298-312. Center for Agricultural Publication and Documentation, Wageningen.

Corlett RT, Westcott DA (2013) Will plant movements keep up with climate change? *Trends in Ecology & Evolution* **28**, 482-488.
Cortes MC, Uriarte M (2013) Integrating frugivory and animal movement: a review of the evidence and implications for scaling seed dispersal. *Biological Reviews* **88**, 255-272.
Croat TB (1978) *Flora of Barro Colorado Island* Stanford University Press, Stanford, California, USA.
Dick CW, Etchelecu G, Austerlitz F (2003) Pollen dispersal of tropical trees (*Dinizia excelsa*: Fabaceae) by native insects and African honeybees in pristine and fragmented Amazonian rainforest. *Molecular Ecology* **12**, 753-764.
Gaino APSC, Silva AM, Moraes MA*, et al.* (2010) Understanding the effects of isolation on seed and pollen flow, spatial genetic structure and effective population size of the dioecious tropical tree species *Myracrodruon urundeuva*. *Conservation Genetics* **11**, 1631-1643.
Ghazoul J (2005) Pollen and seed dispersal among dispersed plants. *Biological Reviews* **80**, 413-443.
Greene DF, Canham CD, Coates KD, Lepage PT (2004) An evaluation of alternative dispersal functions for trees. *Journal of Ecology* **92**, 758-766.
Grivet D, Smouse PE, Sork VL (2005) A novel approach to an old problem: tracking dispersed seeds. *Molecular Ecology* **14**, 3585-3595.
Hamrick JL (2004) Response of forest trees to global environmental changes. *Forest Ecology and Management* **197**, 323-335.
Hanson T, Brunsfeld S, Finegan B, Waits L (2007) Conventional and genetic measures of seed dispersal for *Dipteryx panamensis* (Fabaceae) in continuous and fragmented Costa Rican rain forest. *Journal of Tropical Ecology* **23**, 635-642.
Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters* **9**, 516-525.
Howe H (1989) Scatter-and clump-dispersal and seedling demography: hypothesis and implications. *Oecologia* **79**, 417-426.
Howe HF (1980) Monkey dispersal and waste of a neotropical fruit. *Ecology* **61**, 944-959.
Howe HF (1981) Dispersal of a Neotropical nutmeg (*Virola sebifera*) by birds. *The Auk* **98**, 88-98.
Howe HF (1993) Specialized and generalized dispersal systems: where does the paradigm stand? *Plant Ecology* **107**, 3-13.
Howe HF, Smallwood J (1982) Ecology of seed dispersal. *Annual Review of Ecology and Systematics* **13**, 201-228.
Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography* Princeton University Press, Princeton, NJ.
Hubbell SP, Foster RB, O'Brien ST*, et al.* (1999) Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science* **283**, 554-557.
Hufford KM, Hamrick JL, Rathbun SL (2009) Male reproductive success at three early life stages in the tropical tree *Platypodium elegans*. *International Journal of Plant Sciences* **170**, 724-734.

Jansen PA, Hirsch BT, Emsens WJ*, et al.* (2012) Thieving rodents as substitute dispersers of megafaunal seeds. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 12610-12615.
Janzen DH (1970) Herbivores and the number of tree species in tropical forests. *American Naturalist* **104**, 501-528.
Jones A (2010) Reconciling field observations of dispersal with estimates of gene flow. *Molecular Ecology* **19**, 4379-4382.
Jones F, Chen J, Weng G, Hubbell S (2005) A genetic evaluation of seed dispersal in the neotropical tree *Jacaranda copaia* (Bignoniaceae). *American Naturalist* **166**, 543-555.
Jones FA, Muller-Landau HC (2008) Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. *Journal of Ecology* **96**, 642-652.
Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* **10**, 551-555.
Jordano P (2000) Fruits and frugivory. In: *Seeds: The Ecology of Regeneration in Natural Plant Communities* (ed. Fenner M), pp. 125-166. CABI Publ., Oxon, UK.
Kays R, Jansen PA, Knecht EMH, Vohwinkel R, Wikelski M (2011) The effect of feeding time on dispersal of *Virola* seeds by toucans determined from GPS tracking and accelerometers. *Acta Oecologica-International Journal of Ecology* **37**, 625-631.
Klein EK, Lavigne C, Foueillassar X, Gouyon PH, Laredo C (2003) Corn pollen dispersal: Quasi-mechanistic models and field experiments. *Ecological Monographs* **73**, 131-150.
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.
LePage PT, Canham CD, Coates KD, Bartemucci P (2000) Seed abundance versus substrate limitation of seedling recruitment in northern temperate forests of British Columbia. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* **30**, 415-427.
Levin SA, Muller-Landau HC, Nathan R, Chave J (2003) The ecology and evolution of seed dispersal: A theoretical perspective. *Annual Review of Ecology Evolution and Systematics* **34**, 575-604.
Mallet JB (2001) Gene flow. In: *Insect movement: mechanisms and consequence* (eds. Woiwod IP, Reynolds DR, Thomas CD), pp. 337-360. CABI Publishing, CAB International, Oxon, UK.
Moles AT, Ackerly DD, Tweddle JC*, et al.* (2007) Global patterns in seed size. *Global Ecology and Biogeography* **16**, 109-116.
Moran EV, Clark JS (2011) Estimating seed and pollen movement in a monoecious plant: a hierarchical Bayesian approach integrating genetic and ecological data. *Molecular Ecology* **20**, 1248-1262.
Muller-Landau HC, Wright SJ, Calderon O, Condit R, Hubbell SP (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology* **96**, 653-667.
Nathan R, Horvitz N, He YP*, et al.* (2011) Spread of North American wind-dispersed trees in future environments. *Ecology Letters* **14**, 211-219.

Nathan R, Katul GG, Horn HS*, et al.* (2002) Mechanisms of long-distance dispersal of seeds by wind. *Nature* **418**, 409-413.
Nathan R, Schurr FM, Spiegel O*, et al.* (2008) Mechanisms of long-distance seed dispersal. *Trends in Ecology & Evolution* **23**, 638-647.
Ouborg NJ, Piquot Y, Van Groenendael JM (1999) Population genetics, molecular markers and the study of dispersal in plants. *Journal of Ecology* **87**, 551-568.
Petit RJ, Duminil J, Fineschi S*, et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* **14**, 689-701.
Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annual review of ecology, evolution, and systematics* **37**, 187-214.
Pretzsch H (2009) *Forest dynamics, growth and yield* Springer, Berlin, Heidelberg, Germany.
R Development Core Team (2013) *R: A language and environment for statistical computing, Version 3.0.3* R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.
Regal PJ (1982) Pollination by wind and animals: ecology of geographic patterns. *Annual Review of Ecology and Systematics* **13**, 497-524.
Ribbens E, Silander JA, Pacala SW (1994) Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. *Ecology* **75**, 1794-1806.
Robledo-Arnuncio JJ, Garcia C (2007) Estimation of the seed dispersal kernel from exact identification of source plants. *Molecular Ecology* **16**, 5098-5109.
Russo SE, Portnoy S, Augspurger CK (2006) Incorporating animal behavior into seed dispersal models: implications for seed shadows. *Ecology* **87**, 3160-3174.
Sezen UU, Chazdon RL, Holsinger KE (2005) Genetic consequences of tropical second-growth forest regeneration. *Science* **307**, 891-891.
Sezen UU, Chazdon RL, Holsinger KE (2009) Proximity is not a proxy for parentage in an animal-dispersed Neotropical canopy palm. *Proceedings of the Royal Society B: Biological Sciences* **276**, 2037-2044.
Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* **236**, 787-792.
Stacy EA, Hamrick JL, Nason JD*, et al.* (1996) Pollen dispersal in low-density populations of three Neotropical tree species. *American Naturalist* **148**, 275-298.
Tufto J, Engen S, Hindar K (1997) Stochastic dispersal processes in plant populations. *Theoretical Population Biology* **52**, 16-26.
Uriarte M, Canham CD, Thompson J, Zimmerman JK, Brokaw N (2005) Seedling recruitment in a hurricane-driven tropical forest: light limitation, density-dependence and the spatial distribution of parent trees. *Journal of Ecology* **93**, 291-304.
Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics* **181**, 1579-1594.
Wei N, Detto M, Dick CW (in review) Seed dispersal drives spatial genetic patterns in tropical trees.

Wenny DG, Levey DJ (1998) Directed seed dispersal by bellbirds in a tropical cloud forest. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6204-6207.
Westcott DA, Bentrupperbaumer J, Bradford MG, McKeown A (2005) Incorporating patterns of disperser behaviour into models of seed dispersal and its effects on estimated dispersal curves. *Oecologia* **146**, 57-67.
White GM, Boshier DH, Powell W (2002) Increased pollen flow counteracts fragmentation in a tropical dry forest: An example from *Swietenia humilis* Zuccarini. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 2038-2042.
Williams CG (2010) Long-distance pine pollen still germinates after meso-scale dispersal. *American Journal of Botany* **97**, 846-855.

Table 4.1 Median seed and pollen dispersal distance. The 95% credible intervals are given in the parentheses. GSMi: gene shadow model with immigrant integration; GSM: gene shadow model; SSMi: seed shadow model with immigrant integration; SSM: seed shadow model; CSMi: competing source model with immigrant integration. Model-based estimates of median seed and pollen dispersal distance were obtained with model-specific best-fitted kernel for each species.

Inference method	*Virola*	*Tetragastris*	*Cecropia*	*Triplaris*
Seed dispersal distance				
Parentage	129.5 (110.4, 148.6)	40.2 (25.3, 55.1)	108.4 (96.6, 120.3)	15.8 (9.7, 21.9)
GSMi	274.4 (237.3, 317.6)	42.6 (35.0, 51.8)	144.1 (118.8, 174.3)	25.4 (21.3, 30.3)
GSM	688.8 (193.8, 2428.3)	87.9 (51.6, 150.6)	225.6 (167.5, 304.3)	17.4 (15.4, 19.7)
SSMi	120.1 (87.7, 164.7)	14.5 (12.6, 16.6)	473.9 (471.1, 476.7)	18.3 (16.1, 20.8)
SSM	160.2 (82.8, 307.8)	14.5 (12.5, 16.7)	459.6 (455.9, 463.1)	18.3 (16.2, 20.8)
Pollen dispersal distance				
Parentage	176.9 (146.4, 207.4)	129.9 (97.2, 162.5)	145.9 (120.7, 171.0)	118.7 (103.8, 133.5)
CSMi	305.6 (255.5, 355.3)	411.7 (342.3, 494.6)	386.6 (338.8, 434.9)	360.1 (306.3, 422.9)

Table 4.2 Simulated genetic impacts of near vs. far-tail seed and pollen dispersal. Mean and standard error of three summary statistics are reported for simulations parameterized using observed *Virola* adult population as an example, in which the best-fitted lognormal seed dispersal kernel and Weibull pollen dispersal kernel were used. H_O: observed heterozygosity; F_{IS}, inbreeding coefficient; b, SGS intensity statistic. m_s, rate of immigrant seed dispersal; m_p, rate of immigrant pollen dispersal.

	Scenario	m_s	m_p	H_O		F_{IS}		b	
Seed dispersal	SS	0.05	0.05	0.680	(0.0016)	0.023	(0.0011)	-0.019	(0.0003)
	SC	0.05	0.05	0.699	(0.0014)	0.001	(0.0010)	-0.005	(0.0001)
	SL	0.05	0.05	0.703	(0.0015)	-0.002	(0.0009)	-0.001	(0.0001)
	SS	0.05	0.4	0.713	(0.0009)	0.007	(0.0009)	-0.013	(0.0002)
	SC	0.05	0.4	0.718	(0.0009)	-0.002	(0.0008)	-0.004	(0.0001)
	SL	0.05	0.4	0.720	(0.0010)	-0.005	(0.0010)	-0.001	(0.0001)
Pollen dispersal	PS	0.05	0	0.684	(0.0018)	0.001	(0.0011)	-0.005	(0.0001)
	PC	0.05	0.4	0.717	(0.0011)	-0.001	(0.0009)	-0.004	(0.0001)
	PL	0.05	0.4	0.719	(0.0010)	-0.003	(0.0009)	-0.004	(0.0001)

Scenarios of seed dispersal: SS, short-distance seed dispersal; SC, standard seed dispersal based on the best-fitted seed dispersal kernel; SL, long-distance seed dispersal (see the main text for details). Scenarios of pollen dispersal: PS, short-distance pollen dispersal; PC, standard pollen dispersal based on the best-fitted pollen dispersal kernel; PL, long-distance pollen dispersal (see the main text for details).

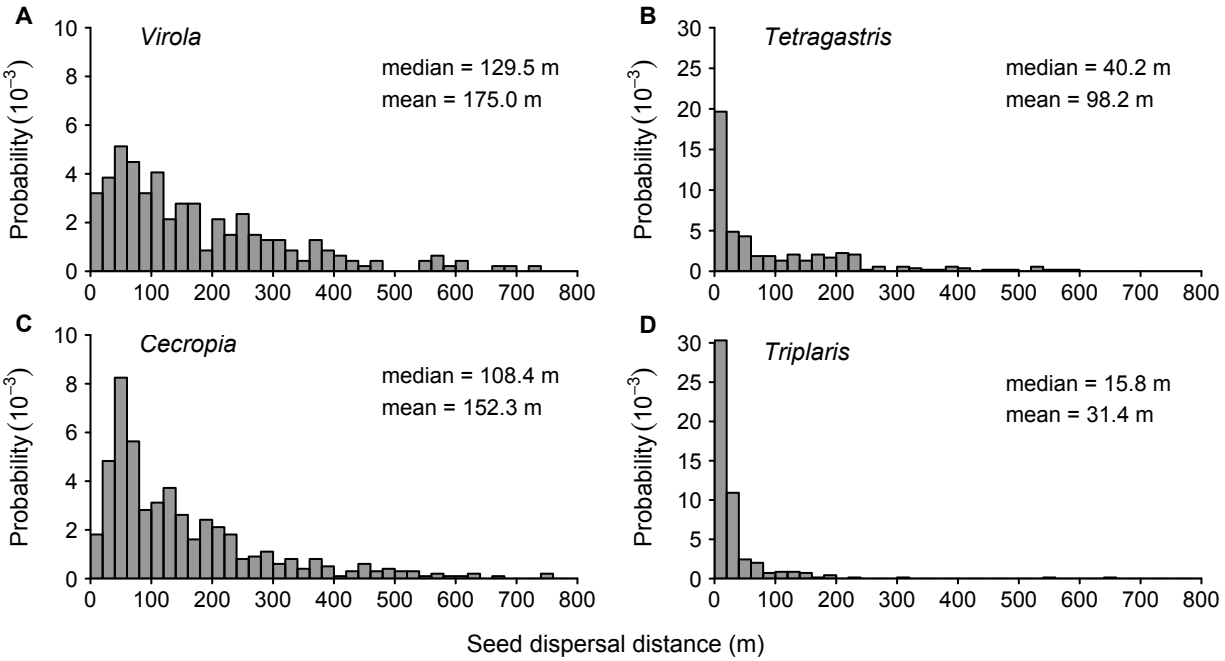

Figure 4.1 Frequency distribution of seed dispersal distances based on parentage inference. Both confident (≥80%) and conservative maternity assignments (see the main text) were used to estimate distances between seedlings and mother trees.

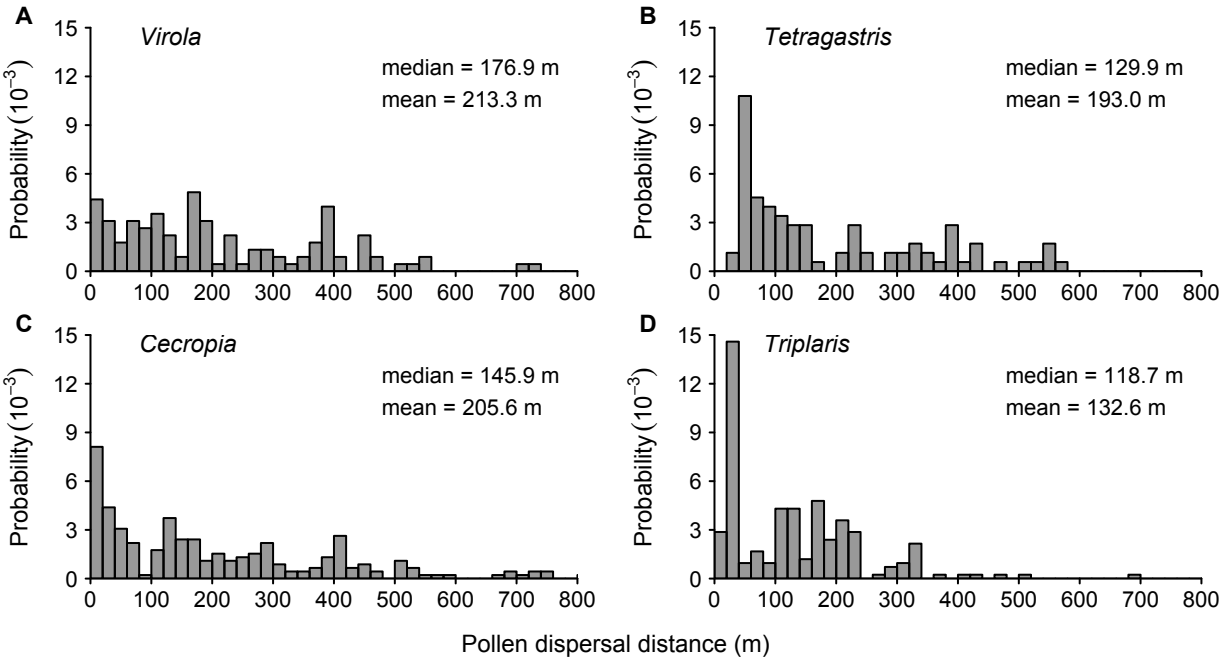

Figure 4.2 Frequency distribution of pollen dispersal distances based on parentage inference. Confident maternity and paternity of individual seedlings were used to estimate distances between mother trees and father trees.

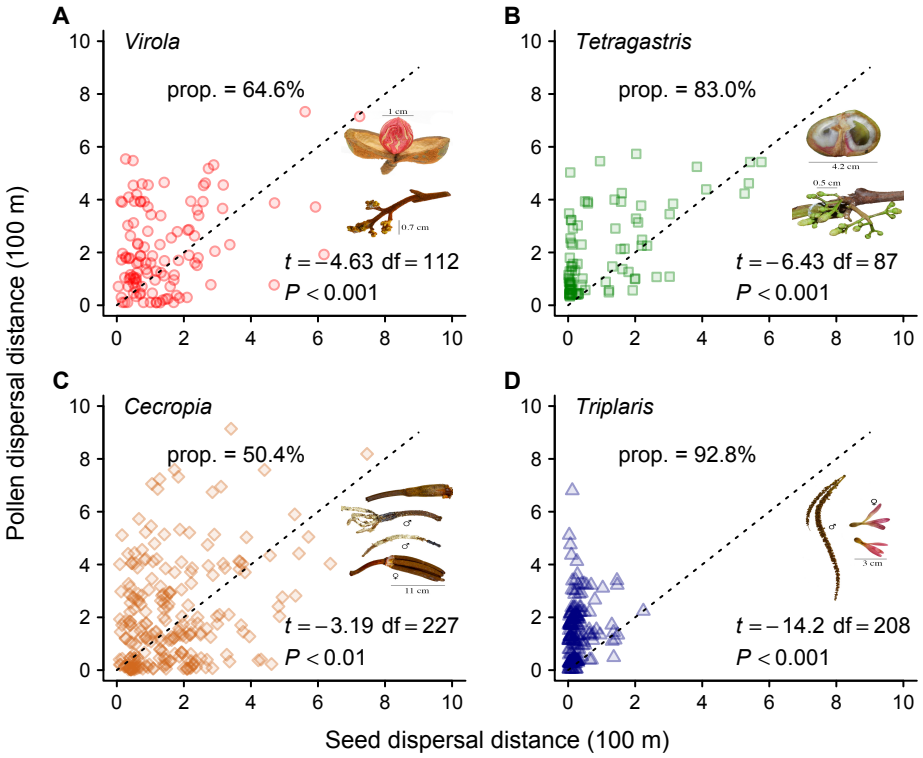

Figure 4.3 Comparisons between seed and pollen dispersal distances. Each data point represents a seedling, of which the mother tree and father tree were identified at a confidence of ≥80%. Dotted lines have a slope equal to 1. The proportion of seedlings, in which pollen dispersal exceeded seed dispersal, is indicated. Paired student *t*-tests were used to compare seed and pollen dispersal distances.

Appendix 4S.1 Genetic and non-genetic inverse models and two-dimensional (2D) seed and pollen dispersal kernels.

GSM and GSMi. In gene shadow models, the expected number of offspring of a focal female tree *i*, in a given area *a* centered at location *t*, $\hat{S}_{it}$, is determined by the tree fecundity F_i and distance to location *t*, d_{it}:

$$\hat{S}_{it} = aF_i p(d_{it}), \quad (1)$$

where $F_i = \exp(\gamma)(\mathrm{dbh}_i/2)^2\pi$, and $p(d_{it})$ is a two-dimensional probability density function. The observed seedlings of a female tree in each 1 × 1 m quadrat (a = 1 m^2), S_{it}, identified by maternity inference were used to derive the likelihood function of GSM over all quadrats or *t* for each female tree *i*, based on a Poisson error distribution:

$$\prod_i \prod_t Poisson(S_{it} \mid \hat{S}_{it}). \quad (2)$$

We also trialed negative binomial error distribution (Muller-Landau *et al.* 2008), with which however models frequently failed to converge. Thus only Poisson error distribution was used. GSM disregards the seedlings in each quadrat (or *t*) that come from outside the plot (i.e. immigrants), $\hat{I}_t$, which can be calculated by integrating over all *x* and *y* coordinates outside the plot:

$$\hat{I}_t = aD_s \iint_{Area\ offplot} p(d_t), \quad (3)$$

assuming seedling production per unit area, $D_s = \sum F_i/A$, consistent within and outside the plot (A = 50 ha). Therefore, the likelihood function of GSMi is defined as

$$\prod_i \prod_t Poisson(S_{it} \mid \hat{S}_{it}) \prod_t Poisson(I_t \mid \hat{I}_t), \quad (4)$$

where I_t is the observed number of immigrant seedlings in each quadrat.

SSM and SSMi. Different from gene shadow models that distinguish seedlings in each quadrat by their genetic sources, seed shadow models only consider the overall number of seedlings in each quadrat centered at location t, $\hat{S}_t$, as the summed reproductive inputs from female trees inside the plot under SSM, or from female trees inside and outside the plot under SSMi. The reproductive input of individual female tree i to location t, as in gene shadow models, depends on its fecundity, F_i, and distance to the location, d_{it}. Under SSM, $\hat{S}_t$ is assumed to come from only on-plot seed dispersal:

$$\hat{S}_t = a\sum\nolimits_i F_i p(d_{it}) . \quad (5)$$

But under SSMi, $\hat{S}_t$ consists of on-plot and off-plot progenies:

$$\hat{S}_t = a\sum\nolimits_i F_i p(d_{it}) + \hat{I}_t , \quad (6)$$

where the off-plot portion $\hat{I}_t$ is defined in equation (3). As seed shadow models require only the coordinates of seedlings and female trees and female tree dbh, we directly counted the observed number of seedlings, S_t, in each quadrat (1 × 1 m) for calculating the likelihood function over all quadrats or t for SSM and SSMi:

$$\prod_t Poisson(S_t \mid \hat{S}_t) . \quad (7)$$

CSMi. In competing source model, on- and off-plot male trees compete to pollinate the flowers of a focal seed tree i. The number of seedlings that an on-plot male tree j produces with seed tree i, $\hat{S}_{ij}$, is determined by the male tree fecundity (F_j) and its distance to seed tree i (d_{ij}):

$$\hat{S}_{ij} = F_j p(d_{ij}) , \quad (8)$$

where $F_j = (\text{dbh}_j/2)^2\pi$. The expected number of seedlings of seed tree i fathered by off-plot male trees, $\hat{I}_i$, follows:

$$\hat{I}_i = aD'_s \iint_{Area\ offplot} p(d_i), \tag{9}$$

where $D'_s = \sum F_j/A$ is assumed constant inside and outside the plot. Aided by maternity and paternity inference, we identified the observed number of seedlings of each seed tree i that were fathered by each on-plot male tree j, S_{ij}, and the observed number of seedlings of seed tree i without identified father within the plot, I_i. The likelihood function of CSMi uses a multinomial distribution,

$$\prod_i Multinomial(S_{i1}, S_{i2}, ..., S_{ij}, I_i \mid \hat{S}_{i1}, \hat{S}_{i2}, ..., \hat{S}_{ij}, \hat{I}_i). \tag{10}$$

In CSMi, female trees that did not produce any seedlings were excluded. In addition, CSMi did not make use of the seedlings that are produced by off-plot seed trees.

The 2D kernels used in this study were lognormal, gamma, Weibull and exponential distribution:

(1) Lognormal kernel

$$p(r) = \frac{1}{2\pi r^2 \sqrt{2\pi} b} e^{-\frac{(\ln r - a)^2}{2b^2}}$$

(2) Gamma kernel

$$p(r) = \frac{b^a}{2\pi r \Gamma(a)} r^{a-1} e^{-br}$$

(3) Weibull kernel:

$$p(r) = \frac{a}{2\pi r b} (\frac{r}{b})^{a-1} e^{-(\frac{r}{b})^a}$$

(4) Exponential kernel:

$$p(r) = \frac{a}{2\pi r} e^{-ar}$$

Appendix 4S.2 Simulation algorithms using different edge-correction methods.

Here we first described the simulation algorithm using reflection edge correction in the main text (algorithm 1). We then continued with the torus edge correction (algorithm 2) and the reflection edge correction that treated offspring genotypes differently (algorithm 3), relative to algorithm 1.

Simulation algorithm 1 using reflection edge correction (see the main text)

1. A female tree i, chosen at random, produces one offspring k that survives to become an adult in the focal population of size N.
2. If female tree i is from outside the plot (with a probability of m_s):
    1) the location (x_k, y_k) of offspring k is random inside the plot of 1000 × 500 m.
        x_k = runif (0, 1000, 1)
        y_k = runif (0, 500, 1)
    2) the genotype (G_k) of offspring k is from background (composed of population allele frequencies).
3. If female tree i is from inside the plot (with a probability of $1 - m_s$):
    i = sample (N, 1)
    1) the location (x_k, y_k) of offspring k is determined by seed dispersal distance d_s, according to specific seed model (SC, SL or SS), and a random angle θ from the location (x_i, y_i) of female tree i.
        $x_k = x_i + d_s \cos(\theta)$
        $y_k = y_i + d_s \sin(\theta)$
        If (x_k, y_k) is outside the plot, a ***reflection edge correction*** is applied.
        $x_k = 0 - x_k$, while $x_k < 0$; $x_k = 1000 - (x_k - 1000)$, while $x_k > 1000$
        $y_k = 0 - y_k$, while $y_k < 0$; $y_k = 500 - (y_k - 500)$, while $y_k > 500$
    2) the genotype (G_k) of offspring k comprises the maternal haplotype from i, according to Mendelian inheritance, and the paternal haplotype from pollen donor j:
        a. If pollen donor j is outside the plot (with a probability of m_p):
            paternal haplotype in k is from background
        b. If pollen donor j is inside the plot (with a probability of $1 - m_p$):
            j is chosen from on-plot male trees, according to specific pollen dispersal model (PC, PL or PS), that has the highest mating probability $p(d_{ij})$, where $p()$ is the best-fitted pollen dispersal kernel and d_{ij} is the distance to i. Paternal haplotype is from the male tree j, according to Mendelian inheritance
4. A dead adult tree l is chosen at random.
    l = sample (N, 1)

5. Offspring k takes the membership and sex status of dead tree l.

Repeating step 1–5 for $N \times 1000$ times (i.e. 1000 generations)

Simulation algorithm 2 using torus edge correction (as a comparison)

1. A female tree i, chosen at random from the adult population N, produces one offspring k that survives to become an adult.
   i = sample (N, 1)
2. The location (x_k, y_k) of offspring k is determined by seed dispersal distance d_s, according to specific seed model (SC, SL or SS), and a random angle θ from the location (x_i, y_i) of female tree i.
   $x_k = x_i + d_s \cos(\theta)$
   $y_k = y_i + d_s \sin(\theta)$
3. If (x_k, y_k) is outside the plot (1000 × 500 m):
   1) a ***torus edge correction*** is applied to the location of offspring k:
      $x_k = x_k + 1000$, while $x_k < 0$; $x_k = x_k - 1000$, while $x_k > 1000$
      $y_k = y_k + 500$, while $y_k < 0$; $y_k = y_k - 500$, while $y_k > 500$
   2) the genotype (G_k) of offspring k comprises a ***background maternal haplotype***, and the paternal haplotype from pollen donor j:
      a. if pollen donor j is from outside the plot (with a probability of m_p):
         paternal haplotype in k is from background
      b. If pollen donor j is inside the plot (with a probability of $1 - m_p$):
         j is chosen from on-plot male trees, according to specific pollen dispersal model (PC, PL or PS), that has the highest mating probability $p(d_{ij})$, where $p()$ is the best-fitted pollen dispersal kernel and d_{ij} is the distance to i. Paternal haplotype is from the male tree j, according to Mendelian inheritance
4. If (x_k, y_k) is inside the plot (1000 × 500 m):
   1) the location of offspring k follows step 2.
   2) the genotype (G_k) of offspring k comprises ***the maternal haplotype*** **from** ***i***, according to Mendelian inheritance, and the paternal haplotype from pollen donor j, following 2) in step 3.
5. A dead adult tree l is chosen at random.
   l = sample (N, 1)
6. Offspring k takes the membership and sex status of dead tree l.

Repeating step 1–6 for $N \times 1000$ times (i.e. 1000 generations)

Simulation algorithm 3 using reflection edge correction but different offspring genotype determination relative to algorithm 1

1. A female tree i, chosen at random from the adult population N, produces one offspring k that survives to become an adult.
   i = sample (N, 1)
2. The location (x_k, y_k) of offspring k is determined by seed dispersal distance d_s, according to specific seed model (SC, SL or SS), and a random angle θ from the location (x_i, y_i) of female tree i.
   $x_k = x_i + d_s \cos(\theta)$
   $y_k = y_i + d_s \sin(\theta)$
3. If (x_k, y_k) is outside the plot (1000 × 500 m):
   1) a ***reflection edge correction*** is applied to the location of offspring k:
      $x_k = 0 - x_k$, while $x_k < 0$; $x_k = 1000 - (x_k - 1000)$, while $x_k > 1000$
      $y_k = 0 - y_k$, while $y_k < 0$; $y_k = 500 - (y_k - 500)$, while $y_k > 500$
   2) the genotype (G_k) of offspring k comprises a ***background maternal haplotype***, and the paternal haplotype from pollen donor j:
      a. if pollen donor j is from outside the plot (with a probability of m_p):
         paternal haplotype in k is from background
      b. If pollen donor j is inside the plot (with a probability of $1 - m_p$):
         j is chosen from on-plot male trees, according to specific pollen dispersal model (PC, PL or PS), that has the highest mating probability $p(d_{ij})$, where $p()$ is the best-fitted pollen dispersal kernel and d_{ij} is the distance to i. Paternal haplotype is from the male tree j, according to Mendelian inheritance
4. If (x_k, y_k) is inside the plot (1000 × 500 m):
   1) the location of offspring k follows step 2.
   2) the genotype (G_k) of offspring k comprises ***the maternal haplotype*** **from** ***i***, according to Mendelian inheritance, and the paternal haplotype from pollen donor j, following 2) in step 3.
5. A dead adult tree l is chosen at random.
   l = sample (N, 1)
6. Offspring k takes the membership and sex status of dead tree l.

Repeating step 1–6 for $N \times 1000$ times (i.e. 1000 generations)

Table 4S.1 Genotyping error rates assumed in parentage inference using COLONY program. Error rate I: the rate of allelic dropout; Error rate II: the composite rate of all other genotyping errors. The magnitude of differences in error rates among loci was evaluated according to their relative ease in accurate allele scoring; for instance, suboptimal DNA or the presence of stuttering may result in a relatively higher genotyping error rate.

Virola[1]			*Tetragastris*[2]			*Cecropia*[3]			*Triplaris*[4]		
Locus	Error rate I	Error rate II	Locus	Error rate I	Error rate II	Locus	Error rate I	Error rate II	Locus	Error rate I	Error rate II
VSE11	0	0.05	Tpan014	0	0.01	CEC_08	0	0.0001	TRI_01	0.001	0.05
VSE30	0	0.05	Tpan015	0	0.01	CEC_10	0.01	0.01	TRI_09	0	0.005
VSE32	0	0.10	Tpan152	0	0.05	CEC_12	0	0.0001	TRI_20	0	0.005
VSE38	0	0.05	Tpan241	0	0.01	CEC_17	0	0.0001	TRI_27	0	0.01
VSE45	0	0.05	Tpan301	0	0.01	CEC_37	0	0.005	TRI_31	0.001	0.05
VSE55	0	0.01	Tpan321	0	0.01	CEC_43	0	0.05	TRI_40	0	0.01
VSE59[5]	0	0.10	Tpan441	0	0.05	CEC_45	0	0.005	TRI_45	0	0.01
VSE68[5]	0	0.05	Tpan681	0	0.10	CEC_46	0	0.005	TRI_49	0	0.01
VSE76[5]	0	0.10	Tpan882	0	0.01	CEC_56	0	0.005	TRI_55	0	0.01
			Tpan893	0	0.01	CEC_61	0	0.0001			
						CEC_64	0	0.0001			

[1]Wei *et al.* (2013); [2]Kenfack and Dick (2009); [3]Wei and Dick (2014a); [4]Wei and Dick (2014b); [5]Wei *et al.* (in review)

Table 4S.2 Estimated parameters of seed dispersal models. As the best-fitted seed dispersal kernel, based on the lowest AIC, was mostly lognormal distribution regardless of species identity, we reported the parameter estimates with lognormal seed dispersal kernel. The mean and 95% confidence interval of model parameters were included. GSMi: gene shadow model with immigrant integration; GSM: gene shadow model; SSMi: seed shadow model with immigrant integration; SSM: seed shadow model. See Appendix 4S.1 for kernel parameter *a* and *b*.

		Fecundity		Dispersal kernel (lognormal)				Dispersal distance	
Species	Model	γ		*a*		*b*		median	mean
Virola	GSMi	-7.980	(-8.083, -7.876)	5.615	(5.470, 5.760)	1.392	(1.262, 1.521)	274.4	371.7
	GSM	-7.480	(-8.137, -6.823)	6.535	(5.274, 7.796)	1.747	(1.276, 2.219)	688.8	405.0
	SSMi	-7.993	(-8.099, -7.888)	4.788	(4.472, 5.104)	1.073	(0.811, 1.335)	120.1	200.5
	SSM	-7.776	(-8.096, -7.456)	5.076	(4.420, 5.732)	1.164	(0.802, 1.526)	160.2	265.7
Tetragastris	GSMi	-7.523	(-7.643, -7.403)	3.752	(3.555, 3.949)	1.596	(1.464, 1.729)	42.6	120.7
	GSM	-7.300	(-7.484, -7.115)	4.477	(3.942, 5.011)	2.013	(1.729, 2.297)	87.9	215.1
	SSMi	-7.554	(-7.674, -7.435)	2.672	(2.530, 2.814)	1.041	(0.928, 1.153)	14.5	24.9
	SSM	-7.554	(-7.674, -7.435)	2.672	(2.530, 2.815)	1.041	(0.928, 1.154)	14.5	24.9
Cecropia	GSMi[1]	-8.340	(-8.429, -8.252)	1.523	(1.366, 1.680)	0.008	(0.007, 0.009)	144.1	180.2
	GSMi[2]	-8.369	(-8.458, -8.281)	4.817	(4.736, 4.899)	0.928	(0.872, 0.983)	123.7	186.5
	GSM	-8.008	(-8.187, -7.828)	5.419	(5.122, 5.716)	1.253	(1.097, 1.409)	225.6	340.0
	SSMi	-7.881	(-7.969, -7.793)	6.161	(6.154, 6.168)	0.015	(0.011, 0.020)	473.9	474.0
	SSM	-7.116	(-7.205, -7.026)	6.130	(6.122, 6.139)	0.041	(0.035, 0.048)	459.6	460.0
Triplaris	GSMi	-8.008	(-8.113, -7.903)	3.236	(3.059, 3.413)	1.500	(1.367, 1.633)	25.4	72.1
	GSM	-8.130	(-8.235, -8.024)	2.858	(2.735, 2.982)	1.103	(1.009, 1.196)	17.4	32.0
	SSMi	-8.079	(-8.181, -7.976)	2.907	(2.783, 3.032)	0.972	(0.874, 1.070)	18.3	29.4
	SSM	-8.079	(-8.181, -7.976)	2.908	(2.783, 3.033)	0.972	(0.873, 1.071)	18.3	29.4

Parameters of GSMi estimated with the best-fitted gamma kernel (AIC = 10737)[1] and with the lognormal kernel (AIC = 10742)[2].

Table 4S.3 Estimated parameters of pollen dispersal model CSMi. The mean and 95% credible interval of model parameters were reported for species-specific best-fitted pollen dispersal kernel(s). See Appendix 4S.1 for kernel parameter *a* and *b* (or *a* alone in exponential distribution).

Species	Best-fitted kernel	Dispersal kernel				Dispersal distance	
		a		*b*		median	mean
Virola[*]	Weibull	1.089	(0.888, 1.290)	428.0	(363.6, 492.4)	305.6	403.4
	Exponential	0.0023	(0.0020, 0.0027)			297.7	406.4
Tetragastris	Lognormal	6.020	(5.836, 6.205)	1.121	(0.983, 1.258)	411.7	473.7
Cecropia[*]	Weibull	1.082	(0.974, 1.189)	542.5	(477.6, 607.4)	386.6	486.1
	Exponential	0.0018	(0.0016, 0.0020)			391.1	490.3
Triplaris	Lognormal	5.886	(5.726, 6.047)	1.362	(1.251, 1.474)	360.1	417.8

[*]For *Virola* and *Cecropia*, Weibull and exponential kernel had similar AIC values ($\Delta AIC < 1$).

Table 4S.4 Simulated genetic impacts of near vs. far-tail seed and pollen dispersal. Simulations followed algorithm 1 in Appendix 4S.2. Mean and standard error of three summary statistics are indicated. Simulations were parameterized using observed information from (a) *Tetragastris* and (b) *Triplaris* adult population respectively. H_O: observed heterozygosity; F_{IS}, inbreeding coefficient; b, SGS intensity. m_s, rate of immigrant seed dispersal events; m_p, rate of immigrant pollen dispersal events.

		Scenario	m_s	m_p	H_O		F_{IS}		b	
(a)	Seed dispersal	SS	0.05	0.05	0.571	(0.0013)	0.008	(0.0010)	-0.014	(0.0003)
		SC	0.05	0.05	0.573	(0.0013)	0.004	(0.0010)	-0.011	(0.0002)
		SL	0.05	0.05	0.577	(0.0012)	-0.001	(0.0010)	-0.002	(0.0001)
		SS	0.05	0.4	0.588	(0.0008)	-0.001	(0.0009)	-0.010	(0.0002)
		SC	0.05	0.4	0.588	(0.0008)	0.000	(0.0009)	-0.009	(0.0001)
		SL	0.05	0.4	0.589	(0.0008)	-0.003	(0.0009)	-0.002	(0.0001)
	Pollen dispersal	PS	0.05	0	0.563	(0.0016)	0.007	(0.0011)	-0.011	(0.0002)
		PC	0.05	0.4	0.587	(0.0008)	0.000	(0.0009)	-0.009	(0.0001)
		PL	0.05	0.4	0.589	(0.0008)	-0.001	(0.0010)	-0.009	(0.0001)
(b)	Seed dispersal	SS	0.05	0.05	0.624	(0.0032)	0.046	(0.0027)	-0.034	(0.0011)
		SC	0.05	0.05	0.638	(0.0029)	0.036	(0.0020)	-0.031	(0.0008)
		SL	0.05	0.05	0.675	(0.0026)	0.001	(0.0016)	-0.008	(0.0003)
		SS	0.05	0.4	0.710	(0.0016)	0.006	(0.0017)	-0.019	(0.0005)
		SC	0.05	0.4	0.708	(0.0018)	0.008	(0.0017)	-0.021	(0.0005)
		SL	0.05	0.4	0.725	(0.0016)	-0.011	(0.0015)	-0.006	(0.0002)
	Pollen dispersal	PS	0.05	0	0.598	(0.0036)	0.035	(0.0022)	-0.029	(0.0007)
		PC	0.05	0.4	0.711	(0.0018)	0.005	(0.0016)	-0.021	(0.0004)
		PL	0.05	0.4	0.711	(0.0016)	0.006	(0.0015)	-0.020	(0.0004)

Scenarios of seed dispersal: SS, short-distance seed dispersal; SC, standard seed dispersal based on the best-fitted seed dispersal kernel; SL, long-distance seed dispersal (see the main text for details). Scenarios of pollen dispersal: PS, short-distance pollen dispersal; PC, standard pollen dispersal based on the best-fitted pollen dispersal kernel; PL, long-distance pollen dispersal (see the main text for details).

Table 4S.5 Simulated genetic impacts of near vs. far-tail seed and pollen dispersal. Simulations followed algorithm 2 in Appendix 4S.2. Mean and standard error of three genetic summary statistics are indicated. Simulations were parameterized using observed information from (a) *Virola*, (b) *Tetragastris* and (c) *Triplaris* adult population respectively. H_O: observed heterozygosity; F_{IS}, inbreeding coefficient; b, SGS intensity. m_s, rate of immigrant seed dispersal events; m_p, rate of immigrant pollen dispersal events.

		Scenario	m_p	H_O		F_{IS}		b	
(a)	Seed dispersal	SS	0.4	0.714	(0.0010)	0.004	(0.0009)	-0.011	(0.0002)
		SC	0.4	0.726	(0.0008)	-0.006	(0.0008)	-0.003	(0.0001)
		SL	0.4	0.727	(0.0008)	-0.006	(0.0009)	-0.0003	(0.0001)
	Pollen dispersal	PS	0	0.718	(0.0010)	-0.003	(0.0009)	-0.003	(0.0001)
		PC	0.4	0.722	(0.0009)	-0.002	(0.0009)	-0.003	(0.0001)
		PL	0.4	0.722	(0.0008)	-0.002	(0.0009)	-0.003	(0.0001)
(b)	Seed dispersal	SS	0.4	0.588	(0.0007)	-0.002	(0.0009)	-0.008	(0.0002)
		SC	0.4	0.590	(0.0007)	-0.001	(0.0009)	-0.009	(0.0001)
		SL	0.4	0.593	(0.0007)	-0.004	(0.0009)	-0.001	(0.0001)
	Pollen dispersal	PS	0	0.576	(0.0012)	0.000	(0.0011)	-0.010	(0.0002)
		PC	0.4	0.590	(0.0008)	-0.002	(0.0010)	-0.008	(0.0001)
		PL	0.4	0.589	(0.0007)	-0.001	(0.0008)	-0.009	(0.0001)
(c)	Seed dispersal	SS	0.4	0.709	(0.0017)	0.004	(0.0015)	-0.014	(0.0007)
		SC	0.4	0.714	(0.0016)	0.003	(0.0015)	-0.020	(0.0004)
		SL	0.4	0.730	(0.0014)	-0.010	(0.0015)	-0.003	(0.0002)
	Pollen dispersal	PS	0	0.608	(0.0036)	0.017	(0.0023)	-0.028	(0.0007)
		PC	0.4	0.711	(0.0016)	0.002	(0.0015)	-0.019	(0.0004)
		PL	0.4	0.716	(0.0015)	0.002	(0.0016)	-0.018	(0.0004)

Scenarios of seed dispersal: SS, short-distance seed dispersal; SC, standard seed dispersal based on the best-fitted seed dispersal kernel; SL, long-distance seed dispersal (see the main text for details). Scenarios of pollen dispersal: PS, short-distance pollen dispersal; PC, standard pollen dispersal based on the best-fitted pollen dispersal kernel; PL, long-distance pollen dispersal (see the main text for details).

Table 4S.6 Simulated genetic impacts of near vs. far-tail seed and pollen dispersal. Simulations followed algorithm 3 in Appendix 4S.2. Mean and standard error of three genetic summary statistics are indicated. Simulations were parameterized using observed information from (a) *Virola*, (b) *Tetragastris* and (c) *Triplaris* adult population respectively. H_O: observed heterozygosity; F_{IS}, inbreeding coefficient; b, SGS intensity statistic. m_s, rate of immigrant seed dispersal events; m_p, rate of immigrant pollen dispersal events.

		Scenario	m_p	H_O		F_{IS}		b	
(a)	Seed dispersal	SS	0.4	0.715	(0.0010)	0.002	(0.0009)	-0.008	(0.0002)
		SC	0.4	0.723	(0.0008)	-0.005	(0.0008)	-0.003	(0.0001)
		SL	0.4	0.724	(0.0007)	-0.005	(0.0008)	-0.001	(0.0001)
	Pollen dispersal	PS	0	0.720	(0.0010)	-0.004	(0.0009)	-0.004	(0.0001)
		PC	0.4	0.722	(0.0008)	-0.002	(0.0008)	-0.003	(0.0001)
		PL	0.4	0.725	(0.0008)	-0.004	(0.0008)	-0.003	(0.0001)
(b)	Seed dispersal	SS	0.4	0.589	(0.0008)	-0.003	(0.0010)	-0.006	(0.0003)
		SC	0.4	0.590	(0.0008)	-0.002	(0.0010)	-0.008	(0.0001)
		SL	0.4	0.592	(0.0007)	-0.003	(0.0009)	-0.001	(0.0001)
	Pollen dispersal	PS	0	0.573	(0.0013)	0.002	(0.0010)	-0.010	(0.0002)
		PC	0.4	0.590	(0.0007)	-0.002	(0.0009)	-0.008	(0.0001)
		PL	0.4	0.590	(0.0007)	-0.002	(0.0010)	-0.008	(0.0001)
(c)	Seed dispersal	SS	0.4	0.714	(0.0016)	-0.004	(0.0016)	-0.012	(0.0010)
		SC	0.4	0.713	(0.0016)	0.002	(0.0015)	-0.017	(0.0004)
		SL	0.4	0.732	(0.0014)	-0.011	(0.0014)	-0.003	(0.0002)
	Pollen dispersal	PS	0	0.612	(0.0036)	0.017	(0.0020)	-0.027	(0.0007)
		PC	0.4	0.714	(0.0016)	0.000	(0.0016)	-0.017	(0.0004)
		PL	0.4	0.715	(0.0015)	0.000	(0.0015)	-0.016	(0.0004)

Scenarios of seed dispersal: SS, short-distance seed dispersal; SC, standard seed dispersal based on the best-fitted seed dispersal kernel; SL, long-distance seed dispersal (see the main text for details). Scenarios of pollen dispersal: PS, short-distance pollen dispersal; PC, standard pollen dispersal based on the best-fitted pollen dispersal kernel; PL, long-distance pollen dispersal (see the main text for details).

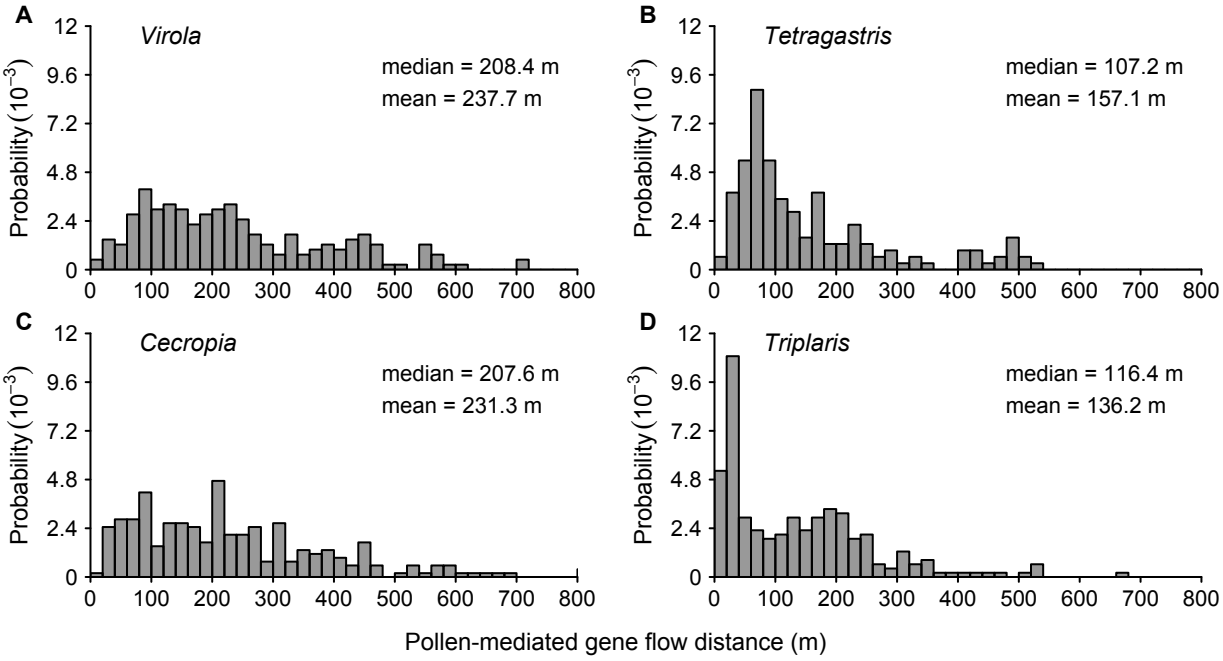

Figure 4S.1 Frequency distribution of pollen-mediated gene dispersal distances. Confident paternity of individual seedlings was used to estimate pollen-mediated gene dispersal from father trees to dispersed seedlings.

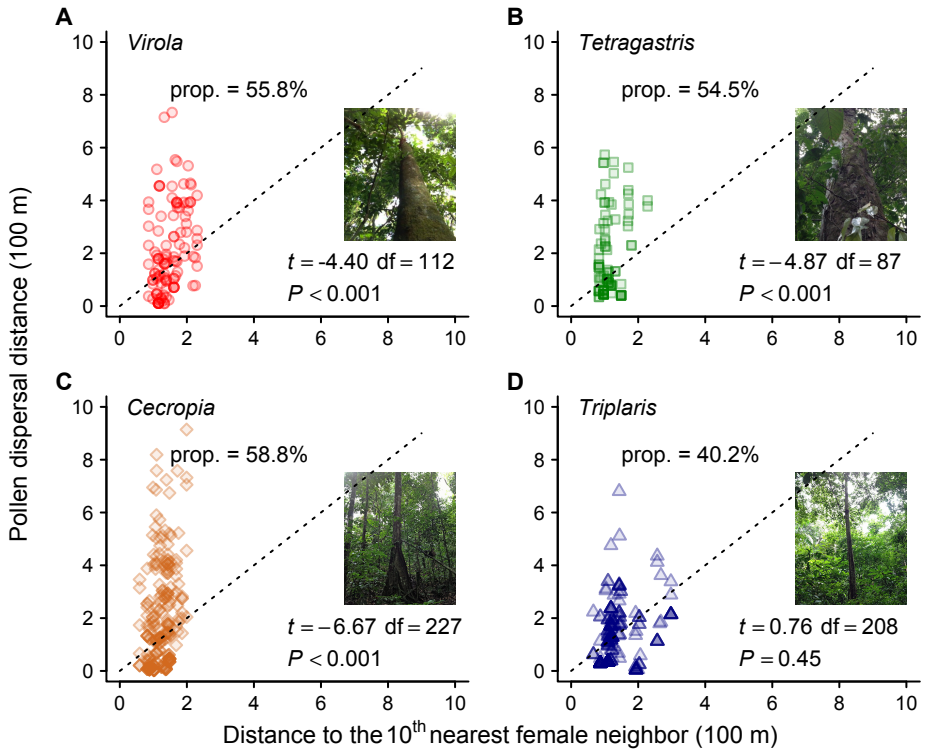

Figure 4S.2 Comparisons between distances of father trees to mother trees and distances to the tenth nearest female neighbors. Each data point represents a seedling, of which both father and mother tree were identified by parentage inference at a confidence of ≥80%. Dotted lines have a slope of 1. The proportion of seedlings, in which pollen dispersal was beyond the defined mating neighborhood size (i.e. the distance to the tenth nearest female neighbor), is indicated. Paired student *t*-tests were employed to compare pollen dispersal distance with near-neighbor distance.

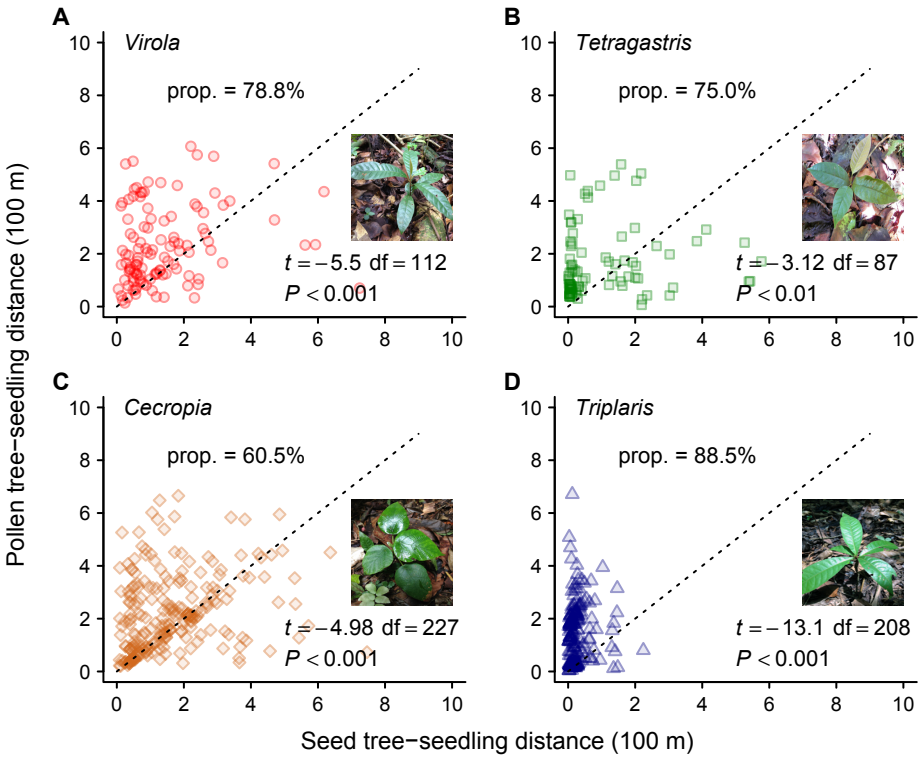

Figure 4S.3 Comparisons between seed and pollen-mediated gene dispersal. Seed-mediated gene dispersal is referred to as distances of dispersed seedlings to their mother trees; pollen-mediated gene dispersal is referred to as distances of dispersed seedlings to their father trees. Each data point represents a seedling, of which both mother tree and father tree were identified by parentage inference at a confidence of ≥80%. Dotted lines have a slope of 1. The proportion of seedlings, in which pollen-mediated gene dispersal was greater than seed-mediated gene dispersal, is indicated. Paired student *t*-tests were employed to compare seed with pollen-mediated gene dispersal.

References

Kenfack D, Dick CW (2009) Isolation and characterization of 15 polymorphic microsatellite loci in *Tetragastris panamensis* (Burseraceae), a widespread Neotropical forest tree. *Conservation Genetics Resources* **1**, 385.

Muller-Landau HC, Wright SJ, Calderon O, Condit R, Hubbell SP (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology* **96**, 653-667.

Wei N, Detto M, Dick CW (in review) Seed dispersal drives spatial genetic patterns in tropical trees.

Wei N, Dick CW (2014a) Characterization of twenty-six microsatellite markers for the tropical pioneer tree species *Cecropia insignis* Liebm (Urticaceae). *Conservation Genetics Resources* **6**, 987-989.

Wei N, Dick CW (2014b) Polymorphic microsatellite markers for a wind-dispersed tropical tree species, *Triplaris cumingiana* (Polygonaceae). *Applications in Plant Sciences* **2**, 1400051.

Wei N, Dick CW, Lowe AJ, Gardner MG (2013) Polymorphic microsatellite loci for *Virola sebifera* (Myristicaceae) derived from shotgun 454 pyrosequencing. *Applications in Plant Sciences* **1**, 1200295.

CHAPTER V

Conclusions

This dissertation research examines the ecological processes and genetic consequences of gene dispersal by seeds and pollen in ecologically important yet relatively under-studied tropical trees. On the basis of integrative approaches, including spatial genetic analysis, parentage inference, modeling and simulations, this dissertation has quantified (1) the respective importance of seed and pollen dispersal on governing spatial genetic patterns, (2) the spatial extent and magnitude of seed and pollen dispersal and (3) population responses to disruptions in seed and pollen dispersal processes, in four tropical lowland rainforest trees. This set of unrelated tropical tree species with varying life-history strategies and seed and pollen dispersal syndromes are representative of many tropical woody plants. Findings based on these species may, therefore, apply to broad taxonomic groups in tropical forests.

In chapter II, I focused on molecular marker development using next-generation sequencing for non-model organisms. The chapter serves as a practical guide of NGS-based marker development for ecological and evolutionary applications, by providing a bioinformatics workflow and a quantitative demonstration of maker development efficiency in relation to sequence attributes. Three different types of sequence simulations were conducted: read length simulations, sequencing error simulations and error trimming simulations. Firstly, read length simulations demonstrated the benefits of long reads for microsatellite maker development. The probability that a sequence contains microsatellites

increases with read length. Longer reads also enhance the ability of detecting problematic sequences that may impair maker amplification. In addition, read length is positively correlated with primer design success. For a given number of reads, longer read lengths are associated with higher microsatellite yields. Secondly, sequencing error simulations showed that marker amplification rate declines with increased sequencing error rate. A per-base error rate of 1% would predict that approximately 60% of the designed primers are amplifiable. Thirdly, error trimming simulations validated the need of quality control to improve marker amplification. Although quality control would shorten read length, the gained benefits outweigh the potential drawbacks. Lastly, combining simulations and the observed sequencing capacity, read length and error rate of the currently available NGS platforms, I found that MiSeq paired-end sequencing is cost efficient and is ideal for large microsatellite projects, whereas PacBio circular consensus sequencing is flexible in scaling sequencing effort and is suitable for fast, small-scale microsatellite projects.

In chapter III, using the microsatellite markers screened and validated in Appendix A–C, I examined the spatial genetic patterns of the four study species, and related the differences in SGS patterns among species to their differences in seed dispersal ability. This chapter focused on the SGS patterns in seedlings banks rather than in adult trees, because post-dispersal mortality-causing processes can modify initial SGS patterns at early life stages and thus play an important role in adult tree SGS, which makes the inference of the effects of gene dispersal challenging. Observed seedling SGS intensity in wind-dispersed *Triplaris cumingiana* and monkey-dispersed *Tetragastris panamensis* was three to four times higher than that in avian-dispersed *Virola sebifera* and *Cecropia insignis*, in which seed dispersal distance might be considerably longer.

An individual-based spatially explicit simulation model was developed here to distinguish the respective effects of seed and pollen dispersal on seedling SGS on the basis of approximate Bayesian computation (ABC). The simulation model was parameterized with species-specific information (e.g. adult population density, spatial structure), but allowed key parameters, that is, seed dispersal, pollen dispersal and individual reproductive success, to change in order to generate seedling SGS that resembles the observed in each species. The best parameter combinations can be used to gauge the relative contributions of seed vs. pollen dispersal to SGS in these species, independent from other confounding factors.

The results showed that the ABC-inferred seed dispersal distances were indeed longer in *V. sebifera* and *C. insignis*, in which seedling SGS intensity was weaker. The median distances of seed dispersal in these two species were four to five times as long as those in *T. panamensis* and *T. cumingiana*. However, the ABC inference of pollen dispersal distance from seedling SGS was less accurate than that of seed dispersal distance. Several potential factors may cause this lack of confidence in inferring pollen dispersal distance from resultant SGS patterns. One possible reason was that the simulation model in use was wrong. To test this 'wrong model' hypothesis, alternative models involving more complex scenarios of pollination events were used; however, model selection still favored the original, simpler model. Another possible reason was that ABC method may have low estimation efficacy; yet, this hypothesis was further rejected, as other processes (i.e. seed dispersal and female tree reproductive variation) could be accurately inferred from SGS patterns. Lastly, the lack of confidence in inferring pollen dispersal was likely because pollen dispersal process may have a weak impact on seedling SGS in these species. To test

this hypothesis, I examined how seedling SGS responds to seed vs. pollen dispersal distance using simulations.

The simulations suggest that only when pollen dispersal is limited between nearby individuals of opposite sex, it has a pronounced effect on seedling SGS. Beyond this local mating neighborhood, seedling SGS intensity declines very slowly with increased pollen dispersal distance. A similar effect of gene flow on genetic structure is predicted theoretically by Wright's (1943) island model. At evolutionary equilibrium, once gene flow (i.e. the effective number of migrants per generation) exceeds a threshold of one, genetic differentiation F_{ST} declines slowly with increasing gene flow, due to the nonlinear relationship between F_{ST} and gene flow (Nm), $F_{ST} = 1/(4Nm + 1)$, where N is effective population size and m is the rate of migration per generation (Whitlock & McCauley 1999; Templeton 2006).

The results suggest that the inference accuracy of one process depends on how large of an effect that process has on SGS. In the cases where pollen dispersal strongly influences seedling SGS, the results showed that it can be accurately inferred. However, it by no means implies that pollen dispersal is not important at an ecological timescale (this chapter) or at an evolutionary timescale (chapter IV). In the next chapter (IV), I showed that long-distance pollen dispersal into the focal population is essential for preventing genetic diversity loss due to drift, despite that it does not affect fine-scale SGS as much as does seed dispersal at an evolutionary timescale. On the other hand, the low accuracy of pollen dispersal inference from seedling SGS here also suggests that pollination exceeded the near neighbors in these species.

A caveat to this study is that a closed system was used in the simulation model. By doing this, ABC-inferred seed dispersal distance was an underestimate. In fact, the median seed dispersal distance inferred from seedling SGS corresponded well with that of the local seed dispersal events estimated in chapter IV. Despite the caveat, this approach provides a powerful way to disentangle the effects of seed vs. pollen dispersal on determining SGS at fine spatial scales. It is particularly useful for understanding the correlates of SGS in a comparative context. Overall, this chapter provided compelling evidence for the dominant role of contemporary seed dispersal in governing SGS in these tropical trees.

In chapter IV, I quantified the magnitude of long-distance seed dispersal and evaluated the hypothesis of an increased importance of gene dispersal by seeds relative to by pollen in vertebrate-dispersed tropical tree species. There is a broad consensus that pollen movement is the dominant manner of gene dispersal in temperate forest trees (Ennos 1994; Ouborg *et al.* 1999; Petit *et al.* 2005), because airborne pollen travels much greater distances than do seeds. However, the latitudinal gradient of biotic interactions reveals increased incidences of pollination and seed dispersal by animals in tropical forests (Regal 1982; Jordano 2000; Schemske *et al.* 2009), which may suggest a different pattern of seed vs. pollen-mediated gene dispersal in tropical taxa (Sezen *et al.* 2005; Hardesty *et al.* 2006).

Parentage inferences between 1518 seedlings and 789 adult trees suggest frequent seed deposition and establishment over 100 m in avian-dispersed *V. sebifera* and *C. insignis* and in monkey-dispersed *T. panamensis*. Seed dispersal distances in these biotically dispersed species were substantially longer than in wind-dispersed *T. cumingiana*, as well as in other wind-dispersed tropical (e.g. Jones & Muller-Landau 2008) and temperate tree species (e.g. Clark *et al.* 1999). The proportion of seed dispersal events exceeding 1 km

was typically above 1% in these animal-dispersed species, and could reach as high as 18% in *V. sebifera*.

Seed dispersal kernel–the probability density function of dispersal distances from source trees (Ribbens *et al.* 1994)–is an important component in modeling applications, such as plant migration and vegetation change in response to climate change (Clark 1998; Corlett & Westcott 2013). Using non-genetic and genetic inverse modeling (Ribbens *et al.* 1994; Jones & Muller-Landau 2008; Muller-Landau *et al.* 2008), I found that the best-fitting seed dispersal kernels principally belonged to the lognormal family. The lognormal kernel is heavy tailed, suggesting the potential of long-distance seed dispersal events. Although the commonly used 2Dt kernel (Clark *et al.* 1999) is also heavy tailed, it places the peak at the center of the bole, and thus does not reflect the patterns of effective seed dispersal in these species, in which density and/or distance-dependent processes may be involved to reduce offspring survival near the bole.

In line with the findings of long-distance pollen dispersal, primarily in the context of habitat fragmentation, in insect-pollinated tropical trees (reviewed in Ashley 2010), the results here in a continuous tropical forest on BCI (Panama) showed that approximately 50% of pollen flow came from outside of the 50-ha Forest Dynamics Plot. By integrating the immigrant pollen flow that cannot be estimated by parentage inferences, the models predicted that median pollen dispersal distances reach 300–400 m and 10–20% of pollen dispersal events could exceed 1 km in both wind and insect-pollinated species in this dissertation. One should note, however, that these estimates only reflect the potential not the actual pollination distances. This is because successful pollination depends on the availability of individuals of opposite sex at the specified distances (Meagher & Vassiliadis

2003). Therefore, despite that pollen can potentially travel long distances by wind or by insects, population distribution determines whether these long-distance events could translate into effective pollination in tropical trees.

Overall, the results agreed with the broad consensus of pollen-dominated gene dispersal in forest trees. However, an increased importance of gene dispersal by seeds was indeed found in tropical species that are dispersed by highly mobile frugivores (e.g. big birds). Simulations provided strong evidence that potential disruptions to seed and pollen dispersal processes can have long-term negative effects on the ecological and evolutionary dynamics of tree populations, such as intensified spatial and genetic aggregation, elevated inbreeding and diminished genetic diversity.

Implications for management

This dissertation research provides quantitative insights into how far seeds and pollen can move and their respective genetic impacts in relatively under-studied tropical tree species. Quantifying the rates of seed and pollen dispersal represents the first step towards our understanding of the potential responses of tropical trees to rapid environmental changes. Determining the respective genetic effects of seed and pollen dispersal in natural populations is essential for predicting the short-term and long-term evolutionary responses to increasingly intensive anthropogenic disturbance (e.g. hunting, fragmentation) in tropical rain forests.

Genetic measures of seed dispersal distance in these tropical tree species and others (Sezen *et al.* 2005; Hardesty *et al.* 2006; Ashley 2010) provide compelling evidence of long-distance seed dispersal in animal-dispersed tropical trees. Heavy-tailed seed dispersal,

allowing >10% of dispersal events above 100 m, was demonstrated to be able to explain rapid migration of trees (100–1000 m/yr) following postglacial warming (Clark 1998). Relative to these temperate trees (Clark 1998; Clark *et al.* 1999), the rates of contemporary seed dispersal in these frugivore-dispersed tropical tree species are even higher (chapter IV). However, whether these rates of seed dispersal are sufficient to allow tropical trees to cope with ongoing climate change is unclear. This is because (1) ongoing climate change occurs at a much faster rate than did in the past (Petit *et al.* 2008; Corlett & Westcott 2013) and (2) the rates of seed dispersal here in a faunally intact tropical forest may not represent the rates in disturbed (e.g. overhunted or fragmented) regions.

Increasingly intensive human disturbance has been seen in tropical rain forests (Redford 1992; Wright 2005). Hunting of seed-dispersing animals, for instance, has negatively influenced seed dispersal, population dynamics and community composition of tropical trees (Wright *et al.* 2007; Terborgh *et al.* 2008; Markl *et al.* 2012; Harrison *et al.* 2013). As seed dispersal governs spatial genetic structure (chapter III), reduced seed dispersal due to anthropogenic loss of frugivores would incur adverse effects on the evolutionary dynamics of tropical tree populations (chapter IV). The outcome could become worse if pollen dispersal process is also disrupted (chapter IV), potentially as a result of habitat fragmentation (Aguilar *et al.* 2006). To maintain the ecological and evolutionary dynamics of tree species in tropical rain forests, increased efforts are required to ensure seed and pollen dispersal processes mediated by mutualistic animal partners.

References

Aguilar R, Ashworth L, Galetto L, Aizen MA (2006) Plant reproductive susceptibility to habitat fragmentation: review and synthesis through a meta-analysis. *Ecology Letters* **9**, 968-980.

Ashley MV (2010) Plant parentage, pollination, and dispersal: how DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences* **29**, 148-161.

Clark J, Silman M, Kern R, Macklin E, HilleRisLambers J (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology* **80**, 1475-1494.

Clark JS (1998) Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. *American Naturalist* **152**, 204-224.

Corlett RT, Westcott DA (2013) Will plant movements keep up with climate change? *Trends in Ecology & Evolution* **28**, 482-488.

Ennos RA (1994) Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72**, 250-259.

Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters* **9**, 516-525.

Harrison RD, Tan S, Plotkin JB*, et al.* (2013) Consequences of defaunation for a tropical tree community. *Ecology Letters* **16**, 687-694.

Jones FA, Muller-Landau HC (2008) Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. *Journal of Ecology* **96**, 642-652.

Jordano P (2000) Fruits and frugivory. In: *Seeds: The Ecology of Regeneration in Natural Plant Communities* (ed. Fenner M), pp. 125-166. CABI Publ., Oxon, UK.

Markl JS, Schleuning M, Forget PM*, et al.* (2012) Meta-analysis of the effects of human disturbance on seed dispersal by animals. *Conservation Biology* **26**, 1072-1081.

Meagher TR, Vassiliadis C (2003) Spatial geometry determines gene flow in plant populations. In: *Genes in environment* (eds. Hails RS, Beringer JE, Godfray HCJ), pp. 76-90. Blackwell Publishers, Malden, Massachusetts, USA.

Muller-Landau HC, Wright SJ, Calderon O, Condit R, Hubbell SP (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology* **96**, 653-667.

Ouborg NJ, Piquot Y, Van Groenendael JM (1999) Population genetics, molecular markers and the study of dispersal in plants. *Journal of Ecology* **87**, 551-568.

Petit RJ, Duminil J, Fineschi S*, et al.* (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Molecular Ecology* **14**, 689-701.

Petit RJ, Hu FS, Dick CW (2008) Forests of the past: A window to future changes. *Science* **320**, 1450-1452.

Redford KH (1992) The empty forest. *BioScience* **42**, 412-422.

Regal PJ (1982) Pollination by wind and animals: ecology of geographic patterns. *Annual Review of Ecology and Systematics* **13**, 497-524.

Ribbens E, Silander JA, Pacala SW (1994) Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersion. *Ecology* **75**, 1794-1806.

Schemske DW, Mittelbach GG, Cornell HV, Sobel JM, Roy K (2009) Is there a latitudinal gradient in the importance of biotic interactions? *Annual Review of Ecology Evolution and Systematics* **40**, 245-269.
Sezen UU, Chazdon RL, Holsinger KE (2005) Genetic consequences of tropical second-growth forest regeneration. *Science* **307**, 891-891.
Templeton AR (2006) *Population genetics and microevolutionary theory* Wiley-Liss, Hoboken, New Jersey, USA.
Terborgh J, Nunez-Iturri G, Pitman NC*, et al.* (2008) Tree recruitment in an empty forest. *Ecology* **89**, 1757-1768.
Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: FST ≠ 1/(4Nm+1). *Heredity* **82**, 117-125.
Wright S (1943) Isolation by distance. *Genetics* **28**, 114-138.
Wright SJ (2005) Tropical forests in a changing environment. *Trends in Ecology & Evolution* **20**, 553-560.
Wright SJ, Hernandez A, Condit R (2007) The bushmeat harvest alters seedling banks by favoring lianas, large seeds, and seeds dispersed by bats, birds, and wind. *Biotropica* **39**, 363-371.