AN ABSTRACT OF THE DISSERTATION OF

Gancho Trifonov Slavov for the degree of Doctor of Philosophy in Forest Science presented on May 20, 2004.
Title: Development and Application of SSR Markers for Measuring Gene Flow in Douglas-fir.

Abstract approved:

Signature redacted for privacy. Signature redacted for privacy.

Wesley T. Adams, Steven H. Strauss, Glenn T. Howe

Gene flow is a major evolutionary force and an important factor in the breeding and conservation of forest trees. I studied the applicability of SSR markers for measuring pollen-mediated gene flow (i.e., pollen flow) in Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco). I developed SSR markers, tested alternative approaches for measuring pollen flow using SSR markers, then measured pollen contamination and characterized within-block mating patterns in one block of a seed orchard complex.

Useful markers were developed from 4.1% of the SSR sequences screened. The 22 markers obtained are among the most informative genetic markers available for Douglas-fir. The observed heterozygosity and the number of alleles per marker averaged 0.855 (SE=0.020) and 23 (SE=1.6), respectively.

Mistyping (i.e., false identification of genotypes) results in overestimating pollen flow. Requiring multiple mismatches for paternity exclusion, while assuring that the probability of detecting immigrant genotypes is high, results in accurate estimates of pollen flow. I developed and made available the Pollen Flow (PFL) computer program,

which performs paternity exclusion and measures pollen flow based on multiple father-offspring mismatches.

Pollen contamination was consistently high in all three years in which seed crops were sampled from the orchard block (mean = 35.3%). Levels of pollen contamination varied substantially among clones, and were significantly higher in clones with early female receptivity (mean = 55.5%) than in those with intermediate (mean = 36.4%) or late (mean = 28.3%) female receptivity. Seeds resulting from self-fertilization were rare (mean = 1.8%). Differences in the relative paternal contributions of the clones in the block were greater than ten-fold, and there was preferential mating among parents with similar floral phenology.

Information from analyses of SSR data can be used to minimize pollen contamination and improve within-orchard mating patterns. Furthermore, SSRs can be used to advance knowledge of gene flow in natural populations. The availability of large sets of highly variable SSRs makes it possible to perform landscape-scale studies of gene flow and better understand the interactions between gene flow and adaptation. These studies will ultimately provide a basis for decisions in breeding and conservation programs.

Development and Application of SSR Markers for Measuring Gene Flow in Douglas-fir

by

Gancho Trifonov Slavov

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Doctor of Philosophy

Presented May 20, 2004

Commencement June 2005

Doctor of Philosophy thesis of Gancho Trifonov Slavov presented on May 20, 2004.

APPROVED:

Signature redacted for privacy.

Co-Major Professor, representing Forest Science

Signature redacted for privacy.

Co-Major Professor, representing Forest Science

Signature redacted for privacy.

Co-Major Professor, representing Forest Science

Signature redacted for privacy.

Head of the Department of Forest Science

Signature redacted for privacy.

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Signature redacted for privacy.

Gancho Trifonov Slavov, Author

ACKNOWLEDGEMENTS

My interactions with a number of people turned working on this dissertation into the most intellectually enriching process that I have experienced. I am greatly indebted to my major professors Tom Adams, Steve Strauss, and Glenn Howe for giving me the chance to work on this project, helping me overcome its challenges, and providing invaluable advice and example for my development as a scientist. Dave Birkes readily suggested solutions to my statistical dilemmas. The thought-provoking questions of Barbara Gartner and Aaron Liston stimulated me to seek understanding of my research in a broad biological context and improve my skills to communicate with scientists from different fields.

Steve DiFazio has been a primary consultant in every phase of my work on this thesis, as well as an inspiring peer and a good friend. Kostya Krutovskii helped me get oriented and learn basic laboratory skills. Jim Smith and other staff members of the Plum Creek Timber Company provided assistance with various field aspects of the project. Christine Lomas, Nuray Kaya, Gokcin Temel, and Santiago González-Martínez all had contributions to the project prior to my arrival. Igor Yakovlev and Jacob Eccles provided laboratory help with developing SSR markers and fingerprinting seed orchard clones. Marilyn Cherry and Joanna Warren helped with collecting tissue samples and phenology data.

This study was funded by the Pacific Northwest Tree Improvement Research Cooperative and was conducted in a seed orchard owned by the Plum Creek Timber Company. I was also supported by an Oregon Sports Lottery Scholarship awarded by Oregon State University, an Alfred N. Moltke Memorial Fellowship and a Mary

McDonald Fellowship provided by the College of Forestry, a Jack Morgan Fellowship, a Henry and Mildred Fowells Fellowship, and an Outstanding PhD Student Award provided by the Department of Forest Science. The Tree Genetics and Biosafety Research Cooperative (formerly Tree Genetic Engineering Research Cooperative) provided space, materials, and logistical support for my laboratory work.

Over the last four years, I received a great deal of moral support from many good acquaintances in Corvallis, and especially from my friends Vicky and Jeff Hollenbeck. Finally, I owe everything that I have accomplished while working on this dissertation to my parents Margarita and Trifon, my brother Slavi, my aunt Penka, and to Aglika, the most important person in my life.

CONTRIBUTIONS OF AUTHORS

Chapter 2

Drs. Gerald Tuskan, Keith Edwards, and John Carlson provided materials, equipment, and logistical support for the development of SSR markers. Dr. Igor Yakovlev was actively involved in screening candidate SSR markers. Dr. Konstantin Krutovskii provided DNA samples and analyzed the data needed to determine the map locations of the SSR markers.

Chapter 3

Dr. David Birkes assisted with designing the simulation program and Aglika Gyaourova translated parts of the data analysis procedure into C code.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX TABLES

To Dr. Peter Zhelev, in recognition of his mentorship and passion for science.

DEVELOPMENT AND APPLICATION OF SSR MARKERS FOR MEASURING GENE FLOW IN DOUGLAS-FIR

Chapter 1. Introduction

RATIONALE

Background

Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco) is a tree species of major ecological and economic importance in North America, and one of the most important exotics grown in Europe, Chile, New Zealand, and Australia (Stein and Owston 2002). Large-scale tree improvement programs have been operating in Oregon and Washington since the 1960s and over 800 ha of Douglas-fir seed orchards have been established in northern California, Oregon, Washington, and British Columbia (Adams et al. 1990). The genetic efficiency of a seed orchard is the degree to which seeds collected from the orchard represent the genetic superiority and diversity of the orchard parents (Friedman and Adams 1982). The genetic efficiency of seed orchards may be adversely affected by pollen contamination, self-fertilization, and unequal representation of orchard parents in seed crops.

Pollen contamination is the pollination of seed orchard parents by pollen from outside of the seed orchard. In seed orchard complexes (i.e., seed orchards with two or more blocks), contaminant pollen in each orchard block can come from (1) natural stands or plantations in the surrounding area or (2) other orchard blocks containing parents from different breeding zones. Pollen contamination reduces the genetic worth of seed orchard crops and may adversely affect the adaptability of the resulting seedlings (Squillace and

Long 1981; Friedman and Adams 1985; Adams and Burczyk 2000; Kang et al. 2001). For example, pollen contamination from natural (i.e., unimproved) stands with a magnitude of 50% is expected to reduce genetic gains by 25%. Few studies have demonstrated that pollen contamination causes maladaptation (Stoehr et al. 1994 and references therein), but maladaptation caused by pollen contamination is a serious concern because seed orchard complexes often consist of blocks that contain parents from different breeding zones. Furthermore, seed orchards are often established far away from these zones. For example, pollen contamination in southern seed orchards consisting of parents from northern latitudes can result in maladaptation of the seed orchard crops to northern growing conditions (Kylmänen 1980; Nikkanen 1982). Previous studies based on allozyme genetic markers demonstrated that pollen contamination in non-isolated, open-pollinated conifer seed orchards (such as most seed orchards of Douglas-fir) often exceeds 30-40% (Adams and Burczyk 2000). Some of these studies also indicated that pollen contamination can vary among trees with different floral phenologies, and that overall levels of pollen contamination can be reduced by applying pollen management techniques (e.g., Wheeler and Jech 1986; El-Kassaby and Ritland 1986b). Quantifying the success of pollen management techniques, however, has been hampered by the low precision of pollen contamination estimates (Adams and Burczyk 2000).

Another factor that reduces the genetic efficiency of seed orchards is self-fertilization. In conifers, self-fertilization generally leads to severe inbreeding depression (Sorensen 1999). Although rates of selfing at the fertilization stage can be substantial in

Douglas-fir (Sorensen 1999), they appear to be low when measured at the developed seed stage (Adams and Birkes 1991; Sorensen 1999).

The genetic efficiency of seed orchards is also reduced when orchard parents contribute unequally to seed crops (Friedman and Adams 1982). Studies based on allozyme markers revealed that male mating success varies dramatically among Douglas-fir seed orchard parents, with distance and floral synchrony among the parents being the best predictors of mating frequency (Erickson and Adams 1989; Adams and Birkes 1991; Adams 1992; Burczyk and Prat 1997). Thus, while inbreeding depression caused by self-fertilization does not appear to be a serious problem for the efficiency of Douglas-fir seed orchards, pollen contamination is typically high and the relative contributions of orchard parents to seed crops may be substantially skewed.

Methods of measuring gene flow

Gene flow is the exchange of genes among populations through the dispersal of propagules. In this section, I review the different methods of measuring gene flow and discuss their applicability for studying contemporary, pollen-mediated gene flow (i.e., pollen flow).

Propagule tracking

The first studies of gene flow in forest trees focused on physically measuring or predicting the distances over which pollen and seeds are dispersed (Lanner 1965; Levin and Kerster 1974). The most common approaches were to (1) establish pollen and seed

traps at certain distances from an isolated source population or tree and (2) track the dispersal of unique morphological markers and allozyme alleles (DiFazio et al. 2004). These studies revealed the great potential for long-distance dispersal of forest tree propagules. However, they provided little information about realized propagule dispersal (i.e., the typical distances traveled by seeds that are successful in establishment or by pollen grains that are successful in fertilization).

Methods based on population genetic models

Most of the available information on gene flow in plants is based on studies of genetic variation and population genetic structure using allozyme marker data. In these studies, gene flow is typically inferred from the degree of genetic differentiation among populations as measured by the fixation index F_{ST} (Wright 1931), or its many extensions and analogs (e.g., G_{ST}, which measures the interpopulation component of total gene diversity; Nei 1973). Other approaches include the "rare allele method" (Slatkin 1985) and coalescent methods (Beerli and Felsenstein 1999). All of these methods are 'indirect' because they apply population genetic models to infer gene flow. Measures of differentiation reflect the complex interactions of all demographic parameters and evolutionary forces acting on populations, and the resulting gene flow estimates should be taken as long-term averages estimated over a large number of populations (Sork et al. 1999). Furthermore, the underlying assumptions of population genetic models are often violated in real populations (Bossart and Prowell 1998; Whitlock and McCauley 1999). Thus, although indirect approaches have provided valuable insights into the forces that

shaped the population genetic structure of forest trees, they cannot be used to estimate contemporary pollen flow.

Direct methods based on parentage analysis and maximum likelihood estimation

Direct methods of measuring gene flow obviate the need for tenuous assumptions about historical conditions. Instead, they provide short-term 'immigration snapshots.' Direct methods generally require that all potential parents in a population be genotyped. Then the proportion of progeny that could *not* have been produced by within-population mating can be estimated. One approach employs simple paternity exclusion, which can be used when the haplotype of the paternally contributed gamete can be determined for each offspring (Smith and Adams 1983). The low variability of allozyme markers, however, greatly limits one's ability to distinguish between local and immigrant genotypes (Adams 1992). To overcome this problem, a number of methods use maximum likelihood to either assign parentage to offspring (Meagher 1986; Devlin et al. 1988; Adams et al. 1992; Smouse and Meagher 1994), or estimate mating parameters that best fit the observed progeny genotypes (Roeder et al. 1989; Adams and Birkes 1991; Burczyk et al. 2002).

The limited variability of the genetic markers available (until recently, almost exclusively allozymes) has restricted the use of direct methods to relatively small, and often discrete, populations. Furthermore, the variance of pollen flow estimates based on allozyme data is often high. Standard errors of pollen contamination estimates, for example, are typically 10-20% of the magnitude of the estimates, even with large sample

sizes (Adams and Burczyk 2000). This is because the standard error of pollen flow estimates is inversely proportional to the probability of detecting immigrant genotypes. Thus, for a population of given size, the standard error is inversely related to the variability of the genetic markers used.

In the 1990s, different types of DNA-based markers became available for parentage analyses. Among these, simple sequence repeats (SSRs) are becoming the markers of choice for parentage analysis because they are highly variable (i.e., highly informative), co-dominant, and abundant in the genomes of most higher organisms (Powell et al. 1996; Parker et al. 1998; Jones and Ardren 2003). Several studies of gene flow in natural tree populations suggest that SSRs have a great potential to increase the precision of pollen contamination estimates and allow the detection of important within-orchard mating patterns (Dow and Ashley 1996, 1998; Streiff et al. 1999; DiFazio 2002). Thus, SSR markers have been developed for most major species of interest to tree breeders, including *Pinus taeda* (Elsik et al. 2000), *P. radiata* (Fisher et al. 1998), *P. sylvestris* (Soranzo et al. 1998), *Picea abies* (Pfeiffer et al. 1997), *P. glauca* (Hodgetts et al. 2001), and *Eucalyptus grandis* and *E. urophylla* (Brondani et al. 1998).

Direct methods based on highly variable markers such as SSRs appear to be the most effective means of measuring contemporary pollen flow and within-population mating parameters. However, estimates can be greatly affected by the presence of non-amplifying (i.e., null) alleles and genotyping errors caused by complex marker phenotypes and equipment imperfections. Both types of error result in overestimating pollen flow and introducing considerable biases in the estimated mating parameters.

Therefore, it is important to treat many of the published pollen flow and mating parameter estimates with caution. Their accuracy should improve over the next several years as applications of molecular technology and statistical methods mature. Furthermore, highly variable markers are available in only a few forest tree species, and their cost and throughput are major limiting factors. Thus, perhaps the most important prerequisites to increasing both the spatial scale and the precision of studies of contemporary gene flow are (1) large sets of highly informative, high-throughput, and low-cost genetic markers and (2) analytical methods that take full advantage of the power of these markers, while also eliminating biases caused by null alleles and genotyping errors.

Combined methods

In the last several years, new methods of quantifying pollen-mediated gene flow have been actively sought (Sork et al. 1998). The "TwoGener" model, for example, combines population structure methods and paternity analysis to estimate the degree of pollen pool heterogeneity among female trees (Φ_{FT}; Smouse et al. 2000). This statistic is then related to the mean pollination distance and the effective number of males mating with each female (Smouse et al. 2000; Austerlitz and Smouse 2001a,b). This approach is promising for measuring pollen flow on a landscape level, but its robustness needs to be verified by large-scale studies of pollen flow that compare estimates obtained using the TwoGener model to those based on some of the relatively assumption-free, direct methods discussed above.

Conclusions

Studies of gene flow in trees have revealed that tree propagules (particularly pollen) have a great potential for long-distance dispersal. While it is clear that pollen flow in small discrete populations is extensive, the precision of pollen flow estimates is relatively low. Increasing the precision of pollen flow estimates and scaling up studies of pollen flow to large, continuous populations hinge on the development of highly informative, cost-efficient genetic markers and appropriate analytical methods.

Benefits of improving methods of measuring gene flow

Increasing the precision of gene flow estimates and the spatial scale at which gene flow can be measured will benefit tree improvement, biotechnology risk assessment, gene conservation, and evolutionary biology. Better methods of measuring pollen contamination and detecting within-orchard mating patterns would help increase the genetic efficiency of seed orchards by providing a means to quantify the success of pollen management techniques and guide their progress. Furthermore, the availability of highly variable SSR markers would allow new, cost-efficient approaches for breeding and testing to be used (Lambeth et al. 2001).

Biotechnology (e.g., genetic engineering) has the potential to accelerate the otherwise slow rate of forest tree domestication (Bradshaw and Strauss 2001). Concerns over the biosafety of transgenic plants, however, have restricted progress in plant genetic engineering in many parts of the world (Wolfenbarger and Phifer 2000; Strauss 2003). In trees, the ability to precisely measure long-distance pollen flow appears to be

indispensable for performing objective transgenic risk assessment (DiFazio 2002). Therefore, better methods of measuring pollen flow are important to the future development and public acceptance of genetic engineering.

Gene flow from domesticated tree populations can potentially interfere with gene conservation efforts. Thus, improved methods of measuring gene flow would provide invaluable information for selecting both *in-situ* gene resource management units and the locations of *ex-situ* gene conservation reserves (Adams and Burczyk 2000). The ability to measure long-distance gene flow would also allow designing powerful experiments that can be used to reveal the nature of interactions between gene flow and local adaptation. Information about these interactions is crucial to understanding evolution and predicting the effects of rapid environmental changes (Stockwell et al. 2003; Morjan and Rieseberg 2004).

THESIS OBJECTIVES AND STRUCTURE

This thesis is part of a project whose goal is to develop effective methods for measuring and managing pollen contamination in seed orchards of Douglas-fir. The objectives of my dissertation were to (1) develop highly variable SSR markers for Douglas-fir, (2) test analytical procedures for precisely measuring pollen contamination using SSR markers, (3) measure pollen contamination in one block of an operational Douglas-fir seed orchard complex, (4) determine how pollen contamination varies among trees with different floral phenologies, and (5) analyze within-block mating patterns with respect to floral phenology.

Chapter 2 describes the development and characterization of 22 highly variable SSR markers in Douglas-fir. Chapter 3 presents the results from a simulation study that tested different approaches for estimating pollen contamination using SSR markers and paternity exclusion. This chapter also provides recommendations for obtaining reliable estimates of pollen flow in a variety of situations. Chapter 4 demonstrates the usefulness of SSR markers and paternity exclusion for measuring pollen contamination and identifying mating patterns in seed orchards of Douglas-fir. Chapter 5 summarizes the most important conclusions of my research.

Chapter 2. Highly Variable SSR Markers in Douglas-fir: Mendelian Inheritance and Map Locations

Gancho T. Slavov, Glenn T. Howe, Igor Yakovlev, Keith J. Edwards,

Konstantin V. Krutovskii, Gerald A. Tuskan, John E. Carlson,

Steven H. Strauss, Wesley T. Adams

Theoretical and Applied Genetics

Springer-Verlag Berlin Heidelberg

2004, Volume 108, pp. 873-880

ABSTRACT

Twenty-two highly variable SSR markers were developed in Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco] from five SSR-enriched genomic libraries. Fifteen PCR primer pairs amplified a single codominant locus, while seven primer pairs occasionally amplified two loci. The Mendelian inheritance of all 22 SSRs was confirmed via segregation analyses in several Douglas-fir families. The mean observed heterozygosity and the mean number of alleles per locus were 0.855 (SE=0.020) and 23 (SE=1.6), respectively. Twenty markers were used in genetic linkage analysis and mapped to ten known linkage groups. Because of their high polymorphism and unambiguous phenotypes, 15 single-locus markers were selected as the most suitable for DNA fingerprinting and parentage analysis. Only three SSRs were sufficient to achieve an average probability of exclusion from paternity of 0.998 in a Douglas-fir seed orchard block consisting of 59 parents.

INTRODUCTION

Highly polymorphic genetic markers such as simple sequence repeats (SSRs) can radically improve the precision of pollen contamination and gene flow estimates. We are developing a paternity exclusion procedure to measure pollen contamination in seed orchards of Douglas-fir [*Pseudotsuga menziesii* (Mirb.) Franco] using SSRs. This class of markers has been used to infer paternity and estimate gene flow through genotypic exclusion in a number of tree species. In each case, only a few SSR loci (4–6) were needed to achieve high exclusionary power (Dow and Ashley 1998; Streiff et

al. 1999; Lian et al. 2001). Unfortunately, the development of SSR markers is still inefficient, time-consuming, and resource-intensive (Zane et al. 2002), particularly in organisms with large and complex genomes, such as conifers. Many attempts to develop SSR markers for conifers have yielded just a handful of useful marker loci (Pfeiffer et al. 1997; Hicks et al. 1998; Soranzo et al. 1998).

Fifty SSR markers for Douglas-fir were reported by Amarasinghe and Carlson (2002). We characterized 22 additional markers, 15 of which produce robust banding patterns and segregate as single codominant loci. We show that this set of 15 single-locus SSRs is a valuable tool for genotype identification, parentage analysis, and genome mapping. The remaining seven SSRs can be used for certain applications, but additional optimization of PCR conditions is needed to obtain clear, single-locus banding patterns for all samples.

MATERIALS AND METHODS

Plant materials and DNA extraction

DNA used to construct genomic libraries was extracted from the needles of a single Douglas-fir seedling using the DNeasy Plant Maxi Kit (QIAGEN, Valencia, Calif.). Seeds for segregation analysis were collected from 18 putatively unrelated Douglas-fir trees growing in a grafted seed orchard in western Oregon, USA. Haploid ($1n$) megagametophyte tissue was obtained by removing the seed coats of each seed, then separating the megagametophyte from the embryo. DNA was extracted from 7–8 megagametophytes of each of the 18 parents using a modified CTAB protocol

(http://www.fsl.orst.edu/pnwtirc/research/CTAB_protocol_Df_seed.pdf). Diploid (2*n*) winter buds were collected from the same 18 parents to confirm the inheritance of SSR alleles identified in megagametophytes. We also collected winter buds from an additional 134 ramets (total=152 ramets from 59 genotypes) in the orchard and from 38 trees in native stands located within one kilometer of the orchard. DNA was isolated from winter buds using a protocol developed at Oregon State University (http://www.fsl.orst.edu/tgerc/dnaext.htm).

The mapping population was a three-generation outbred pedigree described by Jermstad et al. (1994). Ninety-two of the F_2 progeny were genotyped for the SSR markers reported in this paper. Needle tissue from parents, grandparents and progeny was harvested, ground in liquid nitrogen to a coarse powder, stored at –80°C, and then used for DNA isolation as described by Jermstad et al. (1998).

Genomic libraries and isolation of SSRs

Four SSR-enriched genomic libraries were constructed by Genetic Identification Services, Chatsworth, Calif., (GIS; http://www.genetic-id-services.com) using a magnetic bead-capture approach (Peacock et al. 2002). Biotin-$(CA)_{15}$, biotin-$(GA)_{15}$, biotin-$(AAT)_{12}$, and biotin-$(ATG)_{12}$ were used as capture molecules for the four libraries, respectively. A fifth library was constructed at the University of Bristol, UK using membrane hybridization enrichment for a mixture of SSR motifs (Edwards et al. 1996). Vector inserts were amplified using PCR and a subsample of each PCR product was used to determine the length of the insert via electrophoresis in 2%

agarose gels. A second subsample was denatured and a single aliquot was spotted and bound to a positively charged nylon membrane (Roche Boehringer Mannheim Diagnostics, Basel, Switzerland) to measure the relative copy number of each insert sequence in the Douglas-fir genome. Total genomic DNA from the seedling used to construct the genomic libraries was labeled with digoxigenin, then hybridized to the dot-blot membranes using the hybridization conditions described by Pfeiffer et al. (1997). Hybridization signals were quantified using the LabWorks Analysis Software (Ultra-Violet Products, Upland, Calif.).

Plasmids containing low-copy inserts longer than 400 bp were purified using the QIAprep Spin Miniprep Kit (QIAGEN), then sequenced on an ABI Prism 3700 DNA analyzer [Applied Biosystems (ABI), Foster City, Calif.] using the BigDye Terminator v. 3.0 Ready Reaction Sequencing Kit (ABI). Redundant sequences were identified by pairwise BLAST analyses and eliminated from further consideration.

Primer design and detection of putative SSR loci

Primers targeting the SSR flanking sequences were designed using the PRIMER program (version 0.5, Whitehead Institute for Biomedical Research, Cambridge, Mass.). Initial screening of the candidate markers was done by amplifying 10 ng of template DNA in 15 µl of PCR mix including 2.5 mM of $MgCl_2$, 0.67 mg ml^{-1} BSA, 0.17 mM of each dNTP, 0.5 µM of the respective forward and reverse primers, 1× PCR buffer, and 1 unit of *Taq* DNA polymerase (Invitrogen). We added *Taq* polymerase following a hot start at 94°C for 10 min. The program proceeded with

seven cycles of touchdown PCR: 95°C for 30 s, empirically determined optimal annealing temperature (T_a)+7°C for 30 s, then 72°C for 45 s. The T_a was decreased by 1°C for each of the six subsequent touchdown cycles. Following touchdown PCR, the program continued with 32 cycles of 95°C for 30 s, T_a°C for 30 s, 72°C for 45 s, and a final extension of 72°C for 20 min. After electrophoresis in 2% agarose gels, primer pairs that produced variable patterns of bands of the expected size were tested in reactions containing 1.7 μM fluorescent deoxynucleotides (R110 [F]dNTP, ABI), and then detected on an ABI Prism 377 DNA sequencer. DNA samples from one megagametophyte and five adult trees were used for these preliminary screening steps. Forward primers of the best-performing candidate markers were end-labeled with the fluorescent dyes Fam, Hex, or Ned (ABI) and the resulting PCR products were detected on an ABI Prism 3100 Genetic Analyzer. The putative alleles were sized using the GeneScan software (ABI), scored using the Genotyper software (ABI), and then individually verified by visual inspection.

Allelic variability, inheritance, and map locations of the SSRs

After genotyping trees located inside and outside of the orchard, the number of alleles, the frequency of the most common allele, the frequency of null alleles, and the observed and expected heterozygosities for each SSR locus were determined using the CERVUS program (Marshall et al. 1998). This program was also used to calculate the cumulative average probability of exclusion from parentage (PE) provided by the SSR markers. We used a chi-square test to detect deviations from a 1:1 segregation ratio of alleles in megagametophytes from heterozygous mother trees (Adams and Joly 1980). The map locations of the SSRs were determined using the JoinMap software (v. 2.0, Stam and Van Ooijen 1995).

RESULTS AND DISCUSSION

Molecular characterization of SSRs

We screened 1,452 insert-containing colonies from the five SSR-enriched genomic libraries that we obtained (Table 2.1). Inserts were PCR-amplified and their approximate sizes were measured after electrophoresis in agarose gels. We used dot-blot hybridization to determine the relative number of copies of the insert sequences in the Douglas-fir genome. In this assay, a weak hybridization signal suggests that the target sequence has a relatively low genome copy number (Pfeiffer et al. 1997; Scotti et al. 2002). Based on visual assessment of relative hybridization signal intensity, we selected and sequenced 517 low-copy inserts that were longer than 400 bp.

Table 2.1. Development of markers from five SSR-enriched genomic libraries.

Library enriched for SSR motif	Colonies processed	Colonies sequenced	Colonies with an SSR (%)[a]	Primer pairs designed	Markers developed	Repeat units per marker (ave.)	Efficiency[b] (%)
$(CA)_n$[c]	864	322	292 (91)	81	18	44.4	5.6
$(GA)_n$[c]	182	62	58 (94)	17	2	34.5	3.2
$(AAT)_n$[c]	96	35	9 (26)	0	0	-	0.0
$(ATG)_n$[c]	96	50	14 (28)	4	0	-	0.0
$(GA)_n$+$(GT)_n$[d]	214	48	12 (25)	8	1	21	2.0
Total (mean)	**1452**	**517**	**385 (74)**	**110**	**21**[e]	**(33.3)**	**(4.1)**

[a] Percent of sequenced colonies containing an SSR.
[b] Efficiency = 100 × (number of markers developed) / (number of colonies sequenced).
[c] Library developed by Genetic Identification Services.
[d] Library produced at the University of Bristol.
[e] One additional marker, PmOSU_783, was developed using a cDNA sequence downloaded from the GenBank database (accession number AA701783). Thus, the total number of markers developed was 22.

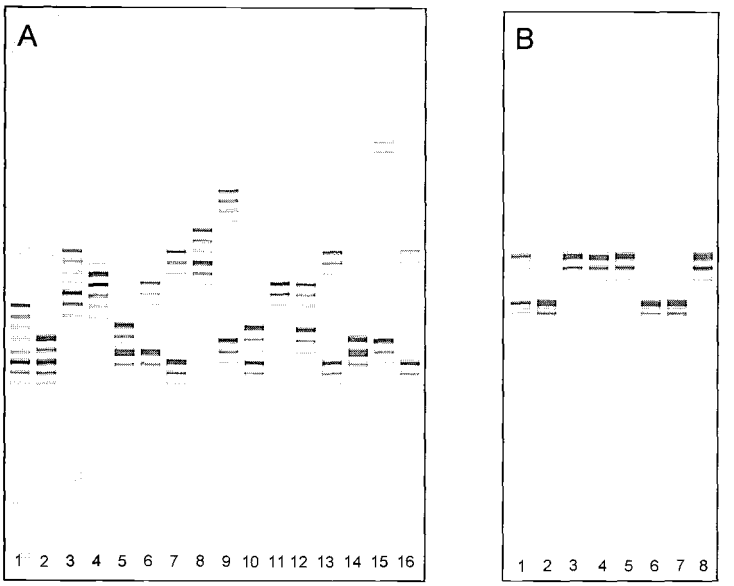


Figure 2.1. Phenotypes demonstrating the variability (A) and the Mendelian inheritance (B) of SSR marker PmOSU_3G9. A. Each of the 16 lanes contains SSR DNA amplified from diploid bud tissue from a different Douglas-fir tree. B. The leftmost lane contains SSR DNA amplified from diploid maternal bud tissue, whereas the remaining lanes contain SSR DNA amplified from seven haploid megagametophytes extracted from seeds of the mother tree.

We also quantified the hybridization signals and compared the mean hybridization signal (as a proportion of the average for each membrane) for the selected colonies with the mean hybridization signal for all colonies. As expected, the mean hybridization signal for the selected colonies was considerably lower (29%) and this difference was statistically significant (one-sided $P<0.001$).

SSRs were found in 385 (74%) of the 517 sequenced inserts. We selected 110 SSRs whose flanking sequences were not redundant and were long enough to design pairs of compatible primers free of sequences capable of forming internal secondary structures. These 110 primer pairs were tested by amplifying template DNA from one megagametophyte ($1n$) and five adult trees ($2n$). Nine primer pairs failed to amplify

products of the expected size, 17 produced a monomorphic banding pattern, 62 produced a complex banding pattern indicating that two or more loci were amplified, and 22 produced a pattern indicative of a single polymorphic locus (e.g., Fig. 2.1A). Fifteen of these 22 primer pairs produced a robust, single-locus pattern, whereas seven (the last seven in Table 2.2) occasionally amplified a second locus or produced phenotypes with excessive band 'stuttering.' The latter seven primer pairs may require different PCR conditions for different samples to achieve uniform quality of data (i.e., obtaining clear, single-locus phenotypes for some samples may require varying the annealing stringency).

The highest efficiency (5.6%) for developing single-locus markers was achieved from the $(CA)_n$-enriched library (Table 2.1). Efficiency could be increased by checking the length of the flanking sequences prior to sequencing each insert as described by Rafalski et al. (1996). For example, we were unable to design primers for 132 of the 385 SSR-containing sequences because the flanking sequences at the 5'-end were too short. Therefore, we recommend selecting only inserts with at least 50 bp of flanking sequence on both sides of the SSR. Oligonucleotides identical to those used for library enrichment could be used in combination with vector primers to amplify and determine the length of the SSR flanks prior to purifying the plasmids. This step would also allow one to confirm the presence of an SSR within each insert (Rafalski et al. 1996).

Table 2.2. Primer sequences and properties of 22 SSR markers in Douglas-fir.

Locus	Forward primer, 5'-3'	Reverse primer, 5'-3'	Repeat motif	Optimal T_a °C (tested range)	N^a	A^b	Allele size (bp)	H_o/H_e^c	f_{max}^d	f_0^e
PmOSU_1C3	CTCCCCTCCAGATTTTACTC	TGGCGTAACAAATAAGAGAAA	$(TC)_{25}(AC)_{32}...(TC)_4$	57 (55-57)	28	28	166-232	0.929/0.968	0.094	0.012
PmOSU_1F9	CCTCATGCATTGGACACTC	GGATTCTTGAGCAGGTAGG	$(AG)_{34}$	55 (52-57)	35	33	201-319	0.943/0.973	0.089	0.008
PmOSU_2C2	TAAATCCGCAGCTCATAGAATC	GGGTGGTGGCTAGGGAAAC	$(AC)_{32}...(CT)_4$	60 (58-62)	38	12	142-200	0.711/0.752	0.487	0.004
PmOSU_2C3	AAAGACAACATTATGAAAGG	GTAATGGTTCGAAAAATAATG	$(TC)_{24}(AC)_{18}$	50 (48-51)	35	25	163-251	0.943/0.955	0.121	0.000
PmOSU_2D4	TTATTGCACATGAGTATTATGA	CAGATGTTGTTTTTTATACCAC	$(AG)_4...(TG)_{18}(AG)_{26}$	50 (48-53)	34	30	108-194	0.912/0.968	0.109	0.022
PmOSU_2D6	GGAAAATATACATCTCACGAC	AAGCATGCGTACTAGGTG	$(AC)_5...(AC)_4...(GC)_8(AC)_{13}...(AC)_7$	54 (N/A)	34	30	162-264	0.912/0.975	0.061	0.026
PmOSU_2D9	TCGATTTACGCTTTTTTCTC	TGTTTATCCCCAGTCTCAAG	$(TC)_{32}(AC)_{15}$	57 (54-57)	16	8	125-181	0.688/0.806	0.312	0.072
PmOSU_2G12	CAAGGACTCATATGGGAAA	AACATCAGTAATAACCTTTT	$(AC)_{11}...(AC)_{19}...(GCAC)_5...(GCAC)_4(AC)_7...(AC)_6$	51 (48-51)	34	16	244-310	0.824/0.914	0.097	0.047
PmOSU_3B2	CTTTGGAGTTCTTAATATAG	GATAATAGCACCCCACCATA	$(TG)_{22}(CG)_7$	49 (46-49)	32	27	88-176	0.938/0.962	0.116	0.005
PmOSU_3B9	TGTGTAAAAATGTCTAATCC	ACTACTATTCGAGGTTTTCT	$(CG)_6(CA)_6...(AC)_6...(AC)_5...AC)_6$	47 (46-49)	30	25	119-223	0.900/0.930	0.204	0.008
PmOSU_3D5	GGCATCCTATTTTTCATTTT	GTGATTACCTAACTTGTGC	$(TG)_{16}(AG)_{26}$	50 (48-51)	35	19	125-193	0.943/0.931	0.172	0.000
PmOSU_3F1	GACTAGATCATCGCAACTT	GGTATTCTTATGGTTTTTAT	$(TG)_6...(TG)_7(AG)_{27}...(AC)_4$	50 (48-52)	27	20	144-246	0.741/0.936	0.159	0.108
PmOSU_3G9	ATTCCTTTTGAGACCTACTT	CTTCAAAAATTCCTACAACA	$(TG)_{12}(AG)_{28}$	51 (48-52)	35	22	110-192	0.857/0.926	0.137	0.034
PmOSU_4A7	TTGTAAAAATTCCCATGTAT	AAGTGGGGGAGTGTGTAAT	$(TG)_5...(TG)_5...(CG)_7(TG)_4...(TG)_{29}...(ATC)_5$	48 (48-54)	34	30	196-340	0.912/0.960	0.142	0.018
PmOSU_4G2	ATTTTTTGTATTGTGCTTG	TGGATATATTTGCATTTTAC	$(AT)_6...(AG)_{29}$	48 (48-51)	30	16	138-168	0.900/0.920	0.143	0.003
PmOSU_2B6[f]	TTGTTGGGTATAATTTTCA	TAATAAAATAGCTCTAACCC	$(TG)_{19}(AG)_{31}...(AT)_4$	49 (48-49)	32	28	134-346	0.813/0.957	0.133	0.075
PmOSU_2G4[f]	ATGCATTCTTGAAAGTAAA	ATAATATGCAAGTGAATCCC	$(TC)_{24}(AC)_{29}...(AC)_{12}$	51(50-51)	27	19	180-272	0.778/0.937	0.177	0.087
PmOSU_3E3[f]	TGCTTCAATTTCATATCTA	TAACATTTCAATCTATTCAC	$(TG)_5...(TG)_4...(TG)_{15}(AG)_{28}$	48 (46-49)	29	31	126-266	0.897/0.969	0.090	0.031
PmOSU_3H4[f]	TTTGCCGTCACATTTTTATTG	GCATCTTTCAGGCATAGTCT	$(GC)_9(AC)_{22}$	55 (48-57)	32	25	170-256	0.875/0.957	0.116	0.035
PmOSU_4E9[f]	GTTGGTTGTGTATATTCAGTTT	GCCTCTTCTTGGTTTGGT	$(AC)_{36}$	54 (48-55)	34	24	120-218	0.853/0.923	0.233	0.024
PmOSU_5A8[f]	CATTTTTGGATTCTGGTTTTG	ATGCCTCAAGCTATGTAATC	$(TG)_{11}...(TG)_{10}$	54 (50-55)	37	7	166-190	0.595/0.805	0.241	0.142
PmOSU_783[f]	GAGCTGATGCCTTGAAGACT	CAAGTCAGTTCACAATTCCT	$(AT)_5...(AT)_5$	57 (56-59)	33	15	205-303	0.939/0.879	0.261	0.000
Mean					**32**	**23**		**0.855/0.923**	**0.163**	**0.036**

Table 2.2. Continued

[a] N is the number of trees genotyped. Because seed orchards may differ in gene diversity from natural populations, we used data for 38 trees sampled in a natural Douglas-fir population adjacent to the seed orchard. Exception is locus PmOSU_2D9, for which we used data from the parents genotyped for segregation analysis.
[b] A is the number of alleles detected in a sample of N trees.
[c] H_o and H_e are observed and expected heterozygosities, respectively.
[d] f_{max} is the frequency of the most common allele.
[e] f_0 is the estimated frequency of null alleles, based on deviations from Hardy-Weinberg equilibrium.
[f] Primers for these loci may amplify two loci and need additional optimization.

Genotyping error is a major concern when molecular markers are used for parentage analysis. If not accounted for, genotyping error may lead to considerable bias in the estimated parameters (SanCristobal and Chevalet 1997; Marshall et al. 1998). This problem is greater when using markers with inherently high variability, such as SSRs (Pemberton et al. 1995; Robinson and Harris 1999). Minimizing the rate of mistyping and avoiding markers with high frequencies of null alleles may be crucial for obtaining unbiased estimates of gene flow and pollen contamination. Although markers which simultaneously detect two or more loci can be useful for some applications (Fisher et al. 1998; Amarasinghe and Carlson 2002), their use in parentage analysis is likely to lead to increased rates of mistyping and false inferences. Therefore, our main goal was to develop SSR markers with strong and consistent single-locus banding patterns, and low frequencies of null alleles.

Prior to developing the 22 markers characterized in this paper, we tested the 49 SSR-amplifying primer pairs for Douglas-fir reported by Amarasinghe and Carlson (2002). Although we experimented with a variety of PCR conditions (e.g., a wide range of T_a and $MgCl_2$ concentrations), and even redesigned the primers for certain SSR sequences, we had no success obtaining markers with the strong and consistent single-locus phenotypes that we desired. The best performing primer pairs (BCPsmAC5, BCPsmAC8, BCPsmAG38 and BCPsmAG39; Table 3 in Amarasinghe and Carlson 2002), were tested with T_a between 48 and 59°C. Only two of these primer pairs (BCPsmAG38 and BCPsmAG39) produced single-locus banding patterns, but the strengths and consistencies of their banding phenotypes were unsatisfactory. In contrast,

the 15 single-locus SSR markers we obtained work well under a range of T_a. We attribute the better performance of our markers to the greater length of the sequences flanking the SSRs that we were able to isolate. Longer flanking sequences provide better chances to design GC-rich, compatible, and highly specific primers, which are not prone to forming internal secondary structures during PCR. The relatively short flanking sequences in the clones isolated by Amarasinghe and Carlson (2002) may have restricted their ability to design such primers far enough from the SSR sequences to enable consistent amplification of single loci.

Mendelian inheritance and polymorphism

Primer sequences, T_a, and other properties of the 22 SSR markers are shown in Table 2.2. We surveyed the allelic variability of the markers by genotyping an average of 32 (range=16–38) of the 38 trees sampled from natural Douglas-fir stands surrounding the seed orchard. The mean number of alleles per locus was 23 (standard error (SE)=1.6), the mean observed heterozygosity was 0.855 (SE=0.020), and the mean expected heterozygosity (H_e) was 0.923 (SE=0.013). The mean frequency of the most common allele was 0.163 (SE=0.079), and the mean frequency of null alleles was 0.036 (SE=0.039). The frequency of null alleles was estimated assuming that deviations from Hardy-Weinberg equilibrium were entirely due to the presence of null alleles (Summers and Amos 1997). The mean number of alleles per locus was 31 (SE=2.00) in a larger sample of trees (mean=78 trees/locus; range=60–95), which included trees located within

the seed orchard and in the surrounding stand that were genotyped for 21 of the SSR markers (Table 2.3).

Table 2.3. Polymorphism and diversity of dinucleotide SSRs in some conifers.

Species	No. of SSRs	N[a]	A[b]	H_o (H_e)[c]	Reference
Pseudotsuga menziesii	21[d]	78	31	0.864	This study, pooled data
	21[d]	46	26	0.864	This study, trees inside the orchard
	21[d]	32	23	0.863	This study, trees outside the orchard
	50	24	8	(0.673)	Amarasinghe and Carlson 2002
Pinus sylvestris	7	13[e]	6.7	(0.850)	Soranzo et al. 1998
Picea abies	7	18	13	(0.789)	Pfeiffer et al. 1997
Pinus halepensis / P. brutia	7	50/47	2.9	0.586	Keys et al. 2000
Picea glauca	15	14	10.2	0.520	Hodgets et al. 2001
Pinus strobus	16	16	5.4	0.515	Echt et al. 1996

[a] N is the mean number of diploid individuals genotyped.
[b] A is the mean number of alleles per locus.
[c] H_o is the observed heterozygosity and H_e is the expected heterozygosity for studies in which observed heterozygosity was not reported.
[d] Data for PmOSU_2D9 were not used because only some trees (N = 16) located within the orchard were sampled.
[e] Megagametophytes were sampled in this study.

Our estimates of heterozygosity and mean number of alleles per locus are among the highest reported for dinucleotide SSRs in conifers (Table 2.3). The markers described in this paper have an H_e that is 37% higher, and a mean number of alleles per locus (based on $\bar{N}$ =32) that is 188% higher than those of the 50 SSRs previously reported for Douglas-fir (Amarasinghe and Carlson 2002). We sampled 33% more individuals, and we may have sampled a more polymorphic population, but the higher levels of polymorphism probably resulted from the longer SSR sequences that we isolated. The average number of dinucleotide repeats for the 50 SSR sequences reported by Amarasinghe and Carlson (2002) (Table 2.3) was 20. In contrast, the average number of repeats for the $(CA)_n$ and $(GA)_n$ markers reported in this study was 39, nearly twice as large. The number of repeats is positively associated with SSR mutation rates and, therefore, with SSR marker polymorphism (reviewed in Estoup and Cornuet 1999). Only two of our 22 markers are based on perfect SSRs (i.e., single uninterrupted repeat motifs), whereas six are based on compound (two or more adjacent SSR domains with different repeat motifs), two on interrupted (two or more SSR domains interrupted by short sequences that do not fit the repeat structure), and 12 on compound interrupted SSRs (Table 2.2). There was a weak, but statistically significant, negative correlation between dot-blot hybridization signal and mean number of alleles per locus for the developed SSR markers (r=–0.477, two-sided P=0.033, n=21).

We confirmed the Mendelian inheritance of all 22 SSRs by analyzing the haploid segregation of alleles in seed megagametophytes from known mothers. In all cases, a bud sample from the parent tree was genotyped and run side-by-side with megagametophyte

samples (Fig. 2.1B). For each SSR, we analyzed 5–8 megagametophytes from 6–17 heterozygous mothers. No significant deviations from a 1:1 segregation ratio were detected after pooling the data over the mother trees (data not shown). Although the small sample sizes within each mother precluded a formal test for segregation heterogeneity among mothers, there were no obvious indications of segregation heterogeneity.

Twenty of the 22 SSR loci also segregated in the progeny ($2n$) of a single controlled cross that was previously used for linkage mapping (Jermstad et al. 1998). One of the parents was heterozygous for 19 of the 20 segregating loci and the other for 17. At the 5% level, there were two statistically significant deviations from a 1:1 segregation ratio (Appendix 1). In both cases, the segregation distortion was limited to only one of the two parents. We detected five null alleles (6.25%) among the 80 alleles (20 loci×2 parents×2 alleles) that we sampled. This finding, although based on a small sample of alleles, confirmed our expectations for the occurrence of a small percentage of null alleles based on population deviations from Hardy-Weinberg equilibrium (reported above and in Table 2.2). All 20 loci were successfully mapped to ten existing Douglas-fir linkage groups at LOD=5 (Table 2.4). The SSRs that we mapped are well dispersed throughout the Douglas-fir genome. All markers mapped at least 10 cM apart, except for loci PmOSU_2G4 and PmOSU_5A8, which mapped 4 cM apart. These mapping distances are unlikely to lead to non-independent association of alleles in multilocus gametes in large outcrossed populations (Epperson and Allard 1987). Therefore, this set of markers is appropriate for fingerprinting and parentage analysis in Douglas-fir.

Table 2.4. Linkage map locations of 20 SSR markers in Douglas-fir (LOD = 5).

Linkage group	SSR[a] markers and terminal[b] markers on the same linkage group	Position (cM)[c]	Distance between adjacent SSRs (cM)
1	rapdUBC_BC_590_975	0.0	
1	**PmOSU_2D9**	68.3	15.6
1	**PmOSU_3B2**	83.9	
1	rflpPmIFG_1246_a	158.3	
4	rflpPtIFG_2413_a	0.0	
4	**PmOSU_2G4**	91.4	4.0
4	**PmOSU_5A8**	95.4	
4	estPtINR_COMT1	146.2	
5	estPmIFG_c519_c	0.0	
5	**PmOSU_1C3**	12.0	
5	estPtIFG_8415_e	157.1	
6	rflpPmIFG_1052_d	0.0	
6	**PmOSU_4E9**	70.1	30.6
6	**PmOSU_783**	100.7	
6	estPmIFG_c572	186.7	
8	rflpPmIFG_1548_a	0.0	
8	**PmOSU_3G9**	17.4	53.4
8	**PmOSU_4G2**	70.8	
8	rflpPmIFG_1591_a	108.8	
11	rflpPmIFG_1278_b	0.0	
11	**PmOSU_2B6**	9.7	59.6
11	**PmOSU_2D4**	69.3	37.5
11	**PmOSU_2G12**	106.8	
11	Ugpp-1	164.7	
13	estPmIFG_73-6-130-E12	0.0	
13	**PmOSU_3F1**	9.2	
13	Idh	29.8	
19	**PmOSU_3H4**	0.0	10.3
19	**PmOSU_4A7**	10.3	21.7
19	**PmOSU_3D5**	32.0	17.9
19	**PmOSU_3E3**	49.9	22.6
19	**PmOSU_2C2**	72.5	
19	estPaTUM_PA66	96.1	
21	**PmOSU_3B9**	0.0	
21	rapdUBC_BC_304_450	18.9	
22	rflpPmIFG_1124_a	0.0	
22	**PmOSU_2D6**	5.9	

[a] SSRs are shown in bold; the complete names of the SSR markers appear in the DENDROME database (http://dendrome.ucdavis.edu) and contain "OSUPCT_ssr" preceding the SSR names used in this paper (e.g., OSUPCT_ssrPmOSU_2B6 is the complete name for marker PmOSU_2B6).

[b] Other markers used to construct the map are described elsewhere (Jermstad et al. 1998; Jermstad et al. 2001a, b).

[c] Kosambi distance in cM from the marker mapped at position 0.0.

Applications in tree improvement

Molecular markers have a variety of applications in tree improvement (Adams 1983; Wheeler and Jech 1992). Because of their high polymorphism, our 15 single-locus SSR markers will be valuable tools in the testing, breeding, and seed orchard phases of tree improvement programs. Compared to using allozyme markers, less effort will be needed to verify genotypes and controlled crosses between selected parents. SSRs should allow pollen and seed contamination in seedlots to be measured cheaply and precisely. It will also be easy to measure the success of seed orchard management techniques such as bloom delay and supplemental mass pollination. Finally, the high polymorphism of these markers can be used to directly determine the relative maternal and paternal contributions in open-pollinated seedlots from seed orchards. Because of their low variability, this is difficult to achieve using allozyme markers. Until recently, maximum likelihood modeling methods were the only feasible way to obtain this information (Adams 1992).

For example, we used our SSR markers to identify parental genotypes in the sampled seed orchard block. For 51 of the genotypes within the block, we sampled 2–3 ramets (total 145). For the remaining seven genotypes, we sampled the only ramet available. We used three of our more variable SSRs (PmOSU_2C3, PmOSU_3B2, and PmOSU_2G12) to genotype all 152 ramets. Assuming Hardy-Weinberg equilibrium, linkage equilibrium, and allele frequencies equal to those estimated from our pooled sample of trees (inside and outside of the seed orchard), no three-locus genotype is expected to occur at a frequency greater than 1.8×10^{-6} (i.e., no three-locus genotype is expected to occur more than once in 555,555 individuals). As expected, all parental

genotypes had distinct three-locus SSR genotypes. In all but one case, the ramets that were labeled as belonging to the same genotype matched for all three loci. The only exception was one ramet whose genotype differed at all three loci from the other two putative ramets of the same genotype. This ramet did not appear to be identical to any of the genotypes in the orchard block. Therefore, it was included as an additional 59th genotype in further analyses. This ramet was probably intended to be included in a different orchard block and was misplaced due to a labeling error made during orchard establishment.

We also evaluated the degree of resolution in parentage analysis provided by the developed set of SSRs. The cumulative average probability of exclusion is the expected proportion of unrelated potential parents that can be excluded from parentage in a finite population using a given set of markers. Along with the number of possible parents, PE is a key determinant of the proportion of offspring that would be assigned unambiguous parentage based on genotypic exclusion (Chakraborty et al. 1988). We calculated PEs for different numbers of loci based on the SSR genotypes of the 59 parental genotypes identified in the orchard block. For example, using the three SSR loci described above, the estimated PE was 0.991 for analyses in which nothing is known about the two parents of a given offspring (e.g., if seeds are collected without keeping track of the mother trees). For cases in which one of the parents could be determined based on other data (e.g., known mother for each seed), the estimated PE was 0.998. Fig. 2.2 shows how PE changes when loci with similar variability are added consecutively. For comparison, 10–15 typical allozyme loci would be needed to exceed a PE of 0.900 (Adams 1992).

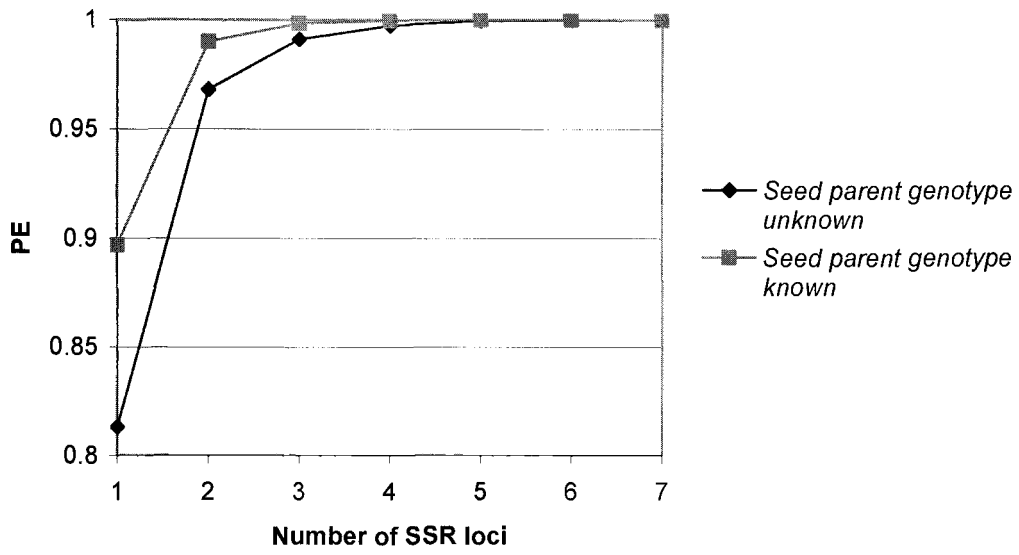


Figure 2.2. Cumulative average probability of exclusion (PE) (defined in text) provided by SSR loci in Douglas-fir. PE was recalculated after adding each of the following seven SSR loci: PmOSU_2C3, PmOSU_3B2, PmOSU_2G12, PmOSU_3G9, PmOSU_2D6, PmOSU_3B9, and PmOSU_4A7. *Seed parent genotype unknown* corresponds to a situation in which seeds are collected without keeping track of the source mothers. *Seed parent genotype known* corresponds to a situation in which the genotype of the mother tree of each seed is known.

The reliability and exclusionary power of our 15 single-locus SSRs make them the most efficient genetic markers available for Douglas-fir. Only three SSRs were enough to measure the success of a supplemental mass pollination experiment in which pollen from a single father tree was used to fertilize three different female parents (unpublished data). Based on preliminary results from computer simulations, we expect that a small number of loci (<10) will provide enough genetic resolution for measuring pollen contamination through paternity exclusion in any currently existing seed orchard of Douglas-fir.

ACKNOWLEDGEMENTS

This study was funded by the Pacific Northwest Tree Improvement Research Cooperative at Oregon State University (OSU). We thank Stephen DiFazio, Craig Echt, and Michele Morgante for valuable advice on developing SSR markers. Materials, laboratory equipment, and technical support were generously provided by OSU's Tree Genetic Engineering Research Cooperative, Oak Ridge National Laboratory, the University of Bristol, and the University of British Columbia.

Chapter 3. Estimating Pollen Flow Using SSR Markers: The Effect of Mistyping on Paternity Exclusion

Gancho T. Slavov, Glenn T. Howe, Aglika V. Gyaourova,

David S. Birkes, Wesley T. Adams

Submitted to:

Molecular Ecology

Blackwell Publishing, 23 Ainslie Place

Edinburgh EH3 6AJ, UK

ABSTRACT

Highly informative genetic markers, such as SSRs, can be used to directly measure pollen flow by parentage analysis. However, mistyping (i.e., false assignment of genotypes caused by the occurrence of null alleles, mutations, and detection errors) can lead to substantial biases in the estimates obtained. We studied the effect of mistyping SSR marker data on estimates of pollen immigration obtained via paternity exclusion. We simulated plant populations of 30 to 600 reproductively mature individuals using SSR markers with 15 to 53 alleles per locus. Mistyping invariably led to inflated estimates of pollen immigration. If ignored, even minor rates of mistyping (1% in the parents and 1.5% in the offspring) resulted in overestimating pollen immigration by up to 150%. When we required at least two mismatches before excluding candidate fathers from paternity, the resulting estimates had small biases for minor or low rates of mistyping (≤4.5%). Requiring at least three mismatches for exclusion was needed to minimize the upward biases of pollen immigration caused by moderate or high rates of mistyping (≤10.5%). The minimum number of SSR loci needed to minimize cryptic gene flow and to obtain reliable estimates of pollen immigration varied from five to seven for a sampling scheme in which pollen gamete haplotypes can be unambiguously determined. This scheme can readily be applied to most gymnosperms. Between five and nine SSR loci were needed for a more general sampling scheme in which only the diploid genotypes of offspring are available (i.e., a scheme applicable to all diploid seed plants). We developed the Pollen Flow (PFL) computer program, which can be used to obtain

unbiased and precise estimates of pollen immigration under a wide range of conditions, including population sizes as large as 600 parents and mistyping rates as high as 10.5%.

INTRODUCTION

Understanding pollen-mediated gene flow (pollen flow) is a priority for plant ecologists, evolutionary biologists, breeders, and biotechnologists (Ouborg et al. 1999; Sork et al. 1999; Ellstrand et al. 1999; Ellstrand 2001; DiFazio 2002). Different methods for measuring pollen flow using genetic markers have been developed and applied (Neigel 1997; Sork et al. 1999; DiFazio et al. 2004). On an evolutionary time scale, mean pollen flow can be inferred from the interpopulation differentiation of allele frequencies for genetic markers based on nuclear and organelle DNA sequences (Ennos 1994). This can be done at low cost and with moderate effort, but relies on a number of tenuous assumptions (Bossart and Prowell 1998; Whitlock and McCauley 1999). Furthermore, we are often interested in estimating current pollen flow, rather than its long-term average. Contemporaneous pollen flow can be estimated directly only by paternity analysis based on extensive sampling and a set of highly informative genetic markers (Adams 1992; Jones and Ardren 2003).

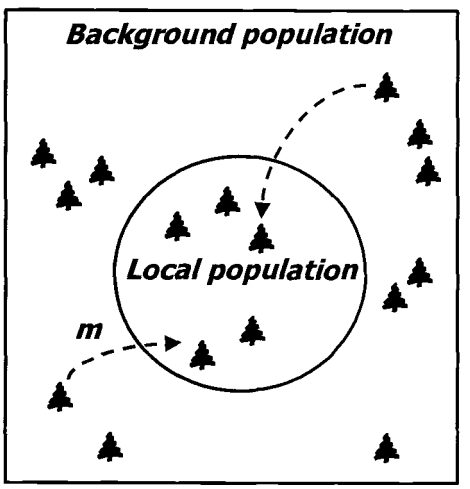
Using paternity analysis, pollen flow and other dispersal parameters can be estimated from the multilocus genotypes of all reproductively mature plants in a population plus a sample of their progeny. Some approaches perform categorical or fractional assignment of paternity by comparing the genotypes of all potential male parents to those of the progeny (reviewed by Jones and Ardren 2003), whereas others

employ maximum likelihood methods to estimate pollen flow and other parameters simultaneously (Roeder et al. 1989; Adams and Birkes 1991).

Paternity exclusion is the most straightforward and assumption-free method to directly measure contemporaneous pollen immigration (i.e., pollen flow into a population). This method is based on testing the genotypic match between the potential male parents in the population and a sample of progeny produced in that population (Fig. 3.1). The progeny whose multilocus genotypes could not have resulted from any cross among the *local parents* (defined in Fig. 3.1) are assumed to result from pollen immigration (Smith and Adams 1983; Devlin and Ellstrand 1990). Ideally, all progeny fathered by immigrant pollen can be identified using this approach. However, achieving this level of genotypic resolution has been practically impossible with allozymes, the traditional markers of choice (Adams 1992).

With low-variability markers, such as allozymes, a substantial proportion of the pollen haplotypes produced in the *background population* (defined in Fig. 3.1) are identical to haplotypes that could be produced in the *local population*, resulting in 'cryptic gene flow' (Devlin and Ellstrand 1990). The proportion of undetectable immigrants is inversely related to the number and variability of the genetic markers used, and increases with the number of *local parents* (Adams 1992). Although procedures that account for cryptic gene flow have been developed (Smith and Adams 1983; Devlin and Ellstrand 1990), the low variability of allozyme markers has limited the application of genotypic exclusion to small populations (i.e., up to 100-200 reproductively mature

individuals; Devlin and Ellstrand 1990; Adams and Birkes 1991; Adams et al. 1997; Pakkanen et al. 2000).



Local population: An arbitrary group of plants for which pollen immigration will be measured

Local parents: All reproductively mature plants in the *local population*

Local pollen haplotypes: Pollen gamete haplotypes that could be produced by the male *local parents*

Background population: The idealized population of all plants that could pollinate any of the *local parents*

Background parents: Reproductively mature plants sampled from the *background population*

Procedure for estimating pollen immigration via paternity exclusion:

1. Determine the multilocus genotypes of all *local parents*, a number of *background parents*, and a sample of offspring produced by mother plants in the *local population*.
2. Infer the paternal gamete haplotypes of the sampled offspring and compare these haplotypes to all *local pollen haplotypes*. The proportion of offspring whose paternal gametes do not match any *local pollen haplotypes* is the observed pollen immigration ($\hat{b}$).
3. Using the allele frequencies estimated from the multilocus genotypes of the *background parents*, account for cryptic gene flow (see text), and obtain a final estimate of pollen immigration ($\hat{m}$).

Figure 3.1. Procedure for estimating pollen immigration via paternity exclusion.

Highly variable PCR-based genetic markers, such as simple sequence repeats (SSRs), revived interest in estimating pollen immigration via genotypic exclusion. This is because five or six polymorphic SSRs can provide exclusionary power that far exceeds anything that could be achieved using allozymes (Dow and Ashley 1996; Streiff et al. 1999). Therefore, larger populations can be analyzed using genotypic exclusion, cryptic gene flow can be minimized, and more precise estimates of pollen immigration can be obtained. Nonetheless, higher rates of mistyping for SSRs (i.e., mutations, null alleles, scoring and other detection errors; Pemberton et al. 1995, Robinson and Harris 1999; Ewen et al. 2000) cause pollen immigration to be overestimated. This is because both mutations and genotyping errors lead to mismatches between some 'true' fathers and their offspring leading to the false classification of these offspring as immigrants. Although there are ways to mitigate or eliminate these upward biases (e.g., requiring more than one mismatch between an offspring and all potential fathers before classifying the offspring as an immigrant), the performance of these approaches under different sampling scenarios and rates of mutation and genotyping error has not been investigated.

Our goal was to develop recommendations for obtaining accurate estimates of pollen immigration using highly variable SSR markers and paternity exclusion. We addressed the following questions (1) To what degree is pollen immigration overestimated when mistyping occurs?; (2) Which analytical approaches minimize this bias?; (3) How do these analytical approaches perform under different scenarios?; and (4) How many highly-variable SSR markers are needed to obtain pollen immigration estimates with little bias and low variance? To answer these questions, we developed a

computer program that generates the multilocus genotypes of a set of parents and their offspring based on SSR allele frequencies, introduces mistyping into these genotypes, then estimates pollen immigration using paternity exclusion and three different approaches for handling mistyping.

MATERIALS AND METHODS

Simulation structure and variables

We conducted our simulations in four phases (Fig. 3.2). First, using allele frequencies for a number of highly variable SSR loci (i.e., having $H_e \geq 0.80$), we generated the multilocus genotypes of the parents in a *local population* of a diploid hermaphroditic plant and those of 100 parents sampled to estimate the allele frequencies in the *background population.* Second, according to predetermined rates of pollen immigration and mutation, we generated the multilocus genotypes of 200 offspring from mothers selected among the *local parents* and introduced mutations in these genotypes. Third, we simulated detection errors (e.g., incorrectly sized alleles, undetected alleles, and cross-contaminated PCR samples) in the multilocus genotypes of the parents and offspring. Finally, we estimated pollen immigration using paternity exclusion with different adjustments for mistyping.

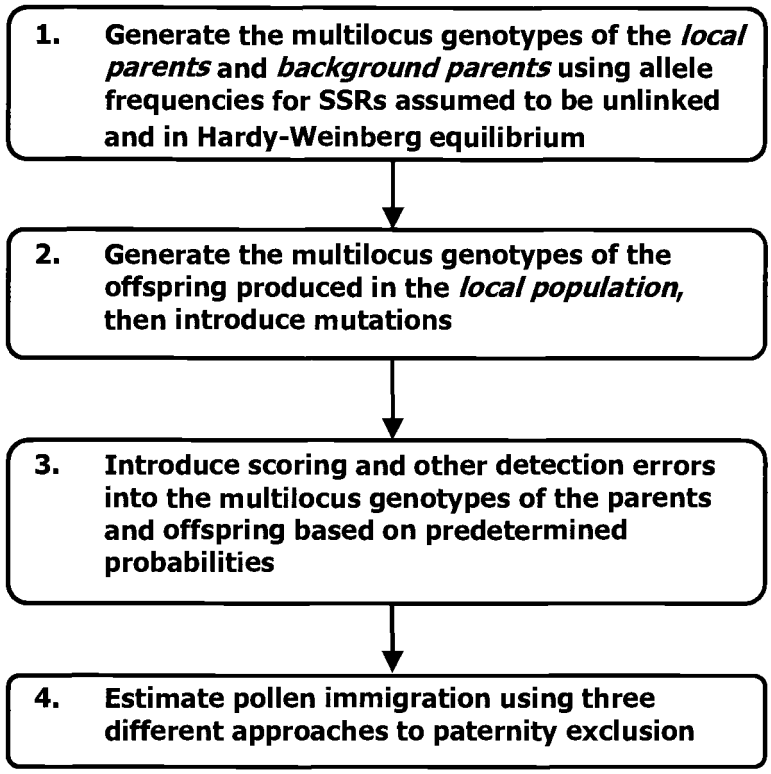


Figure 3.2. Simulation structure.

To develop broadly applicable recommendations, we tested three approaches to paternity exclusion for a range of scenarios (Table 3.1). We evaluated the two most commonly used sampling schemes for estimating pollen immigration in plant populations. The first sampling scheme can only be used with species in which a large amount of haploid megagametophyte tissue is available in developed seeds (e.g., gymnosperms from the *Pinaceae* family; Pichot and El Maataoui 1997). In these species, the haplotype of the paternal gamete can be derived by comparing the diploid multilocus genotype of the seed embryo with the haplotype of the megagametophyte (Müller 1976; Adams et al. 1988). Although this sampling scheme requires two genotyping assays per seed (embryo and megagametophyte), it is often used to estimate pollen immigration into conifer seed orchards based on bulked seed samples (Smith and Adams 1983; Adams et al. 1997; Pakkanen et al. 2000). We will hereafter refer to this sampling approach as the *haploid* scheme. Under the second sampling scheme, open-pollinated progeny arrays are sampled from a number of mothers in the *local population.* The paternal haplotypes of the offspring from each open-pollinated family are inferred by comparing their diploid multilocus genotypes to that of their mother. Using this scheme it is not always possible to infer the complete haplotype of the paternal gamete. For loci at which the mother and the offspring are both heterozygous and have the same genotype, the allele contributed by the father cannot be unambiguously determined and is treated as missing information (Devlin and Ellstrand 1990). We will refer to this sampling scheme, which is applicable to all diploid seed plants, as the *diploid* scheme.

Table 3.1. Key simulation variables used to evaluate the effect of mistyping on pollen immigration estimates obtained by paternity exclusion.

Variable	Symbol	No. of levels	Values or designations of the different levels
Sampling scheme	-	2	*Haploid*; *Diploid*
Number of *local parents*	N_l	5	30; 60; 120; 300; 600
Actual pollen immigration	m	4	0.10; 0.30; 0.50; 0.80
Rate of mistyping in parents and offspring	ε	4	*Minor* ($\varepsilon_{parents}$ = 1.0%, $\varepsilon_{offspring}$= 1.5%) *Low* ($\varepsilon_{parents}$ = 2.5%, $\varepsilon_{offspring}$= 4.5%) *Moderate* ($\varepsilon_{parents}$ = 4.0%, $\varepsilon_{offspring}$= 7.5%) *High* ($\varepsilon_{parents}$ = 5.5%, $\varepsilon_{offspring}$= 10.5%)
Number of SSR loci			
Haploid	-	5	3-7
Diploid	-	8	3-10

We define mistyping as any difference between the observed allele of an individual (parent or offspring) and the allele that would have been observed if there were no null alleles, mutations, or detection errors. To make our generated data realistic, we simulated several sources of mistyping (discussed below). The overall probabilities of mistyping and their subdivision into different sources of mistyping were chosen based on our experience with SSR data and published estimates from paternity analyses (Brinkmann et al 1998; Ewen et al. 2000; Slate et al. 2000). We set the lower bound for mistyping at 1.0%, assuming that estimates of mistyping from human paternity data (0.3-2.4%; Ewen et al. 2000) are a best-case scenario for studies of pollen immigration in plants. We set the upper bound at 10.5% because an alternative set of markers or a different analytical approach would likely be more appropriate if the average per-locus

rate of mistyping substantially exceeded 10%. We assumed that the frequency of mistyping was the same at all loci. In reality, the frequencies of mutations, null alleles, and detection errors vary widely among loci (Brinkmann et al 1998; Ewen et al. 2000; Slate et al. 2000). Nonetheless, for paternity exclusion, an equal rate of mistyping at all loci is the worst-case scenario (i.e., if mistyping frequencies are high at only a few loci, these loci can be easily identified and discarded). We made this conservative assumption because we wanted to determine which analytical approaches work well even under these unfavorable conditions. To confirm the validity of our analytical procedures, we also generated and analyzed data sets without introducing mistyping into the genotypes of the parents or offspring.

In addition to the rates of mistyping, key factors influencing the accuracy of estimates of pollen immigration are the number of *local parents*, the actual level of pollen immigration, and the number of SSR loci analyzed. We varied the number of *local parents* from 30 to 600 and the actual pollen immigration rate from 0.10 to 0.80 (Table 3.1). These ranges span the values reported in previous studies of pollen immigration (e.g., Smith and Adams 1983; Ellstrand and Marshall 1985; Meagher1986; Devlin and Ellstarnd 1990; Schnabel and Hamrick 1995; Dow and Ashley 1996; Streiff et al. 1999). Because the effect of varying the number of offspring is predictable (i.e., pollen immigration estimates get more precise as the number of offspring analyzed is increased), we set that number to 200 in all simulations for both sampling schemes. Sensitivity analyses showed that pollen immigration estimates can be strongly biased if allele frequencies in the *background population* are poorly estimated. Allele frequencies in the

background population are likely to be estimated poorly when too few *background parents* are sampled. Sampling 100 *background parents* appeared to be enough to minimize these biases, whereas increasing this number further had little effect (data not shown). Therefore, we fixed the number of parents sampled in the *background population* at 100. Our analytical procedures were considered adequate if they produced pollen immigration estimates with a (1) mean bias ≤ 0.03 and (2) empirical standard deviation ≤ 0.05. These criteria exceed the precision with which pollen immigration is typically measured (e.g., Adams et al. 1997; Pakkanen et al. 2000). Given the rates of mistyping and the sizes of the local populations that we tested, our preliminary simulations suggested that adequate estimates of pollen immigration can be obtained with seven or fewer highly variable SSR loci in the *haploid* sampling scheme, and ten or fewer loci in the *diploid* sampling scheme. Therefore, we varied the number of SSR loci analyzed from three to seven for the *haploid* scheme, and from three to ten for the *diploid* scheme.

We simulated 400 scenarios for the *haploid* scheme and 640 scenarios for the *diploid* scheme. Within each sampling scheme, these scenarios represented complete factorial experiments with respect to our simulation variables (Table 3.1). For each scenario, we generated ten sets of 200 offspring genotypes for each of ten *local populations* (i.e., ten independently generated sets of genotypes of *local parents*; see Generating parents below). Thus, we generated and analyzed a total of 100 samples of 200 offspring per scenario. This approach was used to minimize the possibility of obtaining biased results because of unusual samples of parental genotypes (i.e., if

parental genotypes were only sampled once per scenario), and also allowed us to assess the empirical variances of estimators of pollen immigration among randomly generated samples of offspring genotypes. We analyzed each simulated set of 200 offspring in three different ways. First, we estimated pollen immigration by classifying offspring as immigrants when their microgamete haplotypes mismatched all *local parents* at one or more loci. Because even a mismatch at a single locus leads to paternity exclusion, this approach assumes that the SSR data are perfect. Second, we obtained a pollen immigration estimate by requiring at least two father-offspring mismatches for paternity exclusion, an approach that is commonly used in human paternity cases to avoid false inferences caused by mistyping (Brenner 2004). Finally, we estimated pollen immigration by excluding local parents from paternity based on at least three mismatches, which is an even more conservative adjustment for mistyping.

Data generation

Generating parents

Allele frequencies for a pool of 20 highly variable dinucleotide SSRs were obtained from previous studies of tree populations (Table 3.2). These allele frequencies were calculated from the diploid genotypes of adult trees and did not include null alleles because the frequency of null alleles cannot be directly estimated without analyzing a large number of families. The frequency of null alleles was set to different levels for different sets of analyses (described below). Alleles at all loci were classified into 2-bp size classes (bins). A set of k SSR loci was drawn at random without replacement and used to generate one

set of *local parents* and ten sets of 200 offspring produced by these parents. Single-locus genotypes for each of the N_l *local parents* and N_b *background parents* were generated by randomly drawing two alleles at each locus. The probability of drawing an allele equaled its frequency. Multilocus genotypes were formed by joining the k independently generated, single-locus genotypes. Thus, we assumed that the *local population* and the *background population* are in Hardy-Weinberg equilibrium and gametic equilibrium for the SSR loci used. Another assumption was that there was no allele frequency differentiation between the *local population* and the *background population*. Violations of this assumption lead to biases in pollen immigration estimates if allele frequencies in the *background population* are poorly estimated (discussed above).

The presence of null alleles often leads to wrongly scoring heterozygous genotypes comprised of one 'visible' allele and one null allele as homozygotes for the visible allele. Thus, null alleles in the *local parents* are a major source of mistyping (Pemberton et al. 1995; Ewen et al. 2000). We introduced null alleles by converting randomly selected visible alleles in the diploid genotypes of the simulated parents into null alleles. This was done with a probability dependent on the overall rate of mistyping (discussed below). Null alleles were passed on from parents to offspring in the same manner as were visible alleles. This approach assumes that null alleles are (1) passed on according to Mendelian laws and (2) present at all loci with approximately equal frequencies. The first assumption is consistent with empirical data (Pemberton et al.1995; Ewen et al. 2000). The rationale for the second assumption was described in the previous section.

Table 3.2. SSR loci used to generate multilocus genotypes.

Locus name	H_e[a]	A[b]
PmOSU_2C3[c]	0.964	45
PmOSU_3B2[c]	0.964	39
PmOSU_2G12[c]	0.928	32
PmOSU_1C3[c]	0.963	31
PmOSU_1F9[c]	0.972	44
PmOSU_2D4[c]	0.964	38
PmOSU_2D6[c]	0.969	50
PmOSU_3B9[c]	0.936	36
PmOSU_3D5[c]	0.931	29
PmOSU_3F1[c]	0.951	34
PmOSU_3G9[c]	0.921	31
PmOSU_4A7[c]	0.964	53
PmOSU_4G2[c]	0.899	16
AG1[d]	0.906	33
P2011[d]	0.821	15
P2156[d]	0.877	21
P2235[d]	0.865	22
P420[d]	0.850	17
P433[d]	0.901	23
P684[d]	0.823	19

[a] H_e is the expected heterozygosity of the markers.
[b] A is the total number of alleles detected at each locus.
[c] SSR locus developed for *Pseudotsuga menziesii* (Chapter 2).
[d] SSR locus developed for *Populus trichocarpa* (DiFazio 2002).

Generating offspring

To make our generated data realistic, we assumed that *local parents* differed in their reproductive success. Each time a new *local population* was generated, coefficients of female and male reproductive success were randomly assigned to all *local parents* based on pseudorandom numbers drawn from two separate but correlated negative exponential distributions. The negative exponential distribution was chosen because it is consistent

with empirically determined patterns of female and male reproductive success (Meagher and Thompson 1987; Smouse and Meagher 1994; Krauss 2000). We also determined that using uniform or normal distributions for assigning coefficients of reproductive success did not affect our estimates of pollen immigration. The correlation between female and male reproductive success coefficients was set to 0.40, which is consistent with published estimates of this correlation in natural populations and seed orchards (Savolainen et al. 1993; Krauss 2000; Kang and El-Kassaby 2002).

Haploid sampling scheme

A megagamete haplotype (megagamete) and a microgamete haplotype (microgamete) were generated for each of the 200 offspring. First, a local female parent was sampled with a probability equal to its reproductive success coefficient. Second, an allele for each locus was drawn at random from the genotype of this female to generate a *k*-locus megagamete. Third, a decision of whether to generate an immigrant or a non-immigrant (locally produced) microgamete was made according to the specified level of pollen immigration (m). Immigrant microgametes were generated by randomly and independently drawing alleles for the k loci based on the source allele frequencies. This approach assumes that the selected loci are in gametic equilibrium. Locally produced microgametes were generated in the same manner as were the megagametes, but the contributing *local parents* were selected based on their coefficients of male reproductive success. Finally, the resulting mega- and microgametes were randomly coupled to form diploid offspring genotypes.

Diploid sampling scheme

Offspring genotypes for the *diploid* sampling scheme were generated using the procedure described above. The only exception was that no female reproductive success coefficients were used because under this sampling scheme, an approximately equal number of offspring is typically collected from pre-determined mothers (Devlin and Ellstrand 1990). In our simulations, 20 *local parents* were randomly chosen to contribute ten offspring to each sample (i.e., $n = 20 \times 10 = 200$). Megagametes were generated by drawing alleles at random from the genotype of the respective contributing mother, whereas microgametes were formed as described above.

Introducing mutations in offspring

We simulated two types of mutations (Table 3.3). SSR sequence mutations were simulated by randomly adding or subtracting 2 bp (i.e., one repeat unit) to the allele sizes in the genotypes of the offspring. Increases and decreases in allele size each occurred with a probability 2×10^{-3}. This approach assumes a stepwise mutation model in which alleles are equally likely to increase or shrink in size by one repeat unit. Although this mutation model may not fit observed allele size distributions in natural populations (Estoup and Cornuet 1999), mutations leading to size alterations by more than a single repeat unit are rare over a single generation, and increases and decreases in allele size appear to be equally likely (Brinkmann et al. 1998; Ewen et al. 2000; Anon. 2002). Visible alleles in the parents can become null alleles in the offspring (i.e., *de-novo* null alleles) because of DNA sequence mutations immediately adjacent to the SSRs (i.e.,

sequences targeted by PCR primers). *De-novo* null alleles were set to occur with probability 10^{-3}. Thus, the overall mutation rate was set at 5×10^{-3} (i.e., $(2+2+1)\times10^{-3}$), a value consistent with estimates from human paternity data (Brinkmann et al. 1998 and references therein; Ewen et al. 2000; Anon. 2002). Empirical data suggest that mutation rates both within and adjacent to SSRs are negligibly small compared to the frequencies of mistyping caused by pre-existing null alleles in the genotypes of the parents or by detection errors (Ewen et al. 2000). In our sensitivity analyses, varying these mutation rates had little effect on the estimates of pollen immigration. Therefore, we set the probabilities of occurrence of mutations in the SSR sequences and in the sequences targeted by PCR primers at levels representative of published estimates and did not vary these probabilities in our simulations.

Introducing genotype detection errors in parents and offspring

In addition to mistyping caused by genetic reasons (i.e., null alleles and mutations), we also introduced mistyping caused by five different types of errors that typically occur during the detection of parent and offspring SSR genotypes (Table 3.3). For parental genotypes, we held the frequencies of the different detection errors constant and set them at a relatively low level (Table 3.3). Thus, the four total rates of mistyping in the genotypes of the parents differed only because of the different frequencies of null alleles, whereas in the genotypes of the offspring, the total rates of mistyping differed because of (1) the different frequencies of null alleles and (2) the different rates of genotype detection error (Table 3.3). We chose this approach because replicated tissue samples or

progeny arrays from the *local parents* are often available and analyzed simultaneously. This helps reduce rates of mistyping in the genotypes of the *local parents* compared to the genotypes of the offspring.

Data analysis

We developed the Pollen Flow (PFL) computer program (Appendix 2) that can be used to perform all analyses described below.

Haploid sampling scheme

When we did not account for mistyping, we used the paternity exclusion approach described by Smith and Adams (1983) and Adams et al. (1997). Using this approach, pollen immigration is estimated as:

$$\hat{m} = \frac{\hat{b}}{\hat{d}} \qquad [1]$$

where $\hat{b}$ is the observed proportion of immigrants and $\hat{d}$ is the detection probability; i.e., the probability that the multilocus haplotype of an immigrant microgamete will be distinguishable from all *local pollen haplotypes* based on the genetic markers available (Smith and Adams 1983). The detection probability can be estimated as:

$$\hat{d} = 1 - \sum_{i=1}^{t} \hat{h}_i \qquad [2]$$

where $\hat{h}_i$ is the frequency of local haplotype i in the background population and t is the total number of distinct *local pollen haplotypes* (Smith and Adams 1983).

Table 3.3. Simulated sources of mistyping and their probabilities (%) per allele.

	Rates of mistyping in parents (%)				Rates of mistyping in offspring (%)				
	Minor	*Low*	*Moderate*	*High*	*Minor*	*Low*	*Moderate*	*High*	Reference
I. Genetic reasons for mistyping									
Pre-existing null alleles[a]	0.500	2.000	3.500	5.000	0.500	2.000	3.500	5.000	Ewen et al. 2000; Anon. 2002
Mutations in SSRs[b]	-	-	-	-	0.400	0.400	0.400	0.400	Brinkmann et al 1998; Ewen et al. 2000
De-novo null alleles[c]	-	-	-	-	0.100	0.100	0.100	0.100	Brohede and Ellegren 1999
II. Mistyping caused by detection errors									
Incorrect sizing[d]	0.100	0.100	0.100	0.100	0.100	0.400	0.700	1.000	Ewen et al. 2000
Missing alleles[e]	0.275	0.275	0.275	0.275	0.275	1.100	1.925	2.750	Ewen et al. 2000
Sample contamination[f]	0.040	0.040	0.040	0.040	0.040	0.160	0.280	0.400	Ewen et al. 2000
Signal leakage[g]	0.005	0.005	0.005	0.005	0.005	0.020	0.035	0.050	Ewen et al. 2000
Sample swaps[h]	0.080	0.080	0.080	0.080	0.080	0.320	0.560	0.800	Ewen et al. 2000
Total mistyping	**1.000**	**2.500**	**4.000**	**5.500**	**1.500**	**4.500**	**7.500**	**10.500**	

[a] Pre-existing null alleles (i.e., null alleles that were present in the genotypes of the *local parents*) occurred in the offspring with the same frequencies as in the *local parents*.
[b] Mutations in SSRs were simulated by adding or subtracting one repeat unit (i.e., 2 bp) to the allele sizes in the genotypes of the offspring. Increasing or decreasing the size of an allele each occurred with a probability 0.2% (total mutation rate = 0.4%), regardless of the overall rate of mistyping.
[c] *De-novo* null alleles are conversions of visible alleles in the parents into null alleles in the offspring because of *de-novo* mutations in the sequences targeted by PCR primers.
[d] Incorrect allele sizing (misclassification to allele bins) was simulated by adding or subtracting 2 bp to the original allele sizes. Increases and decreases in allele size were assumed to be equally likely.
[e] Missing alleles (i.e., failure to record one of the alleles in a heterozygous genotype) was introduced by deleting one of the alleles in a heterozygote, thus changing its genotype to a homozygote. The frequencies in the table represent the average probabilities of occurrence of this type of detection error. Because alleles with larger sizes are more likely to remain undetected, the alleles with greater sizes in heterozygotes were assumed to be 'missed' ten times more often than the alleles with smaller sizes. In the genotypes of the offspring, the paternally contributed alleles were assumed to be missed four times more often than the maternally contributed alleles. This is because the maternal genotype is typically available when scoring offspring genotypes and a simple check would allow missed alleles to be identified.
[f] Sample contamination was simulated by replacing the 'true' allele of a parent or offspring with a randomly selected allele from the same locus from a different, randomly selected parent or offspring.
[g] Signal leakage (i.e., scoring an allele from a locus labeled with a different fluorescent dye in multiplexed PCR products) was simulated by replacing the 'true' allele of an individual (parent or offspring) with a randomly selected allele from the same individual, but from a different randomly selected locus.
[h] Sample swapping was introduced by exchanging the genotypes between two randomly selected parents for a single randomly selected locus. This is a simplification of the more plausible sample swaps in which more than two samples are misplaced (e.g., as a result of inverting a multi-channel pipette during PCR plate preparation), but the final effect on the analysis was assumed to be similar.

An alternative way to estimate d is to generate a large random sample of immigrant multilocus haplotypes (i.e., using the allele frequencies estimated from the genotypes of the *background parents*), and then empirically determine the proportion of these haplotypes that differ from all *local pollen haplotypes*.

A simple approximation for the standard error of $\hat{m}$ is

$$SE(\hat{m}) \approx \frac{1}{\hat{d}}\sqrt{\frac{\hat{b}(1-\hat{b})}{n}} = \hat{m}\sqrt{\frac{1-\hat{b}}{\hat{b}n}} \qquad [3]$$

where n is the number of offspring analyzed. In this approximation, $\hat{d}$ is treated as a constant and its variance is ignored. This leads to underestimating $SE(m)$ when d is substantially lower than one and varies appreciably, which is typical with allozyme markers (Smith and Adams 1983; Adams et al. 1997). We calculated the empirical variance of $\hat{d}$ within sets of parents to test whether this variance is small enough to allow equation [3] to be used with highly variable SSR markers. We also evaluated the accuracy of this approximation by comparing estimates obtained using equation [3] to empirical standard deviations of $\hat{m}$ over ten samples of offspring within each set of *local parents*.

One of the adjustments for mistyping that we tested was to assume that a father-offspring mismatch at one locus could occur simply as a result of mistyping. Following this approach, we determined the observed proportion of immigrants ($\hat{b}$) by classifying offspring as immigrants only when their microgamete haplotypes mismatched all *local pollen haplotypes* at two or more loci. Equation [2], however, can only be used to

estimate the detection probability for exclusion based on a single mismatch. We could not find a straightforward extension of this equation to analyses in which at least two mismatches are required for paternity exclusion. Therefore, we generated 10,000 random immigrant haplotypes using the allele frequencies estimated from the genotypes of the *background parents,* and then estimated *d* as the proportion of these haplotypes that mismatched all *local pollen haplotypes* at two or more loci. Pollen immigration estimates and their standard errors were then estimated using equations [1] and [3]. This approach assumes that *d* is independent of the rate of mistyping. We tested this assumption by comparing estimates of *d* obtained with and without mistyping in the 10,000 random immigrant haplotypes for several sample scenarios, with rates of mistyping varying from *minor* to *high* (Table 3.1). Based on 100 comparisons for each scenario, estimates based on haplotypes without mistyping were consistently lower than estimates based on haplotypes with mistyping. Even for our highest simulated level of mistyping, however, the differences were small (≤ 0.02). Therefore, in our analyses we did not introduce mistyping in the randomly generated immigrant haplotypes used to estimate *d.*

As a more conservative adjustment for mistyping, we also analyzed each dataset requiring at least three mismatches for paternity exclusion. The estimates of pollen immigration and their standard errors were obtained as described above.

Diploid sampling scheme

When the haplotype of the seed megagametophyte is not available, the paternal haplotype can be inferred by comparing the diploid genotype of the mother with that of the

offspring. We treated loci for which the paternally contributed alleles were ambiguous (i.e., when the mother and the offspring were both heterozygous and had identical genotypes) as missing data. Because each mother typically has a different multilocus genotype, the detection probabilities for progeny arrays sampled from different mothers will also be different. Therefore, we obtained separate estimates of d and m for each of the 20 mothers from which offspring were sampled in each simulation run. For each open-pollinated progeny array, we calculated $\hat{b}$ as the proportion of detected immigrants after comparing the inferred microgametes with all *local pollen haplotypes*. To estimate d, we generated 2000 random immigrant microgametes (as described above) and 2000 megagametes from the source mother, randomly coupled the micro- and megagametes into 2000 diploid offspring, and then determined the proportion of offspring whose microgametes mismatched all *local pollen haplotypes* at one or more loci for which the paternally contributed alleles could be determined unambiguously. Finally, we obtained 20 estimates of m (i.e., one from each progeny array) using equation [1] and used their average as an overall estimate of pollen immigration. We used this estimate, as well as the average $\hat{b}$ across the progeny arrays from the 20 source mothers to calculate $\hat{SE}(m)$ using equation [3] (n = 200). This algorithm was repeated to obtain estimates of m and SE(m) when the minimum number of mismatches required for paternity exclusion was set to two and three.

RESULTS

This section summarizes the results from the 1040 scenarios that we simulated. The trends discussed below were consistent across these scenarios, unless otherwise indicated. Results from specific scenarios are available from G. T. Slavov upon request.

Mistyping results in substantial overestimates of pollen immigration

Simple paternity exclusion (i.e., requiring one or more mismatching loci for exclusion) resulted in unbiased estimates of pollen immigration when no mistyping was present in the genotypes of parents and offspring (Fig. 3.3A,C). When mistyping was present, however, this approach resulted in considerable upward biases in pollen immigration estimates (Fig. 3.3B,D). Within the range of scenarios that we tested (Table 3.1), these biases appeared to increase linearly with the number of SSR loci used in the analysis. This was true for situations in which the complete paternal haplotype could be inferred (Fig. 3.3B) and when this was not possible (Fig. 3.3D). Although Fig. 3.3 only shows results for pollen immigration set at 10% (0.10), mistyping led to substantial overestimates for all values of m that we tested (0.10, 0.30, 0.50, 0.80). Even the minor rates of mistyping (i.e., $\varepsilon_{offspring}$= 1.5%) resulted in overestimates of pollen immigration by up to 150% (e.g., 0.25 estimated vs. 0.10 actual; Fig. 3.3D). Higher rates of mistyping led to even greater biases in the pollen immigration estimates, and for a given rate of mistyping, biases were greater at lower levels of actual pollen immigration. For example, when ignored, mistyping resulted in upward biases as high as 0.43 when mistyping was

high and actual gene flow was low ($\varepsilon_{offspring}$= 10.5%; m = 0.10), and up to 0.10 when both mistyping and actual gene flow were high ($\varepsilon_{offspring}$= 10.5%; m = 0.80).

Cryptic gene flow is negligible when 5-8 highly variable SSR loci are used

With a probability equal to 1-d an immigrant offspring will have a paternal haplotype that is identical to at least one of the *local pollen haplotypes*. Estimates of m based only on observed immigrants (i.e., not adjusted for d) will underestimate actual pollen immigration (Devlin and Ellstrand 1990; Dow and Ashley 1996).

Although the number of loci required to reach a given value of $\hat{d}$ increases with the number of local parents (Fig. 3.4A, B), only 5-6 highly variable SSR markers were necessary to achieve $\hat{d}$ > 0.99 with 30-600 local parents using the *haploid* sampling scheme (Fig. 3.4A). For the *diploid* scheme, 7-8 loci were sufficient for $\hat{d}$ to exceed 0.99 (Fig. 3.4B).

For a fixed number of *local parents*, detection probabilities decrease with increasing the number of mismatching loci required for paternity exclusion. Approximately two additional highly variable markers are needed to compensate for the loss of detection probability caused by each additional mismatching locus required for exclusion (Fig. 3.5). For example, the number of loci needed to achieve $\hat{d} \geq 0.80$ with 120 *local parents* was four when exclusion was based on a single mismatch, six when two or more mismatching loci were required for exclusion, and eight when at least three mismatches were required for exclusion (Fig. 3.5).

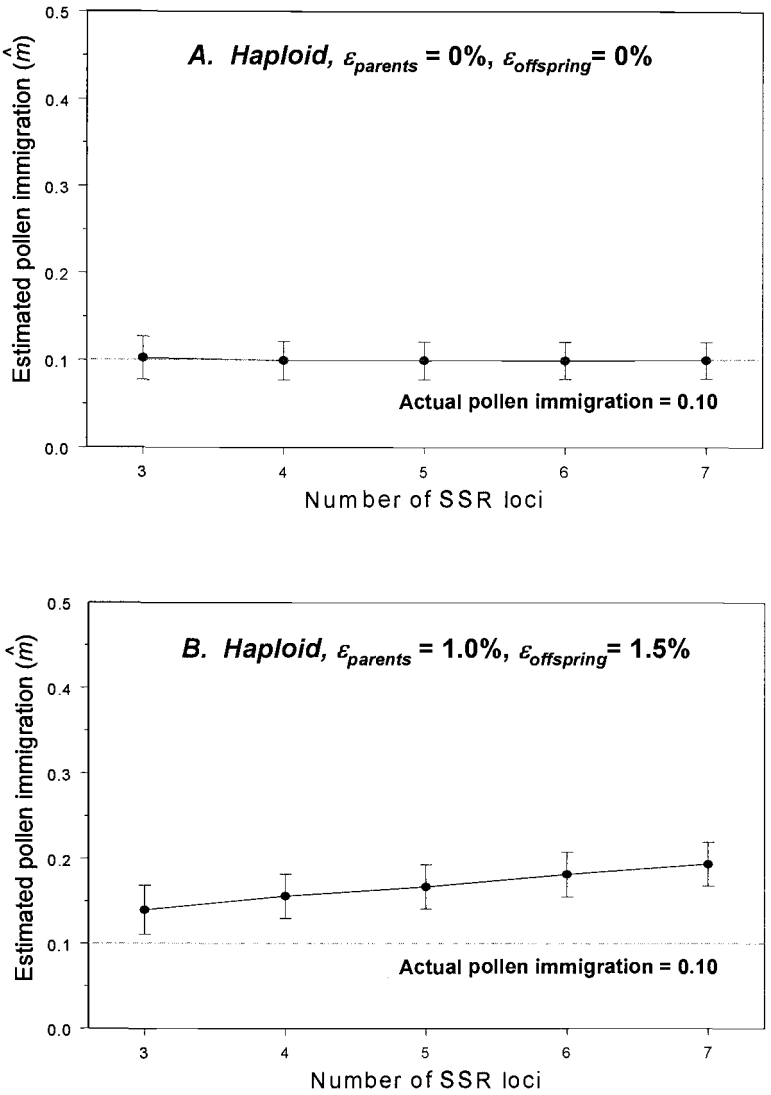
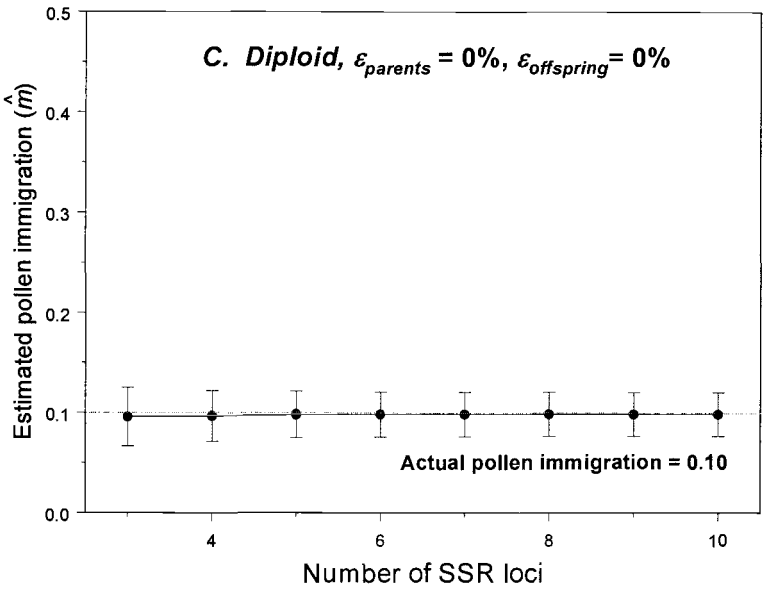


Figure 3.3. Pollen immigration estimated via paternity exclusion without accounting for mistyping (actual $m = 0.10$, $N_l = 120$). Each data point is the mean pollen immigration estimate over 100 samples of 200 offspring (i.e., 10 sets of *local parents* × 10 samples of 200 offspring within each set of *local parents*). The error bars represent the mean empirical standard deviations of pollen immigration estimates (i.e., the empirical standard deviations over 10 samples of offspring within a set of *local parents*, averaged over 10 sets of *local parents*). (A) Paternal gamete haplotypes were unambiguously inferred for all loci (*haploid* scheme), and there was no mistyping. (B) Same as A but with the *minor* rate of mistyping. (C) Paternal gamete haplotypes could not be inferred for loci at which the mother and the sampled offspring were heterozygous and had the same alleles (*diploid* scheme), and there was no mistyping. (D) Same as C but with the *minor* rate of mistyping.



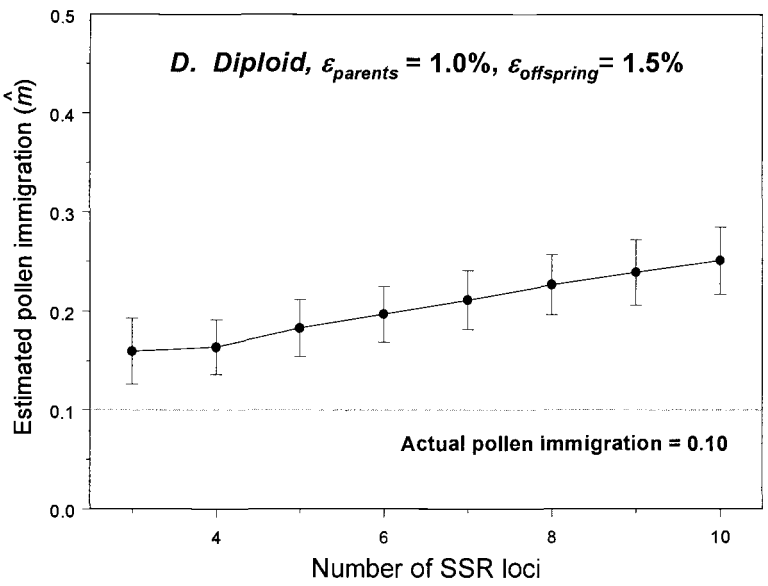


Figure 3.3. Continued

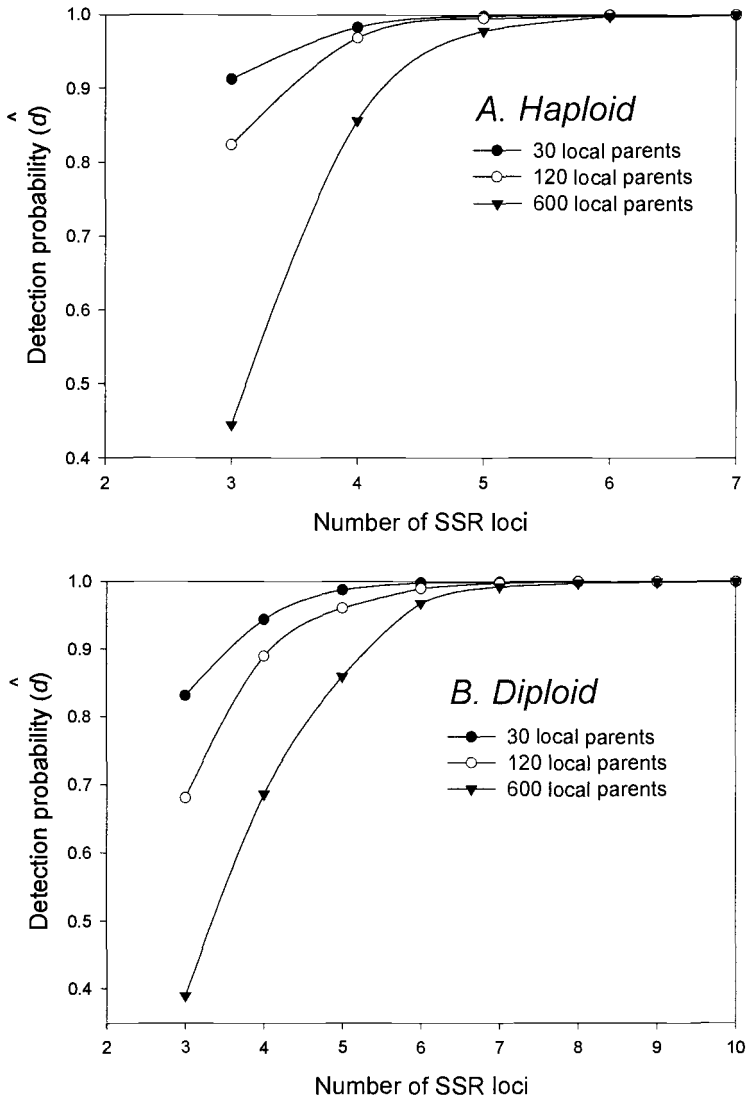


Figure 3.4. Detection probabilities when there is no mistyping. Each data point is an average of ten replicates of 200 simulated offspring samples generated within each of ten sets of 30, 120, or 600 simulated *local parents* (i.e., a total of 10 × 10 = 100 samples of 200 offspring per data point). (A) Paternal gamete haplotypes were unambiguously inferred for all loci (*haploid* scheme). (B) Paternal gamete haplotypes were unambiguously inferred for all loci, except those at which the mother and the sampled offspring were heterozygous and had the same alleles (*diploid* scheme).

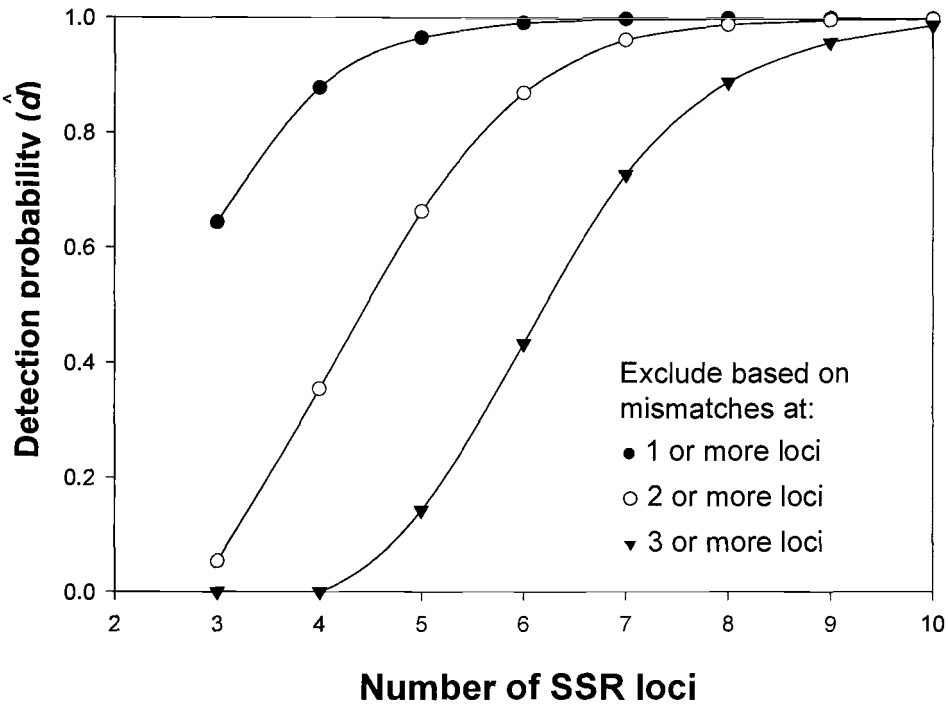


Figure 3.5. Detection probabilities when father-offspring mismatches are required at multiple loci for exclusion. Each data point is the average of ten sets of 200 simulated offspring genotypes generated within ten sets of simulated parent genotypes under the *diploid* sampling scheme ($N_I = 120$).

Requiring multiple mismatches for exclusion results in accurate estimates of pollen immigration

In all scenarios, estimates of *m* with acceptable biases and variances could be obtained when $\hat{d}$ was sufficiently high (i.e., $\hat{d} \geq 0.80$) and when an appropriate adjustment for mistyping was applied. For *minor* rates of mistyping ($\varepsilon_{offspring}$=1.5%), requiring two or more mismatches for exclusion resulted in estimates of pollen immigration with little or no bias (Fig. 3.6A). The same was true for *moderate* rates of mistyping ($\varepsilon_{offspring}$=4.5%) (data not shown). Given $\hat{d} > 0.80$, pollen immigration estimates with acceptable biases (i.e., bias ≤ 0.03) could be obtained for all scenarios by requiring three or more mismatches for exclusion (e.g., Fig. 3.6B).

When $\hat{d} \geq 0.80$, estimates of *m* had empirical standard deviations of less than 0.05 (e.g., Fig. 3.6C) and the empirically determined variances of $\hat{d}$ were negligibly small ($\mathrm{Var}(\hat{d}) \leq 2 \times 10^{-4}$) regardless of which scenario was simulated. With $\hat{d} > 0.80$, the difference between the empirical standard deviation and the approximated standard error$[SE(\hat{m})]$ was also small (e.g., mean difference across 240 scenarios = 0.003, standard deviation = 0.005).

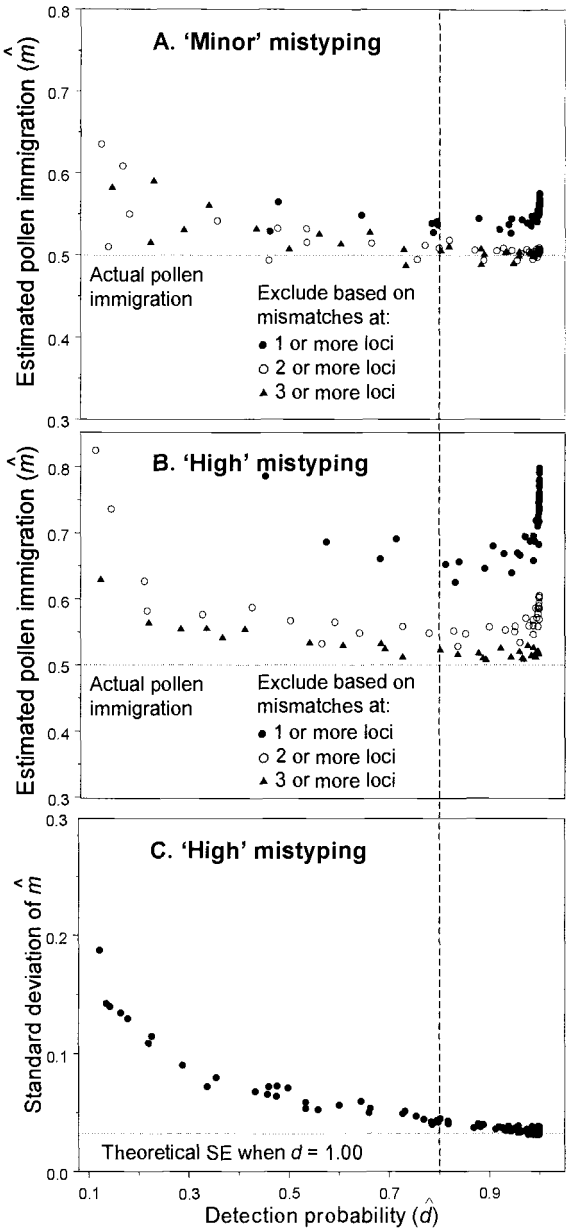


Figure 3.6. Pollen immigration estimates and their empirical standard deviations under the *diploid* sampling scheme, actual $m = 0.50$. (A) *Minor* mistyping (Table 3.1.). Each data point represents the average pollen immigration over ten sets of 200 simulated offspring genotypes generated within ten sets of N_f (30, 60, 120, 300, or 600) simulated *local parents* using 3-10 SSR loci and requiring at least 1-3 mismatching loci for exclusion. (B) Same as in A but with the *high* rate of mistyping. (C) Average empirical standard deviations of the pollen immigration estimates shown in B. The horizontal dashed lines mark the theoretical standard error of m when $d = 1.0$ and $n = 200$ (i.e., SE(m) = 0.035 as calculated using equation [3]). For detection probabilities greater than 0.80, the empirical standard deviations approach their theoretical minimum.

DISCUSSION

Mistyping results in substantial overestimates of pollen immigration

If ignored, even minor rates of mistyping will lead to inflated pollen immigration estimates obtained by paternity exclusion (Figs. 3.3, 3.6). Thus, mistyping should always be estimated when SSRs are used to measure pollen immigration. Mistyping can be estimated by analyses of known pedigrees or comparisons of duplicate samples (Ewen et al. 2000). Furthermore, planning to account for father-offspring mismatches caused by mistyping should be a common practice when determining the number of SSR loci needed to reliably estimate pollen immigration via paternity exclusion.

The biases in $\hat{m}$ caused by mistyping were larger at low levels of pollen immigration and when more loci were analyzed. The former result is not surprising because the number of false exclusions due to mistyping increases with the number of offspring resulting from mating between local parents. The latter relationship occurred because we assumed equal rates of mistyping for all loci. Therefore, the probability of a spurious mismatch (i.e., a mismatch caused by mistyping) between a true father and its offspring increased linearly as more loci were added. Thus, pollen immigration estimates are particularly sensitive to mistyping when the actual pollen immigration is low and many loci are used in the analysis.

Cryptic gene flow is negligible when 5-8 highly variable SSR loci are used

Cryptic gene flow is inversely related to the detection probability, which depends on (1) the number and variability of the marker loci, (2) the number of *local parents*, and (3) the number of mismatching loci required for paternity exclusion (Figs. 3.4, 3.5). The last of these factors depends on one's assessment of the quality of the genotypic data (i.e., the estimated value of ε). Thus, the number of loci needed to minimize cryptic gene flow after adjusting for mistyping will vary widely among different scenarios. Using the PFL program (Appendix 2), detection probabilities for specific sets of data can be estimated from the multilocus genotypes of the *local parents*, estimated allele frequencies for the *background population*, and assumptions about the rate of mistyping in the data. Within the parameter space that we studied (Table 3.1), proceeding with analyses via paternity exclusion will result in reliable estimates of pollen immigration if: (1) the value of $\hat{d}$ after requiring multiple mismatches for exclusion exceeds 0.80, and (2) the rate of mistyping assumed by the user is accurate or conservative (i.e., if the user is not substantially underestimating the rate of mistyping) (Fig. 3.7).

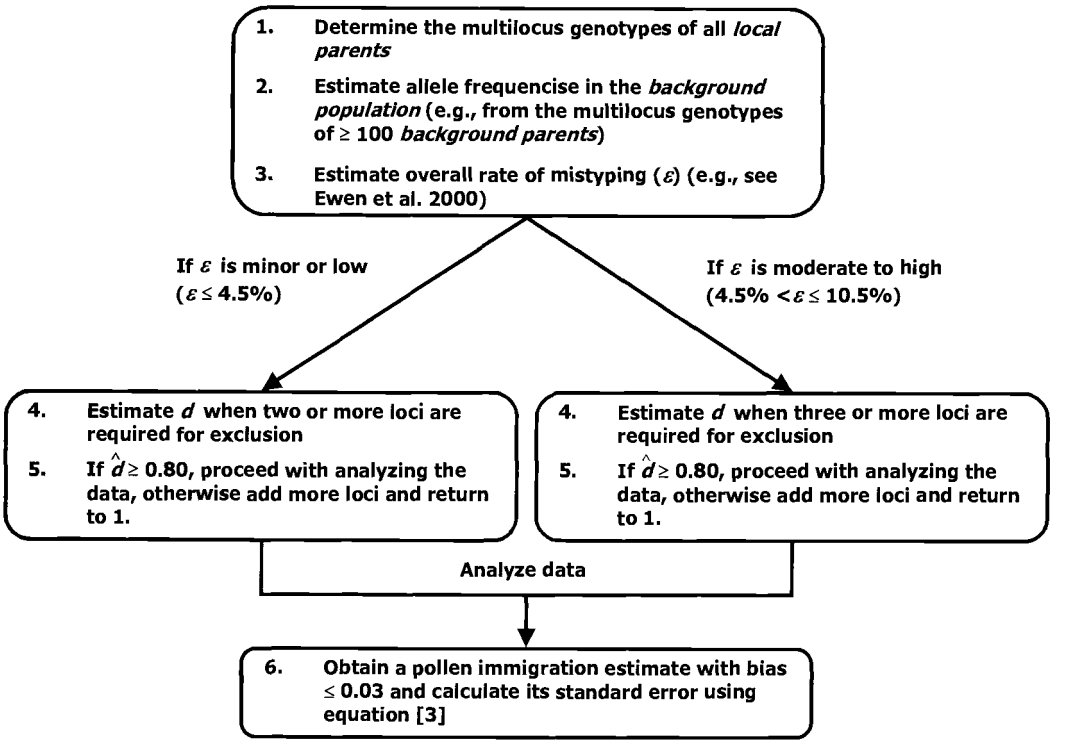


Figure 3.7. Recommended procedure for determining the minimum number of SSR loci needed to obtain a pollen immigration estimate with a bias ≤ 0.03 and a standard error that can be approximated using equation [3]. Detection probabilities reflecting exclusion based on mismatches at multiple loci can be estimated using the PFL program (Appendix 2).

Our method of estimating detection probabilities is based on two assumptions: (1) the background population is an infinitely large population in Hardy-Weinberg equilibrium and gametic equilibrium, and (2) the frequencies of immigrant multilocus haplotypes can be predicted using the allele frequencies estimated from the genotypes of the *background parents*. If a few highly fertile parents in the *background population* contribute most of the immigrant pollen and their multilocus genotypes are identical or very similar to those of some local parent(s), analyses under our assumptions will result in inflated estimates of d, and underestimates of m. This scenario, however, becomes increasingly unlikely as the number of SSR loci analyzed increases. An alternative to sampling *background parents* is to jointly estimate pollen immigration and allele frequencies in the *background population* from the paternal haplotypes of the offspring using a neighborhood mating model (Burczyk et al. 2002; Burczyk and Chybicki 2004).

Requiring multiple mismatches for exclusion results in accurate estimates of pollen immigration

In our simulations, adjustments for mistyping produced consistently accurate and precise estimates of pollen immigration only when $\hat{d}$ exceeded 0.80 (Fig. 3.6A, B). We identified two reasons for the poor performance of our adjustments for mistyping when $\hat{d}$ was low. First, with low detection probability, the observed proportion of pollen immigrants is low. Thus, $\hat{b}$ is more sensitive to false exclusions caused by mistyping, which leads to greater upward biases in pollen immigration estimates (Fig. 3.6A, B). Second, the variance of pollen immigration estimates is inversely related to $\hat{d}$ (equation [3]; Fig. 3.6C).

The effects of *minor* to *low* rates of mistyping (i.e., $\varepsilon_{offspring} \leq 4.5\%$) can be minimized by requiring mismatches with the genotypes of all *local parents* at two or more loci before classifying offspring as immigrants (Fig. 3.6A). Biases caused by *moderate* and *high* rates of mistyping (i.e., $4.5 < \varepsilon_{offspring} \leq 10.5\%$) can only be mitigated by a more conservative correction for mistyping, such as requiring mismatches at three or more loci for paternity exclusion (Fig. 3.6B).

Extrapolating from our simulations, an even more conservative adjustment for mistyping will probably be necessary to account for per-locus rates of mistyping substantially higher than the *high* rate that we simulated (Table 3.1). For example, four or more mismatching loci may need to be required for exclusion to obtain pollen immigration estimates with acceptable biases and variances. We expect, however, that reliable estimates of pollen immigration would be obtained only if enough loci are available to achieve detection probabilities of at least 0.80. Minimizing the per-locus rate of mistyping through rigorous pre-screening of the available SSR loci or by combining SSRs with other less variable but more reliable genetic markers (e.g., allozymes) is probably a more appropriate approach in these situations.

The minimum number of highly variable SSR loci needed to obtain accurate estimates of pollen immigration depends on (1) whether a *haploid* or a *diploid* sampling scheme is used, (2) the number of local parents, and (3) the average rate of mistyping (Table 3.4). Given our criteria of mean bias ≤ 0.03 and empirical standard deviation ≤ 0.05, the number of loci needed varied between five, for 30 *local parents* and *minor* to

low rates of mistyping in the *haploid* scheme, and nine, for 600 local parents and *moderate* to *high* rates of mistyping in the *diploid* scheme.

Table 3.4. Minimum number of SSR loci needed to obtain estimates of pollen immigration with a bias ≤ 0.03 and a standard error ≤ 0.05, based on 200 offspring analyzed. The numbers in parentheses indicate the minimum numbers of loci at which father-offspring mismatches are required for exclusion.

Number of local parents	Rate of mistyping[a]			
	Minor	*Low*	*Moderate*	*High*
Haploid				
30	5 (2)	5 (2)	6 (3)	6 (3)
60	5 (2)	5 (2)	7 (3)	7 (3)
120	5 (2)	5 (2)	7 (3)	7 (3)
300	6 (2)	6 (2)	7 (3)	7 (3)
600	6 (2)	6 (2)	7 (3)	7 (3)
Diploid				
30	5 (2)	5 (2)	7 (3)	7 (3)
60	6 (2)	6 (2)	7 (3)	7 (3)
120	6 (2)	6 (2)	8 (3)	8 (3)
300	6 (2)	6 (2)	8 (3)	8 (3)
600	7 (2)	7 (2)	9 (3)	9 (3)

[a] See Table 3.1 for values corresponding to the rates of mistyping.

Recommendations

The minimum number of SSR loci needed to obtain reliable estimates of pollen immigration must be determined on a case-by-case basis. First, using the PFL program, researchers can evaluate the adequacy of their data for paternity exclusion. Second, we recommend that pollen immigration estimates obtained via SSR-based paternity exclusion always be adjusted for mistyping. In our simulated populations, analyses based on detection probabilities exceeding 0.80 after an appropriate adjustment for mistyping

(Fig. 3.7) consistently lead to nearly unbiased estimates of pollen immigration (i.e., bias ≤ 0.03) and standard errors that could be approximated using equation [3]. Finally, we recommend sampling at least 100 *background parents* to estimate the allele frequencies in the *background population*. This sample should include reproductively mature plants that are likely sources of immigrant pollen.

ACKNOWLEDGEMENTS

This study was supported by the Pacific Northwest Tree Improvement Research Cooperative and the Department of Forest Science at Oregon State University. We thank Jaroslaw Burczyk, Igor Chybicki, Randy Johnson, and Steve DiFazio for their helpful comments on an earlier version of this manuscript.

Chapter 4. Pollen Contamination and Mating Patterns in a Douglas-fir Seed Orchard as Measured by SSR Markers

Gancho T. Slavov, Glenn T. Howe, Wesley T. Adams

To be submitted to:

Canadian Journal of Forest Research

NRC Research Press, National Research Council of Canada

Ottawa, ON K1A 0R6, Canada

ABSTRACT

Pollen contamination is detrimental to the genetic quality of seed orchard crops. Highly variable SSR markers make it possible to accurately measure pollen contamination and to characterize patterns of within-orchard mating by directly identifying the male and female parent of each seed produced in the orchard. We used nine SSR markers to measure pollen contamination and characterize mating patterns based on seed samples collected in three years (1999, 2000, and 2003) from one block of a non-isolated, open-pollinated seed orchard of Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco) in western Oregon. Pollen contamination was consistently high across the three years (mean = 35.3%, range = 31.0-41.3%), and appeared to result primarily from cross-pollination among the orchard blocks. Levels of pollen contamination varied substantially among clones, and were higher in clones with early female receptivity (mean = 55.5%) than in those with either mid (mean = 36.4%) or late (mean = 28.3%) female receptivity. We detected low rates of self-pollination (mean = 1.8% per clone), and over ten-fold differences in the relative paternal contributions of the clones. There was a clear pattern of positive assortative mating with respect to floral phenology. This study illustrates that SSR markers are a powerful tool for characterizing seedlots and helping improve the design and management of Douglas-fir seed orchards.

INTRODUCTION

Methods of measuring and managing pollen contamination (i.e., pollination of seed orchard parents from non-orchard sources) and within-orchard mating patterns are important to tree breeders (Squillace and Long 1981; Webber and Painter 1996). Pollen contamination is expected to reduce the genetic worth and adversely affect the adaptability of seed orchard crops (Squillace and Long 1981; Friedman and Adams 1985; Wheeler and Jech 1986; Adams and Burczyk 2000; Kang et al. 2001). Self-fertilization and unequal contributions of parents to seed crops also reduce the genetic efficiency of seed orchards (i.e., the degree to which seed crops represent the genetic superiority and diversity of orchard clones; Friedman and Adams 1982).

A simple, indirect way to estimate pollen contamination in wind-pollinated species is to use pollen traps to compare the abundance of pollen produced within the seed orchard throughout the pollination period with the background abundance of pollen produced by nearby stands of the same species using pollen traps (Greenwood and Rucker 1985; Wheeler and Jech 1986). Pollen contamination is subsequently estimated as the ratio of the average pollen catch in the background to that in the orchard. Early studies that used this method suggested that up to 88% of the seeds produced in non-isolated, open pollinated seed orchards is fertilized by non-orchard pollen (Greenwood and Rucker 1985; Wheeler and Jech 1986).

A more sophisticated approach, which is applicable to both wind- and animal-pollinated species, is to determine the proportion of seeds produced in the orchard that have been fathered by non-orchard trees. This can be accomplished by using genetic

markers to compare the multilocus genotypes of all orchard parents to the genotypes of a sample of their seeds, a method called paternity analysis (Adams 1992; Jones and Ardren 2003). Paternity analyses based on monoterpene and allozyme markers confirmed that pollen contamination in non-isolated, open-pollinated conifer seed orchards often exceeds 30-40% (Squillace and Long 1981; Smith and Adams 1983; Friedman and Adams 1985; Wheeler and Jech 1986; Adams and Burczyk 2000; Pakkanen et al. 2000, but see also El-Kassaby and Ritland 1986a). Furthermore, this approach allowed pollen contamination estimates ($\hat{m}$) to be partitioned into different sources; i.e., cross-pollination among orchard blocks within the same orchard complex versus pollination by non-orchard pollen (Smith and Adams 1983; Wheeler and Jech 1986).

Because allozymes have low allelic diversity of (i.e., effective number of alleles per locus; Hamrick and Godt 1989), it is impossible to directly detect all seeds that have been fertilized by non-orchard pollen. This is because some of these seeds have marker genotypes that are indistinguishable from the genotypes that could be produced by the males within the orchard. In addition, with allozymes it is generally infeasible to characterize within-orchard mating patterns by unambiguously assigning paternity for seeds resulting from crosses among orchard parents (Adams 1992). In lieu of complete paternity assignment, probabilistic and maximum likelihood models have been developed to estimate the proportion of undetected contaminants and the effect of factors related to male reproductive success (Smith and Adams 1983; Adams and Birkes 1991; Burczyk et al. 2002). The results of applying these models to empirical data from both seed orchards and natural stands of several conifer species indicate that pollen contamination (or gene

flow) is often substantial, and that rates of self-fertilization are typically low. Furthermore, these models revealed that distance and floral synchrony among trees are the best predictors of mating frequency, with mating success varying by orders of magnitude among males (Erickson and Adams 1989,1990; Adams and Birkes 1991; Adams 1992; Burczyk et al. 1996; Burczyk and Prat 1997). Nonetheless, the effective number of males mating with each female appears to be large (>10) in both seed orchards and naturals stands (Adams 1992; Burczyk et al. 1996; Burczyk and Prat 1997).

The standard error of pollen contamination estimates is inversely related to the detection probability (i.e., the probability that an immigrant pollen haplotype would differ from all pollen haplotypes that can be produced in the seed orchard; Smith and Adams 1983). Because detection probabilities are typically low with allozymes (i.e., 0.10-0.50), standard errors of $\hat{m}$ are often as high as 15-20% of the estimated pollen contamination levels. (Adams et al. 1997; Adams and Burczyk 2000). Thus, the precision of these pollen contamination estimates obtained may be too low to detect small but important differences in m, such as responses to pollen management techniques or differences in pollen contamination among clones with different floral phenologies (Webber 1995; El-Kassaby and Ritland 1986b).

Highly variable, PCR-based genetic markers, such as simple sequence repeats (SSRs), can substantially increase the precision of pollen contamination estimates by increasing detection probabilities (i.e., making the proportion of undetected contaminants negligibly small; Dow and Ashley 1996; 1998; Gerber et al. 2000). These markers have already been used to measure pollen contamination and selfing rates, directly quantify the

proportion of seeds fathered by each seed orchard parent, and test for deviations from random mating with respect to distance between orchard trees (Stoehr and Newton 2002; Chaix et al. 2003). SSR markers are currently available for many economically important tree species (Elsik et al. 2000; Brondani et al. 1998; Table 2.3), and their use for studying the genetic efficiency of seed orchards is likely to increase.

Our goal was to test the usefulness of SSRs for measuring pollen contamination and characterizing mating patterns in Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco) seed orchards. The objectives of this study were to (1) estimate pollen contamination in three seed crops of one seed orchard block, (2) test whether pollen contamination levels vary among clones with different floral phenologies, (3) determine the relative paternal contributions of the clones in the block, and (4) test for assortative mating with respect to floral phenology.

MATERIALS AND METHODS

Study orchard

We studied a clonal, open-pollinated Douglas-fir seed orchard complex in western Oregon, located approximately 50 miles from the Pacific Ocean in the Willamette Valley. The orchard consists of five blocks, each containing parents from a different breeding (i.e., geographical) zone in the Oregon Coast Range (Fig. 4.1). On average, the distance between the orchard complex and the breeding zones is approximately 70 miles and the climate in the orchard is considerably warmer and dryer than in these zones. The seed

orchard complex was established in 1973. All ramets had reached sexual maturity when the first seed collections for this study were made.

Within each block, ramets are arranged according to a systematic design with no ramets of the same clone adjacent to each another. All blocks are subdivided into three sections of approximately equal size (Fig. 4.1). Each year, all ramets in one of the three sections are fertilized and partially girdled to stimulate heavy flowering the following year. Thus, sections are stimulated and harvested once every three years. Because of this stimulation strategy, pollen production is approximately the same in all blocks each year. There is essentially no spatial isolation between adjacent blocks or between the orchard complex and surrounding mature natural stands of Douglas-fir (Fig. 4.1). Other than floral stimulation, no special treatments to control within-block mating or to reduce pollen contamination are applied.

We characterized pollen contamination and mating patterns for one of the five blocks in the orchard (i.e., the Test Block; Fig. 4.1). Based on the labels attached to each tree, we initially believed that the Test Block contained 342 ramets from 58 parental clones (mean number of ramets per clone = 6, range = 1-19). Our SSR analyses (discussed below) revealed the presence of one additional parental genotype, which brought the total number of clones to 59.

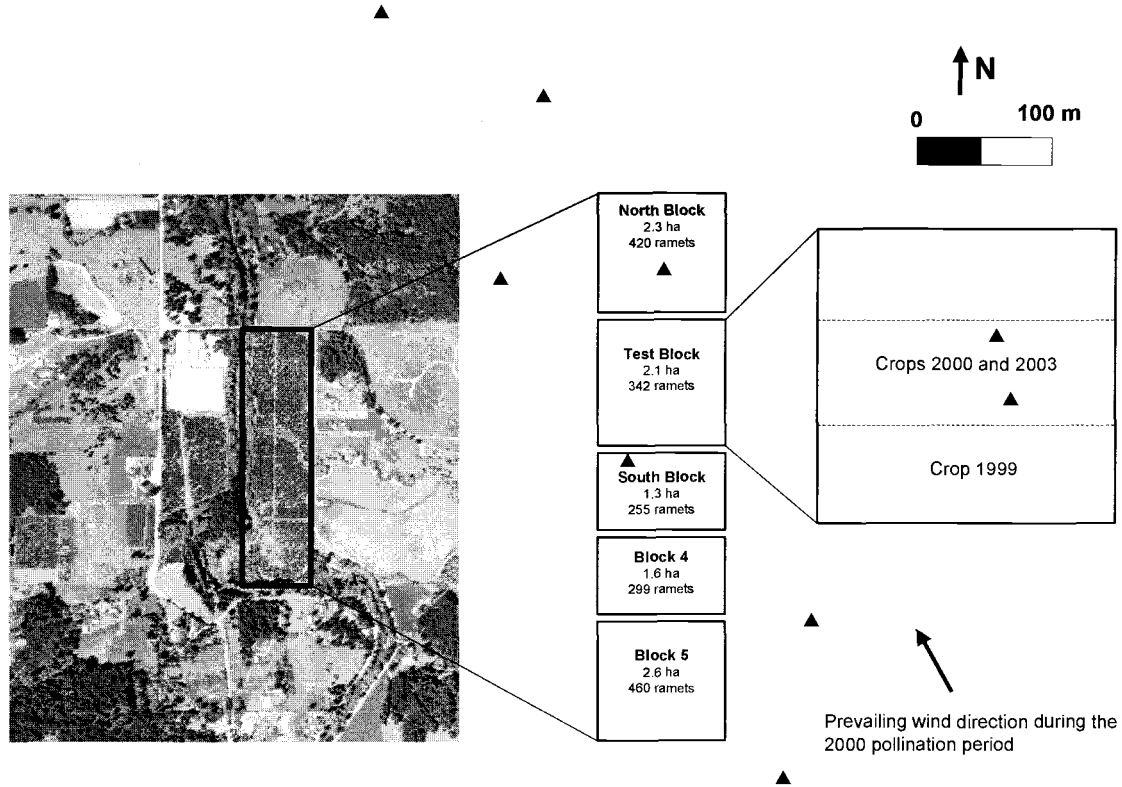


Figure 4.1. Douglas-fir seed orchard complex in western Oregon. Aerial photograph: locations of the seed orchard complex (area in rectangle) and nearby mature natural stands of Douglas-fir. Schematic diagram: each seed orchard block is subdivided into three sections as indicated by the dashed lines shown for the Test Block. Within each block, a different section is stimulated and harvested every year. Triangles indicate the approximate locations of pollen traps.

Data collection

We sampled seeds from the 1999, 2000, and 2003 crops harvested in the Test Block. We used two different sampling methods and replicated each of them in two years. Two 'bulk' samples (1999 and 2000) were constructed by mixing an approximately equal number of seeds from each ramet for which cones were operationally harvested. We used the bulk samples to estimate mean pollen contamination per ramet and evaluate the relative paternal contributions of the clones in the Test Block. Furthermore, floral phenology data were collected and two individual-ramet samples (2000 and 2003) were collected from a total of 24 ramets to compare pollen contamination levels among ramets with different timing of female cone receptivity.

To estimate pollen contamination and evaluate within-block mating patterns, we needed to know the multilocus genotypes of the seeds we sampled and those of all clones in the Test Block, as well as the allele frequencies in the background population (i.e. outside of the Test Block). Therefore, we sampled diploid vegetative tissue from all ramets within the Test Block and a number of trees outside of the Test Block. We isolated DNA from all vegetative tissue and seed samples and genotyped each sample using highly variable SSR markers. We also used pollen traps to measure the relative abundances of pollen in the Test Block, its adjacent blocks, and outside of the orchard throughout the pollination period in 2000.

Plant materials and DNA isolation

To test the accuracy of clonal identification in the Test Block (Adams 1983; Adams et al. 1988), we aimed at sampling winter buds from three ramets of each clone (i.e., a total of $58 \times 3 = 174$ samples). Because some clones had fewer than three ramets, however, the total number of ramets sampled was 152. We extracted DNA from the buds using a protocol developed at Oregon State University (http://www.fsl.orst.edu/tgerc/dnaext.htm), and genotyped them for three highly variable SSRs. With one exception, putative ramets of the same clone had identical three-locus genotypes. The exception was one ramet whose three-locus genotype was not identical to any of the 58 clones in the block (Chapter 2). This additional genotype was included in all analyses of pollen contamination and within-block mating patterns, bringing the total number of parental clones in the test Block to 59. To further investigate potential ramet mislabeling, we collected needle tissue from all ramets in the block, extracted DNA using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA) from tissue samples pooled from up to six putative ramets of the same clone, then genotyped the resulting pooled DNA sample using the same three highly variable SSR loci. In test experiments, this procedure was 100% successful in detecting heterogeneous tissue samples (G. T. Slavov, unpublished). When a heterogeneous tissue sample was detected, DNA was isolated from each individual ramet included in the pooled sample and genotyped using the same three SSR loci. The three-locus genotype of each ramet was then compared to the genotypes of the 59 clones in the Test Block that were previously identified. Using this approach, we found four additional mislabeled ramets whose genotypes matched perfectly a different

clone in the Test Block. Thus, we surveyed all ramets in the Test Block and determined their clonal identity.

To estimate the allele frequencies in the background population, we sampled winter buds and genotyped 104 reproductively mature Douglas-fir trees located outside of the Test Block. Sixty of these trees were sampled from the other four blocks of the seed orchard, whereas 44 were sampled from the native stands of Douglas-fir near the orchard complex. We analyzed 192 seeds (the capacity of two 96-well PCR plates) from each of the bulk seed samples. DNA was extracted separately from the haploid megagametophyte and the diploid embryo of each seed using the DNeasy Plant Mini Kit. This was done because the haplotype of the paternal gamete of each seed can be inferred by comparing the diploid genotype of the embryo to the haplotype of the megagametophyte (Müller 1976; Adams et al. 1988).

We also collected and analyzed seeds from eight ramets in each of three female receptivity classes (early, mid, and late; described below). The individual-ramet sample in 2000 included seeds from one ramet from each of 10 clones (two early-, four mid-, and four late-receptivity ramets). The individual-ramet sample in 2003 included 14 ramets from 13 clones (six early-, four mid-, and for late-receptivity ramets). Across the two years, 16 clones were represented (six early-, five mid-, and five late-receptivity clones). Twenty-four seeds per ramet were analyzed (i.e., 8 ramets × 3 classes × 24 seeds = 576 seeds, or 8 × 24 = 192 seeds per female receptivity class). We extracted DNA only from the embryos of these seeds using the DNeasy Plant Mini Kit.

Genetic markers

We used the SSR markers PmOSU_2C3, PmOSU_3B2, PmOSU_2G12, PmOSU_3F1, PmOSU_1F9, PmOSU_3G9, PmOSU_4A7, PmOSU_2C2, and PmOSU_4G2 and the PCR conditions described in Chapter 2. We genotyped all clones in the Test Block, the 104 trees sampled outside of the Test Block, and the embryos from the individual-ramet samples for these nine SSR loci. Megagametophytes and embryos from the bulk seed samples were genotyped using only the first seven loci from the list. This was done because the ability to infer the paternal gamete haplotype of each seed by comparing the genotype of the embryo to the haplotype of the megagametophyte makes it possible to obtain reliable estimates of pollen contamination with fewer SSR loci (Chapter 3). The mean expected heterozygosity (H_e) and the mean number of alleles per locus (A) were calculated using the CERVUS computer program (Marshall et al. 1998). The degree of allele frequency differentiation (F_{ST}) between the Test Block and the background population was estimated using the GENETIX computer program (Belkhir et al. 2004).

Floral phenology

In 2000, we made observations on the timing of pollen receptivity of female cones and pollen release by male cones for all 126 ramets of the 42 clones (1-8 ramets per clone) in the section of the Test Block that received stimulation treatments the previous year. Phenological observations were made every other day during the flowering period in the Test Block (i.e., between March 30 and May 1). A female strobilus was considered receptive when it was exposed halfway out of the bud scales (Webber and Painter 1996).

Using binoculars, the receptivity of each tree was scored numerically based on the percentage of female strobili in the upper third of the crown that were receptive: 0 = 0% receptive, 1 = 1-25% receptive, 2 = 26-50% receptive, 3 = 51-75% receptive (peak receptivity), 4 = 76-99% receptive, and 5 = 100% receptive. We divided the clones into three phenology classes based on the mean timing of peak female receptivity (PFR) of their ramets. The rationale of this classification was to include approximately 25% of the ramets in the stimulated section of the Test Block in each of the extreme phenology classes (early and late). Ramets of clones with mean PFR prior to April 15 fell into the early receptivity class (11 clones, mean PFR April 12), the late-receptivity class included ramets of clones with mean PFR after April 18 (nine clones, mean PFR April 20), and ramets of all other clones were included in the mid-receptivity class (22 clones, mean PFR April 16).

The timing of pollen release in 2000 was determined for the same 126 ramets based on observations made on a marked reachable branch on the south side of each tree. Male strobili were classified as releasing pollen when pollen sacs were beginning to split and tapping the strobili released pollen. The stage of pollen shed was scored numerically using the same scoring system as for female receptivity and clones were divided into three male floral phenology classes (early, mid, and late) as described above .

To evaluate clonal consistency in timing of female receptivity over years, we collected additional floral phenology data in 2003 when the trees for which phenology observations were made in 200 were stimulated again. Because of time and resource limitations, our phenology observations in 2003 were made on fewer trees and only on a

single day midway through the pollination period. We recorded the stage of female cone development of 51 ramets from 26 clones. Using binoculars, we obtained a score between 1 and 5 for each tree based on the predominant developmental stage (Webber and Painter 1996) of the female strobili in the upper third of the crown (i.e., 1 = most cone buds not swollen, 2 = most cone buds swollen and elongated, 3 = most cone buds beginning to burst, 4 = most cones receptive, and 5 = most cones hanging down in the post-receptive stage). As in the data set for 2000, approximately 25% of the ramets in the stimulated section of the Test Block were included in each of the extreme phenology classes (early and late). Ramets of clones with a mean score ≥ 4.0 were included in the early-receptivity class. The late-receptivity class included ramets of clones with mean score ≤ 2.5. All other ramets were included in the mid-receptivity class. The classifications of clones into classes with respect to the timing of female receptivity (based on the mean score per clone) in 2000 and 2003 were strongly correlated ($r = 0.93$, $P<0.0001$) and only two clones fell into different classes (mid versus late) in the two years.

Pollen abundance

In 2000, we measured pollen abundances in the Test Block, the two blocks immediately adjacent south and north of the Test Block (Fig. 4.1), and outside of the orchard complex using pollen traps consisting of microscope slides with double-coated Scotch tape oriented towards the wind (Greenwood and Rucker 1985). We established two pollen traps in the Test Block, one in the South Block, one in the North Block, and five in open spaces outside of the orchard (approximate locations are shown on Fig. 4.1). Microscope

slides were collected daily from each pollen trap from April 1 to April 28. Pollen abundance was quantified by counting the number of pollen grains per cm^2 of slide area using a microscope and a square grid (4 mm × 4 mm).

Data analysis

Pollen contamination

We estimated pollen contamination from the bulk and individual-ramet seed samples using the paternity exclusion approach developed by Smith and Adams (1983). Using this method, pollen immigration is estimated as:

$$\hat{m} = \frac{\hat{b}}{\hat{d}} \qquad [1]$$

where $\hat{b}$ is the proportion of seeds whose pollen gamete haplotypes differ from all pollen gamete haplotypes that can be produced by the orchard parents (i.e., observed pollen immigrants), and $\hat{d}$ is the probability that an immigrant pollen gamete haplotype will differ from all pollen gamete haplotypes that can be produced by the orchard parents (i.e., detection probability).

Mistyping is the false identification of genotypes caused by the occurrence of null alleles, mutations, and detection errors. When ignored, even relatively low rates of mistyping result in substantial upward biases in pollen contamination estimates (Chapter 3). Rates of mistyping can be considerable for highly variable markers such as SSRs (Slate et al. 2000; Ewen et al. 2000; DiFazio 2002). We developed the Pollen Flow (PFL) computer program for estimating pollen contamination (Appendix 2), which extends the

approach of Smith and Adams (1983) to cases in which mistyping cannot be ignored and to scenarios in which the complete paternal haplotype cannot be unambiguously determined (e.g., when only seed embryos are available for analysis). The program makes adjustments for mistyping by requiring mismatches at multiple loci between an observed pollen gamete haplotype and all pollen gamete haplotypes that can be produced by the trees in the orchard before classifying a seed as a contaminant (Chapter 3). We tested the performance of this analytical approach in numerous computer simulations, including scenarios that were nearly identical to this study (Chapter 3).

We estimated pollen contamination from the two bulk seed samples using the "*Haploid*" option of PFL, which applies to situations in which the complete paternal gamete haplotypes are available for analysis. The data required to estimate pollen contamination were the seven-locus genotypes of the (1) 59 clones in the block, (2) 104 trees that we sampled to estimate the allele frequencies outside of the Test Block, and (3) inferred pollen gametes (i.e., paternal gamete haplotypes) of the sampled seeds. A seed was classified as an observed pollen contaminant only if its paternal haplotype mismatched all haplotypes that could be produced by the 59 clones in the Test Block at three or more loci. Our simulations (Chapter 3) showed that using seven highly variable SSR loci and requiring at least three mismatching loci for exclusion would result in reliable estimates of pollen contamination, given the observed rate of mistyping (discussed below). The detection probability for analyses of bulk samples was 0.93. We estimated the standard error of $\hat{m}$ using the following approximation:

$$SE(\hat{m}) \approx \frac{1}{\hat{d}}\sqrt{\frac{\hat{b}(1-\hat{b})}{n}} = \hat{m}\sqrt{\frac{1-\hat{b}}{\hat{b}n}} \qquad [2]$$

where n is the number of offspring analyzed (Smith and Adams 1983; Chapter 3). To certify the origin of the bulk seed samples that we obtained and to assess the representation of the clones included in these samples, we also compared the megagametophyte (i.e., the maternally contributed) haplotype of each seed to the genotypes of the 59 clones in the Test Block. The analysis of each megagametophyte haplotype resulted in the assignment of a unique mother or failure to find a genotypically compatible mother (i.e., observed seed contamination resulting from mishandling during cone processing or seed extraction). Observed seed contaminants were not used in further analyses of pollen contamination and within-block mating patterns.

Pollen contamination from individual-ramet seed samples was estimated using the "*Diploid*" option of PFL (Chapter 3; Appendix 2). In this case, paternal gamete haplotypes are inferred by comparing the diploid genotype of each seed embryo to that of its mother. For loci at which the mother and the embryo are both heterozygous and have the same genotype, alleles contributed by the fathers cannot be unambiguously determined and are treated as missing data (Chapter 3). Because of this loss of information, 1-2 additional highly variable loci are needed to obtain reliable estimates of pollen contamination compared to the *Haploid* option (Chapter 3). Therefore, we used the nine-locus diploid genotypes of the 59 clones in the Test Block, the 104 trees that we sampled outside of the Test Block, and 24 seed embryos to calculate $\hat{b}$ and $\hat{d}$ for each of

the 24 progeny arrays comprising the individual-ramet samples. We subsequently estimated pollen contamination and its standard error for each female receptivity class using the average $\hat{b}$ and $\hat{d}$ over the progeny arrays from ramets in the same receptivity class and equations [1] and [2]. As in the *Haploid* option, seeds were classified as pollen contaminants only if their inferred paternal haplotypes mismatched all haplotypes that could be produced by the 59 clones in the Test Block at three or more loci. The detection probability for analyses of individual-ramet samples was 0.97.

To estimate mistyping, we compared the maternally contributed haplotype (i.e., the megagametophyte in bulk samples) or the diploid genotype of the embryo (in individual-ramet samples) of each seed to the diploid genotype of its known or assigned mother using the PFL program (Chapter 3). A mismatch at a given locus was scored when the offspring haplotype or genotype did not contain any of the maternal alleles. The observed frequency of mismatches was 6.2% (range across loci = 2.5-7.3%). This represents only a minimal estimate of mistyping, however, because this method does not detect all sources of mistyping (Ewen et al. 2000). In the analyses of both bulk, and individual-ramet samples, we applied a conservative adjustment for mistyping when estimating pollen contamination (i.e., we required three or more mismatching loci for exclusion). This approach assumes that the true rate of mistyping might have been up to 50% higher than 6.2% (Chapter 3). This high level of mistyping is not unusual for highly variable SSR markers (Slate et al. 2000; DiFazio 2002).

Factors affecting pollen contamination

Using a Chi-square contingency test (Wackerly et al. 2002), we determined that differences in pollen contamination among female receptivity classes were independent of the year in which individual-ramet samples were collected ($\chi^2 = 3.40$, df $= 2$, $P = 0.183$). Therefore, we used analyses of variance to detect differences in mean pollen contamination levels among the three female receptivity classes after pooling data across 2000 and 2003. We also performed specific comparisons (i.e., early versus mid, early versus late, and mid versus late) and controlled for the experimentwise Type I statistical error by using critical values based on the Bonferroni inequality (Wackerly et al. 2002).

In addition to testing for differences in pollen contamination among female receptivity classes, we also wanted to study the relationship between relative timing of female receptivity and pollen contamination. To equalize the scales of the variables used to measure the timing of female receptivity across the two years (day of peak female receptivity in 2000 versus stage of female cone development in 2003), we expressed data from both years as standard deviates. Mean pollen contamination levels per clone (data pooled across years) were then correlated to mean receptivity standard deviates.

Within-block mating patterns

Paternity exclusion was used to assign paternity to each seed assumed to result from within-block mating. This was done by excluding all but one male from paternity and assuming that this was the true father. Paternity assignment was performed using the PFL computer program that recorded the genotypically compatible father(s) for each seed analyzed, given that three mismatching loci were required for exclusion. The main limitations of this approach are that (1) only a small proportion of the potential fathers (i.e., including those in the background population) is typically sampled, and (2) paternity cannot be assigned unambiguously for some seeds because their pollen gamete haplotypes are compatible with the genotypes of two or more parents. These limitations can lead to biases in the mating parameters estimated (Adams 1992; DiFazio et al. 2004). In our analyses of within-block mating patterns, however, the probability that paternity is falsely assigned to a parent in the Test Block when the actual father is in the background population is low. Because this probability equals 1 - $\hat{d}$, its value in our analyses was 0.07 in bulk seed samples, and 0.03 in individual-ramet seed samples. Furthermore, we identified more than one possible father for only 22 out of 557 (3.9%) seeds that were assumed to result from within-block mating. These 22 seeds were not considered in further analyses of relative paternal contributions. Thus, it is unlikely that paternity exclusion introduced more than minimal biases in our estimates of relative paternal contributions to seed crops in the Test Block. The effective number of male parents in the Test Block (N_{ep}) was calculated based on the relative paternal contributions of the 59 clones, as proposed by Burczyk et al (1996).

To test whether floral synchrony influences mating patterns within the Test Block, we compared the observed counts of crosses among trees within the same floral phenology class and among trees from different phenology classes to the counts expected under random mating. The statistical significance of the deviations was evaluated using a goodness-of-fit Chi-square test (Wackerly et al. 2002).

RESULTS

SSR markers

We used seven highly variable SSR loci to analyze bulk seed samples and additional two loci to analyze individual-ramet seed samples. The mean expected heterozygosities for both sets of loci exceeded 90%, and the mean number of alleles per locus was over 30. The allele frequency differentiation between the Test Block and the background population was low in both cases (Table 4.1). The detection probabilities for both types of analysis exceeded 0.90, even when three or more mismatching loci were required for paternity exclusion (Table 4.1).

Table 4.1. Summary statistics for the SSR markers used for analyses of pollen contamination and within-block mating patterns (estimates are based on the genotypes of the 59 parents in the Test Block and the 104 parents sampled outside of the Test Block).

Type of analysis	No. of SSR loci used	H_e [a]	A [b]	d [c]	F_{ST} [d]
Haploid (bulk seed samples)	7	0.95	39.6	0.93	0.0018
Diploid (individual-ramet seed samples)	9	0.92	34.4	0.97	0.0022

[a] H_e is the mean expected heterozygosity.
[b] A is the mean number of alleles per locus.
[c] d is the detection probability when three or more mismatching loci are required for exclusion.
[d] F_{ST} is a measure of the allele frequency differentiation between the Test Block and the background population.

Seed contamination

We detected seed contamination in both bulk samples. Two seeds (1.0%) from the 1999 sample and 90 seeds (46.9%) from the 2000 sample did not match any of the parents in the Test Block (Table 4.2). Observed seed contaminants were not used in later analyses of pollen contamination from bulk samples. No seed contamination was detected in the individual-ramet samples.

Table 4.2. Seed and pollen contamination in one block of a Douglas-fir seed orchard.

Year	Type of seed collection	No. of seeds analyzed	Observed seed contamination (%)	Pollen contamination[a] (%) ± SE
1999	Bulk	192 (190[b])	1.0	31.0±3.5
2000	Bulk	192 (102[b])	46.9	36.8±5.2
2000	Individual-ramet	240	0	32.0±3.2
2003	Individual-ramet	336	0	41.3±2.8
Mean				**35.3±2.4**

[a] The expected pollen contamination for a randomly selected ramet in the Test Block. For bulk seed samples, pollen contamination was estimated using equation [1]. Individual-ramet seed samples were not representative of the population of ramets harvested in a given year. Therefore, for individual-ramet seed samples pollen contamination was estimated as the average pollen contamination of three female receptivity classes weighted by the relative number of ramets that were harvested in the respective year in each receptivity class.
[b] Number of seeds used to estimate pollen contamination after accounting for seed contamination.

Pollen contamination

Mean pollen contamination per ramet was high in all four seed samples and showed little variation among the three years in which seed samples were collected (31.0-41.3%; Table 4.2). We obtained similar results from the bulk and the individual-ramet samples collected in 2000.

Synchrony of pollen shed and female cone receptivity in 2000

The period of maximum pollen abundance in the Test Block coincided with the time when most of the clones in the flower-stimulated section of the Test Block reached peak female receptivity (Fig. 4.2 A, B). There was a moderate correlation between pollen abundance in the Test Block on a given day and the number of clones that reached peak pollen shed on that day ($r = 0.74$, $P<0.001$). The early-flowering clones in the Test Block (i.e., those that reached peak female receptivity prior to April 15; Fig. 4.2A) were receptive during maximum pollen abundance in the South Block (Fig. 4.2C) and before pollen abundance in the Test Block had reached a stable maximum (Fig. 4.2B). On average, pollen abundance prior to April 15 was eight times higher in the South Block than in the Test Block. The pattern of pollen abundance in the North Block was similar to that of the Test Block (Fig. 4.2D). Throughout the flowering period, the pollen abundance measured outside of the orchard was substantially lower than in any of the seed orchard blocks in which pollen traps were established (Fig. 4.2E). Based on the relative pollen abundances measured in 2000, pollen contamination from the natural stands of Douglas-fir surrounding the orchard was only 6.4% (i.e., almost six times lower than the total rate of pollen contamination estimated using SSR markers).

Influence of floral phenology on pollen contamination

Pollen contamination estimates averaged over 2000 and 2003 for the early-, mid-, and late-receptivity classes were 55.5, 36.4, and 28.3%, respectively (Fig. 4.3). Both comparisons involving the early-receptivity class (early-mid, early-late) were statistically

significant ($P = 0.014$ and 0.002, respectively), but the difference between the mid- and the late classes was not ($P = 0.179$). There was a moderate, negative correlation between pollen contamination per clone and the time of peak female receptivity ($r = -0.63$, $P = 0.008$; Fig. 4.4).

Within-block mating patterns

Selfing

After assigning paternity for 96% of the non-immigrant seeds, we detected few seeds resulting from self-fertilization. The selfing rate was 2.1% in the 1999 bulk sample and 1.0% in the 2000 bulk sample. We detected seeds resulting from selfing in four of the 16 clones included in the individual-ramet samples. Selfing rates per clone, averaged over 2000 and 2003, varied between 0 and 8% (overall mean = 1.8%).

Relative parental contributions

Thirty of the 39 clones from which seeds were harvested in 1999 were identified as mothers in our analyses of the 1999 bulk seed sample. Twenty-four of the 31 clones harvested in 2000 were represented as mothers in the analyses of the 2000 bulk seed sample.

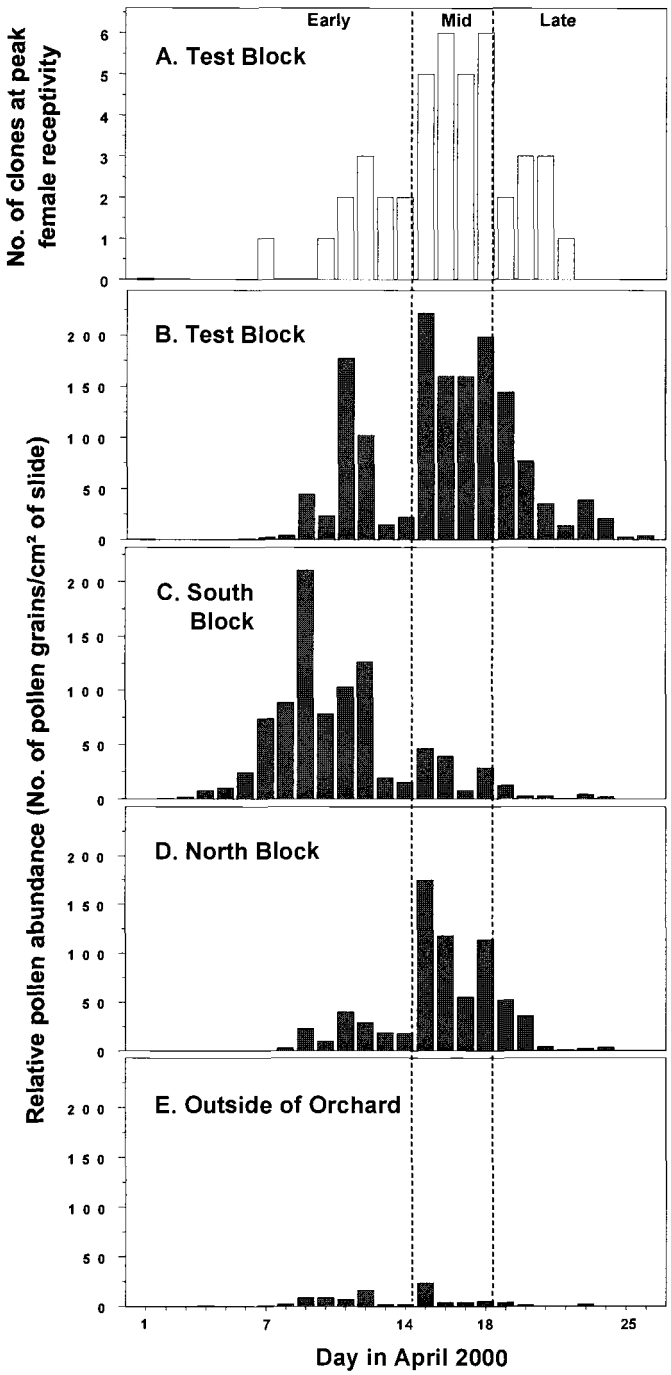


Figure 4.2. Synchrony between female cone receptivity and pollen shed in 2000: Timing of peak female receptivity of 42 clones from the Test Block (A) and relative abundance of pollen measured in the Test Block (B), South Block (C), North Block (D), and outside of the seed orchard (E) (See Fig. 4.1 for relative locations of the seed orchard blocks). The dashed lines indicate the borders between the early-, mid-, and late-receptivity classes (see text). The sudden drop in pollen abundance on April 13 and 14 was probably caused by 43 mm of precipitation recorded for these two days.

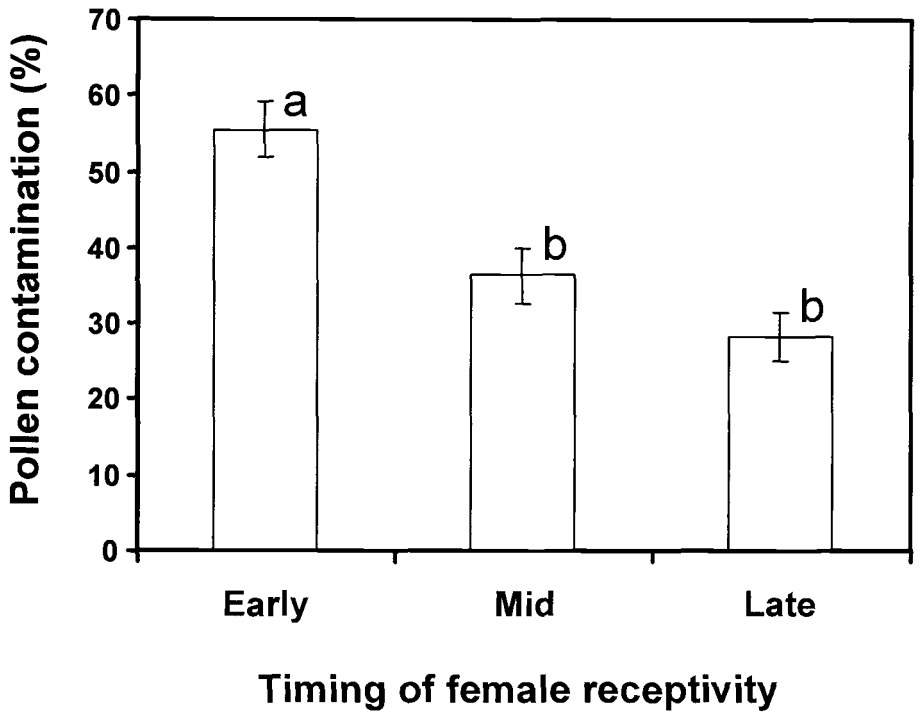


Figure 4.3. Mean pollen contamination for parents with early, mid and late female receptivity in one block of a Douglas-fir seed orchard. Estimates are based on data combined from the 2000 and 2003 individual-ramet seed samples. Error bars indicate standard errors calculated using equation [2]. Bars marked with the same letter correspond to means that are not significantly different at the 0.05 experimentwise significance level.

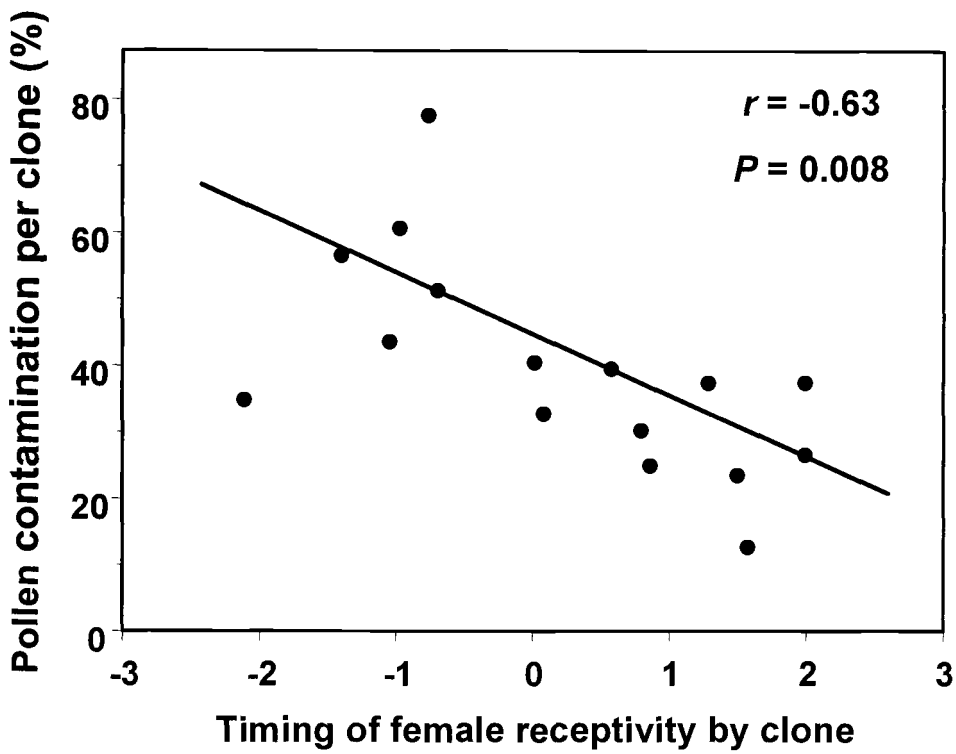


Figure 4.4. Relationship between pollen contamination per clone and the timing of peak female receptivity in standard deviations from the mean receptivity within each year. Results are combined from the 2000 and 2003 individual-ramet seed samples.

Based on the bulk-sample seeds assumed to result exclusively from mating among parents in the Test Block (n = 198 pooled over 2000 and 2003), paternity was assigned to a total of 40 clones. The relative paternal contributions of these clones, however, were uneven, with the number of seeds fathered in the samples ranging from one to 17 (Fig. 4.5A). Despite this wide range in paternal success, the effective number of male parents within the Test Block (N_{ep}) contributing to bulked seed samples was relatively high (mean N_{ep} = 20.0, 17.8 in 1999, and 22.2 in 2000). Based on pooled data from the 2000 and 2003 individual-ramet seed samples, N_{ep} was substantially lower in clones with early (mean = 6.2) than in clones with mid (mean = 12.4) and late female receptivity (mean = 15.3). There was a moderate, positive correlation between the paternal contribution per clone and the number of ramets per clone in the Test Block (r = 0.63, $P < 0.0001$; Fig. 4.5B).

Assortative mating with respect to the timing of female receptivity

Crosses among clones from the same phenology class (e.g., early female receptivity parent × early pollen shed parent) were more frequent than expected under the assumption of random mating (Fig. 4.6). This deviation was statistically significant for all phenology classes, and appeared to be stronger for Early × Early (Fig. 4.6A) and Late × Late (Fig. 4.6C) than for Mid × Mid crosses (Fig. 4.6B). We did not observe crosses among parents from the two extreme phenology classes (i.e., Early × Late or Late × Early; Fig. 4.6A, C).

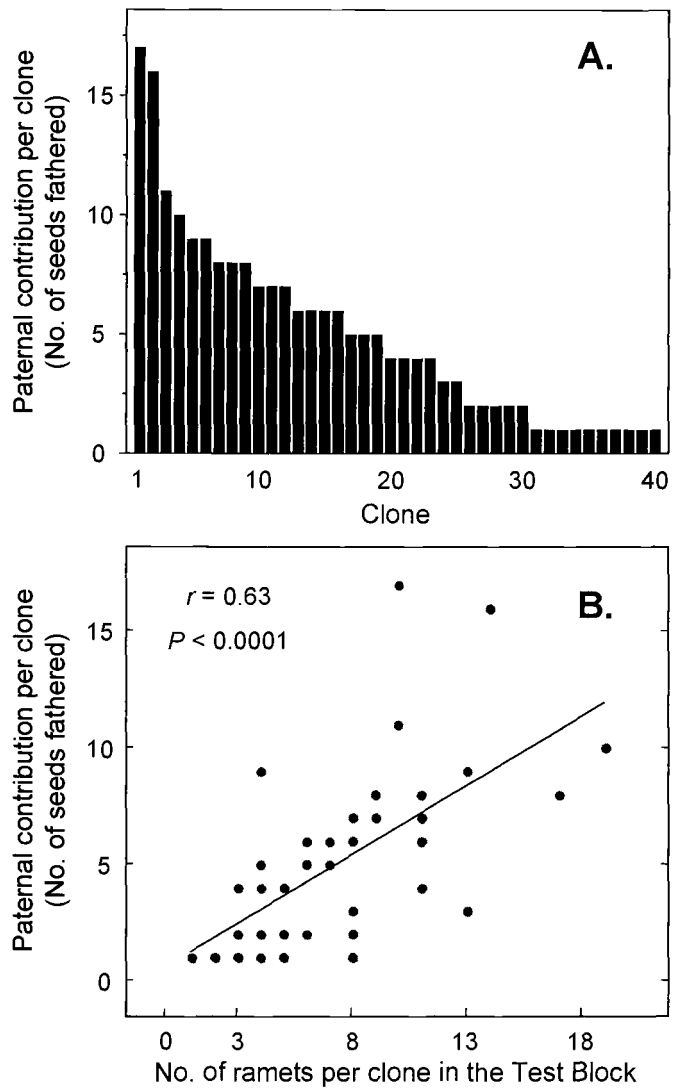


Figure 4.5. Within-block mating patterns. (A) Paternal contributions of the 40 clones in the Test Block that fathered one or more seeds in the 1999 and 2000 bulk samples. (B) Relationship between the paternal contribution per clone and the number of ramets per clone in the Test Block. Results are combined from the 1999 and 2000 bulk seed samples (total number of seeds across the two years = 198).

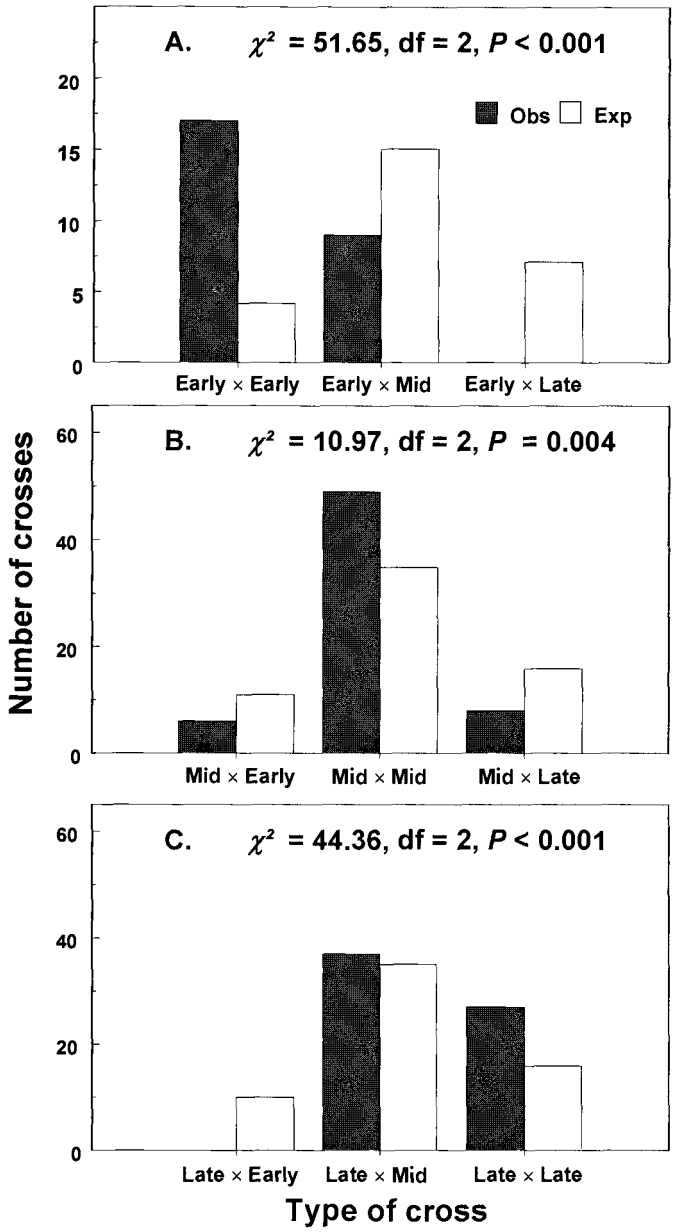


Figure 4.6. Goodness-of-fit-tests for observed and expected number of crosses within and among three floral phenology classes. (A) Female parent has early female receptivity. The statistically significant deviation from random-mating indicates that clones with early female receptivity mate preferentially with clones with early pollen shed. (B) Female parent is receptive during the time when most clones reach peak receptivity (mid receptivity class). There is statistically significant excessive mating among clones with mid female receptivity and clones with mid pollen shed, but the resulting deviation is not as strong as in A and C. (C) Female parent has late female receptivity. The statistically significant deviation from random-mating indicates that clones with late female cone receptivity mate preferentially with clones with late pollen shed.

DISCUSSION

SSR markers

The high variability of the SSR markers that we used allowed us to precisely estimate pollen contamination and unambiguously assign paternity for the vast majority of the seeds assumed to result from mating among the parents in the Test Block. The detection probabilities for the analyses of both bulk and individual-ramet seed samples (Table 4.1) were substantially higher than in any published study based on allozymes (e.g., Smith and Adams 1983; Adams et al. 1997; Pakkanen et al. 2000). Consequently, the standard errors of our estimates of pollen contamination (Table 4.2) were approximately three times lower than those reported in a study in which similar numbers of seeds were analyzed using 11 allozyme loci (Adams et al. 1997).

In our analyses of the two individual-ramet samples, pollen contamination was estimated precisely and within-block mating patterns were characterized without genotyping seed megagametophytes to infer the paternal haplotype of each seed. Two additional loci were needed to compensate for the fact that paternal alleles could not be inferred for loci at which the mother and the seed were heterozygous and had the same alleles. In our experience, however, reducing the number of genotyping assays in half by not genotyping seed megagametophytes far outweighs the extra cost incurred by the need to use two additional SSR loci. Thus, individual-ramet seed samples appear to be more cost-efficient than bulk seed samples for measuring pollen contamination and characterizing within-orchard mating patterns when highly variable SSR loci are used.

Seed and pollen contamination

Seed contamination

Low levels of seed contamination have been reported previously in Douglas-fir seed orchards (e.g., Adams et al. 1997). The high seed contamination in the bulk sample from 2000 probably resulted from mixing the seedlot produced in the Test Block with a seedlot from a different orchard block during cone processing or seed extraction. The two seeds in the 1999 bulk sample for which no genotypically compatible mothers could be found, probably also resulted from seed contamination, but it is possible that mistyping prevented us from identifying their mothers among the 59 clones in the Test Block. Using highly variable SSR markers, seed contamination can be easily detected and the subsequent deployment of maladapted seeds avoided.

Pollen contamination

Pollen contamination was high in all four seed samples that we analyzed. This result is consistent with pollen contamination levels reported earlier for seed orchards of Douglas-fir and other conifers (Adams and Burczyk 2000; Pakkanen et al. 2000). In the absence of substantial spatial isolation (>1-2 km) from other orchard blocks or stands of the same species, pollen contamination in open-pollinated conifer seed orchards can be minimized only by the effective implementation of pollen management techniques (e.g., supplemental mass pollination and bloom delay; Wheeler and Jech 1986; El-Kassaby and Ritland 1986b; Adams and Burczyk 2000).

Synchrony of pollen shed and female cone receptivity

Pollen abundance outside of the orchard was low relative to that in the Test Block throughout the pollination period. This suggests that most of the pollen contamination that we detected in the Test Block resulted from fertilization by clones in the other four orchard blocks. Similar conclusions were reached in a study of relative pollen abundance and pollen contamination in four blocks of a Douglas-fir seed orchard located in western Washington (Wheeler and Jech 1986). Conversely, two studies in a Douglas-fir seed orchard in western Oregon suggested that most of the pollen contamination resulted from fertilizations by pollen coming from outside of the orchard (Smith and Adams 1983; Adams et al. 1997). In four loblolly pine seed orchards in North Carolina, Georgia, and Arkansas, pollen abundance outside of the orchard was 31-88% of that in the orchard (Greenwood and Rucker 1985). Based on the limited empirical data, no general pattern is apparent in the relative amounts of pollen contamination resulting from cross-pollination among orchard blocks compared to that caused by pollen coming from outside of the orchard. This not surprising because this pattern is influenced by the relative abundances of pollen within and outside of the orchard, and the floral synchrony among blocks and surrounding stands. Both of these factors are bound to vary tremendously in different situations. Although measurements of pollen abundance within and outside of the seed orchard are relatively simple to make and could be used routinely to obtain crude measures of pollen contamination, these data are of limited use for identifying sources of pollen contamination.

In 2000, the South Block reached maximum pollen production six days before the Test Block (Fig. 4.2B, C), a pattern that is consistent with the long-term phenology observations of the seed orchard staff. The period of elevated pollen abundance in the South Block (April 6-12) coincided with the time in which seven clones in the Test Block (approximately 17% of the clones for which phenology observations were made) reached peak female receptivity (Fig. 4.2A, C). During that period, pollen abundance in the North Block was an order of magnitude lower than in the Test Block. Thus, it is more likely that the excessive pollen contamination in the early receptivity class was caused by pollen coming from the South Block than from the North Block or outside of the orchard.

Influence of floral phenology on pollen contamination

It is expected that pollen contamination will be higher in seedlots from clones whose flowering is out of synchrony with the majority of clones from the same block or orchard. In both 2000 and 2003, pollen contamination was higher in clones that became receptive early than in clones that became receptive around and past the mean time for the Test Block (Figs. 4.3, 4.4). This pattern is partly consistent with the results from a similar study of pollen contamination in a Douglas-fir seed orchard in British Columbia in which El-Kassaby and Ritland (1986b) detected the highest levels of pollen contamination in the clones that became receptive early. In the same study, however, supplemental mass pollination was applied to all cone-bearing trees from the mid-receptivity class and this class had the lowest mean level of pollen contamination. When seedlots from individual clones of the Test Block are deployed, it should be considered that seeds from clones that

become receptive early have 30-100% higher pollen contamination (absolute difference = 9.7-34.8%) than seeds from clones with mid or late receptivity.

Within-block mating patterns

In addition to pollen contamination, the genetic efficiency of seed orchards can be affected negatively by (1) high occurrence of seeds resulting from self-pollination, (2) unequal contributions of the orchard clones to seed crops, and (3) substantial departures from random mating (Friedman and Adams 1982).

Low selfing rates are typically detected in seed orchards of Douglas-fir and other conifers (Adams and Birkes 1991; Stoehr et al. 1998; Stoehr and Newton 2002). Our results agree with these findings. Although self-pollination can be as high as 50% in Douglas-fir, severe inbreeding depression in seed development reduces selfing at the developed seed stage to levels that do not appear to be a serious problem for the production of seedlots with high genetic quality (Sorensen 1999).

The way in which bulk seed samples were formed in this study (i.e., by taking an approximately equal number of seeds from each ramet harvested in a given year) made it trivial to analyze the relative maternal contributions of parents in the Test Block to seed crops. The fact that each seed megagametophyte was successfully assigned a unique mother, however, demonstrates that SSRs allow these analyses to be performed in other studies in which all harvested seeds are bulked and the relative maternal contributions of orchard parents are of practical interest.

Clones with a higher number of ramets tended to have higher relative paternal contributions to seed crops (Fig. 4.5B). Nine clones (15% of the total number of clones in the Test Block) fathered 96 (48%) of the bulk-sample seeds assumed to result from within-block mating, whereas 19 clones (32% of the total number of clones in the Test Block) did not father any of the bulk-sample seeds that we analyzed (Fig. 4.5A). Thus, the relatively high N_{ep} for the bulk samples can be explained by the high number of clones in the Test Block ($N = 59$), rather than by their even paternal contributions. Differential paternal success has been reported in other seed orchards of Douglas-fir and other conifers (Stoehr et al. 1998; Stoehr and Newton 2002; Goto et al. 2002). Thus, both pollen contamination and the relative gametic contributions of seed orchard parents should be taken into account when calculating the genetic worth of seedlots (Woods et al. 1996; Xie and Yanchuk 2003).

Mating among clones from the same floral phenology class tended to occur more often than expected under the assumption of random mating. This pattern is of particular concern for clones with extremely early or late floral phenologies because these clones have little chance to mate with the majority of the other clones within the same orchard block. In the Test Block, for example, clones from the early-receptivity class were characterized by 52% higher pollen contamination and 50% lower N_{ep} than clones from the mid-receptivity class.

Implications for seed orchard management

Our results have several practical implications. First, we confirmed that pollen contamination in non-isolated, open-pollinated conifer seed orchards can be high. Pollen contamination in seed orchards of Douglas-fir can be reduced using pollen management techniques, such as bloom delay and supplemental mass pollination (Wheeler and Jech 1986; El-Kassaby and Ritland 1986b). The ultimate solution to this problem, however, may be the transition to (1) establishing seed orchards in areas isolated by at least a few kilometers from non-orchard sources of contaminant pollen and choosing appropriate regimes of flower stimulation, or (2) alternative seed orchard designs allowing the effective application of pollen management techniques, including controlled pollination (Webber and Painter 1996).

Second, only a slight reduction in pollen contamination can be expected if seedlots from clones with extreme floral phenology are not included in bulk seed crops. In the Test Block, for example, we estimated that if seeds from clones with early female receptivity had been excluded from bulk crops, the overall pollen contamination would have been reduced from 32.0 to 30.2% in 2000 and from 41.3 to 35.9% in 2003. Thus, variation in pollen contamination among clones with different phenologies is only practically important if individual-clone seedlots are to be deployed and if some of the clones are receptive at times when little pollen is produced in the orchard block.

Third, our results suggest that the higher pollen contamination detected in early-receptive clones probably resulted from pollen produced in the South Block (Fig. 4.2). Furthermore, it appeared that most of the pollen contamination in all receptivity classes

resulted from immigrant pollen produced in the other four blocks of the seed orchard, rather than outside of the orchard. If these hypotheses are correct, pollen contamination in the Test Block could be reduced by changing the stimulation regime in the orchard. For example, by stimulating all trees in only one block, or two widely-separated blocks each year, pollen contamination would be decreased in two ways. First, within-block pollen abundance would be maximized, and second, the main sources of pollen contamination (i.e., other orchard blocks) would produce less pollen.

Cross-pollination between seed orchard blocks serving different breeding zones may adversely affect the adaptability of the resulting seedlots (Kylmänen 1980; Nikkanen 1982; Stoehr et al. 1994). For example, extensive fertilization of clones in the Test Block by pollen from the South Block could have a negative impact on the adaptability of seedlots produced in the Test Block. The South Block includes parents native to stands located more than 1° of latitude south of the native stands of the parents included in the Test Block. Compared to the provenance represented in the Test Block, parents from the provenance represented in the South Block are characterized by earlier vegetative bud flush and bud set, and seedlings are substantially more susceptible to fall frost damage (Campbell and Sorensen 1973; B. St. Clair, unpublished). If practical reasons mandate that seed crops are harvested yearly from multiple blocks, the risk of compromising the adaptability of seed crops can be minimized by simultaneously stimulating blocks that serve ecologically similar breeding zones.

Finally, this study illustrates that SSR markers are useful for directly measuring the factors affecting the efficiency of open-pollinated seed orchards. Fewer than ten SSRs

were needed to (1) measure pollen contamination and selfing rates, (2) measure the relative paternal success of the clones in the Test Block, and (3) detect deviations from random mating with respect to floral phenology. Because SSRs provide a way to measure genetic efficiency parameters with high accuracy, they will be a useful tool for the future improvement of seed orchard design and management.

ACKNOWLEDGEMENTS

This study was funded by the Pacific Northwest Tree Improvement Research Cooperative and the Department of Forest Science at Oregon State University (OSU). The Tree Genetic Engineering Research Cooperative at OSU provided laboratory equipment and support. Christine Lomas collected the phenology data in 2000. Jim Smith provided logistical help throughout the study. Konstantin Krutovskii, Nuray Kaya, Santiago González-Martínez, and Gokcin Temel helped with various aspects of the study in its early stages.

Chapter 5. Conclusions

In this thesis, I studied the applicability of SSR markers for measuring contemporary pollen flow in Douglas-fir. I developed SSR markers, tested alternative analytical approaches to measuring pollen flow using SSR markers, then used the best of these approaches to measure pollen contamination and characterize within-block mating patterns for three seed crops from one block of an operational, open-pollinated seed orchard complex. Furthermore, I demonstrated that SSR markers are an effective genetic fingerprinting and parentage analysis tool that can be used to enhance tree improvement and gene conservation efforts, and improve our knowledge of the evolutionary biology of Douglas-fir.

The development of SSR markers in species with large and repetitive genomes (such as Douglas-fir) is a costly and inefficient process. Even though I used genomic libraries that were highly enriched for SSRs, and performed several screening steps to increase the efficiency of SSR development, only 4.1% of the sequences that I obtained resulted in a useful SSR marker. Despite the low efficiency of SSR development, the 22 markers that I obtained are among the most informative genetic markers available in Douglas-fir. The mean observed heterozygosity and the mean number of alleles per marker were 0.855 and 23, respectively. I verified the Mendelian inheritance of all 22 markers and determined the genetic map locations for 20 of them. The polymorphism of these SSR markers is among the highest reported in conifers. Fifteen markers have a robust single-locus pattern and are suitable for parentage analysis, whereas the remaining

seven need further optimization, but can be used for genetic fingerprinting and genome mapping.

Although SSR markers are considered the most appropriate markers for parentage analysis, they have substantially higher rates of mistyping than allozyme markers, for which most analytical methods have been developed. Mistyping is the false identification of genotypes caused by the occurrence of null alleles, mutations, and detection errors. Using computer simulations, I studied the effects of mistyping on estimates of pollen flow obtained via paternity exclusion based on SSR markers. If not accounted for, mistyping can result in substantial upward biases in pollen flow estimates. I also evaluated different ways of accounting for mistyping when estimating pollen flow. Requiring multiple mismatches for exclusion, while assuring that detection probability is high, results in pollen flow estimates with low biases and predictable variances.

I developed the PFL computer program, which performs paternity exclusion based on multiple father-offspring mismatches and has a user-friendly interface. PFL can be used to obtain unbiased and precise estimates of pollen flow for any diploid seed plant species and under a wide range of conditions. This program is potentially useful for measuring pollen flow at large spatial scales.

To demonstrate the usefulness of SSR markers, I employed nine of the best-performing markers to measure pollen contamination and characterize mating patterns in one block of a non-isolated, open-pollinated seed orchard complex of Douglas-fir. I analyzed seed samples collected in three years (1999, 2000, and 2003). Pollen contamination was consistently high in all three years (mean = 35.3%, range = 31.0-

41.3%) and appeared to result primarily from pollen flow from other orchard blocks, rather than from the surrounding natural stands. The standard errors of these estimates were approximately 2-3 times smaller than those in a study that used allozyme markers and similar sample sizes (Adams et al. 1997). Levels of pollen contamination varied substantially among clones, and were significantly higher in clones with early female receptivity (mean = 55.5%) than in those with intermediate (mean = 36.4%) or late (mean = 28.3%) receptivity. This result is consistent with the expectation that pollen contamination will be higher in clones that become receptive when few clones in the block or orchard shed pollen.

I was able to assign paternity to 96% of the seeds pollinated by the parents in the block that I studied. Seeds resulting from self-fertilization were rare (mean = 1.8%). There were over ten-fold differences in the relative paternal contributions of the clones in the block. I also detected strong assortative mating with respect to floral phenology, which was explained by excessive mating among phenologically similar parents. Thus, I demonstrated that SSRs can provide specific information that can be used to evaluate the need for applying pollen management techniques to minimize pollen contamination and optimize mating patterns among orchard parents.

In summary, this study illustrates that SSR markers are a powerful tool for characterizing seed crops and helping improve the design and management of seed orchards. I have shown that SSRs can be used to (1) identify genotypes and test the accuracy of ramet labeling in seed orchards, (2) measure seed and pollen contamination with high precision, and (3) characterize within-orchard mating patterns. SSRs can also

be used to reduce the costs and increase the benefits of tree improvement in Douglas-fir by providing an effective means of minimizing testing costs, and guiding the progress of pollen management techniques.

Finally, by developing SSRs and appropriate methods of their application, we can advance our knowledge about gene flow in natural populations of Douglas-fir and other forest trees. The availability of large sets of highly variable SSRs will allow us to perform large-scale studies of gene flow that will help us better understand the interactions between gene flow and adaptation, a prerequisite to our ability to perform environmental risk assessment and predict the implications of global environmental changes.

Bibliography

Adams, W.T. 1983. Application of isozymes in tree breeding. *In*: Tanksley, S.D., and Orton, T.J. (eds.) Isozymes in plant genetics and breeding, Part A. Elsevier Science Publishers BV, Amsterdam, pp. 381-400.

Adams, W.T. 1992. Gene dispersal within forest tree populations. New Forests. **6**: 217-240.

Adams, W.T., and Joly, R.J. 1980. Genetics of allozyme variants in loblolly pine. J. Hered. **71**: 33-40.

Adams, W.T., and Birkes D.S. 1991. Estimating mating patterns in forest tree populations. *In* Biochemical markers in the population genetics of forest trees (eds. Fineschi S, Malvolti ME, Cannata F, Hattemer HH), pp. 157-172. S.P.B. Academic Publishing bv, The Hague, The Netherlands.

Adams, W.T., and Burczyk, J. 1993. GENFLOW: a computer program for estimating levels of pollen contamination in clonal seed orchards. Release 1. Department of Forest Science, Oregon State University, Corvallis, USA.

Adams, W.T., and Burczyk, J. 2000. Magnitude and implications of gene flow in gene conservation reserves. *In*: Forest conservation genetics: principles and practice. Young, A., Boshier, D. and Boyle, T. (eds.). CSIRO Publishing, Collingwood, Victoria, Australia, pp. 215-224.

Adams, W.T., Neale D.B., and Loopstra, C.A. 1988. Verifying controlled crosses in conifer tree-improvement programs. Silvae Genet. **37**: 147-152.

Adams, W.T., Griffin, A.R., and Moran, G.F. 1992. Using paternity analysis to measure effective pollen dispersal in plant populations. Am. Nat. **140**: 762-780.

Adams, W.T., Hipkins, V.D., Burczyk, J., and Randall, W.K. 1997. Pollen contamination trends in a maturing Douglas-fir seed orchard. Can. J. of For. Res. **27:** 131-134.

Adams, W.T. Johnson, G., Copes, D.L., Daniels, J., Quam, R.G., Heaman, J.C., and Weber, J. 1990. Is research keeping up with the needs of Douglas-fir tree improvement programs. West. J. Appl. For. **5**: 135-137.

Amarasinghe, V., and Carlson, J.E. 2002. The development of microsatellite markers for genetic analysis in Douglas-fir. Can. J. For. Res. **32**: 1904-1915.

Anonymous. 2002. Annual report summary for testing in 2001. American Association of Blood Banks, Parentage Testing Program Unit. http://www.aabb.org/About_the_AABB/Stds_and_Accred/ptannrpt01.pdf

Austerlitz, F., and Smouse, P.E. 2001a. Two-generation analysis of pollen flow across a landscape. II. Relation between Φ_{FT}, pollen dispersal, and interfemale distance. Genetics. **157**: 851-857.

Austerlitz, F., and Smouse, P. 2001b. Two-generation analysis of pollen flow across a landscape. III. Impact of adult population structure. Genet. Res. Cambridge. **78**: 271-280.

Beerli, P., and Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. **152**: 763-773.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., and Bonhomme, F. 2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France). http://www.univ-montp2.fr/%7Egenetix/genetix/genetix.htm

Bossart, J.L., and Prowell, D.P. 1998. Genetic estimates of population structure and gene flow: limitations, lessons and new directions. Trends Ecol. Evol. **13**: 202-206.

Bradshaw, H.D.,Jr., and Strauss, S.H. 2001. Breeding strategies for the 21st century: domestication of poplar. *In*: Dickmann, D.I., Isebrands, J.G.T., Eckenwalder, J.E., and Richardson, J. (eds.). Poplar Culture in North America, Part B. NRC Research Press, National Research Council of Canada, Ottawa, Canada. pp. 383-394.

Brenner, C.H. 2004. DNA VIEW. User's Manual. http://dna-view.com/papers.htm

Brinkmann, B., Klintschar, M., Neuhuber, F., Huehne, J., and Rolf, B. 1998. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62**: 1408-1415.

Brohede, J., and Ellegren, H. 1999. Microsatellite evolution: Polarity of substitutions within repeats and neutrality of flanking sequences. Proceedings of the Royal Society of London - Series B: Biological Sciences. **266**: 825-833.

Brondani, R.P.V., Brondani, C., Tarchini, R., and Grattapaglia, D. 1998. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla.* Theor. Appl. Genet. **97**: 816-827.

Burczyk, J. and Prat, D. 1997. Male reproductive success in *Pseudotsuga menziesii* (Mirb.) Franco: the effects of spatial structure and flowering characteristics. Heredity **79**: 638-647.

Burczyk, J. and Chybicki, I.J. 2004. Cautions on direct gene flow estimation in plant populations. Evolution **58**: 956-963.

Burczyk, J., Adams, W.T., and Shimizu, J.Y. 1996. Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuata* Lemmon.) stand. Heredity **77**: 251-260.

Burczyk, J., Adams, W.T., Moran, G.F. and Griffin, A.R. 2002. Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. Mol. Ecol. **11**: 2379-2391.

Campbell, R.K., and Sorensen, F.C. 1973. Cold-acclimation in seedling Douglas-fir related to phenology and provenance. Ecology. **54**: 1148-1151.

Chaix, G., Gerber, S., Razafimaharo, V., Vigneron, P., Verhaegen, D., and Hamon, S. 2003. Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. Theor. Appl. Genet. **107**: 705-712.

Chakraborty, R., Meagher, T., and Smouse, P. 1988. Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. Genetics. **118**: 527-536.

Devlin, B., and Ellstrand, N.C. 1990. The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. Evolution. **44**: 248-259.

Devlin, B., Roeder, K., and Ellstrand, N.C. 1988. Fractional paternity assignment: theoretical development and comparison to other methods. Theor. Appl. Genet. **76**: 369-380.

DiFazio, S.P. 2002. Measuring and Modeling Gene Flow from Hybrid Poplar Plantations: Implications for Transgenic Risk Assessment. PhD Dissertation, Oregon State University, Corvallis, OR, USA.

DiFazio, S.P., Slavov, G.T., Burczyk, J., Leonardi, S., and Strauss, S.H. 2004. Gene flow from tree plantations and implications for transgenic risk assessment. In: *Plantation forest biotechnology for the 21st century* (eds. Walter, C., and Carson, M.), **in press**. Research Signpost. Trivandrum, India.

Dow, B.D., and Ashley, M.V. 1996. Microsatellite analysis of seed dispersal and parentage of saplings of bur oak, *Quercus macrocarpa*. Mol. Ecol. **5**: 615-627.

Dow, B.D. and Ashley, M.V. 1998. High levels of gene flow in Bur oak revealed by paternity analysis using microsatellites. J. Heredity. **89**: 62-70.

Echt, C.S., May-Marquardt, P., Hseih, M., and Zahorchak, R. 1996. Characterization of microsatellite markers in eastern white pine. Genome. **39**: 1102-1108.

Edwards, K.J., Barker, J.H.A., Daly, A., Jones, C., and Karp, A. 1996. Microsatellite libraries enriched for several microsatellite sequences in plants. BioTechniques. **20**: 758-760.

El-Kassaby, Y.A., and Ritland, K. 1986a. Low levels of pollen contamination in a Douglas-fir seed orchard as detected by allozyme markers. Silvae Genet. **35**: 224-229.

El-Kassaby, Y. A., and Ritland, K. 1986b. The relation of outcrossing and contamination to reproductive phenology and supplemental mass pollination in a Douglas-fir seed. Silvae Genet. **35**: 240-244.

Ellstrand, N.C. 2001. When transgenes wander, should we worry? Plant Phys. **125**: 1543-1545.

Ellstrand, N.C., and Marshall, D. 1985. Interpopulation gene flow by pollen in wild radish, *Raphanus sativus*. Am. Nat. **126**: 606-616.

Ellstrand, N.C., Prentice, H.C., and Hancock, J.F. 1999. Gene flow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics. **30**: 539-563.

Elsik C.G., Minihan, V.T., Hall, S.E., Scarpa, A.M., and Williams C.G. 2000. Low-copy microsatellite markers for *Pinus taeda* L. Genome. **43**: 550-555.

Ennos, R.A. 1994. Estimating the relative rates of pollen and seed migration among plant populations. Heredity. **72**: 250-259.

Epperson, B.K., and Allard, R.W. 1987. Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics. **115**: 341-352.

Erickson, V.J., and Adams, W.T. 1989. Mating success in a coastal Douglas-fir seed orchard as affected by distance and floral phenology. Can. J. For. Res. **19**: 1248-1255.

Erickson, V.J., and Adams, W.T. 1990. Mating system variation among individual ramets in a Douglas-fir seed orchard. Can. J. For. Res. **20**: 1672-1675.

Estoup, A., and Cornuet, J.-M. 1999. Microsatellite evolution: inferences from population data. In: *Microsatellites: Evolution and Applications* (eds. Goldstein, D.B., and Schlötterer, C.), pp. 49-65. Oxford University Press, UK.

Ewen, K.R., Bahlo, M., Treloar, S., Levinson, D., Mowry, B., Barlow, J., and Foote, S. 2000.Identification and analysis of error types in high-throughput genotyping. Am. J. Hum. Genet. **67**: 727-736.

Fisher, P.J., Richardson, T.E., and Gardner, R.C. 1998. Characteristics of single- and multi-copy microsatellites from *Pinus radiata*. Theor. Appl. Genet. **96**: 969-979.

Friedman, S.T., and Adams, W.T. 1982. Genetic efficiency in loblolly pine seed orchards. *In* Proceedings of the 16th Southern Forest Tree Improvement Conference, pp. 213-220. Blacksburg, VA, USA.

Friedman, S.T., and Adams, W.T. 1985. Estimation of gene flow into two seed orchards of loblolly pine (Pinus taeda L.). Theor. Appl. Genet. **69**: 609-615.

Gerber, S., Mariette, S., Streiff, R., Bodénès, C., and Kremer, A. 2000. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol. Ecol. **9**: 1037-1048.

Goto, S. , Miyahara, F., and Ide, Y. 2002. Monitoring male reproductive success in a Japanese black pine clonal seed orchard with RAPD markers Can. J. of For. Res. **32**: 983-988.

Greenwood, M., and Rucker, T. 1985. Estimating pollen contamination in loblolly pine seed orchards by pollen trapping. *In*: Proc. 18th South. For. Tree Improv. Conf., Gulfport, MS, pp. 179-186.

Hamrick, J.L., and Godt, M.J. 1989. Allozyme diversity in plants species. *In*: Plamt population, genetics, breeding, and genetic resources (eds. Brown, A.H.D., Clegg, M.T., Kahler, A.L., and Weir, B.S.) Sinauer Associates, Sunderland, MA, pp.43-63.

Hicks, M., Adams, D., O'Keefe, S., MacDonald, E., and Hodgetts, R. 1998. The development of RAPD and microsatellite markers in lodgepole pine (*Pinus contorta* var. *latifolia*). Genome. **41**: 797-805.

Hodgetts, R.B., Aleksiuk, M.A., Brown, A., Clarke, C., Macdonald, E., Nadeem, S., and Khasa, D. 2001. Development of microsatellite markers for white spruce (*Picea glauca*) and related species. Theor. Appl. Genet. **102**: 1252-1258.

Jermstad, K.D., Reem, A.M., Henifin, J.R., Wheeler, N.C., and Neale, D.B. 1994. Inheritance of restriction fragment length polymorphisms and random amplified polymorphic DNAs in coastal Douglas-fir. Theor. Appl. Genet. **89**: 758-766.

Jermstad, K.D., Bassoni, D.L., Wheeler, N.C., and Neale, D.B. 1998. A sex-averaged linkage map in coastal Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco var '*menziesii*') based on RFLP and RAPD markers. Theor. Appl. Genet. **97**: 762-770.

Jermstad, K.D., Bassoni, D.L., Jech, K.S., Wheeler, N.C., and Neale, D.B. 2001a. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. I. Timing of vegetative bud flush. Theor. Appl. Genet. **102**: 1142-1151.

Jermstad, K.D., Bassoni, D.L., Wheeler, N.C., Anekonda, T.S., Aitken, S.N., Adams, W.T., and Neale, D.B. 2001b. Mapping of quantitative trait loci controlling adaptive traits in coastal Douglas-fir. II. Spring and fall cold-hardiness. Theor. Appl. Genet. **102**: 1152-1158.

Jones, A.G., and Ardren, W.R. 2003. Methods of parentage analysis in natural populations. Mol. Ecol. **12**: 2511-2523.

Kang, K.-S., and El-Kassaby, Y.A. 2002. Considerations of correlated fertility between genders on genetic diversity: The *Pinus densiflora* seed orchard as a model. Theor. Appl. Genet. **105**: 1183-1189.

Kang, K. S., Lindgren, D., and Mullin, T. J. 2001. Prediction of genetic gain and gene diversity in seed orchard crops under alternative management strategies Theor. Appl. Genet. **103**: 1099-1107.

Keys, R.N., Autino, A., Edwards, K.J., Fady, B., Pichot, C., and Vendramin, G.G. 2000. Characterization of nuclear microsatellites in *Pinus halepensis* Mill. and their inheritance in *P. halepensis* and *Pinus brutia* Ten. Mol. Ecol. **9**: 2157-2159.

Krauss, S. 2000. Patterns of mating in *Persoonia mollis* (Proteaceae) revealed by an analysis of paternity using AFLP: Implications for conservation. Australian Journal of Botany. **48**: 349-356.

Kylmänen, P. 1980. Preliminary results concerning usability of North Finland×South Finland hybrid seed born in young Scots pine seed orchards. Folia For. **423** (Summary in English).

Lambeth, C., Lee, B.-C., O'Malley, D., and Wheeler, N. 2001. Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. Theor. Appl. Genet. **103**: 930-943.

Lanner, R.M. 1965. Needed: A new approach to the study of pollen dispersion. Silvae Genet. **15**: 50-52.

Levin, D.A. and Kerster, H.W. 1974. Gene flow in seed plants. *In*: Evolutionary Biology 7. Plenum Press, New York. pp. 139-220.

Lian, C., Miwa, M., and Hogetsu, T. 2001. Outcrossing and paternity analysis of *Pinus densiflora* (Japanese red pine) by microsatellite polymorphism. Heredity. **87**: 88-98.

Marshall, T.C., Slate, J., Kruuk, L.E., and Pemberton, J.M. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7**: 639-655.

Meagher, T. 1986. Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents. Am. Nat. **128**: 199-215.

Meagher, T., and Thompson, E. 1987. Analysis of parentage for naturally established seedlings of *Chamaelirium luteum* (Liliaceae). Ecology. **68**: 803-812.

Morjan, C.L., and Rieseberg, L.H. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Mol. Ecol. **13**: 1341-1356.

Müller, G. 1976. A simple method of estimating rates of self-fertilization by analyzing isozymes in tree seeds. Silvae Genet. **25**: 15-17.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci. USA **70**: 3321-3323.

Neigel, J.E. 1997. A comparison of alternative strategies for estimating gene flow from genetic markers. Annual Review of Ecology and Systematics. **28**: 105-128.

Nikkanen, T. 1982. Survival and height growth of North Finland×South Finland hybrid progenies of Scots pine in intermediate areas. Folia For. **527** (Summary in English).

Ouborg, N.J., Piquot, Y., and Groenendael, J.M. 1999. Population genetics, molecular markers and the study of dispersal in plants. J. Ecology. **87**: 551-568.

Pakkanen, A., Nikkanen, T., and Pulkkinen, P. 2000. Annual variation in pollen contamination and outcrossing in a *Picea abies* seed orchard. Scand. J. For. Res. **15**: 399-404.

Parker, P. G., Snow, A.A., Schug, M.D., Booton, G.C., and Fuerst, P.A. 1998. What molecules can tell us about populations: choosing and using a molecular marker. Ecology. **79**: 361-382.

Peacock, M., Kirchoff, V.S., and Merideth, S.J. 2002. Identification and characterization of nine polymorphic microsatellite loci in the North American pika, *Oncotona princeps*. Mol. Ecol. Notes. **2**: 360-362.

Pemberton, J., Slate, J., Bancroft, D.R., and Barrett, J.A. 1995. Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Mol. Ecol. **4**: 249-252.

Pichot, C., and El Maataoui, M. 1997. Flow cytometric evidence for multiple ploidy levels in the endosperm of some gymnosperm species. Theor. Appl. Genet. **94**: 865-870.

Pfeiffer, A., Olivieri, A.M., and Morgante, M. 1997. Identification and characterization of microsatellites in Norway spruce (*Picea abies* K.). Genome. **40**: 411-419.

Powell, W., Machray, G.C., and Provan, J. 1996. Polymorphism revealed by simple sequence repeats. Trends in Plant Sci. **1**: 215-222.

Rafalski, J.A., Vogel, J.M., Morgante, M., Powell, W., Andre, C., and Tingey, S.V. 1996. Generating and using DNA markers in plants. *In*: Birren, B., and Lai, E. (eds.). Nonmammalian genomic analysis – a practical guide. Academic Press, San Diego, CA, USA, pp. 75-134.

Robinson, J.P., and Harris, S.A. 1999. Amplified fragment length polymorphism and microsatellites: A phylogenetic perspective. In: *Which DNA Marker for Which Purpose?* Final Compendium of the Research Project Development, Optimization and Validation of Molecular Tools for Assessment of Biodiversity in Forest Trees. http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm

Roeder, K., Devlin, B., and Lindsay, B.G. 1989. Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. Biometrics. **45**: 363-379.

SanCristobal, M., and Chevalet, C. 1997. Error tolerant parent identification from a finite set of individuals. Genet. Res. **70**: 53-62.

Savolainen, O., Karkkainen, K., Harju, A., Nikkanen, T., and Rusanen, M. 1993. Fertility variation in *Pinus sylvestris*: a test of sexual allocation theory. Am. J. Bot. **80**: 1016–1020.

Schnabel, A., and Hamrick, J.L. 1995. Understanding the population genetic structure of *Gleditsia triacanthos* L.: the scale and pattern of pollen gene flow. Evolution. **49**: 921-931.

Scotti, I., Paglia, G.P., Magni, F., and Morgante, M. 2002. Efficient development of dinucleotide microsatellite markers in Norway spruce (*Picea abies* Karst.) through dot-blot selection. Theor. Appl. Genet. **104**: 1035-1041.

Slate, J., Marshall, T., and Pemberton, J. 2000. A retrospective assessment of the accuracy of the paternity inference program CERVUS. Mol. Ecol. **9**: 801-808.

Slatkin, M. 1985. Rare alleles as indicators of gene flow. Evolution. **39**: 53-65.

Smith, D.B., and Adams, W.T. 1983. Measuring pollen contamination in clonal seed orchards with the aid of genetic markers. *In* Proceedings of the 17th Southern Forest Tree Improvement Conference, pp. 64-73. Athens, GA, USA.

Smouse, P., and Meagher, T. 1994. Genetic analysis of male reproductive contributions in *Chamaelirium luteum* (L.) Gray (Liliaceae). Genetics. **136**: 313-322.

Smouse, P., Dyer, R.J., Westfall, R.D., and Sork,V.L. 2001. Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. Evolution. **55**: 260-271.

Soranzo, N., Provan, J., and Powell, W. 1998. Characterization of microsatellite loci in *Pinus sylvestris* L. Mol. Ecol. **7**: 1260-1261.

Sorensen, F.C. 1999. Relationship between self-fertility, allocation of growth, and inbreeding depression in three coniferous species. Evolution. **53**:417-25.

Sork, V.L., Nason, J., Campbell, D.R., Fernandez, J.F. 1999. Landscape approaches to contemporary gene flow in plants. Trends Ecol. Evol. **14**: 219-224.

Sork, V.L., Campbell, D., Dyer, R., Fernandez, J.F., Nason, J., Petit, R., Smouse, P., and Steinberg, E. 1998. Proceedings from a Workshop on gene flow in fragmented, managed, and continuous populations. National Center for Ecological Analysis and Synthesis, Santa Barbara, California. Research Paper No. 3. http://www.nceas.ucsb.edu/nceas-web/projects/2057/nceas-paper3

Squillace, A.E., and Long, E.M. 1981. Proportion of pollen from non-orchard sources. *In* E.C. Franklin (ed.) Pollen management handbook. U.S.D.A. Agriculture Handbook 587. Washington, D.C. pp. 15-19.

Stam, P., and Van Ooijen, J.W. 1995. JoinMap^{TM} version 2.0: Software for the calculation of genetic linkage maps. Centre for plant breeding and reproduction research (CPRO-DLO), Wageningen, the Netherlands.

Stein, W.I., and Owston, P.W. 2002. *Pseudotsuga* Carr. In: Woody plant seed manual. F.T. Bonner (tech. coord.) and RG. Nisley (managing ed.), USDA-Forest Service, Washington , DC. http://wpsm.net/Pseudotsuga.pdf

Stockwell, C.A., Hendry, A.P., and Kinnison, M.T. 2003. Contemporary evolution meets conservation biology. Trends Ecol. Evol. **18**: 94-101.

Stoehr, M.U., and Newton, C.R. 2002. Evaluation of mating dynamics in a lodgepole pine seed orchard using chloroplast DNA markers. Can. J. of For. Res. **32:** 469-476.

Stoehr, M.U., Webber, J.E., and Painter, R.A. 1994. Pollen contamination effects on progeny from an off-site Douglas-fir seed orchard. Can. J. of For. Res. **24**: 2113-2117.

Stoehr, M.U., Orvar, B., Vo, T., Gawley, J., Webber, J. and Newton, C. 1998. Application of a chloroplast DNA marker in seed orchard management evaluations of Douglas-fir. Can. J. For. Res. **28**: 187-195.

Strauss, S.H. 2003. Genomics, genetic engineering, and domestication of crops. Science. **300**: 61-62.

Streiff, R., Ducousso, A., Lexer, C., Steinkellner, H., Gloessl, J., and Kremer, A. 1999. Pollen dispersal inferred from paternity analysis in a mixed stand of *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. Mol. Ecol. **8**: 831-841.

Summers, K., and Amos, W. 1997. Behavioral, ecological, and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. Behav. Ecol. **8**: 260-267.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. 2002. Mathematical statistics with applications. Sixth edition. Duxbury, USA.

Webber, J.E. 1995. Pollen management for intensive seed orchard production. Tree Physiology. **15**: 507-514

Webber, J.E., and Painter, R.A. 1996. Douglas-fir pollen management manual. Second Edition. Res. Br., B.C. Min. For., Victoria, B.C. Work. Pap. 02/1996.

Wheeler, N., and Jech, K. 1986. Pollen contamination in a mature Douglas-fir seed orchard. *In* Proc. IUFRO Conference on Breeding Theory, Progeny Testing, and Seed Orchards, Williamsburg, Virginia, pp. 13-17.

Wheeler, N., and Jech, K. 1992. The use of electrophoretic markers in seed orchard research. New For. **6**: 311-328.

Whitlock, M.C., and McCauley, D.E. 1999. Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. Heredity. **82**: 117-125.

Wolfenbarger, L.L., and Phifer, P.R. 2000. The ecological risks and benefits of genetically engineered crops. Science. **290**: 2088-2093.

Woods, J.H., Stoehr, M.U., and Webber, J.E. 1996. Protocols for rating seed orchard seedlots in British Columbia. Res. Rep. 06, B.C. Min. For.,Victoria. 26 p.

Wright, S. 1931. Evolution in Mendelian populations. Genetics. **16**: 97-159.

Xie, C.-Y., and Yanchuk, A. D. 2003. Breeding values of parental trees, genetic worth of seed orchard seedlots, and yields of improved stocks in British Columbia. West. J. Appl. For. **18**: 88-100.

Zane, L., Bargelloni, L., and Patarnello, T. 2002. Strategies for microsatellite isolation: A review. Mol. Ecol. **11**: 1-16.

APPENDICES

APPENDIX 1

Table A.1. Segregation of SSR alleles in megagametophytes of Douglas-fir (data pooled over 6-17 mother trees).

Locus	N^a	Number of megagametophytes with:		$\chi^{2\,b}$	P^c
		longer allele	shorter allele		
PmOSU_1C3	9	32	27	0.42	0.52
PmOSU_1F9	14	40	51	1.33	0.25
PmOSU_2B6	14	48	44	0.17	0.68
PmOSU_2C2	14	47	47	0.00	1.00
PmOSU_2C3	14	51	41	1.09	0.30
PmOSU_2D4	7	20	26	0.78	0.38
PmOSU_2D6	10	32	29	0.21	0.65
PmOSU_2D9	9	36	22	3.38	0.07
PmOSU_2G4	10	23	36	2.86	0.09
PmOSU_2G12	17	47	53	0.36	0.55
PmOSU_3B2	17	55	59	0.14	0.71
PmOSU_3B9	15	45	38	0.59	0.44
PmOSU_3D5	12	44	35	1.03	0.31
PmOSU_3E3	10	33	29	0.26	0.61
PmOSU_3F1	13	51	48	0.09	0.76
PmOSU_3G9	13	45	43	0.05	0.83
PmOSU_3H4	13	41	42	0.01	0.91
PmOSU_4A7	15	50	52	0.04	0.84
PmOSU_4E9	12	41	38	0.11	0.74
PmOSU_4G2	12	42	43	0.01	0.91
PmOSU_5A8	7	22	26	0.33	0.56
PmOSU_783	6	22	17	0.64	0.42

[a] N is the number of heterozygous trees with 5-8 megagametophytes per tree genotyped.
[b] χ^2 is a chi-square test statistic for the expected 1:1 ratio of segregating alleles (Adams and Joly 1980).
[c] P is the one-sided P-value based on a chi-square test with null hypothesis of no deviation from the expected 1:1 ratio.

Table A.2. Segregation of SSR alleles in the diploid progeny of a controlled cross used for linkage mapping in Douglas-fir.

Locus	N^a	Female parent		$\chi^{2\,b}$	Male parent		$\chi^{2\,b}$
		longer allele	shorter allele		longer allele	shorter allele	
PmOSU_1C3	86	42	44	0.1	42	44	0.1
PmOSU_2B6	90	47	43	0.2	33	57[c]	6.4**
PmOSU_2C2	92		homozygous		43	49	0.4
PmOSU_2D4	84	38	46[c]	0.8	46	38	0.8
PmOSU_2D6	92	39	53	2.1	40	52	1.6
PmOSU_2D9	91	49	42	0.5		homozygous	
PmOSU_2G4	90	44	46	0.0	38	52	2.2
PmOSU_2G12	92	53	39	2.1	52	40	1.6
PmOSU_3B2	91	36	55	4.0*	42	49	0.5
PmOSU_3B9	92	46	46	0.0	45	47	0.0
PmOSU_3D5	90	44	46[c]	0.0	52	38	2.2
PmOSU_3E3	91	41	50	0.9	48	43	0.3
PmOSU_3F1	89	46	43	0.1	36	53	3.3
PmOSU_3G9	91	40	51	1.3		homozygous	
PmOSU_3H4	86	46	40	0.4	42	44[c]	0.1
PmOSU_4A7	90	44	46	0.0	40	50	1.1
PmOSU_4E9	83	40	43	0.1	45	38	0.6
PmOSU_4G2	90	54	36	3.6*		homozygous	
PmOSU_5A8	90	48	42	0.4	44	46[c]	0.0
PmOSU_783	89	43	46	0.1	44	45	0.0

[a] N is the number of progeny genotyped.
[b] χ^2 is a chi-square test statistic for the expected 1:1 ratio of segregating alleles (Adams and Joly 1980).
* P <0.05.
** P <0.01.
[c] Null-allele.

APPENDIX 2

User's guide to PFL, a computer program for estimating pollen flow using paternity exclusion and SSR markers

Access and installation

PFL compiled for Windows will be available on the web at the Pacific Northwest Tree Improvement Research Cooperative site (http://www.fsl.orst.edu/pnwtirc/research). It can also be obtained from Gancho T. Slavov (gancho@lifetime.oregonstate.edu). To install the program, download and decompress the file pfl.zip.

Overview

PFL estimates pollen immigration ($\hat{m}$) and its standard error using paternity exclusion as described in the *Data analysis* section of Chapter 3. PFL allows the user to require multiple mismatches between the pollen gamete haplotype of an offspring and all pollen gamete haplotypes that can be produced in the population for which pollen immigration is estimated before classifying that offspring as an observed immigrant. This is done in order to eliminate upward biases of $\hat{m}$ caused by mistyping. Specific recommendations on obtaining reliable estimates of pollen immigration using PFL are discussed in Chapter 3 and presented visually in Fig. 3.7. PFL is designed to handle two general types of data. The *Haploid* option can be applied in conifers in which the complete pollen gamete haplotype of each offspring can be unambiguously inferred (i.e., when both haploid seed megagametophytes and diploid embryos are genotyped). The *Diploid* option can be

applied to any diploid seed plant species but it requires that the mother of each offspring is known and genotyped. Figure 3.1. should be referred to for terminology.

Input

A. *Haploid* option

Three input data files are required to run PFL and obtain estimates of $\hat{m}$ and $SE(\hat{m})$:

1) A file with the multilocus genotypes of all *local parents*. The following rules apply to the format of this file (also, see sample file hapar.txt):
   - Each genotype must be on a separate line, no blank lines are allowed.
   - There must be exactly 2N integer numbers on each line, where N is the number of loci used for genotyping the *local parents*. Non-digit symbols will generate error, use zeros for missing data.
   - The integer numbers must be separated by white spaces.
2) A file with the multilocus genotypes of the *background parents* or a file with allele frequencies in the *background population*. If the former option is chosen, all formatting rules listed in 1) apply (see sample file haback.txt). Otherwise if the allele frequencies in the *background population* are already estimated, use sample file hapalfs.txt as a formatting template and pay attention to:
   - Separate loci using "$", without leaving any blank lines.
   - Separate the alleles from their frequencies using white spaces.
   - Include only integers in the allele column and only floating point numbers in the frequency column.

- Make sure that allele frequencies at each locus sum up to exactly one.

3) A file with the multilocus pollen gamete haplotypes of all offspring. Follow the formatting rules listed in 1) but include N integers on each line (see sample file hagams.txt).

B. *Diploid* option

Three input data files are required to run PFL and obtain estimates of $\hat{m}$ and $SE(\hat{m})$ for each progeny array (i.e., sample of offspring from a given mother):

1) A file with the multilocus genotypes of all *local parents*. The formatting requirements are the same as in the *Haploid* option, see sample file dipar.txt.
2) A file with the multilocus genotypes of the *background parents* or a file with allele frequencies in the *background population*. The formatting requirements are the same as in the *Haploid* option, see sample files diback.txt and dipalfs.txt.
3) A file with the multilocus diploid genotypes of all offspring and their respective mothers. Follow the formatting rules given in 1) and put the maternal genotype on the first line. If progeny arrays from multiple mothers need to be analyzed, see sample file dipoff.txt and pay attention to:
   - Separate progeny arrays from different mothers using "$".
   - Begin each genotypic array with the genotype of the mother.

Running the program and examples

To start the program, double-click on pfl.exe in the folder containing all decompressed files. In the main menu, select "1" to execute the *Haploid* option or "2" to execute the *Diploid* option.

Example 1. *Haploid* option, allele frequencies in the *background population* estimated from the genotypes of *background parents*.

1. Select "1" from the main menu.
2. Enter "7" when asked for number of loci.
3. Enter "hapar.txt" as the name of the file containing the genotypes of the *local parents*.
4. Select "1" to estimate allele frequencies from genotypes.
5. Enter "haback.txt" as the name of the file containing the genotypes of *background parents*.
6. Enter "3" as a number of mismatches required for exclusion.
7. Select "Yes" to continue with obtaining an estimate of $\hat{m}$ (if "No" is selected, the program will only estimate the detection probability).
8. Enter "hagams.txt" as the name of the file containing pollen gamete haplotypes.
9. If no error message is printed, the results will be saved in file output.dat. This file should look almost identical to the sample output file outex1.dat provided with the program. The only differences should be in the values of d, m, and $SE(m)$. These differences should be very small; they result from the different sets of random numbers used to estimate d.

Example 2. *Diploid* option, allele frequencies in the *background population* entered directly.

1. Select "2" from the main menu.
2. Enter "10" when asked for number of loci.
3. Enter "dipar.txt" as the name of the file containing the genotypes of the *local parents*.
4. Select "2" to enter allele frequencies directly.
5. Enter "dipalfs.txt" as the name of the file containing allele frequencies in the *background population*.
6. Enter "3" as a number of mismatches required for exclusion.
7. Enter "dipoff.txt" as the name of the file containing the genotypes of mothers and offspring (dipoff.txt contains the genotypes of two mothers, with ten offspring per mother).
8. Select "Yes" to continue with obtaining estimates of $\hat{m}$ for each progeny array (if "No" is selected, the program will only estimate the detection probability for each progeny array).
9. If no error message is printed, the results will be saved in file output.dat. This file should look almost identical to the sample output file outex2.dat provided with the program. The only differences should be in the values of d, m, and SE (m) for each of the two progeny arrays. These differences should be very small; they result from the different sets of random numbers used to estimate d.