



Evolutionary Dynamics of a Multiple-Ploidy System in Arabidopsis Arenosa

Citation

Arnold, Brian. 2015. Evolutionary Dynamics of a Multiple-Ploidy System in Arabidopsis Arenosa. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467222>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Evolutionary dynamics of a multiple-ploidy system in *Arabidopsis arenosa*

A dissertation presented

by

Brian John Arnold

to

The Department of Organismic and Evolutionary Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Biology

Harvard University

Cambridge, Massachusetts

May, 2015

© 2015 – Brian Arnold

All rights reserved.

EVOLUTIONARY DYNAMICS OF A MULTIPLE- PLOIDY SYSTEM IN *ARABIDOPSIS ARENOSA*

ABSTRACT

Whole-genome duplication (WGD), which leads to polyploidy, has been implicated in speciation and biological novelty. In plants, many species have experienced historical bouts of WGD or exhibit extant ploidy variation, which is likely representative of an early stage in the evolution of new polyploid lineages. To elucidate the evolutionary dynamics of autopolyploids and species with multiple ploidy levels, I develop population genetic theory in Chapter 2 that I use in Chapter 4 to extract information about the evolutionary history of *Arabidopsis arenosa*, a European wildflower that has diploid and autotetraploid populations. Chapter 3 involves a separate project exploring the ascertainment bias in restriction site associated DNA sequencing (RADseq). In Chapter 2, I develop coalescent models for autotetraploid species with tetrasomic inheritance and show that the ancestral genetic process in a large population without recombination may be approximated using Kingman's standard coalescent, with a coalescent effective population size $4N$. Using this result, I

was able to use existing coalescent simulation programs to show in Chapter 4 that, in *A. arenosa*, a widespread autotetraploid race arose from a single ancestral population. This autopolyploidization event was not accompanied by immediate reproductive isolation between diploids and tetraploids in this species, as I find evidence of extensive interploidy admixture between diploid and tetraploid populations that are geographically close.

To draw these conclusions about population history in Chapter 4, I used a reduced representation genome-sequencing approach based on restriction digestion. However, I was bothered by the possibility that sampling chromosomes based on restriction digestion may introduce a bias in allele frequency estimation due to polymorphisms in restriction sites. To explore the effects of this nonrandom sampling and its sensitivity to different evolutionary parameters, we developed a coalescent-simulation framework in Chapter 3 to mimic the biased recovery of chromosomes in RADseq experiments. We show that loci with missing haplotypes have estimated diversity statistic values that can deviate dramatically from true values and are also enriched for particular genealogical histories. These results urge caution when applying this technique to make population genetic inferences and helped me tailor analyses in Chapter 4 to accommodate for this particular method of DNA sequencing.

TABLE OF CONTENTS

Title Page	i.
Copyright Page	ii.
Abstract	iii.
Table of contents	v.
Statement of contributions	vii.
Acknowledgements	ix.
Dedication	xi.
Chapter 1 – Introduction	1.
Chapter 2 – Extending Coalescent Theory to Autotetraploids	7.
2.1 Abstract	7.
2.2 Introduction	7.
2.3 Theory	10.
2.4 Discussion	30.
Chapter 3 - RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling	36.
3.1 Abstract	36.
3.2 Introduction	37.
3.3 Results	40.
3.4 Discussion	52.
3.5 Methods	59.
Chapter 4 - Single geographic origin of autotetraploid <i>Arabidopsis arenosa</i> followed by interploidy admixture	65.
4.1 Abstract	65.
4.2 Introduction	66.
4.3 Results	69.
4.4 Discussion	86.
4.5 Methods	92.
Chapter 5 - Conclusions and Future Directions	99.
Supplementary material for Chapter 2	102.
Supplementary Text	102.
Supplementary Tables	104.
Supplementary material for Chapter 3	118.
Supplementary Figures	118.
Supplementary material for Chapter 4	124.
Supplementary Text	124.

Supplementary Figures	127.
Supplementary Tables	141.
Bibliography	155.

CONTRIBUTIONS

Chapter 1

Contributions: Brian Arnold (BA) wrote the text.

Chapter 2

Contributions: BA and John Wakeley (JW) designed the project. BA performed the math. BA, JW, and Kirsten Bomblies (KB) wrote the manuscript. Chapter 2 has previously been published with minor changes as the following paper:

Arnold B, Bomblies K, Wakeley J. 2012. Extending coalescent theory to autotetraploids. <i>Genetics</i> . 192: 195-204.
--

Chapter 3

Contributions: BA, Russ Corbett-Detig (RCD), and KB designed the project. BA and RCD made the scripts. BA analyzed the data. BA, RCD, KB, and Dan Hartl wrote the manuscript. Chapter 3 has previously been published with minor changes as the following paper:

Arnold B, Corbett-Detig R, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. <i>Mol Ecol</i> 22: 3179-3190.
--

Chapter 4

Contributions: BA designed the project, collected and analyzed the data. Sang-Tae Kim provided some data. BA and KB wrote the manuscript.
Chapter 4 has been recently accepted to the journal of Molecular Biology and Evolution.

Chapter 5

Contributions: BA wrote the text.

ACKNOWLEDGEMENTS

The past six years have been incredible, both personally and academically. Though I experienced ups and down and saw the limits of my sanity, I could not have created a better set of experiences myself, and I will be forever thankful to the people who enabled me to grow into the young adult that I am today. To me, this is the most important section of my thesis.

Although I like to think my work ethic and perseverance are responsible for my successes, the unconditional love and support of my parents is the primary reason I have had amazing opportunities in life. They served as unwavering role models, sent me to good schools, passively encouraged academic excellence, and allowed me to develop and explore any curiosity I had. All of my beneficial attributes are some product of their parenting. I would also be far less interesting if it were not for my two amazing brothers.

Harvard University has been an incredible place with even better people. I am not worthy of the intelligent, considerate, and creative advisors I have had the pleasure of working under. Kirsten Bomblies has been a remarkable advisor and a great friend. Over the years she has always been encouraging and nurturing, and only from these cultivating interactions was I able to develop a semblance of academic self-confidence: the most important quality I have acquired in graduate school. My scientific interests have dramatically wandered, and while she kept me on track, she also allowed me to pursue ideas even if it meant working with someone else (which is a lot to ask of a pre-tenure faculty in my opinion). The working environment she created, which has cohesiveness unparalleled by any other lab, has brought me years of happiness. She provided me with the coolest biological system to work on that has challenged and developed my ability to think about evolution. Without her I would not be the enthusiastic, curious scientist I am today, and my writing would also be far less intelligible.

During my second year of graduate school, John Wakeley asked me to help teach his course on Coalescent Theory, trusting my abilities despite never having taken the course. However, John is inspiringly generous with his time and met with me weekly to answer any basic questions I had about theoretical population genetics. To this day, I frequently use this theory to build my intuition of how evolution affects patterns of genetic variation, and this theoretical framework not only allowed me to answer questions I have been passionate about but also effectively market myself to a wide variety of postdoc advisors.

After a lively meeting with Nancy Kleckner to discuss the evolution of tetraploid meiosis and how it could potentially be modeled, she proposed a fascinating “small” project I could carry out in her lab. She taught me the basics of experimental molecular biology and her rigorous logic of reconstructing molecular processes from mutant analyses. From my interactions with her, I have undoubtedly become a better, more thorough scientist. At first, I was not accustomed to her standards of precision, and the first group meeting I gave in her lab was more demanding than my qualifying exam.

However, I quickly began to cherish her constructive criticism and emerged enlightened and excited from every conversation we had. I am also extremely grateful to Liangran for the time he spent training me, and for encouraging me to always keep trying despite failures.

I am also extremely grateful for the two other incredible professors that served on my committee: Hopi Hoekstra and David Reich. Their insightful input on my projects dramatically changed the course of my PhD in all good ways and motivated me to continue working hard.

My completion of graduate school was largely due to the love and support of a close friend. Perhaps the best part of the past six years of my life was meeting Shane Campbell-Staton. A lot major life changes can happen throughout a six-year time span, and Shane was invariably there for me every single time. I have never had a deeper, more fulfilling friendship with someone. Shane and I have had extremely different upbringings, and being able to try to see the world through his eyes has been the most educational experience of my graduate school career. Thanks for all the good times; I love you buddy.

My experience at Harvard would not have been nearly as fun without the people in the best, most cohesive lab on campus. Pierre has been an amazing friend, has helped me through tough times, and has also converted me into a bit of a Francophile. Among the most memorable experiences I have had involve me, him, and some type of electronic music. Jesse has been an inspiration both personally and academically, and to this day I strangely catch myself mimicking some of his mannerisms. Collecting plants across Central Europe with him is a memory I will treasure forever. Ben has also enriched my life with his amazing, uncommon (or perhaps just British) sense of humor and lots of booze. Julie has been one of the most considerate (yet sassy) people I have ever met, and I have long lost count how many times she has brightened my day with her generosity. Kevin and Andrew have been amazing to spend time with and talk science. Lastly, I would like to thank Yanniv and Kristin, the two most unique friends I've ever had; they both have benefitted my life in ways I continue to discover.

I am also extremely thankful to the Harvard Ballroom Dance Team for introducing me to a hobby that has brought me so much happiness and has been the most efficient stress reliever. I would like to especially thank Christie for being an amazing, diligent dance partner and a great friend.

Lastly I would like to show my utmost gratitude to my sources of funding: Herchel Smith, James Mills Peirce, and the NSF for a graduate research fellowship and a doctoral dissertation improvement grant.

Dedicated to my advisors Kirsten Bomblies, John Wakeley, and Nancy Kleckner, for their wisdom and support.

CHAPTER 1 – INTRODUCTION

Whole-genome duplication (WGD) has occurred in many organisms across eukaryotic kingdoms and has profoundly shaped genome evolution (Kellis et al. 2004; Dehal and Boore 2005; Jiao et al. 2011). These large-scale genomic events are implicated in increased genomic complexity and are associated with adaptive radiations of major lineages throughout the tree of life (Dehal and Boore 2005; Jiao et al. 2011). WGD is particularly frequent in plants; ancient WGD events are estimated to have occurred in 30-100% of angiosperm lineages (Stebbins 1950; Grant 1981; Masterson 1994; Cui et al. 2006). Thus, many of the diploid genomes we observe today are in fact derivatives of polyploidy ancestors, making the evolutionary dynamics of WGD an important component of genome architecture across the tree of life.

Although WGD has occurred numerous times throughout the evolutionary history of plants, many extant plant species are known to have multiple ploidy levels (see for review Ramsey and Schemske 1998; Soltis et al. 2010), showing that polyploidy remains an active force in plant evolution. The probability neopolyploids persist and give rise to stable populations depends on many factors including the rate at which polyploids arise, fertility defects of higher ploidy individuals, and the degree to which ploidy levels are reproductively isolated (Thompson and Lumaret 1992). Species with multiple ploidies provide a glimpse of the early evolutionary processes by which polyploids arise from diploid populations and form demographically stable populations.

There are two major classes of polyploids: autopolyploids, which form from within-species WGD and generally randomly segregate homologs, and allopolyploids, which have a hybrid origin and usually diploid-like inheritance (Ramsey and Schemske 1998; Parisod et al. 2010; Bomblies and Madlung 2014). These two types of polyploids differ dramatically from one another. Both types involve an increase in the haploid number of chromosomes per cell, but allopolyploidy is accompanied by the numerous effects of hybridization between two species, while autopolyploids are much more similar to the diploid parents from which they arose (Doyle et al. 2008). Autopolyploids were once thought vanishingly rare compared to allopolyploids, but they are much more common than previously suspected (Soltis et al 2007). Historically, autopolyploids may not have been recognized because they look morphologically similar to their diploid progenitor, and many autopolyploids that display diploid-like pairing of chromosomes during meiosis have been misclassified as allopolyploids (Soltis et al 2007).

Autotetraploids are an extremely common form of autopolyploid and arise via unreduced gamete formation within diploid populations. These tetraploids may come directly from the union of unreduced gametes, or they may arise from triploid intermediates, which arise from unreduced gametes fusing with haploid gametes and may produce some tetraploid progeny (Ramsey and Shemske 1998). Autotetraploids likely depend on their diploid progenitor for a source of mates during early stages of demographic establishment, and later on as an occasional source of genetic variation. Support for the degree to which autopolyploids may admix with their diploid progenitor comes from numerous data that shed light on the ability of autotetraploids to exchange

genetic information with diploids (Thórsson et al. 2001, Ståhlberg 2009, Husband and Sabara 2011, Oberle et al. 2012, Sonnleitner et al. 2013, Clark et al. 2015). The ability of autopolyploids to exchange genes with their diploid ancestor, either through semi-fertile interploidy crosses or the continued production of unreduced gametes in diploids, may enhance the probability spontaneously-arisen autopolyploids give rise to a stable population. If true, then it should be no surprise that numerous studies document evidence for interploidy gene flow in species with multiple ploidies (Petit et al. 1999). Thus, either nascent autopolyploidy does not serve as a particularly strong reproductive barrier to gene flow, or there is selection for multiploidy systems with incomplete isolation between ploidy levels.

However, autopolyploids may be self-reliant in multiple ploidy systems that are self-compatible or if whole-genome duplication is coupled with a break down in the self-incompatibility (SI) system. There is no clear consensus as to whether autopolyploidy is associated with self-compatibility, as some studies group auto- and allo-tetraploids (Mable 2004, Barringer 2007), but SI need not necessarily breakdown in neoautopolyploids that were derived from diploids with an intact SI system since these polyploids contain a single SI locus as in the diploid progenitor (Pandey 1977). Regardless of what the consensus may be, there will invariably be exceptions to the rule as there are numerous examples of autopolyploids with intact and perturbed SI systems, suggesting flexible evolutionary paths to stable autopolyploid populations. The presence of an SI system has important consequences for autopolyploid evolution in demographic

establishment, gene flow with its diploid ancestor, and the long-term evolution of genomic load in populations.

Autotetraploids often exhibit tetrasomic inheritance (arising from random segregation of all four homologs), which presents an intriguing problem that has important implications for population genetic analysis of genomic data (Bever and Felber 1992). Compared to diploid populations, autotetraploid populations contain twice as many homologs per individual, enabling higher effective sizes and greater amounts of genetic diversity (Moody et al. 1993, Arnold et al. 2012). However, we may not necessarily observe this in nature as nonequilibrium demography can greatly diminish levels of genetic variation. For example, since autotetraploids arise from a rare mutational process in which diploids produce diploid gametes (nondisjunction), entire populations may arise from relatively few individuals (Thompson and Lumaret 1992, Ramsey and Schemske 1998). This population bottleneck would reduce levels of genetic variation that would only recover over longer periods of evolutionary time.

Another important difference between individuals of different ploidy is heterozygosity. For example, if a particular gene has two alleles segregating in a population, A and a , potential autotetraploid genotypes include $aaaa$, $Aaaa$, $AAaa$, $AAAa$, or $AAAA$, three of which are heterozygous (Bever and Felber 1992). Diploids have only three possible genotypes: aa , Aa , and AA . Thus, not only do autotetraploids have the potential for giving rise to populations with greater effective sizes, but they also have more possibilities for diversity within individuals. This could be advantageous if diversity within an individual is beneficial, such as at loci involved in pathogen resistance

or at loci in which deleterious recessive alleles segregate. In the latter case, more possibilities for heterozygosity ensure more individuals have at least one perfectly functional allele.

However, the enhanced ability of autotetraploids to mask deleterious recessives is a double-edged sword. Although more individuals have at least one functional allele, deleterious recessives in autotetraploid populations effectively experience relaxed selection, which acts mostly on homozygotes (i.e. *aaaa*). Population genetic theory predicts deleterious recessive mutations to have higher frequencies at equilibrium in autotetraploid populations when compared to diploid populations (Ronfort 1999). Consequently, more deleterious mutations accumulate in autotetraploid genomes because they may reach higher frequencies at mutation-selection balance and persist for longer periods of evolutionary time before going extinct (Ronfort 1999)..

Despite the evolutionary significance and intriguing nature of autotetraploids, surprisingly little genomic data exists for these species. Thus, I have dedicated my thesis to generating new theory, original genomic analyses, and novel genomic datasets to enhance our understanding of the evolutionary dynamics of autopolyploidy. In Chapter 2, I extend a body of theory, called the “Coalescent”, to autotetraploid organisms. Coalescent theory is frequently used in population genetics as a means to interpret patterns of genetic variation and reconstruct the evolutionary history of populations. The results presented in this chapter enable the use of coalescent theory with genomic data from autotetraploids. Chapter 3 does not directly study the genomics of autotetraploid evolution but involves a quantification of the ascertainment biases

intrinsic to a genomic technology I later use in Chapter 4. I discovered that this genomic technology called RADseq (Restriction-site Associated DNA sequencing), an extremely popular sequencing method among biologists studying non-model-organism genomics, underestimates genetic diversity. The degree to which this ascertainment bias affects diversity statistics is locus-specific, so these results helped tailor my analyses of RADseq datasets. Finally in chapter 4, using genomic data that I generated, I reconstructed the history of a species with multiple ploidy levels, diploid and autotetraploid. This study represents the first reconstruction of evolutionary history in a multiple ploidy system to illuminate how autotetraploids evolve from and interact with diploid ancestors.

CHAPTER 2 – EXTENDING COALESCENT THEORY TO AUTOTETRAPLOIDS

2.1 Abstract

We develop coalescent models for autotetraploid species with tetrasomic inheritance. We show that the ancestral genetic process in a large population without recombination may be approximated using Kingman's standard coalescent, with a coalescent effective population size $4N$. Numerical results suggest that this approximation is accurate for population sizes on the order of hundreds of individuals. Therefore, existing coalescent simulation programs can be adapted to study population history in autotetraploids simply by interpreting the timescale in units of $4N$ generations. We also consider the possibility of double reduction, a phenomenon unique to polysomic inheritance, and show that its effects on gene genealogies are similar to partial self-fertilization.

2.2 Introduction

Polyploidy, which results from whole-genome duplication, is a significant evolutionary force throughout the tree of life. It is particularly widespread in higher plants but also occurs in fishes, amphibians, reptiles, insects, and even a mammal (Leggat and Iwama 2003, Gregory and Mable 2005, Sexton 1980, Gallardo et al. 1999). In plants, estimates of the proportion of angiosperm species that have experienced

genome doubling during their evolutionary history varies from 30% to 100% (Stebbins 1950, Grant 1981, Masterson 1994, Ciu et al 2006), and polyploidy is thought to be a potent mechanism of sympatric speciation (Wood et al 2009).

Polyploids can arise via interspecific hybridization (allopolyploids) or intraspecific genome duplication (autopolyploids), e.g. through the fusion of unreduced gametes (Stebbins 1947). Allopolyploids, born from the union of distinct genomes, frequently exhibit bivalent pairing and disomic inheritance. As a result, the duplicated chromosome sets are only partially homologous (homeologous) and follow separate evolutionary paths if they do not pair and recombine. Conversely, autotetraploids contain four non-diverged sets of chromosomes that are fully homologous. During meiosis, these homologs may form multivalents or bivalents with random chromosome pairing, resulting in tetrasomic inheritance in both cases.

Autotetraploids were once thought vanishingly rare compared to allopolyploids, but they are much more common than previously suspected (Soltis et al 2007). Numerous wild plant species have been demonstrated to exhibit tetrasomic inheritance with random chromosome pairing, despite forming only bivalents at meiosis I (Table S2.1 and references therein, Soltis et al. 2007). However, allele segregation in autopolyploids can become much more complex than that of diploids when chromosomes form multivalents from crossing over with more than one homolog during meiosis, as this may lead to double reduction. Double reduction occurs when multivalents are resolved such that segments of sister chromatids migrate to the same pole at meiosis I, allowing, for example, an ABBB genotype to produce AA gametes

(Haldane 1930, Mather 1935). Among cytologically characterized natural autotetraploid plants, multivalent formation is less common than bivalent pairing but is still present in enough species to merit attention in theoretical models.

The body of literature on theoretical population genetics of polyploids has grown but is still very small in comparison to the work done on diploids (Bever and Felber 1992, Otto and Whitton 2000). Studies have characterized equilibrium genotype proportions (Haldane 1930), genetic drift and levels of genetic variation (Wright 1938, Moody et al. 1993), as well as population structure of autotetraploid populations (Ronfort et al 1998, Luo et al 2006). A few studies have considered the effects of double reduction on the equilibrium frequencies of neutral and deleterious alleles (Crow 1954, Butruille and Boiteux 1999). The gene genealogical or coalescent approach to population genetics (Hudson 1983, Tajima 1983) has proven a useful framework for interpreting genetic variation. Coalescent models have been applied to data from tetraploid species, but the justification for this has not been elucidated. The ability to extract all the information from a set of DNA sequences collected from natural populations will help answer questions about autopolyploid evolution. For example, do most autopolyploids experience severe bottlenecks from the formation event? How many independent formation events do most autopolyploid species experience? Is there gene flow between ploidy levels? How old are these autopolyploid populations and how evolutionarily stable is tetrasomic inheritance? What does genome structure look like in natural populations, in terms of evolutionarily important parameters such as the population mutation and recombination rates?

Despite the increasing awareness of autopolyploidy, the burgeoning advances in DNA sequencing technology, and the utility of these data for studying evolution, few studies have analyzed nuclear DNA sequence variation in natural populations of autotetraploids (St. Onge et al. 2012, Jørgensen et al. 2011, Tiffin and Gaut 2001). Here we develop a coalescent model for autotetraploids in order to facilitate the analysis of DNA sequence data. Specifically, we derive the coalescent effective population size (Sjödín et al 2005) for autotetraploid species. We consider both double reduction and the possibility of partial selfing. Our mathematical results hold in the limit as the population size tends to infinity, but we show that they are numerically very accurate when the population size is only moderately large (in the hundreds). These results provide a mathematical framework to explicitly model DNA sequence evolution in autotetraploid populations, which may be employed to estimate mutation and recombination rates, infer ancestral demography, and detect various types of selection from DNA sequence data sets of diploids. In short, we show how standard coalescent models or simulations may be applied to autotetraploids with only minor modification, thus allowing for detailed predictions to be made about patterns of genetic variation and for population history and the evolutionary forces acting on natural populations to be inferred from DNA sequence data.

2.3 Theory

Kingman (1982a, 1982b) gave a formal proof of the existence of what he called the '*n*-coalescent'—now simply coalescent or Kingman's coalescent—as the ancestral

genetic process for a sample from a large haploid population. Specifically, for a genetic locus at which all variation is selectively neutral and there is no intra-locus recombination, the genetic ancestry of a sample of size n may be modeled as a process in which each pair of lineages ancestral to the sample ‘coalesces’ with rate equal to one. Derivations of the coalescent begin with the computation of single-generation probabilities (i.e. of coalescence) then proceed by taking a limit as the population size (N) tends to infinity, rescaling time in units proportional to N generations so that a coalescence rate of one per pair of lineages is obtained. Coalescent models have been extended to include population subdivision and migration, changes in population size over time, recombination, and natural selection. Efficient software is available to simulate samples of genetic data (Hudson 2002, Ewing and Hermisson 2010).

Kingman’s derivation of the coalescent is valid only for haploid population models (this includes the diploid monoecious Wright-Fisher model, because it can be reduced to a haploid model). In particular, Kingman assumed that genetic lineages are ‘exchangeable’ as in the haploid population models of Cannings (1974). The general formalism for treating diploids or other (non-exchangeable) population structures was developed by Möhle (1998a, 1998b, 1998c). It has been applied in a variety of situations to show that Kingman’s coalescent is robust to deviations from Kingman’s original assumptions, but also to describe an augmented set of ancestral genetic processes that are closely related to Kingman’s coalescent; see Wakeley (2008) for a review. Here we use Möhle’s technique to derive the ancestral genetic process at a single neutral locus without recombination in an autotetraploid species.

Sample size of two chromosomes

Consider a Wright-Fisher population of N autotetraploid, monoecious individuals that create gametes via tetrasomic inheritance with no recombination (i.e. no double reduction, which is considered later). That is, an individual with four distinct alleles at a locus produces $\binom{4}{2} = 6$ kinds of gametes, each with equal frequency. Generations are non-overlapping. Each of the N offspring that form the next generation is created by the union of two gametes sampled randomly with replacement from all possible gametes of the parental generation. Thus each individual is produced by selfing with probability $1/N$. Due to the added structure in which each gamete contains two distinct parental copies of each locus, and in contrast to the diploid monoecious case, samples of chromosomes from a tetraploid species are not exchangeable. In particular, the conditional probability of coalescence depends on whether two genetic lineages are within the same individual or in separate individuals. For example, without recombination and double reduction, two lineages cannot coalesce in the immediately previous generation if they are within the same individual and came from the same gamete.

We can construct a single-generation transition matrix \mathbf{P} for an ancestral process that accounts for the specific details of tetrasomic inheritance by setting up an absorbing Markov chain that has three states for a sample size of $n=2$: (1) two distinct lineages within the same individual, (2) two lineages in separate individuals, and (3) a

single lineage that is the common ancestor of the two lineages. As we trace them back in time, the two lineages may travel between states one and two until they ultimately coalesce. Transition probabilities are calculated by conditioning on two pieces of information: whether or not lineages came from the same gamete in the previous generation and the particular pattern of ancestry.

For instance, if both lineages are in the same individual (state 1), they coalesce in the previous generation if they came from different gametes (probability $2/3$) produced by the same parent (probability $1/N$) and trace back to the same chromosome in that parent (probability $1/4$), (See Figure 2.1). Thus, $P_{1,3} = 1/6N$. This is in agreement with Wright 1938, who showed that the “proportion of unlike pairs of genes” decays by a factor of $5/6$ for self-fertilizing autotetraploids. If two lineages are in separate individuals (state 2) the probability of coalescence must be computed for all possible patterns of shared ancestry that potentially result in identity by descent (Table S2.2), with the marginal probability of coalescence being $P_{2,3} = 1/4N$ after simplification.

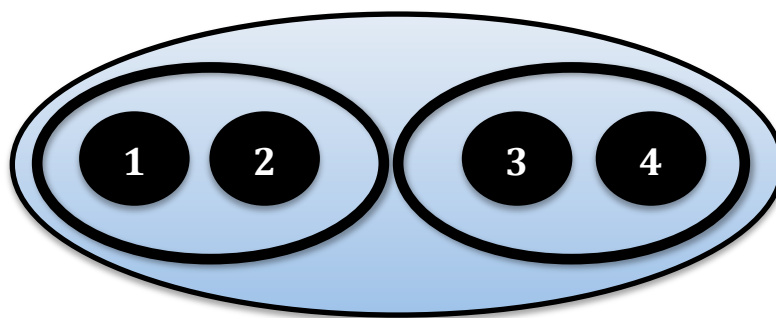


Figure 2.1 A diagram of a tetraploid individual, showing the two diploid gametes that united to create it. If chromosomes are sampled without replacement, after the first lineage is sampled, with probability $2/3$ we sample the second lineage from a different gamete, which came from the same parent with probability $1/N$ (where N is the population size) assuming random mating. Conditional on these events, these two sampled lineages coalesce with probability $1/4$.

Applying the same logic to calculate the probabilities of other possible transitions, we get the following 3-by-3 single-generation transition matrix \mathbf{P} with states 1, 2, and 3 represented by rows one, two, and three respectively:

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} + \frac{1}{2N} & \frac{2}{3}\left(1 - \frac{1}{N}\right) & \frac{1}{6N} \\ \frac{3}{4N} & \left(1 - \frac{1}{N}\right) & \frac{1}{4N} \\ 0 & 0 & 1 \end{bmatrix}.$$

This matrix describes the exact, discrete-time ancestral process for the two lineages. As in Kingman's coalescent, we seek a continuous-time approximation which will be accurate when the population size is large. Specifically, we take the limit $N \rightarrow \infty$, with time rescaled by N , in order to test whether the ancestral limit process converges to Kingman's coalescent.

Möhle (1998a) obtained a useful convergence result for Markov processes with two timescales such as the one described by the matrix above. Since \mathbf{P} contains terms that become increasingly different in the limit $N \rightarrow \infty$, we can use Möhle's result to construct a continuous-time approximation. We rewrite \mathbf{P} in three parts, such that

$$\mathbf{P} = \mathbf{F} + \frac{\mathbf{S}}{N} + o\left(\frac{1}{N}\right) \quad (1)$$

where $\mathbf{F} = \lim_{N \rightarrow \infty} \mathbf{P}$ and $\mathbf{S} = \lim_{N \rightarrow \infty} N(\mathbf{P} - \mathbf{F})$. Matrix \mathbf{F} contains the “fast” events that occur frequently on the timescale of the original discrete-time process (i.e. the movement of lineages from within to between gametes), whereas matrix \mathbf{S} contains the “slow” events (i.e. coalescence). The terms in \mathbf{S}/N become very small as N tends to infinity. These events occur on the timescale of N generations. While Möhle’s result allows for additional terms, of $o(1/N)$, which tend to zero more quickly than $1/N$, in our case these terms are equal to zero.

If the matrix $\mathbf{E} = \lim_{t \rightarrow \infty} \mathbf{F}^t$ exists, the continuous-time approximation to our ancestral process involves the “fast” events instantaneously reaching their equilibrium (\mathbf{E}), after which they enter the slower process of coalescence described by rate matrix $\mathbf{G} = \mathbf{E}\mathbf{S}\mathbf{E}$ (Möhle 1998a). More formally, Möhle’s Theorem 1 (Möhle 1998a) states that the rescaled ancestral process converges to $\lim_{N \rightarrow \infty} \mathbf{P}^{[Nr]} = \mathbf{E}e^{t\mathbf{G}}$. Together matrices \mathbf{E} and \mathbf{G} describe the rescaled ancestral process, with time measured in units of N generations.

Applying this theorem to our discrete-time matrix above, we have

$$\mathbf{F} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} \frac{1}{2} & -\frac{2}{3} & \frac{1}{6} \\ \frac{3}{4} & -1 & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}$$

such that

$$\mathbf{E} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} 0 & -\frac{1}{4} & \frac{1}{4} \\ 0 & -\frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}.$$

Here, we see that if lineages start out in state 1 (within the same individual), they may remain in that state if they came from the same gamete ($\mathbf{F}_{1,1}$) or different gametes from the same parent ($\mathbf{S}_{1,1}$). However, our continuous-time approximation for large populations shows that the number of generations the lineages remain in state 1 is negligible on the timescale of N generations, so they immediately travel to separate individuals (from state 1 to state 2 since $\mathbf{E}_{1,2}=1$). Once in state 2, the pair of lineages enter the coalescence process given by \mathbf{G} , which is a simple exponential process in which coalescence occurs with rate equal to $1/4$ (on the timescale of N generations), which agrees with forward-time models of autotetraploid populations (Moody et al. 1993). If time is rescaled again, by the constant factor four, so that one unit of time is equal to $4N$ generations, then the rate of coalescence is one, just as Kingman's coalescent. That is, we have shown for a sample size of two that the coalescent process for an autotetraploid species without double reduction converges to the same limiting ancestral process as a haploid population model when the coalescent effective size is defined as $N_e = 4N$.






Extension to larger samples

Convergence to Kingman's coalescent for a sample of size two does not guarantee convergence for larger samples. It must also be true that pairwise coalescent events dominate the limiting ancestral process (Möhle and Sagitov 2001). In the case of an autotetraploid without double reduction, we must check that multiple coalescent events between lineages within a single individual are negligible in the limit $N \rightarrow \infty$. For example, four lineages within a single individual will coalesce in two pairs in the immediately previous generation with probability $1/6N$, which is of the same order of magnitude as the rate of pairwise coalescence. In fact, such events become negligible in the limit because four lineages in a single individual will be overwhelmingly more likely to be descended from two pairs of lineages in two different individuals. We show this by briefly repeating our previous analysis, using Möhle's (1998a) technique, or eq. (1), but for a sample of $n=4$.

We construct this more complicated ancestral process by defining a new absorbing Markov chain with seven states which include all possible configurations of four lineages. Transient states 1 through 5 account for all possible ways lineages can be distributed within and between individuals (Table 2.1). States 6 and 7 are both absorbing, with the former representing single, pairwise coalescence events and the latter defined to include all possible multiple coalescent events. States 6 and 7 absorbing in the sense that here we are concerned only with the process during which there are four ancestral lineages. However, the coalescent process resumes on the remaining ancestral lineages, if more than one remain, with a new transition matrix

appropriate for the smaller sample size. As in the derivation for $n=2$, transition probabilities between states can be calculated by conditioning on patterns of ancestry. All possible patterns for four lineages, along with their associated conditional transition probabilities, are shown in Table S2.3.

Table 2.1 Markov states that account for all possible configurations of four lineages in an autotetraploid population, with two types of absorbing states to assess the relative probabilities of single and multiple coalescence events.

Configuration	State	Description
	1	All 4 lineages within the same individual
	2	3 lineages within the same individual, 1 lineage in separate individual
	3	2 pairs of lineages, each within the same individual
	4	2 lineages within the same individual, other 2 each in separate individuals
	5	All 4 lineages in separate individuals
	6	Three lineages remain after single coalescence event
	7	One or two lineages remain after multiple coalescence event

We obtain a new \mathbf{P} which is now the 7-by-7 single-generation transition matrix for a sample of size $n=4$ (not shown). The continuous-time approximation for a large population is obtained by applying Möhle's theorem (1998a), decomposing \mathbf{P} into three separate matrices that contain events which occur on different timescales, as done above. In the limit as $N \rightarrow \infty$, with time rescaled in N generations, we obtain matrices \mathbf{E} and \mathbf{G} which describe the ancestral process:

$$\mathbf{E} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \\ 0 & 0 & 0 & 0 & -3/2 & 3/2 & 0 \end{bmatrix}.$$

Similarly to the process for $n=2$, there is an instantaneous adjustment of the sample to a state in which all lineages are in separate individuals (here state 5), independent of the starting state. The sample then enters the continuous-time process given by \mathbf{G} , in which the rate of single coalescence events, on the timescale of N generations, is six times greater than in the case of $n=2$, or $6 \cdot (1/4) = 3/2$. This is identical to Kingman's coalescent, in which coalescence times are exponentially-distributed and occur with rate equal to the number of pairs of lineages: $\binom{4}{2}$.

Our result is analogous to the one obtained by Möhle (1998b) for a diploid dioecious population. In the diploid case, even though two lineages in a single individual cannot coalesce in the previous generation (thus violating exchangeability assumption of Kingman's coalescent), lineages quickly assume a state in which each is in a separate individual, and the long-term coalescence rate is equal to one per pair of lineages when time is measured in units of $2N$ generations. In the tetraploid monoecious case, although lineages may travel together in gametes for some number of generations without coalescing, they are overwhelmingly more likely to become scattered such that

each is in a separate individual, and the long-term coalescence rate is equal to one per pair of lineages when time is measured in units of $4N$ generations.

Accuracy for finite N

Although the continuous-time approximation applies in the limit as population size tends to infinity, we would like to know the validity of this approximation for smaller, finite populations. All the information about the ancestral process is contained in the single-generation transition matrix \mathbf{P} , so we can analyze the dynamics of this discrete-time process for a range of population sizes and compare them to the continuous-time approximation that only allows single coalescent events. We investigate the accuracy of two key features of the limiting ancestral process: that coalescence events occur predominantly between pairs of lineages, and that the majority of the ancestral process before coalescence is spent in state 5 (i.e. with all lineages in separate individuals, making them exchangeable).

We can do this using standard theory of absorbing Markov chains, for example in Chapter 11 of Grinstead and Snell (1997), by writing the transition matrix in canonical block form,

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

in which \mathbf{Q} is a 5-by-5 matrix of transition probabilities among transient (non-coalescent) states, \mathbf{R} is a 5-by-2 matrix of transition probabilities from transient to absorbing (coalescent) states, $\mathbf{0}$ is a 2-by-5 zero matrix, and \mathbf{I} is the identity matrix (in this case 2-by-2). Then, for each starting state, the 5-by-5 matrix inverse $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$ contains the expected numbers of generations spent in each transient state before absorption and the 5-by-2 matrix product \mathbf{NR} contains the probabilities of absorption in each absorbing state.

In terms of how genetic data are typically sampled, there are two extreme starting states for the process: state 1 or state 5, in which a biologist samples all chromosomes from a single autotetraploid individual or one chromosome from each of four individuals. Given these starting states, the probability of a single coalescence event (absorbing to state 6 rather than state 7) is plotted in Figure 2.2 for a range of population sizes, indicating rapid convergence to an ancestral process involving predominantly pairwise coalescence events. Likewise, Figure 2.3 shows that the fraction of time spent in state 5 (when all lineages in separate individuals) approaches one quickly as the population size increases. In Figure 2.3 the sample is assumed to start in state 1, which may be considered the farthest from the transient-equilibrium (state 5) of the limiting ancestral process. From Figure 2.2 and Figure 2.3, we conclude that the limiting ancestral process for a sample of size $n=4$ should be quite accurate as long as the population size is greater than about 100.

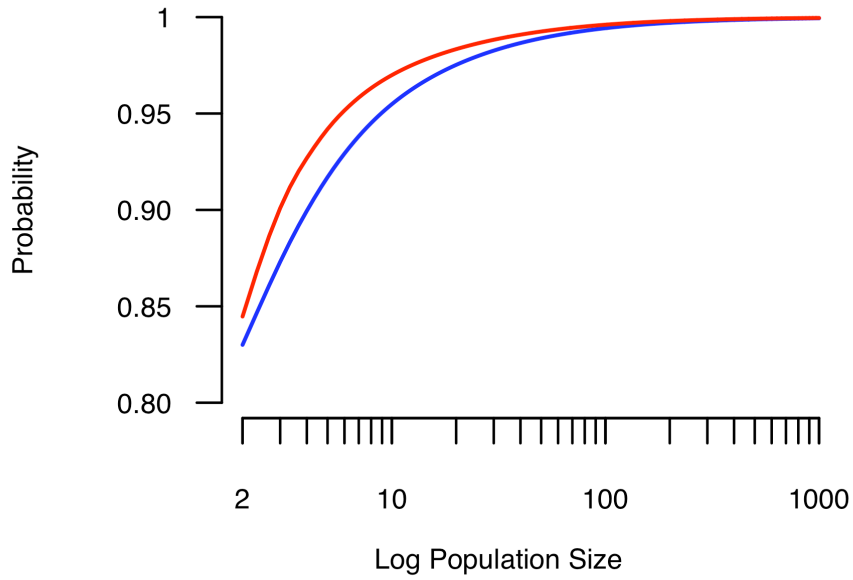


Figure 2.2 The probability of a single coalescence event, conditional on an absorbing event, given the process started in state 1 (blue) or state 5 (red) for N up to 1000 autotetraploid individuals.

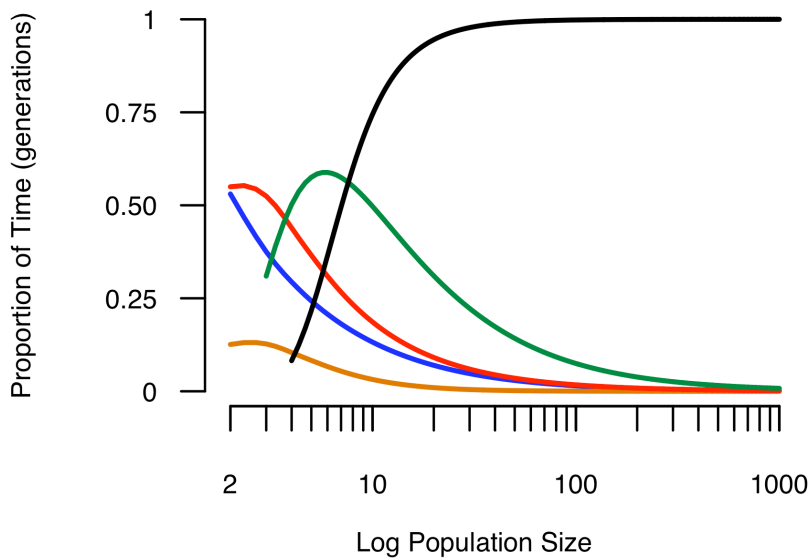


Figure 2.3 The expected proportion of generations the Markov chain spends in state 1 (blue), state 2 (orange), state 3 (red), state 4 (green), or state 5 (black) for a range of population sizes. Here the process starts in state 1 but spends a vast majority of its time in state 5, even for small populations. These proportions were calculated from matrix \mathbf{N}

(see text). Note that state 4 is possible only when $N \geq 3$ and state 5 is possible only when $N \geq 4$.

Extension to arbitrary sample size n

An exact calculation for larger sample sizes is not practical given complexity of the model with $n=4$. However, a strong heuristic derivation can be made based on the fact that when $n \ll N$, events that occur with probability $O(1)$ or $O(1/N)$ per generation will dominate the ancestral process. In short, if n lineages are in fewer than n individuals, the most probable events are $O(1)$ and these send the lineages into different individuals. If instead each lineage is in a separate individual, the most probable events are $O(1/N)$ and these bring lineages into the same individual, possibly to coalesce. Due to the $O(1)$ transition probabilities, the chance that n lineages will ever be in fewer than $n-1$ individuals (other than when they are originally sampled) is negligible if $n \ll N$.

Thus, starting with n lineages in n individuals, we can accurately summarize the process using a Markov chain with just three states: (1) n lineages are within $n-1$ individuals, (2) n lineages remain in n individuals, and (3) $n-1$ lineages arrived at by a coalescent event (see Table S2.4). We have

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} + O\left(\frac{1}{N}\right) & \frac{2}{3} + O\left(\frac{1}{N}\right) & O\left(\frac{1}{N}\right) \\ \left(\binom{2n}{2} - n\right) \frac{3}{16N} + O\left(\frac{1}{N^2}\right) & 1 + \left(n - \binom{2n}{2}\right) \frac{1}{4N} + O\left(\frac{1}{N^2}\right) & \left(\binom{2n}{2} - n\right) \frac{1}{16N} + O\left(\frac{1}{N^2}\right) \\ 0 & 0 & 1 \end{bmatrix}$$

Applying Möhle's theorem (1998a), we obtain matrices **E** and **G**:

$$\mathbf{E} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{G} = \begin{bmatrix} 0 & -\binom{n}{2}\frac{1}{4} & \binom{n}{2}\frac{1}{4} \\ 0 & -\binom{n}{2}\frac{1}{4} & \binom{n}{2}\frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}$$

which demonstrates that the ancestral process, starting either in state 1 or state 2, corresponds to Kingman's standard '*n*-coalescent' if time is measured in units of $4N$ generations.

An ancestral process with double reduction

Autotetraploids contain four sets of homologous chromosomes, and crossovers may potentially occur among any of them. When more than one crossover arises per chromosome, they may involve different pairing partners and create multivalents at metaphase I. Depending on how chromosomes segregate, double reduction may occur at a particular locus if there is a crossover between this locus and the centromere (Mather 1935, 1936, Crow 1954). Retrospectively, these lineages are automatically identical by descent (i.e. they coalesce) in the previous generation. This is qualitatively similar to the case of partial selfing studied by Nordborg and Donnelly (1997) and Möhle (1998a), in which lineages within a single individual may coalesce in the immediately previous generation if the individual was produced by selfing.

Multivalent formation is a necessary precursor for double reduction (Mather 1935). Most established autotetraploids form bivalents at meiosis (see Table S2.3) almost certainly because multivalents are associated with aneuploid gametes and thus reduced fitness of progeny (reviewed in Comai 2005). Double reduction is thus a phenomenon primarily of newly formed autotetraploids that have not adapted to a genome-doubled state, but individuals capable of correctly segregating multivalents could theoretically be selected for (Comai 2005). Depending on how meiotic mechanisms evolve in a particular autotetraploid species, double reduction may not occur over long enough periods of evolutionary time (i.e. on the coalescent timescale of N generations) to have a significant effect. Nonetheless, the presence of double reduction in at least some natural autopolyploids means that it merits consideration. We will use Möhle's technique here as well, with $n=2$, to study the effects of double reduction on the coalescent process.

Consider a locus at some distance from the centromere, such that recombination may occur between them. For simplicity, we will assume that recombination does not occur within the locus under consideration. Following Stift et al. (2008), the frequency of double reduction (α) has a theoretical maximum value of $1/6$, assuming that chromosomes always form quadrivalents and one crossover occurs between the locus and centromere. With probability $1/3$, the recombined chromosomes migrate to the same pole during meiosis I (assuming all chromosome pairs are equally likely to segregate), and the probability that the two sister chromatids also migrate to the same pole during meiosis II is $1/2$. Double reduction can be less than this maximum if

recombination between the locus and centromere does not occur in every multivalent association (i.e. the locus is distal to centromere) or if chromosomes form bivalents with some probability at meiosis I. Thus, α may range from 0 to 1/6 and will differ between loci.

To study an ancestral process for $n=2$ that includes double reduction, we define an absorbing Markov chain with the same three states used previously. The resulting matrix looks very similar to the one for tetrasomic inheritance:

$$\mathbf{P} = \begin{bmatrix} \frac{1-\alpha}{3} + \frac{1}{2N} & \frac{2}{3} - \frac{2}{3N} & \frac{\alpha}{3} + \frac{1}{6N} \\ \frac{3}{4N} & 1 - \frac{1}{N} & \frac{1}{4N} \\ 0 & 0 & 1 \end{bmatrix}.$$

This shows that double reduction increases the probability of coalescence from state 1 by an amount $\alpha/3$. Since double reduction does not affect lineages that came from separate gametes, the transition probabilities for state 2 are unchanged. When in state 1, if lineages came from the same gamete in the previous generation, which occurs with probability 1/3, there is a chance α that they were sister chromatids in the immediately previous generation and thus coalesce. Importantly, the chance of this novel transition when lineages are in state 1 does not depend on the population size N . There is no reason to suppose that α is small, and we assume that it is a constant when we take

the limit $N \rightarrow \infty$. Using Möhle's theorem (1998) to obtain a continuous-time approximation on the timescale of N generations, we have

$$\mathbf{E} = \begin{bmatrix} 0 & \frac{2}{2+\alpha} & \frac{\alpha}{2+\alpha} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and}$$

$$\mathbf{G} = \begin{bmatrix} 0 & \frac{3}{(\alpha+2)^2} - \frac{2}{\alpha+2} & \frac{3\alpha}{2(\alpha+2)^2} + \frac{1}{2(\alpha+2)} \\ 0 & \frac{3}{2(\alpha+2)} - 1 & \frac{1}{4} + \frac{3\alpha}{4(\alpha+2)} \\ 0 & 0 & 0 \end{bmatrix}.$$

Unlike the previous models above, coalescence is now possible in both the “fast” and “slow” processes. On the limiting timescale of N generations, lineages may instantaneously coalesce instead of moving directly to state 2. This will create an association of alleles within individuals. Once in state 2, lineages will remain in this state for a majority of their ancestry but coalesce at a rate faster than $1/4$ when time is measured in units of N generations. Thus double reduction decreases the long-term coalescent N_e . In addition, a single exponentially-distributed rate of coalescence (i.e. the Kingman coalescent) is not sufficient to capture all the dynamics of an ancestral process

with double reduction; with time rescaled in units proportional to N generations, lineages may coalesce instantaneously if they were sampled from the same individual.

As previously mentioned, these effects are also observed in the coalescent process with partial self-fertilization (Nordborg and Donnelly 1997, Möhle 1998b). However, the two models are not identical for tetraploids. In the Supplementary Text, we extend our model of coalescence for autotetraploids without double reduction to include partial selfing, with probability s . We find that the maximum rate of double reduction ($\alpha = 1/6$) produces the same ancestral process as one with $s = 1/4$, but that there are slight quantitative differences between the two models. Figure 2.4 shows the relationship between s and α , in terms of the parameter values that give the same coalescent N_e . The relationship is only slightly nonlinear; though mechanistically distinct, selfing and double-reduction have very similar effects on patterns and levels of genetic variation.

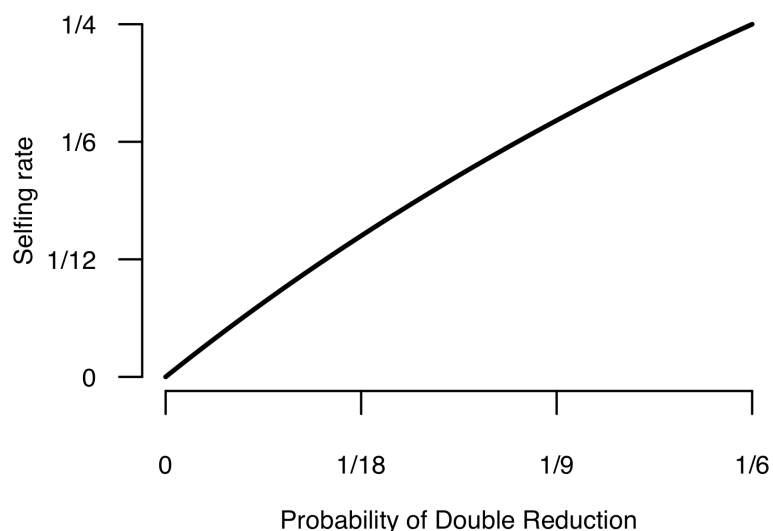


Figure 2.4 The slightly non-linear relationship between double reduction and self-fertilization. The maximum probability of double reduction ($1/6$) corresponds to a selfing rate of $1/4$.

The coalescent with double reduction for arbitrary n

Consider a sample of n alleles from a population such that k individuals contain two alleles, l individuals contain three alleles, m individuals contain four alleles, and $n-2k-3l-4m$ alleles are in separate individuals. Unlike the extension to arbitrary n for the model above, coalescence events may occur with $O(1)$ if double reduction is possible. Thus, the instantaneous adjustment of the sample involves both the movement of lineages from within to between individuals and coalescence events. After this instantaneous adjustment, the remaining lineages are in separate individuals and coalesce with probability $O(1/N)$.

The number of instantaneous coalescence events depends on the sample configuration, i.e. the number of lineages within an individual. Following Nordborg and Donnelly 1997, for each of the k individuals that contain two lineages, the number of instantaneous coalescence events is $X \sim \text{Binomial}(k, \alpha/(\alpha+2))$, with the probability of coalescence calculated above from the analysis with $n=2$. For the l individuals that contain three lineages, the number of instantaneous coalescence events is $Y \sim \text{Binomial}(l, 3\alpha/(\alpha+2))$: two of the lineages must come from the same gamete and coalesce with probability α , but if they do not coalesce with probability $1-\alpha$, then they coalesce in previous generations from double reduction with probability $(\alpha/(\alpha+2))$. Thus, the overall probability of coalescence is $\alpha + (1-\alpha)(\alpha/(\alpha+2)) = 3\alpha/(\alpha+2)$. Applying the same logic to the m individuals with four lineages (see Supplementary Text), the number of coalescence events is $\mathbf{Z}=(Z_1,Z_2,Z_3) \sim \text{Multinomial}(m, \mathbf{p}=(p_1,p_2,p_3))$. Here, Z_1 is the number

of single coalescence events that occur with probability $p_1 = \frac{12\alpha(1-\alpha)}{(\alpha+2)^2}$, Z_2 is the number

of double coalescence events that occur with probability $p_2 = \frac{9\alpha^2}{(\alpha+2)^2}$, and Z_3 is the

number of times no lineages coalesce (i.e. four lineages within an individual get

separated into four distinct individuals) with probability $p_3 = \frac{4(\alpha-1)^2}{(\alpha+2)^2}$. The number of

lineages that remain after this instantaneous adjustment, $n-X-Y-Z_1-2Z_2$, are in separate

individuals and enter the “slow” process of coalescence with rate

$\binom{n-X-Y-Z_1-Z_2}{2} \left(\frac{1}{4} + \frac{3\alpha}{4(\alpha+2)} \right)$ if time is measured in units of N generations. The

coalescence rate calculated here is the same as above for $n=2$ but applied to each pair of remaining lineages.

2.4 Discussion

Autotetraploids with tetrasomic inheritance have long been considered vanishingly rare but now are increasingly recognized as a common phenomenon in plant evolution (Soltis et al. 2007). Many established polyploid species or populations have been shown to have tetrasomic inheritance (Table S2.1). Nuclear DNA sequence data will provide invaluable insight into many unknown aspects of autopolyploid evolution, such as the process of formation and establishment from relatively few individuals that are at least partially reproductively isolated from their diploid progenitor. Here, we extend the coalescent, a widely-used model in population genetics and phylogenetics,

to autotetraploid populations to aid in the analysis of DNA sequence data sets from these species.

Our results show that, although tetrasomic inheritance creates additional configurations of lineages that have unique probabilities of coalescence compared to those for diploid organisms, the ancestral process for autotetraploids without double reduction converges to Kingman's haploid coalescent model with time rescaled by $4N$ generations. Intuitively, this convergence occurs because in a large, panmictic population ancestral lineages quickly get separated to different individuals such that the sample spends a majority of its history in a configuration in which lineages are exchangeable.

Simulating data with the tetrasomic inheritance model without double reduction would produce genealogies like those generated from diploid and haploid models, such as Hudson's *ms* (Hudson 2002), with the exception that the timescale of the process must be interpreted differently. The time to the most recent common ancestor (T_{MRCA}) of a pair of lineages in an autotetraploid population is exponentially distributed with a mean of one when time is measured in units of $4N$ generations. Thus a given value of T_{MRCA} is interpreted as $4N \times T_{\text{MRCA}}$ rather than $2N \times T_{\text{MRCA}}$ as in diploid coalescent models. From this it follows that, as has been previously demonstrated (Moody et al. 1993), autotetraploids are expected to have twice the levels of genetic variation as diploids for a given demographic size. However, many demographic or biological factors may lead to departures from this expectation, such as population history, the distribution of offspring per individual, or nonrandom mating.

A useful application of the coalescent is to infer ancestral demography. Little is known about the characteristics of the bottlenecks associated with the formation of an autotetraploid lineage, which can arise through the union of unreduced gametes produced by diploids. Since autotetraploid formation events are relatively rare, but potentially on the order of the genic mutation rate (Ramsey and Schemske 1998), entire populations may be founded by a small number of individuals, resulting in a severe genetic bottleneck. However, the severity of this bottleneck likely varies among cases as certain environmental factors and alleles may greatly affect the rate of unreduced gamete formation in diploids (Ramsey and Schemske 1998); higher rates can lead directly to repeated autotetraploid formation or promote gene flow from diploids to tetraploids. Multiple formation events or gene flow from diploid gene pools may increase the effective population size during autotetraploid formation and reduce the severity of the bottleneck. Several studies have documented multiple origins in autotetraploids (Soltis et al. 1989, Wolf et al. 1990, Ptacek et al. 1994). Alternatively, gene flow among ploidy levels may occur via inter-ploidy hybrids (i.e. triploids) if they have non-zero fertility and produce some euploid gametes (Felber and Bever 1997, Husband 2004).

The utility of the models developed here depends on the specific demographic history of the autotetraploid population. If the present day autotetraploid race traces back to relatively few individuals (i.e. ~ 10), then the sample size n , or number of ancestral lineages, is not much smaller than the population size N . Since $n \ll N$ is a critical assumption of the Kingman coalescent, the continuous-time models developed

here in the context of large population size may not be applicable. Simulated gene genealogies produced by standard coalescent programs such as Hudson's *ms* may not look like the actual pattern of ancestry that contains multifurcations, a likely gene tree structure of small populations that affects predicted patterns of genetic variation. Special simulation programs may be developed that can accommodate extreme population size crashes, for example as in the metapopulation model of Wakeley and Aliacar (2001). Alternatively, the distribution of pairwise coalescence and linkage disequilibrium can be used to quantify the severity of the bottleneck, such as Li and Durbin's (2011) pairwise coalescent model, if multiple mergers are a likely feature of the underlying genealogy. Simulations should be done to see if these models are robust to such large population size reductions.

If mating is not random (i.e. self-fertilization occurs) or if the autotetraploids form multivalents and exhibit double reduction, the ancestral process does not converge to Kingman's simple model. In these cases, Kingman's coalescent does not fully capture the relatively complicated ancestry of the sample of lineages because the coalescent process cannot be described by a single exponential distribution; coalescent events may occur instantly with a nonzero probability even as population size tends to infinity. For sample sizes greater than two, simultaneous multiple mergers become an issue. For instance, if four lineages are within the same individual that was created by a selfing event, a simultaneous multiple merger occurs in the previous generation with probability $s/6$, where s is the selfing rate (see Supplementary Text for details). This probability is independent of population size and thus does not tend to zero as $N \rightarrow \infty$.

For a coalescent model with double reduction and four lineages within the same individual, as $N \rightarrow \infty$ gametes likely come from different parents unlike the selfing model. However, a simultaneous multiple merger occurs at the locus of interest if both gametes that formed the individual were produced by double reduction events with probability α^2 . Thus, these ancestral processes are more complex than the Kingman coalescent since coalescent events may happen on two separate timescales and simultaneous multiple mergers likely occur.

The coalescent model with double reduction is similar to the many-demes model, with its “scattering phase” and “collecting phase” (Wakeley 1999) and to the coalescent with partial selfing (Nordborg and Donnelly 1997, Mohle 1998a). It is interesting to note, however, that a maximum of two $O(1)$ coalescence events may occur in the instantaneous adjustment for four lineages sampled from within a single tetraploid individual. This results from the fact that with probability of $O(1)$ the two gametes that form each individual came from different parents, and only each pair that came from the same gamete may coalesce via double reduction. For finite N , more coalescence events are possible if gametes originated from the same parent, but this has probability $O(1/N)$ and thus does not happen in the instantaneous adjustment. Tetrasomic inheritance thus creates a genetic structure that has a mathematically distinct ancestral process.

However, there are two observations which suggest that, at least in plants, self-fertilization or double reduction may rarely affect results: first, of the established autotetraploid plant species that have been documented and studied, multivalent

formation (and thus the possibility of double reduction) and self-fertilization are both rare. For example, among 24 species for which tetrasomic inheritance has been molecularly confirmed, chromosome associations have also been examined in 11. Among these, only two closely related species commonly form multivalents. Two other species form multivalents only rarely, while the remainder show exclusively bivalent associations at meiosis (Table S2.1). Thus in most of these examples, double reduction would be rare or absent. Self-fertilization is also rare. Though selfing can in theory promote the establishment of tetraploids by helping avoid minority cytotype exclusion (Rodriguez 1996), it is very rare among polyploid species in nature (Stebbins 1947). The majority of plant species we identified in the literature that have tetrasomic inheritance are obligately outcrossing (see Table S2.1 notes for references). Selfing rates are therefore generally zero for most autotetraploids. Thus, the simple tetrasomic coalescent model may be widely applicable.

In conclusion, our results demonstrate that Kingman's coalescent is robust to tetrasomic inheritance, making existing coalescent models applicable for analyzing population genomic data collected from natural autotetraploid populations that exhibit this mode of inheritance. These models will greatly facilitate the study of the evolutionary forces acting on these organisms. However, standard coalescent simulators cannot be used to interpret these data if the autotetraploids self-fertilize at an appreciable rate or if some loci experience double reduction, the latter being verified by cytology and segregation studies.

CHAPTER 3 - RADSEQ UNDERESTIMATES DIVERSITY AND INTRODUCES GENEALOGICAL BIASES DUE TO NONRANDOM HAPLOTYPE SAMPLING

3.1 Abstract

Reduced representation genome-sequencing approaches based on restriction digestion are enabling large-scale marker generation and facilitating genomic studies in a wide range of model and non-model systems. However, sampling chromosomes based on restriction digestion may introduce a bias in allele frequency estimation due to polymorphisms in restriction sites. To explore the effects of this nonrandom sampling and its sensitivity to different evolutionary parameters, we developed a coalescent-simulation framework to mimic the biased recovery of chromosomes in restriction-based short-read sequencing experiments (RADseq). We analyzed simulated DNA sequence datasets and compared known values from simulations with those that would be estimated using a RADseq approach from the same samples. We compare these "true" and "estimated" values of commonly used summary statistics π , θ_w , Tajima's D , and F_{ST} . We show that loci with missing haplotypes have estimated summary statistic values that can deviate dramatically from true values and are also enriched for particular genealogical histories. These biases are sensitive to non-equilibrium demography, such as bottlenecks and population expansion. *In silico* digests with 102 completely sequenced *D. melanogaster* genomes yielded results similar to our findings

from coalescent simulations. Though the potential of RADseq for marker discovery and trait mapping in non-model systems remains undisputed, our results urge caution when applying this technique to make population genetic inferences.

3.2 Introduction

High-throughput sequencing technology has revolutionized evolutionary genetics, enabling biologists to generate massive amounts of genomic data to address diverse questions in ecology and evolution. Importantly, new techniques allow high-throughput identification of variable sites (e.g. single nucleotide polymorphisms; SNPs), even in species whose genomes are prohibitively large for sequencing or for which a reference genome is unavailable. In these situations, it is often preferable to eschew whole-genome sequencing in favor of a reduced-representation approach that can be used to sample a fraction of the genome across many individuals at the same loci. A promising new technology, restriction-associated DNA (RADseq), is becoming popular for reducing genomic complexity in DNA libraries to sequence a small portion of the genome across many individuals (reviewed in Davey *et al.* 2011). Hundreds of indexed RAD libraries can be easily and inexpensively constructed and sequenced to characterize levels and patterns of genetic variation throughout the genome, even for non-model organisms. RADseq has already been employed in studies of population structure and biogeography (Hohenlohe *et al.* 2010, Emerson *et al.* 2010, Gompert *et al.* 2010), allele frequency estimation (Van Tassel *et al.* 2008), association studies (Parchman *et al.* 2012), genetic mapping (Baird *et al.* 2008, Andolfatto *et al.* 2011, Pfender *et al.* 2011),

selection and introgression (Hohenlohe *et al.* 2011, Gompert *et al.* 2012), and linkage disequilibrium (Hohenlohe *et al.* 2012).

RADseq differs from other genome-sequencing approaches in that DNA fragments for construction of a library of sequences are generated by digesting genomes with a restriction enzyme, as opposed to random DNA shearing. Enzyme digestion results in nonrandom cleavage that ensures primarily the same regions are sampled across individuals. While powerful, the RADseq approach may be affected by numerous, largely uncharacterized biases. Potential problems arising from PCR bias in library construction, sequencing errors, and inaccurate genotyping with lower sequencing depths have been recognized previously (Rokas and Abbot 2009), but these biases are expected to affect all resequencing projects. RADseq has an additional ascertainment bias whose effects have not been explored extensively: some recognition sequences will themselves be polymorphic, resulting in missing data for some chromosomes and thus nonrandom sampling of lineages in a sample (Figure 3.1).

How does non-randomly missing data affect estimation of levels and patterns of genomic variation necessary for population genetic inference? Here we address this question by developing a coalescent-simulation framework to mimic the biased recovery of haplotypes (hereafter genealogical bias) in RAD libraries. Our work is consistent with but extends beyond that of Gautier *et al.* (2012) who also studied how missing data biases estimates of expected heterozygosity and F_{ST} . We analyze our simulations with additional commonly used summary statistics (π , θ , Tajima's D, F_{ST} , and

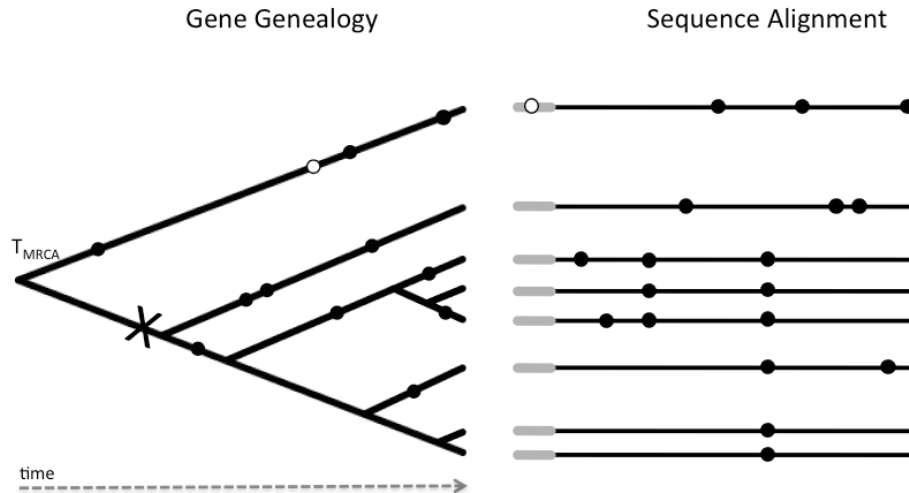


Figure 3.1 An example of a DNA sequence alignment (horizontal lines at right) along with the underlying genealogy of the locus (left). Dots represent segregating mutations in the sequence and where in the genealogy they occurred. The wider gray portion of the sequence alignment represents the recognition sequence and a white dot indicates a mutation in the recognition sequence. Haplotypes are not observed in a RADseq experiment if mutations occur within this region. In this example, the true time to most recent common ancestor (T_{MRCA}) of the sample is lost since a mutation occurred within the recognition sequence in the most divergent haplotype; the genealogy is thus truncated to point "X" and results in incomplete sampling that is biased against recovery of the most divergent haplotype(s).

the complete allele frequency spectrum) that are used to study demographic history and detect selection. We explore how RADseq affects genome-wide estimates of these statistics and how it impacts outlier analyses.

We show that RADseq nonrandomly subsamples the genome in two ways. First, within a locus, variants in a recognition sequence result in missing data and therefore truncate genealogies relative to the complete sample at these loci. This truncation results in underestimates of commonly used diversity statistics π and θ_w . Estimates of Tajima's D are also less accurate, but F_{ST} is relatively robust. Second, certain genealogies are more likely to result in missing haplotypes than others, such that RADseq samples a

biased subset of all genealogies. For example, loci with intermediate amounts of missing data are more polymorphic than the simulation average and more likely have genealogies with deeper divergences. We show with *in silico* digests of 102 completely sequenced *D. melanogaster* genomes that our coalescent simulations capture the major features of RADseq's genealogical bias. We discuss our findings and provide general guidelines for using RADseq for population genetic inference.

3.3 Results

RADseq Underestimates Polymorphism:

We generated simulated datasets for 100 haploid individuals and analyzed them mimicking a RADseq protocol (see methods for details). In comparing “true” values of summary statistics (π_t , θ_{wt} , D_t) with “estimated” values (π_e , θ_{we} , D_e , calculated from the data using only chromosomes that would be sampled by RADseq), it is apparent that the RADseq protocol results in systematic underestimation of polymorphism (Figure 3.2A). Not surprisingly, increasing amounts of missing data exacerbates this bias, and there is a strong positive correlation between chromosome sampling depth and estimates of polymorphism (Figure 3.2A). Fortunately, a majority of loci have all chromosomes sampled, especially for lower parameter values of θ in the simulations (Figure 3.2B). We found that π_t is more sensitive to missing data than θ_{wt} . Recombination decreases this sensitivity and brings values of both π_t and θ_{wt} closer to the simulation parameter value of θ (Figure S3.1). The difference between estimated and true values is greater for loci from simulated data sets with higher input values of θ (Figure S3.2), though increasing

the recombination rate tends to decrease this difference. This is because recombination decreases correlations between variants in the recognition sequence and those in the flanking sequence.

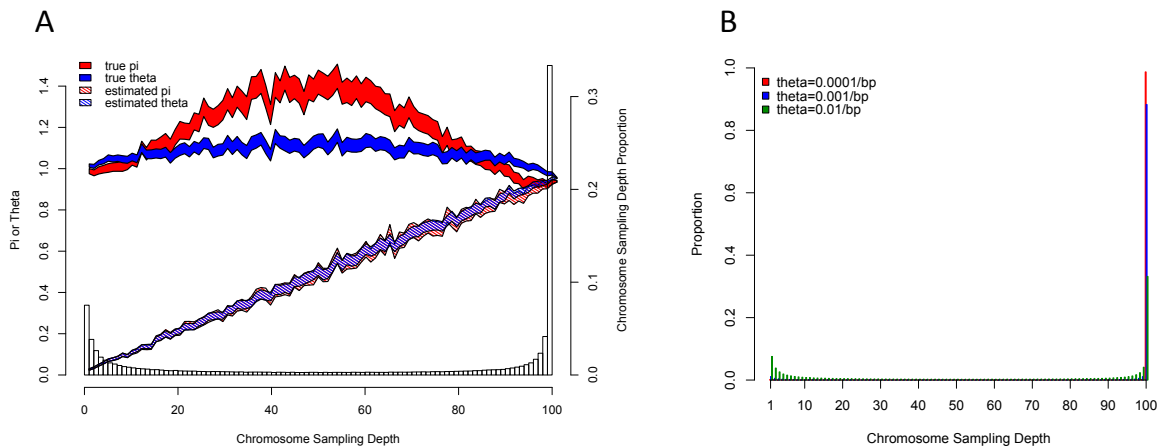


Figure 3.2 (A) True and estimated values of π (red) and θ_w (blue) from *in silico* RADseq as a function of chromosome sampling depth for $\theta = 0.01/\text{bp}$ without recombination. Here, the simulation average of θ is 1 per 100bp sequence read. Shaded regions show the 95% bootstrap percentile confidence intervals (1000 simulations) for the mean of true values of π (solid red) and θ_w (solid blue) and estimated values of π (shaded red) and θ_w (shaded blue) from *in silico* RADseq. "Chromosome sampling depth" refers to the number of chromosomes that are actually sampled (have intact restriction sites) in the *in silico* experiment, and "true" values are those calculated using the complete data for the same markers. The histograms in **A** (no recombination) and **B** (with recombination, $\rho = \theta$) show the proportion of each chromosome sampling depth in the data and indicate that most markers are highly sampled with these simulation parameters, especially for lower values of θ (**B**).

Simulations of the double digest RADseq protocol (Peterson *et al.* 2012) produce similar results. However, relative to the standard RADseq protocol, loci that have higher chromosome sampling depths are less frequent in the double-digest protocol (Figure S3.3) and have true and estimated values of π and θ_w that are even lower than

simulation averages (Figure S3.4). As in the standard RADseq protocol, a lower population mutation rate mitigates this effect (Figure S3.5).

Although by definition true and estimated summary statistics are identical when all chromosomes are sampled, loci with complete data still tend to have lower polymorphism than simulation averages (Table 3.1, Figure 3.2, Figure S3.2), particularly for the double-digest protocol (Table 3.1, Figure S3.4, S5). This bias is exacerbated in the double digest simulation when longer fragments were selected (350-450 instead of 150-250bps). Thus, while completely sampled loci are not biased individually, they will not capture the true genome-wide distribution of values. For simulations with higher polymorphism and no recombination, estimates of means and variances are further reduced below true simulation averages.

Table 3.1 Comparison of estimated values of summary statistics (θ_{we} or π_e) when all chromosomes are sampled to true simulation averages (θ_{wa} or π_a).

Protocol	θ per bp	Mean				Variance			
		Recombination		No Recombination		Recombination		No Recombination	
		θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a	θ_{we}/θ_{wa}	π_e/π_a
Standard	0.0001	0.994	0.995	0.991	0.990	0.994	0.996	0.990	0.990
	0.001	0.987	0.982	0.988	0.984	0.988	0.980	0.988	0.979
	0.01	0.956	0.933	0.940	0.909	0.941	0.901	0.904	0.837
Double Digest	0.0001	0.835	0.836	0.838	0.837	0.836	0.836	0.839	0.836
	0.001	0.858	0.851	0.829	0.823	0.857	0.841	0.830	0.815
	0.01	0.829	0.797	0.811	0.772	0.812	0.737	0.771	0.684

Notes: Results from two different simulation parameters of θ are shown. When recombination is present, $\rho = \theta$. Results are given for both the standard and double digest RADseq protocols.

Chromosome Sampling Depth is Correlated with Particular Genealogies:

Since the underlying genealogy of a sample of chromosomes at a locus provides information about its evolutionary history, we examined how genealogies vary with chromosome sampling depth using the allele frequency spectrum (AFS). The true AFS present in the sequence flanking a restriction site, conditioning on the chromosome sampling depth recovered in a RADseq experiment, shows that each respective sampling depth has a unique AFS and thus contains a nonrandom subset of the “true” genealogies (Figure 3.3A). Although recombination reduces this effect, a strong correlation between the frequencies of polymorphisms within a read and frequencies of the recognition sequence remains apparent in the AFS (Figure 3.3B). This is consistent with empirical observations of significant LD on the scale of a 100-bp sequencing read observed in many natural populations (*e.g.* Miyashita and Langley 1988, Hohenlohe *et al.* 2012, Langley *et al.* 2012, Pool *et al.* 2012). Lastly, in agreement with their higher values of π_t , loci with intermediate amounts of missing data in a RADseq experiment have genealogies with a greater time to common ancestry (T_{MRCA} 's, not shown) relative to the simulation average.

Non-equilibrium demography and population subdivision affects true and estimated summary statistics

Non-equilibrium demographic processes can affect the AFS. Therefore we asked what effect the introduction of a RADseq capture method can have on estimates of summary statistics for populations not at equilibrium. To this end, we simulated data

under two commonly used demographic models: a population bottleneck and exponential growth. For the standard RADseq protocol, a population bottleneck

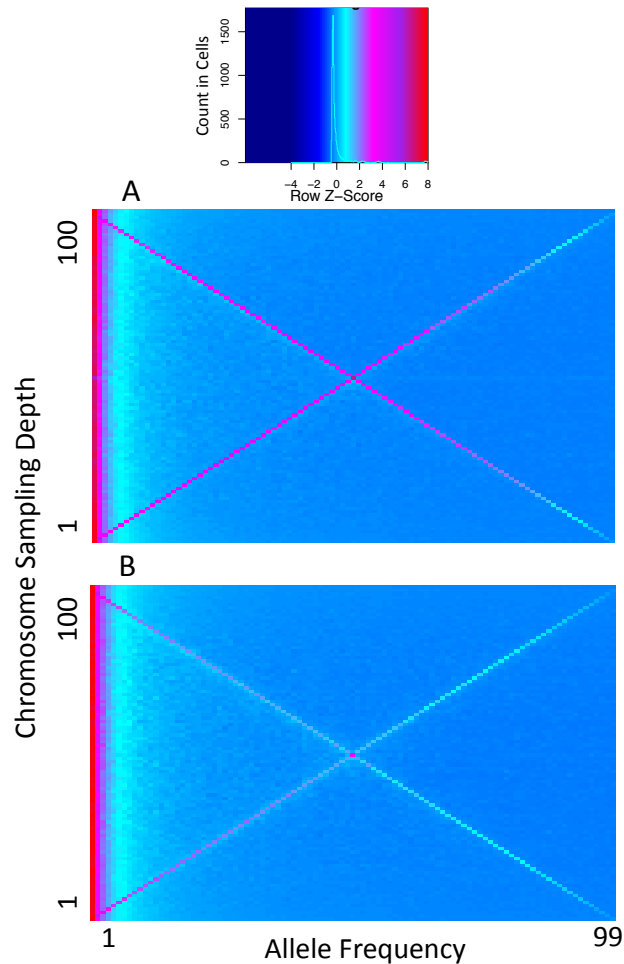


Figure 3.3 Density plot of true allele frequency spectra (AFS) for loci with different chromosome sampling depths **(A)** without and **(B)** with recombination. Each row represents the AFS for a particular chromosome sampling depth with the density of a particular allele frequency indicated as a heat map. The Z score fits a normal distribution to the entries in each row, and each cell is colored based on this fitting. This shows that loci with complete sampling (top of each graph) have an AFS characterized by abundant low-frequency polymorphisms, whereas loci with more missing data have greater proportions of intermediate frequency variants.

followed by growth slightly decreases the effect missing data has on estimating true summary statistic values by slightly increasing the correlation between estimated and

true values of θ_w and D (Figure 3.4). However, bottlenecks have little effect on estimates of π . Exponential population growth greatly reduces the sensitivity of π and θ_w to missing data and causes loci at all sampling depths to have estimated values of summary statistics that more closely resemble their true values (Figure 3.4). Both of these scenarios mitigate the effect missing data has on estimation of summary statistics because effective population sizes are reduced (relative to an equilibrium population of equal present size), particularly for the exponential growth model.

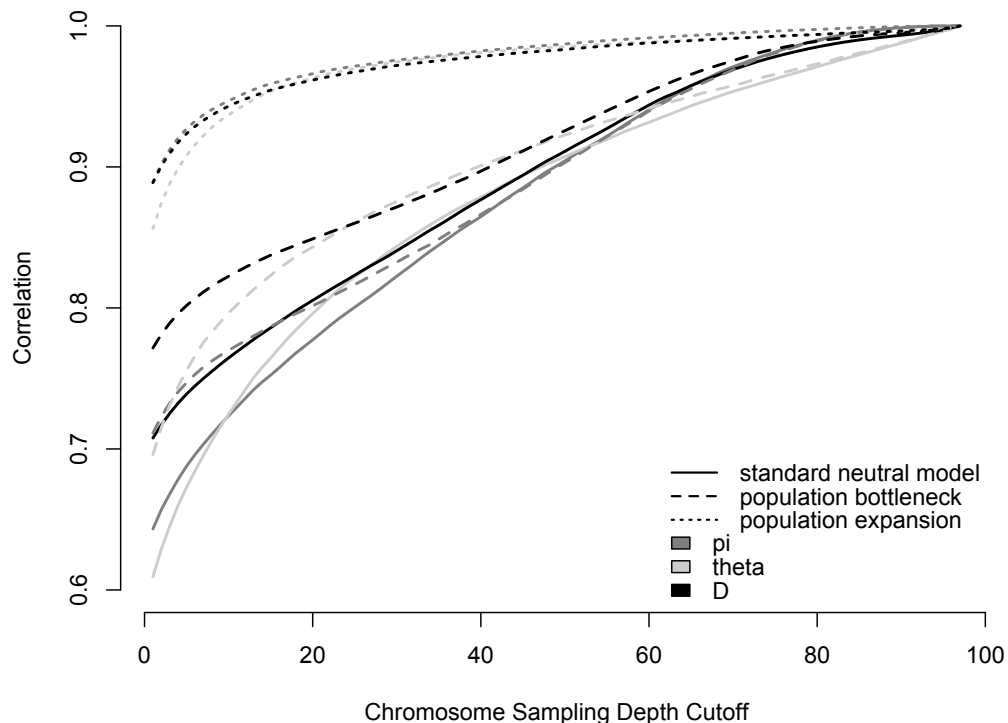


Figure 3.4 Correlations between true and estimated values of summary statistics are sensitive to non-equilibrium demography and chromosome sampling depth cutoffs. Values for π (gray), θ_w (light gray), and Tajima's D (black) under different demographic models are plotted (solid lines = standard neutral model, dashed lines = bottleneck, dotted lines = population expansion). The Y-axis is the correlation between true and estimated values for loci that satisfy a given chromosome sampling depth cutoff (*i.e.* with at least a minimum number of chromosomes with intact recognition sequences).

A common goal of population genetic analyses is to detect and study population structure and differentiation. To explore the effects of RADseq on a common metric of genetic differentiation, F_{ST} , we simulated two populations at demographic equilibrium that exchange migrants at a constant rate per generation (described in methods, performed only for the standard RADseq protocol). Unlike the results for metrics that summarize the AFS within a population, the distribution of estimated F_{ST} for loci with all chromosomes sampled is nearly identical to the true distribution (Figure S3.6) for effective migration rates of $Nm = 10$ and $Nm = 1$. This strong concordance breaks down when populations exchange one migrant every ten generations ($Nm = 0.1$; Figure S3.6C). Importantly, including loci with missing data biases the estimated F_{ST} distribution, since missing data tends to inflate estimates of F_{ST} (Figure 3.5). This is consistent with the results of Gautier *et al.* (2012) who considered biases of RADseq using a slightly different population subdivision demographic model.

F_{ST} , θ_w , π , and D outliers are sensitive to missing data

Although the levels and patterns of genetic variation in neutral loci that are linked to locally adapted alleles will depend on demographic and selective circumstances, it is interesting to consider outliers in the distributions of summary statistics as potential metrics for detecting positive selection and local adaptation. In particular, high F_{ST} may indicate that a locus is in linkage disequilibrium with locally adapted alleles. However, we show that missing data may inflate F_{ST} values, and rates of false positives quickly increase as the chromosome sampling depth cutoff decreases, especially when chromosome sample sizes among populations are allowed to vary as

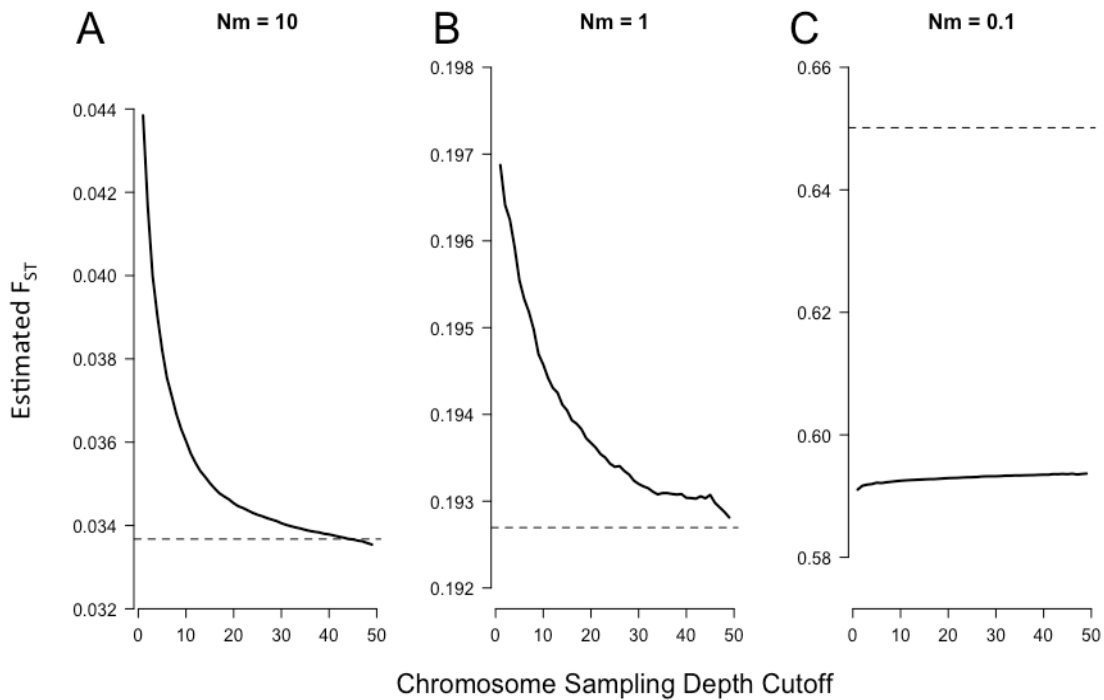


Figure 3.5 Estimated F_{ST} (solid line) as a function of chromosome sampling depth cutoff per population (each consisting of 50 chromosome total) for three different migration rates: $Nm=10$ (A), $Nm=1$ (B), and $Nm=0.1$ (C). The dashed line is the true simulation average. Here, we condition on sample sizes being the same in both populations to avoid inflated estimates of F_{ST} . Note that the Y-axes do not start at zero to more clearly illustrate differences between true and estimated values.

little as 20% (Figure 3.6). Thus, it may be wise to constrain analyses to loci with complete chromosome sampling, but of loci in the upper 5% tail of true F_{ST} distribution, only 13%, 11% and 5% have complete chromosome sampling in both populations for $Nm = 10, 1,$ and 0.1 respectively.

Within a population, genomic regions with low nucleotide diversity and left-skewed site frequency spectra may indicate the presence of a recent selective sweep via

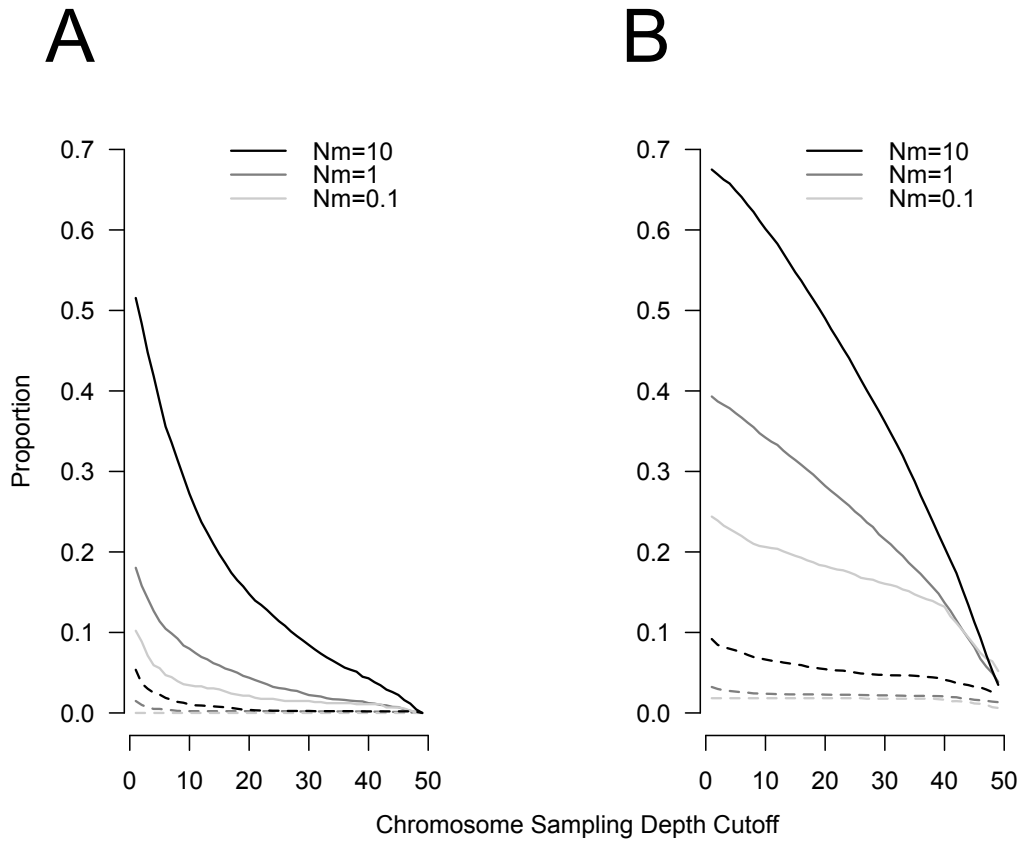


Figure 3.6 Proportion of estimated F_{ST} 5% outlier loci that are false positives (solid lines) or false negatives (dashed lines) relative to the true distribution for different chromosome sampling depth cutoffs (50 chromosomes per population in complete sampling). Three different rates of migration are represented: $Nm=0.1$ (light gray), $Nm=1$ (gray), and $Nm=10$ (black). If missing data is present, analyses were performed on loci for which chromosome sample sizes are exactly the same in both populations (**A**) or allowed to vary by 20% (**B**).

the hitchhiking effect (Maynard-Smith and Haigh 1974), or strong purifying selection (Charlesworth 1993). We explored the effect of missing data on outlier analyses involving the commonly used diversity statistics θ_w and π . Specifically, we examined the lower 5% tail of the distributions of these statistics to assess how missing data affects false positive and false negative rates. Using different sampling depth cutoffs, rates of false positives and false negatives increase with the inclusion of loci with missing data

for both the standard RADseq protocol (Figure 3.7) and the double digest protocol (Figure S3.7). Similar analyses with lower values of θ (0.001/bp and lower) were not possible since the 5% quantile of summary statistics contained the majority of loci due to low levels of polymorphism.

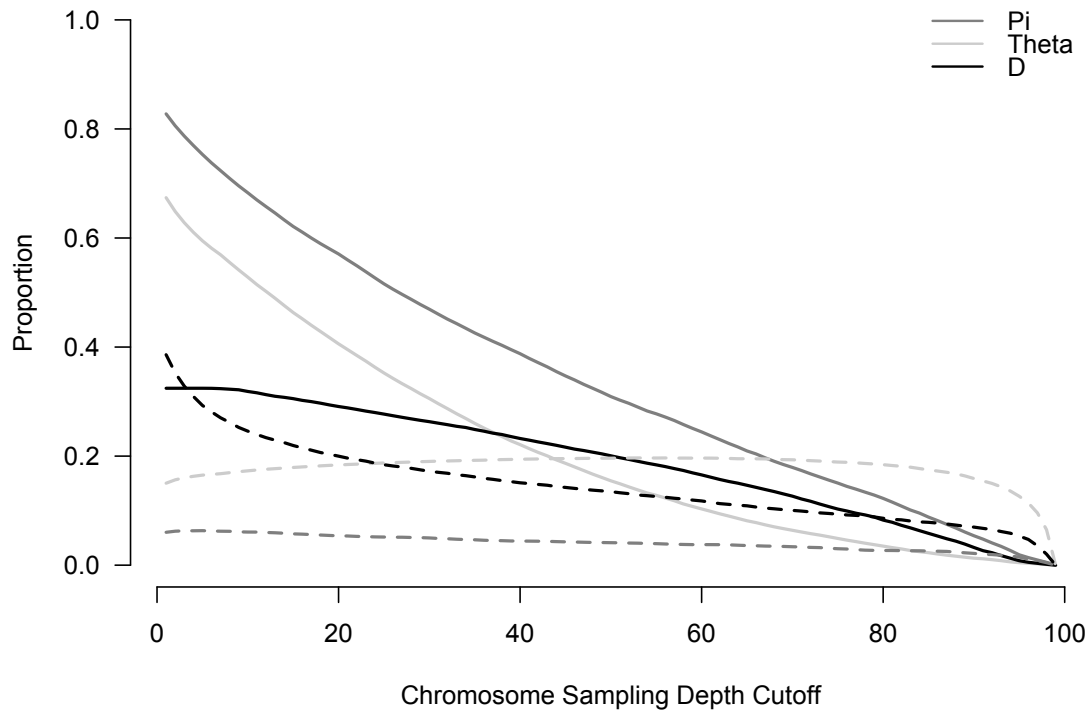


Figure 3.7 Proportion of estimated π (gray), θ_w (light gray), or D (black) outliers that are false positives (solid lines) or false negatives (dashed lines) for inclusion in the lower 5% tail for different chromosome sampling depth cutoffs.

Since loci with missing data have more false positives and negatives, a possible solution is to limit outlier analyses to loci with complete chromosome sampling. If outliers were evenly distributed across loci irrespective of missing data, 5% of loci in each sampling depth category would be outliers. However, in agreement with the results presented in Table 3.1, loci with complete sampling have slightly decreased

diversity and are more likely to fall within the lower 5% tail of the true distribution of π and θ_w and less likely to fall within the upper 5% tail (Table 3.2). Thus limiting analyses to completely sampled sites may inadvertently enrich for loci that have experienced recent positive selection or are highly constrained by strong purifying selection.

Table 3.2 Loci with complete sampling are more likely to fall within the lower 5% tail of the true distribution of π and θ_w .

Protocol	Tail	Ratio	
Standard	Lower tail	1.19	1.17
	Upper tail	0.76	0.74
Double Digest	Lower tail	1.76	1.95
	Upper tail	0.45	0.37

Notes: Shown are the ratios of the proportion of loci with complete chromosome sampling depth that are true outliers to the proportion of true outliers in the entire simulated data set.

In silico digestion of *Drosophila melanogaster* genomes

In order to test whether our framework captures the major biases associated with RADseq, we performed *in silico* digests of 102 recently released *Drosophila melanogaster* genome assemblies (Pool *et al.* 2012) using the standard RADseq protocol (Baird *et al.* 2008). The choice of restriction enzyme greatly affects which features of the genome are sampled (Figure 3.8A). The GC-rich recognition sequence of EagI samples exons more frequently than loci sampled at random and much more frequently than the AT-rich AseI, which disproportionately samples intronic and intergenic regions. EcoRI, which has an intermediate base composition, samples genomic regions at frequencies similar to their abundance in the genome. Likely owing to different levels of

polymorphism in different parts of the genome (e.g. due to stronger purifying selection in exonic versus intergenic sequences), choice of restriction enzyme results in different estimates of nucleotide diversity (Figure 3.8B).

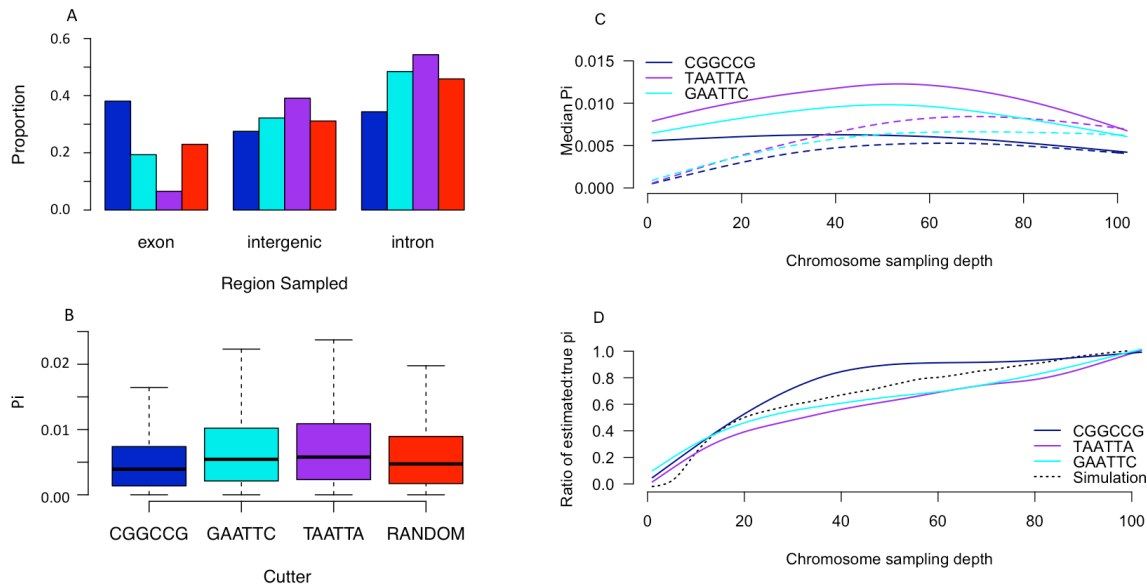


Figure 3.8 Results for the *in silico* digests 102 *D. melanogaster* genomes. **(A)** Proportion of sites located in distinct regions of the genome when *in silico* digests are performed with different enzyme recognition sequences. GC-rich recognition sequences sample more exons, whereas AT-rich recognition sequences sample comparatively more introns and intergenic regions. “Random” values are calculated from fragments selected at random throughout the genome. **(B)** Box plots of true π for regions sampled by enzymes with different recognition sequences. **(C)** The median true π (solid line) and estimated π (dashed line) as a function of chromosome sampling depth for three different recognition sequences. **(D)** Median of the ratio of estimated π to true π as a function of the number of sampled chromosomes. Dark blue, purple, and cyan lines represent the three different restriction enzymes used in the *in silico* digest of the *D. melanogaster* genomes, and the dotted black line is from simulations with $\rho=0.1/\text{bp}$, $\theta=0.01/\text{bp}$.

Similar to the simulation results, in regions of the genome where π_t is higher, it is more common for an intermediate number of chromosomes to be sampled (Figure 3.8C), which is consistent with the results of our simulations (above). This difference in π_t between loci with different chromosome sampling depths changes depending on the

recognition sequence of the restriction enzyme used and increases as more polymorphic regions of the *D. melanogaster* genome are sampled (i.e. with an enzyme with an AT-rich recognition sequence). Again, we observe similar patterns in our simulation framework (above), suggesting that our simulations accurately reflect much of the bias associated with RADseq.

To compare our framework to the *D. melanogaster* data, we ran simulations with an increased recombination rate ($\rho=0.1/\text{bp}$, $\theta=0.01/\text{bp}$); $\rho = 10*\theta$ has been used previously in demographic inference of this species (e.g. Thornton 2009). We then recorded the median of the ratio expected to true π (π_e/π_t) for each locus with a particular number of sampled chromosomes (Figure 3.8D). While our simulation appears to accurately model the majority of genealogical bias, we did not perfectly capture the dynamics of loci that have < 10 sampled chromosomes, perhaps as a result of violations of the infinite sites mutation-model (see Discussion).

3.4 Discussion

RADseq provides a simple and inexpensive means of collecting genome-wide sequence data from diverse non-model organisms (e.g. Emerson *et al.* 2010, Hohenlohe *et al.* 2011, Gompert *et al.* 2012, Parchman *et al.* 2012). This approach is increasing in popularity as a means of population genomic inference, but the effects of the ascertainment bias associated with polymorphism in recognition sites have not been extensively explored (but see Gautier *et al.* 2012). Biases can arise from mutations segregating in the recognition sequence such that haplotypes are nonrandomly sampled

for loci linked to these polymorphisms. Though it may seem comparatively rare for a mutation to occur within a recognition sequence, these variants are frequent enough to enable detailed population genetic analyses (e.g. Restriction Fragment Length Polymorphisms, Botstein *et al.* 1980). Consequently, a thorough examination of RADseq bias is essential for enabling detailed and accurate population genetic analyses based on this methodology.

Our coalescent simulations model two separate RADseq protocols (Baird *et al.* 2008, Peterson *et al.* 2012) and show that in both cases true and estimated values of π and θ_w vary with the amount of missing data that would occur in a RADseq experiment. Loci with higher π_t and θ_{wt} generally have fewer sampled chromosomes. These loci also have distinct frequency spectra and deeper divergence times. These patterns indicate that certain genealogies are particularly prone to missing data in RADseq experiments. Both π_e and θ_{we} and their correlations with true values decrease systematically as a function of the chromosome sampling depth, making loci with higher diversity the most strongly underestimated. Tajima's D is also sensitive to missing data. One potential solution might be to limit analysis to loci for which one can be certain of complete sampling. However, while this will reduce bias from sampling particular branches of the genealogy, it is important to remember that loci where RADseq samples all chromosomes are also a non-random subset of genome-wide π and θ_w distributions. Underestimated polymorphism has been previously observed in RADseq but was attributed to conservative SNP calling (Hohenlohe *et al.* 2010).

Loci with complete sampling for the double digest protocol have further decreased estimates of diversity (compared to the true genome-wide estimate) than the standard protocol because missing data may arise not only from mutations within recognition sequences but also from novel restriction sites that cause some haplotypes to be outside of the size-selection range. Indeed, this problem is exacerbated for the simulation in which longer fragments were selected since there is a larger region within which novel restriction sites may occur. In reality, segregating insertions or deletions may also contribute to missing data by changing the length of sequences between cut sites to outside the range of size selection, but this additional source of bias was not modeled in this study.

Importantly, inclusion of loci with incomplete sampling may actually invert relative estimates of π and θ , such that loci that are in reality more diverse will have lower estimates for these parameters than loci with complete sampling that are taken from less diverse regions. In practice, for a particular locus with incomplete chromosome sampling depth, it may not be feasible to determine if chromosomes were not sequenced from polymorphism in the restriction site or from low sequencing depth.

The correlations between estimated and true values of summary statistics are also sensitive to non-equilibrium demography. Both population bottlenecks and expansions increase correlations between true and estimated values. The greater correlations presumably occur because both demographic scenarios decrease the effective population size and therefore reduce genetic diversity, so fewer loci have missing data and thus inaccurate estimates of summary statistics. Since natural

populations likely have complex evolutionary histories, summary statistics may be affected by a combination of multiple demographic events in addition to the population mutation and recombination rates. Having estimates of these parameters *a priori* for a given study system help predict how frequently loci will contain missing data and how sensitive estimated values of summary statistics are to missing data.

We also explored the ability of RADseq datasets to detect population structure and differentiation by calculating F_{ST} between two populations at demographic equilibrium that exchange migrants at a constant rate per generation.

The distribution of estimated F_{ST} values for loci with all chromosomes sampled is very similar to the true distribution. The relative robustness of F_{ST} to the RADseq protocol suggests that this methodology is perhaps well suited to estimating rates of migration between populations.

Since outliers in the distributions of summary statistics are frequently used as metrics for detecting selection, we explored the sensitivity of F_{ST} , θ_w , and π outliers to missing data. We find that rates of false positives and false negatives increase for F_{ST} , θ_w , and π as chromosome sampling depth decreases, since missing data biases estimates. This has important implications for outlier analyses as tests for selection or local differentiation and indicates that empirical outliers obtained from RADseq experiments where complete chromosome sampling cannot be established with certainty should be interpreted with caution. Again, a potential solution is to restrict analyses to loci with complete chromosome sampling depth, but with this correction a vast majority of true F_{ST} outliers would be missed since many true outliers have incomplete sampling.

Moreover, since many investigators sequence diploid organisms, it may be difficult to quantify the amount of missing data and the sample size variation among populations, both of which would inflate estimated F_{ST} values.

Our *in silico* RADseq analyses of 102 *D. melanogaster* genomes were largely consistent with the results of our simulations, in that polymorphism is underestimated, especially for more diverse genomic regions. Although undoubtedly the populations from which these samples are derived are experiencing non-equilibrium selective and demographic processes that we did not model (Pool *et al.* 2012; Corbett-Detig and Hartl 2012), the overall congruence of our simulations with the *Drosophila* data suggest our basic simulation framework captures the major biases that affect RADseq. One possible explanation of the poor fit of our model at low chromosome sampling depths is that the real data includes violations of the infinite-sites mutation model, such that mutations recur within nascent recognition sequences on different haplotypes. This would effectively inflate diversity relative to infinite-sites assumptions of the coalescent simulations. Nonetheless it is clear that even though our simulations are relatively simplistic, we have identified a major potential bias inherent to the RADseq methodology.

The nucleotide composition of the recognition sequence affects which features of the genome are sampled and this suggests an appealing means of tuning RADseq for the specific goals of each respective study. For example, for the purpose of SNP discovery, one may prefer to select an enzyme with an AT-rich recognition sequence; conversely if the goal is to study genetic differentiation between divergent populations,

GC-rich recognition sequences will generally access a higher proportion of conserved regions of the genome and may increase the overlap in sampled loci between populations. However, such choices must still be considered with appropriate caveats, for instance, in species with DNA-methylation, CG sites are known to mutate at significantly higher rates than the genomic average (Cooper *et al.* 1995). In this case, using an enzyme which cuts sequences that contain these motifs may increase the amount of missing data, and violate a tacit assumption of our model that the per-site mutation rate in the recognition sequence is identical to that in the sequenced read. We thus emphasize that because each restriction enzyme will access different genomic regions, which may not have identical allele frequency spectra, the choice of restriction enzyme will also affect population genetic inferences.

Our results are also consistent with those of Gautier *et al.* (2012), but our interpretation of how RADseq affects estimates of diversity is different. In their study, Gautier *et al.* state that RADseq results in overestimates of heterozygosity because they only consider segregating sites that are observed after the *in silico* digest of simulated fragments. This effect occurs because mutations in linked recognition sequences more likely arise on the major allele haplotype, thus inflating minor allele frequencies and estimates of heterozygosity. Here, we examine the effect that RADseq has on commonly used diversity statistics per site and thus account for both observed and unobserved segregating sites. Because variants in a recognition sequence truncate genealogies relative to the complete sample at these loci, some true variants are not observed, overall resulting in underestimates of π and θ_w .

RADseq is an important emerging methodology, and is likely to see increased use; it is therefore important to identify biases and where possible to develop a means of accounting for them. In general, the assumption of identical per-site mutation rates in the cut-site and sequenced read is likely to be reasonable. Given this assumption, it may be possible to account for the genealogical bias of RADseq using our (or a similar) coalescent simulation modification framework. That is, standard coalescent simulations can be performed, and the resulting sequence digested and analyzed as we describe. If the resulting biased summary statistics are then compared with empirically-obtained RADseq summary statistics (e.g. using approximate Bayesian computation software such as ABCreg; Thornton 2009), it may be possible both to directly account for this source of bias in population genetic analyses and to recover unbiased estimates of the true distributions of relevant summary statistics.

This study can serve as a useful guide for investigators using RADseq for population-genomic analyses. From our simulations and empirical *in silico* digests, loci with missing data give inaccurate estimates of summary statistics and may increase the rate of false positives in outlier analyses. Thus identifying and pruning loci with incomplete sampling will be important in any RADseq experiment aimed at accurately estimating commonly used summary statistics. Since RADseq will generally produce thousands or tens of thousands of markers throughout the genome, pruned datasets that retain only loci with complete sampling will still be substantial (Figure 3.2). However, if RADseq is to be used for demographic inference, it remains important to recognize that ignoring loci with missing data, which are enriched for particular

genealogical structures, will also affect estimation of evolutionary parameters and may not accurately represent a "genome average" value. If many loci have high sequencing depths such that sites with missing data can easily be detected by differences in coverage, RADseq provides a powerful way to estimate genome-wide divergence among populations to describe biogeographic patterns. Thus, though our findings urge caution, with careful consideration of experimental design, data use, and interpretation, RADseq will likely continue to develop as a powerful technique for addressing questions in evolutionary biology.

3.5 Methods

Coalescent Simulations

We used Hudson's *ms* (Hudson 2002) to simulate 10kb DNA fragments for 100 haploid individuals with different population mutation and recombination rates (i.e. $\theta=4N_e\mu$ and $\rho=4N_er$, with μ and r being the mutation and recombination rates, respectively). Three values of θ were used for simulations (0.0001, 0.001, or 0.01 per bp), with either $\rho = 0$ and $\theta = \rho$. We first simulated a single population at demographic equilibrium under each set of parameters above. To explore the effect of demographic history on RADseq, we modeled a bottleneck in which the population shrunk to 25% of the original size for $0.1N_e$ generations, $0.1 N_e$ generations before present, after which it recovered to its original size. We also modeled an exponential growth scenario in which the population grows exponentially from 10% of its present day size over $0.2 N_e$ generations. Simulations were repeated 100,000 times for each parameter set.

To explore the ability of RADseq to effectively detect population subdivision using a common metric of genetic differentiation (F_{ST}) we simulated two populations at demographic equilibrium that exchange migrants at a constant rate per generation. We simulated varying levels of population structure with 50 haploid individuals, or chromosomes, per population with migration rates (Nm) of 10, 1, or 0.1, and $\theta = \rho = 0.01/\text{bp}$.

In silico RADseq experiment

Using custom Perl scripts, we performed an *in silico* digest by searching these simulated fragments for a specific recognition sequence. Since Hudson's ms (Hudson 2002) models DNA sequences with zeroes and ones, we used recognition sequences consisting of 12 zeroes and ones. Assuming equal nucleotide base composition, this motif occurs as frequently as a 6-base DNA restriction enzyme site (about 2.8 times per 10kb). Fragments that contained no recognition sequences were not analyzed. After the *in silico* digest, we analyzed the sequence 100 bp to the right of each recognition sequence to model the standard RADseq protocol (Baird *et al.* 2008). This length was chosen because it is currently a commonly used read length in Illumina sequencing. We compared "true" summary statistic values (before digest) with "estimated" ones (after digest, using only chromosomes that would have been recovered in a RADseq experiment). Here we focus exclusively on biases induced by restriction site polymorphism, which ignores other potential sources of bias arising from sequencing and alignment, such as other sources of nonrandom sampling of haplotypes, sequencing

errors, and reference bias (reviewed in Rokas and Abbot 2009, Pool *et al.* 2010). These are expected to be general issues for most or all resequencing projects and are not addressed here.

Our simulation framework models the biased recovery of haplotypes in the RADseq protocol due to restriction site polymorphism. At a particular locus, a chromosome may not be sampled for two reasons: (1) a cut site, which is polymorphic in the population, is not present on that chromosome or (2) a recognition sequence is present within 100 bp to the right of another recognition sequence, resulting in a fragment that is removed in the size-selection step and thus not sampled. As a result, the number of chromosomes sampled to the right of a particular recognition sequence, hereafter referred to as “chromosome sampling depth,” varies among loci and may be less than the total 100 simulated DNA sequences. To demonstrate the effect of missing data due to the RADseq protocol, in the results below, we either binned loci by chromosome sampling depth or imposed cutoffs such that only loci with at least a minimum number of sampled chromosomes are analyzed.

After the *in silico* digest of each fragment, we calculated the allele-frequency spectrum (AFS) for the 100 bp to the right of each recognition sequence using all simulated chromosomes (the “true” AFS). We also calculated the AFS using only chromosomes that have the correct recognition sequence and would therefore be sampled by a RADseq protocol (the “estimated” AFS). We then used these to calculate typical summaries of the data such as average number of pairwise differences (π , Tajima 1983), Watterson’s θ (θ_w , Watterson 1975), Tajima’s D (Tajima 1989), and F_{ST} (Weir and

Cockerham 1984). As above, true values for these summary statistics (π_t, θ_{wt}, D_t) were calculated using all chromosomes at the locus, and estimated values (π_e, θ_{we}, D_e) were calculated only for chromosomes that would be sampled in a RADseq experiment.

For the simulations with population subdivision, for any one locus, chromosomes are sampled according to criteria described above to mimic the RADseq protocol. F_{ST} can be inflated when one population has greater sampling depth, which may occur if a recognition-site mutation rises to a higher frequency in one population than the other, and this may confound inferences based on F_{ST} . Thus, for our analyses, we condition on sample sizes being the same for both populations to avoid these artifacts that inflate estimates of F_{ST} .

Double digest RADseq

We modified our framework to explore how summary statistics are affected by another RADseq protocol recently developed by Peterson *et al.* (2012), which relies on double digests. Briefly, this method requires first digesting the genome with two restriction enzymes and then selecting those fragments that fall within a defined size interval. We mimicked this process by sampling only fragments that were flanked by the same two complete recognition sequences of 6 zeroes and ones that were either within 150-250 or 350-450bps of each other. The length of restriction sequence was chosen to make the overall size of the mutational target associated with each chromosome at a locus the same as the standard RADseq protocol mentioned above. We further required that no additional cut sites be present in between that cause the fragment to be shorter

than the selected size. We then sampled the 100bp immediately adjacent to the left recognition sequence and analyzed this as described above for the standard RADseq method. Although a double digest would normally involve sampling fragments flanked by two distinct recognition sequences, we only use a single recognition sequence for this *in silico* digest (repeated twice). However, since the sampling properties are the same for any arbitrary sequence of a specified length, we still refer to this modified framework as a “double digest.” All analyses presented for the double digest protocol used the size selection with shorter fragments (150-250bps) unless otherwise stated.

Empirical confirmation with *Drosophila melanogaster*

To confirm whether the predictions of our simulation framework reflect biases that could arise in an actual RADseq experiment, we performed *in silico* digests of 102 fully-sequenced hemizygous (i.e. only one chromosome is sampled) *D. melanogaster* individuals (Pool *et al.* 2012). We acquired genome assemblies in fastq format from www.dpgp.org and subsequently translated these to fasta format requiring a minimum nominal base quality of 30. We masked regions of putative identity-by-descent, described in Pool *et al.* (2012), using the conversion/masking script provided by www.dpgp.org. We selected three different recognition sequences representing distinct base compositions, AseI (TAATTA), EcoRI (GAATTC), and EagI (GCCGGC), to digest the assemblies *in silico*. Digests were performed as described above for coalescent simulations mimicking the standard RADseq protocol (Baird *et al.* 2008). In brief, we digested each genome with a specific recognition sequence and considered that

chromosome to be sampled if there was not an additional recognition sequence within 100 bp to the right. In cases where there was missing data in the recognition sequence (i.e. due to masked low-quality base calls, and not due to high-quality variants in the recognition site), we excluded those chromosomes from calculations of both true and estimated π . Each recognition site with at least one observed chromosome was considered for downstream analysis if at least 100 of the chromosomes in the original genome assemblies were covered by quality 30 or greater sequence through the entire region spanned by the recognition sequence.

CHAPTER 4 - SINGLE GEOGRAPHIC ORIGIN OF AUTOTETRAPLOID *ARABIDOPSIS ARENOSA* FOLLOWED BY INTERPLOIDY ADMIXTURE

4.1 Abstract

Whole-genome duplication (WGD), which leads to polyploidy, has been implicated in speciation and biological novelty. In plants, many species exhibit ploidy variation, which is likely representative of an early stage in the evolution of new polyploid lineages. To understand the evolution of such multiploidy systems, we must address questions such as whether polyploid lineage(s) had a single or multiple origins, whether admixture occurs between ploidies, and the timescale over which ploidy variation affects the evolution of populations. Here we analyze three genomic datasets using nonparametric and parametric analyses, including coalescent-based methods, to study the evolutionary history of *Arabidopsis arenosa*, a new model system for understanding the molecular basis of autopolyploid evolution. Autotetraploid *A. arenosa* populations are widely distributed across much of Northern and Central Europe, while diploids occur only in Eastern Europe and along the southern Baltic coast; the two ploidies overlap in the Carpathian Mountains. We find that the widespread and variable autotetraploid likely arose from a single ancestral population ~11,000-30,000 generations ago in the Northern Carpathians, where its closest extant diploid relatives are found today. Afterward, the tetraploid population split into at least four major

lineages that colonized much of Europe. Reconstructions of population history suggest substantial levels of interploidy admixture occurred in both directions, but only among geographically proximal populations. We find two cases in which gene flow was likely followed by selection on an introgressed locus, suggesting persistent interploidy gene flow has a local influence on patterns of genetic variation in *A. arenosa*.

4.2 Introduction

Whole-genome duplication (WGD) has occurred in many organisms across eukaryotic kingdoms and has profoundly shaped genome evolution (Kellis et al. 2004; Dehal and Boore 2005; Jiao et al. 2011). These large-scale genomic events are implicated in increased genomic complexity and are associated with adaptive radiations of major lineages throughout the tree of life (Dehal and Boore 2005; Jiao et al. 2011). WGD is particularly frequent in plants; ancient WGD events are estimated to have occurred in 30-100% of angiosperm lineages (Stebbins 1950; Grant 1981; Masterson 1994; Cui et al. 2006). WGD is also implicated in speciation as it is one of the few processes that may instantaneously give rise to reproductive isolation due to the lower success of interploidy crosses, and may thus serve as a mechanism of sympatric speciation (Wood et al. 2009).

Many extant plant species are known to have multiple ploidy levels (see for review Ramsey and Schemske 1998; Soltis et al. 2010), showing that polyploidy remains an active force in plant evolution. It has also been suggested that the establishment of new autotetraploid populations may be affected by interploidy gene flow (Ramsey and

Schemske 1998). Thus for species with ploidy variation, it is important to understand the dynamics of polyploid formation and establishment, interploidy gene flow, and the evolutionary timescale over which multiple ploidies coexist and shape species-wide patterns of genetic variation. New genomic approaches and methods to reconstruct polyploid history (Arnold et al. 2012) hold promise to enable new detailed understanding of evolutionary dynamics in multiploidy systems.

There are two major classes of polyploids: autopolyploids, which form from within-species WGD and generally randomly segregate homologs, and allopolyploids, which have a hybrid origin and usually diploid-like inheritance (Ramsey and Schemske 1998; Parisod et al. 2010; Bomblies and Madlung 2014). Of these, autopolyploids have received less attention in the evolutionary genetics literature, though in several species previous studies documented that autopolyploids arose multiple times and/or from more than one individual (Soltis et al. 1989; Brochmann and Elven 1992; Van Dijk and Bakx-Schotman 1997; Seagraves et al. 1999; Yamane et al. 2003; Yang et al. 2006; Luo et al. 2014). There is now good evidence that autopolyploids are more common than was previously appreciated (Soltis et al. 2010). Their often tetrasomic mating system (arising from random segregation of all four homologs) presents an intriguing problem and has important implications for autopolyploid population genetics.

Here, we reconstruct the evolutionary history of autotetraploid *Arabidopsis arenosa*. This species is newly being developed as a model for understanding the molecular basis of autopolyploid evolution (Hollister et al. 2012; Yant et al. 2013); for this it is particularly important to understand the evolutionary history of the polyploid

lineage and the degree of interploidy admixture in more detail. This species is an obligate outcrosser closely related to *A. lyrata* and the widely used model *A. thaliana* (Al-Shehbaz and O’Kane 2002), both of which have sequenced and annotated genomes (The Arabidopsis Initiative 2000; Hu et al. 2011). Autotetraploid *A. arenosa* has high genetic diversity (Hollister et al. 2012; Schmickl et al. 2012; Hohmann et al. 2014), and populations are widely distributed through much of Central and Northern Europe, while diploids are found in the Balkans, Eastern Europe and along the southern Baltic Coast in Poland; the two types overlap in the Carpathian Mountains (Schmickl et al. 2012; Kolar et al. 2015). Previous work suggested that the tetraploids experience gene flow from diploids, but not the reverse (Jørgensen et al. 2011). However, the age of the polyploid lineage is not known, nor is it known whether it arose once or has multiple origins.

We use both parametric and nonparametric analyses of three distinct genomic sequencing datasets to infer the number and timing of autotetraploid origins in *A. arenosa*, estimate the geographic location of origin(s), and quantify the extent and direction(s) of interploidy gene flow. We find that the populations we sampled of the widespread autotetraploid *A. arenosa* likely arose from a single ancestral population, probably in the Northern Carpathians approximately ~11,000-30,000 generations ago. Thereafter, the tetraploid lineage split into at least three major lineages that colonized much of Europe. We find evidence that geographically proximal diploid and tetraploid populations experienced ancient bidirectional interploidy admixture, and in rare cases, introgressed haplotypes may have come under selection in the recipient population. The

methods we use and develop will be applicable in a wide range of species with ploidy variation.

4.3 Results

Genomic Data and assessment of ploidy and inheritance

We collected seed samples from 6 diploid and 14 tetraploid *A. arenosa* populations from across much of its European range (Figure 4.1) and used three different genome datasets (summarized in Table S4.1) to analyze their population history. First, we generated a Restriction-Associated DNA sequencing (RADseq; Peterson et al. 2012) dataset for 358 plants. We complemented this with two additional genome datasets with overlapping population samples: (1) whole-genome sequencing of population pools (PoolSeq) which sampled 89 of these plants (Wright et al. 2014), and (2) a previously generated whole-genome sequencing dataset (IndSeq) that sampled a subset of 16 of these plants (Yant et al. 2013). The IndSeq dataset, though it samples fewer plants, serves as a standard, since it does not suffer from ascertainment biases potentially present in RADseq and PoolSeq data (Cutler and Jensen 2010; Arnold et al. 2013; Gautier et al. 2013) and has higher sequencing depths per chromosome, allowing more accurate single-nucleotide polymorphism (SNP) calls. We determined genotypes using the GATK (McKenna et al. 2010), which accommodates diploid and tetraploid samples, and only considered biallelic sites with sequencing depth cutoffs of 8 or higher per individual. All three datasets produce similar estimates of allele frequencies, though estimates of genetic diversity differ; relative to the IndSeq dataset, RADseq

underestimates diversity, while PoolSeq overestimates it (Tables S4.2 and S4.3, Supplementary Methods). Both of these follow expected trends based on previously identified biases in such datasets (Cutler and Jensen 2010; Arnold et al. 2013).

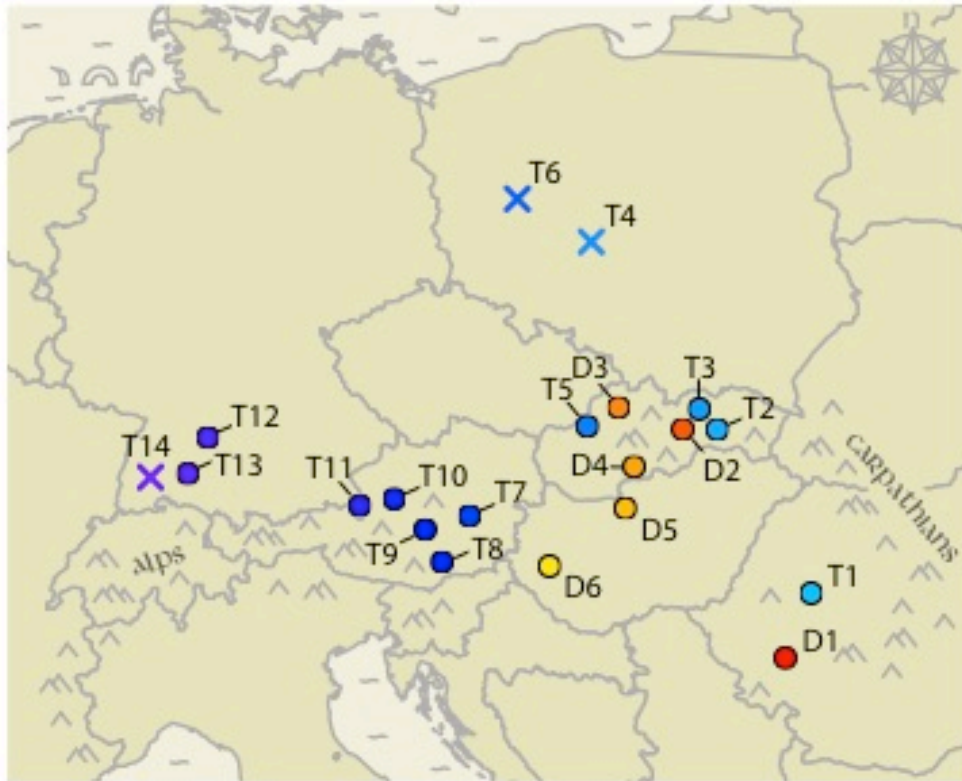


Figure 4.1 Map of central Europe with sample collection sites. Diploids are labeled D1 through D6 and colored in hues of red to yellow. Tetraploids are labeled T1 through T14 and colored in hues of light blue to purple. Both number and coloring schemes correspond to a longitudinal gradient from East to West. Tetraploids collected from railway habitats are labeled with X's. Our sampling includes collection sites within Romania (D1, T1), Hungary (D5, D6), Slovakia (T2, T3, T5, D2, D3), Poland (T4, T6), Austria (T7-T10), and Germany (T11-T14).

We previously assessed ploidy for several of these *A. arenosa* populations by flow cytometry (Hollister et al. 2012), but not all individuals were sampled. Therefore, we assessed the ploidy of each individual sampled here bioinformatically using the

RADseq dataset and the simple logic that, at polymorphic sites, raw SNP count data for diploid and autotetraploid samples should be different. Specifically, the distribution of non-reference base counts for all polymorphic sites within a single diploid individual should resemble a binomial distribution with a mean of 0.5, while the same distribution for an autotetraploid should be trimodal, due to an amalgamation of three distinct binomial distributions with means of 0.25, 0.5, and 0.75, as autotetraploids have three types of heterozygotes. Using the RADseq dataset, limiting ourselves to filtered heterozygous sites within an individual that have sequencing depths of at least 30, counting the number of non-reference base calls and comparing them to simulated expected distributions using a *G* statistic (see Materials and Methods), we easily discriminated between samples of different ploidy (Figure S4.1, Table S4.4). With the exception of one putatively tetraploid individual found in an otherwise diploid population (and excluded from subsequent demographic analyses), all samples from a collection site were of the same ploidy. This is consistent with previous findings in *A. arenosa* (Schmickl et al. 2012, Kolar et al. 2015). The one tetraploid we found in a diploid population is a potentially spontaneous neotetraploid in population D5 (as opposed to a migrant from a tetraploid population), as it is genetically similar to diploids sampled from the same population (Figure S4.2). Although two additional diploid samples did not have simple binomial non-reference base count distributions, they were likely diploid (see Materials and Methods).

An important assumption when modeling autotetraploid data using the coalescent is that chromosomes are exchangeable (Arnold et al. 2012). This assumption

would be violated if there were restricted recombination to certain chromosome pairs, as would occur if chromosomes have pairing partner preferences. Structure between duplicated chromosome sets due to pairing preferences should create an enrichment of alleles at 50% frequency relative to an allele frequency spectrum (AFS) of a population with unstructured chromosomes (Hollister et al. 2012). Neither the IndSeq nor the RADSeq datasets display an excess of alleles at 50% frequency (Figure S4.3A), and tetraploid genotype proportions closely resemble those expected under Hardy-Weinberg equilibrium for tetrasomic inheritance (Figure S4.3B). These data confirm that the assumption of random chromosome assortment is not violated and that *A. arenosa* populations retain fully tetrasomic inheritance (as previously shown for a smaller set of samples in Hollister et al. (2012)).

Principal component analysis suggests a single geographic origin of the tetraploids

To study the genetic relatedness of sampled *A. arenosa* populations, we used principal component analysis (PCA), a non-parametric approach that allows for multiple ploidies. For this analysis we used the RADseq dataset, which included the largest number of individuals sampled. Since PCA is sensitive to sample sizes (Novembre and Stephens 2008), we used a subsampling approach to control for the disparity in diploid and tetraploid representation within the RADseq dataset. A PCA of only diploids (10 individuals per population, 11,758 SNPs) shows there are two distinct groups within our diploid samples (Figure 4.2A), one found in the Carpathian Mountains in Romania and Slovakia (D1-D3) and another in the biogeographically distinct Pannonian Basin in

Southern Slovakia and Central Hungary (D4-D6). To study the relationship of the tetraploids with these diploid gene pools, we added a subsample of 30 tetraploids from across the *A. arenosa* range (6,117 SNPs, Figure 4.2B) and found that all tetraploids group with Northern Carpathian diploids (D2 and D3).

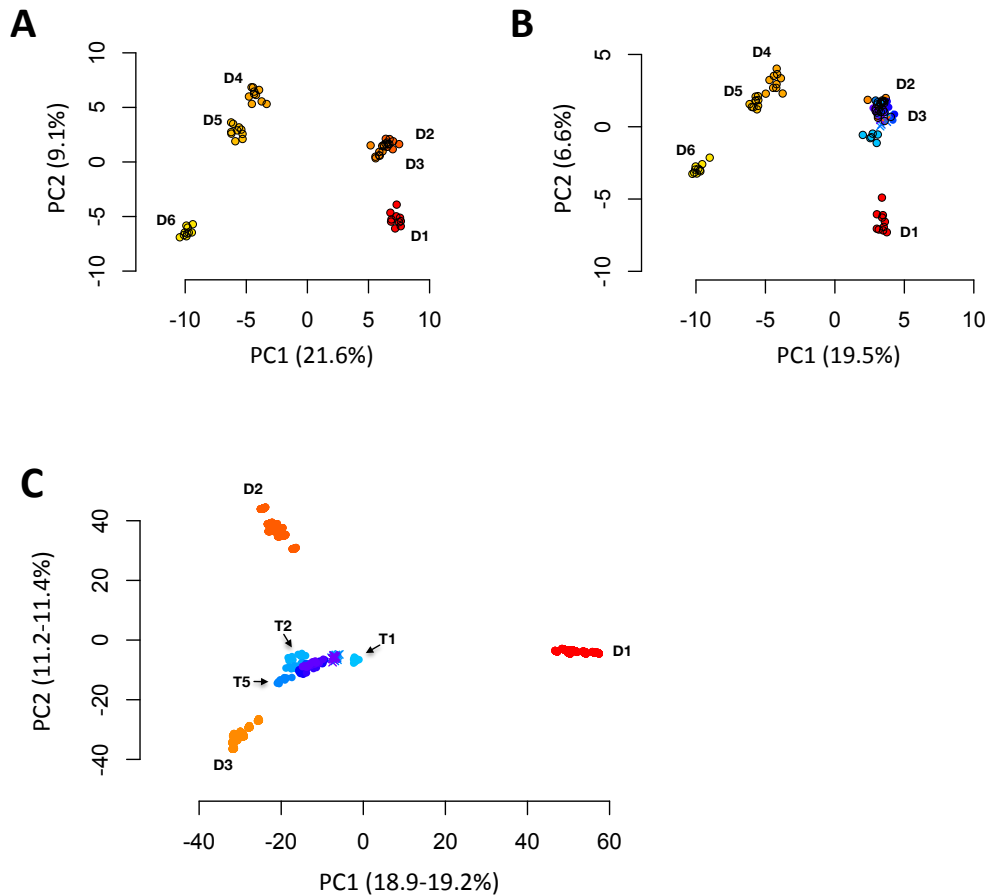


Figure 4.2 PCA of diploid and tetraploid *A. arenosa*. **(A)** PCA of all diploid populations separates groups D1-D3 and D4-D6 on PC1. PC2 primarily corresponds with latitude. **(B)** PCA as in **(A)**, but including a subsample of tetraploids. Axes in **(A)** and **(B)** are labeled with the proportion of the total variance explained by that principal component. **(C)** PCA of single tetraploid individuals with diploid populations D1-D3. A separate PCA was performed for each tetraploid, 10 individuals from each of 14 tetraploid populations, and superimposed onto the same PC axes. Populations T1, T2, and T5, which have admixed with populations D1, D2, and D3, respectively, are labeled to show how they radiate out from the tetraploid group towards the diploid population with which they have exchanged alleles. Axes are labeled with the range of the percent of the total variance explained by that principal component across the 140 PCAs represented.

These PCA results contain two important pieces of information: (1) the tetraploids, despite their high genetic diversity, are all comparatively closely related, and (2) within our sampling the tetraploids are most closely related to Northern Carpathian diploids. To better visualize the relationship between the tetraploids and the closest diploid relatives we sampled, we performed another PCA using just Carpathian diploids (D1-D3). To circumvent the large differences in sample sizes between diploids and tetraploids, we used 30 diploids (10 per population) and a single tetraploid individual to elucidate how each tetraploid relates to the principal component space of diploid genetic variation. We repeated this analysis 140 times, sampling 10 individuals from each of 14 tetraploid populations, and superimposed the results onto the same pair of principal component axes (415,718 SNPs, Figure 4.2C). The tetraploids cluster together between the three diploid populations, suggesting there is a single tetraploid gene pool within our sample (which was sampled broadly across the tetraploid *A. arenosa* range). As before, the tetraploid gene pool is more closely related to the Northern Carpathian diploids than the Southern Carpathian diploid population we sampled. Three tetraploid populations radiate from the central cluster towards each of the three diploid populations, suggesting there may have been admixture. In each case, the populations where admixture is suggested are the most geographically proximal to each respective diploid. We further explore the possibility of admixture below using coalescent analyses. A single PCA with these 30 Carpathian diploids and 30 tetraploids from across the range yields similar results (Figure S4.4).

Demographic modeling affirms a single geographic tetraploid origin with ancient interploidy admixture

The above analyses suggested autotetraploid populations arose from a single ancestral population, with potential admixture among geographically proximal diploid and tetraploid populations. To verify this result and quantify admixture proportions, we explicitly modeled the history of these populations using the coalescent. We modeled groups of three populations in each case, in an approach we call “trio analyses.” In each analysis, we constructed models of one diploid and two tetraploid populations, and tested which of two models better fits observed polymorphism data: (1) a single tetraploid origin allowing for subsequent interploidy admixture between geographically proximal populations (model A in Figure 4.3) or (2) two independent tetraploid origins with potential admixture between tetraploid populations (model B in Figure 4.3). While it would be possible to generate increasingly complex models, limiting analyses to trios avoids the problem of excess empty categories in the multidimensional allele frequency spectrum (AFS) that would occur if more populations were included in each analysis.

For each trio analysis, we always included one tetraploid that is geographically distant from diploids and displayed no evidence of interploidy admixture according to simple demographic models (i.e. T7 and T13, Figure S4.5). In addition to a one-time, bidirectional admixture event in which populations are allowed to potentially exchange a larger proportion of genetic lineages, low levels of equilibrium migration among demes is also allowed. For these analyses, we used 4-fold degenerate sites (coding sites

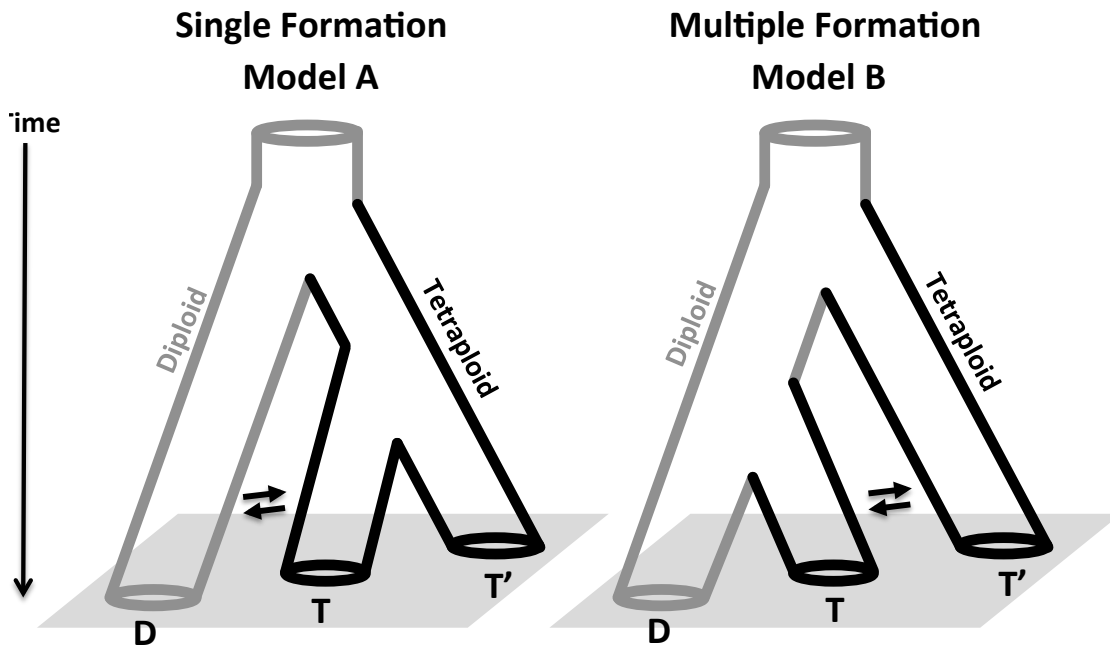


Figure 4.3 Models of population trios used to infer the number of times the tetraploid arose from a diploid ancestor. A present-day tetraploid population (T) may share more genetic variation with a geographically proximal diploid population (D) than a second, geographically distant tetraploid population (T') because of interploidy admixture (Model A) or because of a second, independent origin of a tetraploid population from the diploid ancestor (Model B). In addition to equilibrium migration rates between all populations, a one-time, bidirectional admixture event was allowed (black arrows).

where mutations to any base will not alter amino acid sequence) of populations from our RADseq dataset, since these data produced similar model parameter estimates as the IndSeq data (Table S4.5, See Materials and Methods below).

We find that among our samples the single tetraploid origin model is unambiguously supported in all trio analyses (Table S4.6). In cases involving geographically proximal tetraploids and diploids, we also find evidence of subsequent interploidy admixture. For example, tetraploids T1, T2, T5, and the railway tetraploids (T4, T6, T14) share more genetic variation with Carpathian diploid populations than

tetraploid populations further within the tetraploid range (i.e. T7 and T13; Figure 4.2C, Figure S4.5). Nevertheless, population trio analyses show these admixed tetraploid populations diverged from the same ancestral population as T7 and T13 (Table S4.6A,B), and that allele sharing thus reflects subsequent admixture, not independent origins. Analyses of the same populations present in the PoolSeq dataset validated these results (Table S4.7).

Unidirectional gene flow from diploid to tetraploid *A. arenosa* was previously reported (Jørgensen et al. 2011). However, in our models, maximum likelihood estimates (MLEs) of bidirectional admixture proportions strongly suggest that interploidy gene flow occurred among geographically proximal populations in both directions (Table 4.1). This result holds for 4-fold degenerate as well as noncoding sites (Table S4.8) and for both RADseq and PoolSeq datasets. Moreover, this result is robust to higher sequencing depths and thus greater genotype-calling accuracy (Table S4.9). Similar models that only allow for unidirectional admixture from diploids to tetraploids invariably have significantly lower likelihoods than those that allow bidirectional admixture (Table S4.10).

Interploidy admixture appears to be a local effect, as we estimated admixture proportions using tetraploid populations that are geographically distant from any diploid and found these are near zero (Table S4.11). For these populations, models that do not allow for any interploidy admixture fit the data significantly better than those that do allow admixture (Table S4.10B). The PoolSeq data again gives the same results (Table S4.10C). Major interploidy admixture events (as opposed to ongoing background

Table 4.1 MLEs of model parameters using population trios.

Parameter	Population Trio			
	D3,T5,T7 63,669 sites	D1,T1,T7 93,914 sites	D2,T2,T7 75,373 sites	D1,T4,T7 82,122 sites
Adm_{DT}	0.22 (0.08, 0.41)	0.25 (0.12, 0.39)	0.33 (0.13, 0.44)	0.24 (0.09, 0.34)
Adm_{TD}	0.44 (0.25, 0.54)	0.09 (0.06, 0.19)	0.22 (0.14, 0.35)	0.10 (0.02, 0.18)
T_{Adm}	6877 (4546, 9153)	9317 (7354, 12162)	5023 (4129, 8439)	6995 (4253, 9920)
D_1	8318 (6100, 12106)	12629 (9312, 16276)	7077 (5731, 11076)	7814 (4762, 11674)
D_2	33837 (20046, 37053)	55298 (35567, 68114)	40911 (30388, 70942)	49139 (30707, 61092)
N_D	79142 (52419, 93859)	60119 (44793, 72179)	43930 (34351, 65234)	50600 (38264, 66971)
N_T	45256 (33292, 59610)	118047 (86096, 137344)	48340 (37030, 76297)	42253 (27038, 64637)
$N_{T'}$	96011 (70612, 122310)	96725 (71698, 121498)	77614 (59388, 122849)	59750 (38873, 92965)

Notes: MLEs were obtained using the RADseq dataset, with 95% parametric bootstrap CIs shown below each number. Shown are the admixture proportions from diploids to tetraploids (Adm_{DT}) and tetraploids to diploids (Adm_{TD}) going backwards in time, the time of admixture (T_{Adm}), the divergence time between tetraploids (D_1) and the ancestral tetraploid and diploid (D_2), and the population sizes of the diploid (N_D), admixed tetraploid (N_T), and the outgroup tetraploid ($N_{T'}$). Divergence times are expressed in generations, population sizes are in haploid number of chromosomes, and the number of sites used in each analysis is listed below the trio.

levels of admixture) may also be relatively ancient; for all trio analyses involving populations with interploidy admixture, the 95% parametric bootstrap CIs for the timing (in generations) of large-scale admixture events overlap with the CIs for the divergence time of the two tetraploid populations in the trio (Table 4.1).

The ancestral tetraploid is most closely related to Northern Carpathian diploids

We modified the trio analysis used above to confirm the likely geographic origin of the autotetraploid. Here, we used two diploid populations and one tetraploid population in each analysis to test whether tetraploids are more closely related to a particular diploid gene pool within our sample, while again accounting for interploidy admixture when present (Figure S4.6). Since we already established above that the tetraploids are more closely related to Carpathian diploids (D1-D3) than to Pannonian diploids (D4-D6, Figure 4.2), we only used Carpathian diploid populations. When the Romanian diploid (D1) is included, the tetraploid population used is consistently more closely related to one of the Slovakian diploid populations (either D2 or D3; Table S4.12). This explicit modeling of population history agrees with results from PCA (Figure 4.2B) and suggests that the ancestral tetraploid population is derived from a diploid lineage whose closest extant relatives (within our sample) are found in the Northern Carpathian Mountains today. This area corresponds to what Schmickl et al. (2012) called the “cradle of speciation” for the *A. arenosa* species complex.

Age of the tetraploid lineage

To estimate the age of the tetraploid, we constructed a model for coalescent analyses using a population from the oldest tetraploid split (T1 in Romania) and the closest diploid relative in our sample (D3 in Slovakia). The estimate of the oldest tetraploid divergence in our sample serves as a lower bound to the age of the tetraploid, assuming this split occurred soon after the ancestral tetraploid arose in the Northern Carpathians. Likewise, the estimate of the divergence between the ancestral tetraploid

and the closest diploid relative serves as an upper bound to the age of the tetraploid. We estimated these divergence times by constructing a model of four populations (T1,T5,D1,D3) that accounts for interploidy admixture (Figure 4.4). The MLE for the divergence between the Romanian and Slovakian tetraploids (T1 and T5) is ~15,000 generations, while the MLE for the divergence between the ancestral tetraploid and Slovakian diploid D3 is ~19,000 generations. Thus, the ancestral tetraploid arose ~15,000-19,000 generations ago (or ~11,000-28,000 generations using the 95% parametric bootstrap confidence intervals, Figure 4.4).

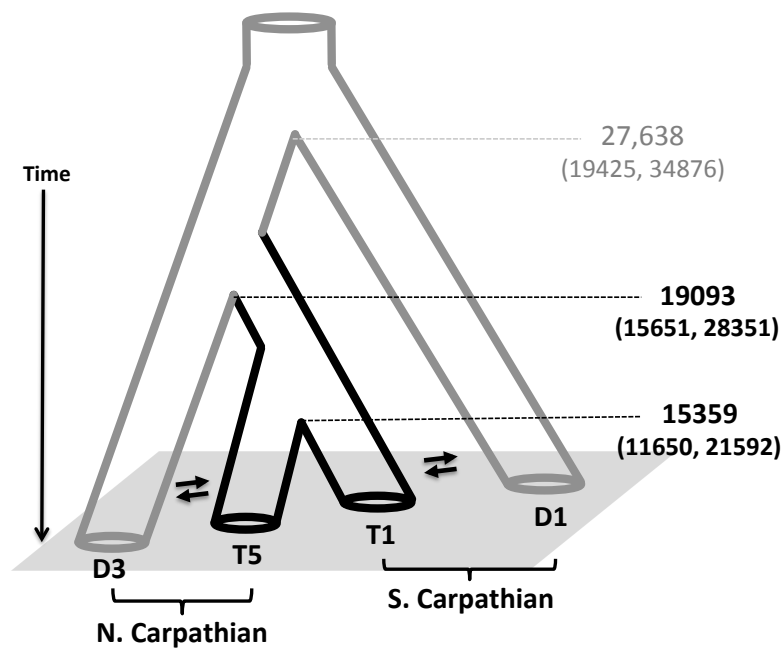


Figure 4.4 Coalescent model to estimate the age of the tetraploid. Divergence times (in generations) are shown for the oldest tetraploid split (T5 and T1) and for the split between the ancestral tetraploid and its potential diploid progenitor (D3 and the ancestor of T5 and T1). 95% parametric bootstrap confidence intervals of divergence times are listed below estimates in parentheses. This model accounts for interploidy admixture between geographically close diploids and tetraploids (bidirectional black arrows among N. Carpathian and S. Carpathian populations).

Genetic structure within tetraploids reveals distinct clades and multiple migration routes

To better understand the demographic history of the tetraploid lineage after it arose, we conducted three separate analyses using only tetraploids: STRUCTURE, *Treemix*, and PCA (Figure 4.5). STRUCTURE results indicate that tetraploid populations in our sample fall into five major clades that roughly correspond to geographic origin. The exception is that samples collected from railroads defy the general trend of isolation-by-distance (Figure 4.5A). Population graph analysis with *Treemix*, a program that constructs a population tree and allows admixture, supports the STRUCTURE results of five tetraploid clades, although bootstrap support for a few nodes are low (Figure 4.5B). We also performed coalescent analyses using one population from each of the four non-railroad clades, and this showed that the tree topology in Figure 4.5B fits the data significantly better than all other possible topologies (Table S4.14). PCA broadly agrees with STRUCTURE and *Treemix* results (Figure 4.5C), and the patterning of individuals within the first two principal components resembles the null expectation of a stepping-stone model across Europe (Novembre and Stephens 2008). However, principal component three (Figure 4.5D) highlights differences between Southwest German (Swabian) and Alpine clades. All three analyses in Figure 4.5 show that the geographically diffuse, panmictic network of railroad tetraploids has admixed with a population from the Alpine clade, T11, which grows on a railway in the Alps. Nearby population T10 was also collected near a railway and exhibits low levels of admixture (Figure 4.5A). Although *Treemix* results suggest admixture between T1 and the railroad

tetraploids, this may be an artifact of not including D1 with which they have admixed (Table 4.1, Figure 4.2C, Figure S4.4).

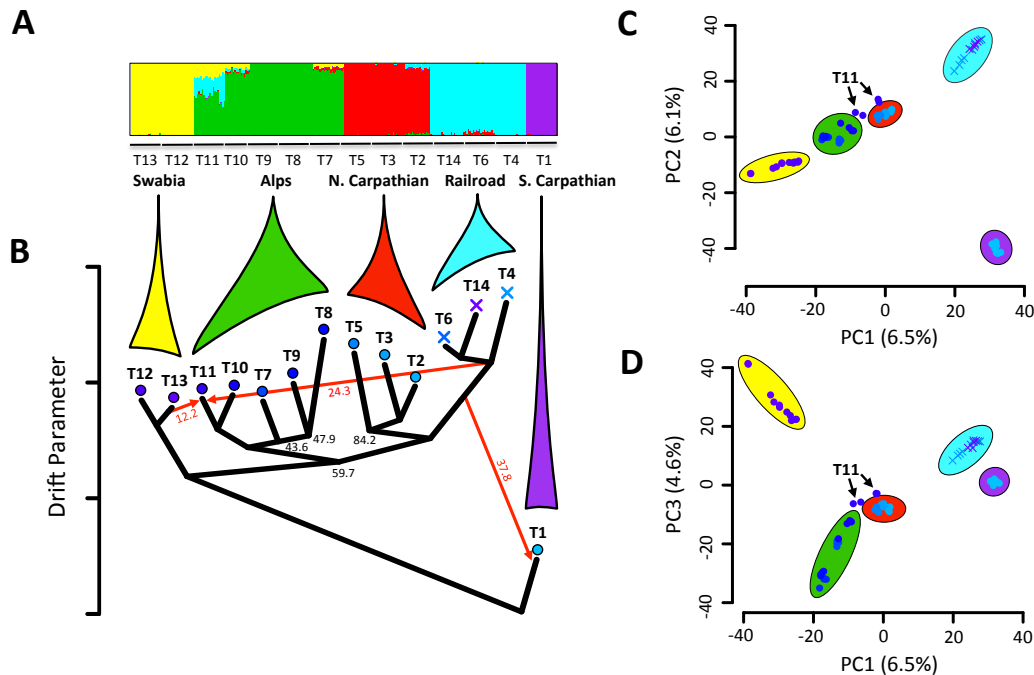


Figure 4.5 Genetic structure within the tetraploid. **(A)** STRUCTURE groups tetraploids into five major clades that correspond to geographic regions, with the exception of tetraploids collected from railroads, which cluster together irrespective of geography. There has been extensive admixture between these railway tetraploids and a population from the Alpine clade, T11. **(B)** Population graph analysis with *Treemix* supports STRUCTURE results of five clades with admixture and reveals the evolutionary relationships among clades. Migration edges (red arrows) indicate evidence for admixture, with numbers indicating migrant ancestry percentages. Bootstrap values under 90% are shown. **(C)** PCA using the first two principal components shows individuals cluster according to geographic region (colored according to STRUCTURE results) with the exception of the admixed T11. **(D)** The third principal component reveals the genetic structure between the Swabian (SW German) and Alpine clades. PCA axes are labeled with percent of the total variance explained by that principal component.

From *Treemix* analysis (Figure 4.5B) and corroborating coalescent simulations we can infer the oldest split within our sample separates the Southern Carpathian

tetraploid from other populations, suggesting an early migration event likely along the Carpathian Mountains into Romania (T1), while the second oldest split involved a lineage that we sampled from the Swabian Alb in Southwestern Germany (T12, T13). Tetraploids sampled from the Alps (mostly in Austria) are more closely related to Slovakian tetraploids from the Northern Carpathian Mountains (Figure 4.5B), suggesting they may represent a single colonization route along the Carpathian mountains into the Alps. The PCA in Figure 4.5D agrees with the interpretation that the Swabian and Alpine lineages represent separate radiations out of an ancestral Slovakian clade.

Interploidy admixture introduced alleles that came under selection

Among several populations, we found evidence of bidirectional admixture. This may increase levels of genetic variation in both ploidies and raises the possibility that gene flow could introduce adaptive alleles. We thus looked for regions of the genome in which proximal diploid and tetraploid populations have experienced selection on the same set of genetic variants where these were likely transferred by gene flow (rather than representing shared ancestral variation). We identified candidate events using admixed populations in the PoolSeq dataset. We scanned both admixed population pairs for regions in which both ploidies displayed evidence of selection on similar sets of geographically unique SNPs. Specifically, we identified loci with significantly low values of Fay and Wu's H (Fay and Wu 2000) that also display an excess of high-frequency, geographically unique shared variants compared to genome-wide patterns. We required that both populations display evidence of selection because we do not know the

direction of admixture for particular loci. Furthermore, finding haplotypes under positive selection in one population, but not in the other, could also be explained by selection preceding neutral gene flow. This method is thus conservative and will not detect all loci that may have experienced selection after admixture.

We find only two examples where there is evidence by our criteria of selective events following admixture in likely introgressed genomic regions. These signals are not due to fluctuations in sequencing depth, as local depths are similar to genome-wide averages (Figure S4.7). First, in populations D1 and T1, only one region has 5% outlier low values of Fay and Wu's H in both populations, indicating an excess of high-frequency derived variants (Figure 4.6A). Using only shared SNPs unique to D1 and T1 relative to all other sampled populations, we calculated θ_H , a metric sensitive to high-frequency derived variants (Fu 1995). Elevated θ_H in this genomic region suggests it is enriched for geographically unique, high frequency shared variants in both populations (Figure 4.6B). These polymorphisms are closely linked (Figure 4.6C,D). This metric contrasts with the version of θ_H used in Fay and Wu's H , which uses all SNPs. We find no evidence of selection in other populations at this locus, suggesting the allele found in these two populations may be locally adaptive (Figure S4.8). Two genes within this region have many high-frequency derived SNPs (relative to the *A. lyrata* reference and other *A. arenosa* populations) that are shared among D1 and T1, one of which causes an amino acid change (Table S4.15). The two genes are orthologs of *A. thaliana* genes AT3G63330 and AT3G63340, both of which encode protein phosphatases of otherwise unknown function (Table S4.16).

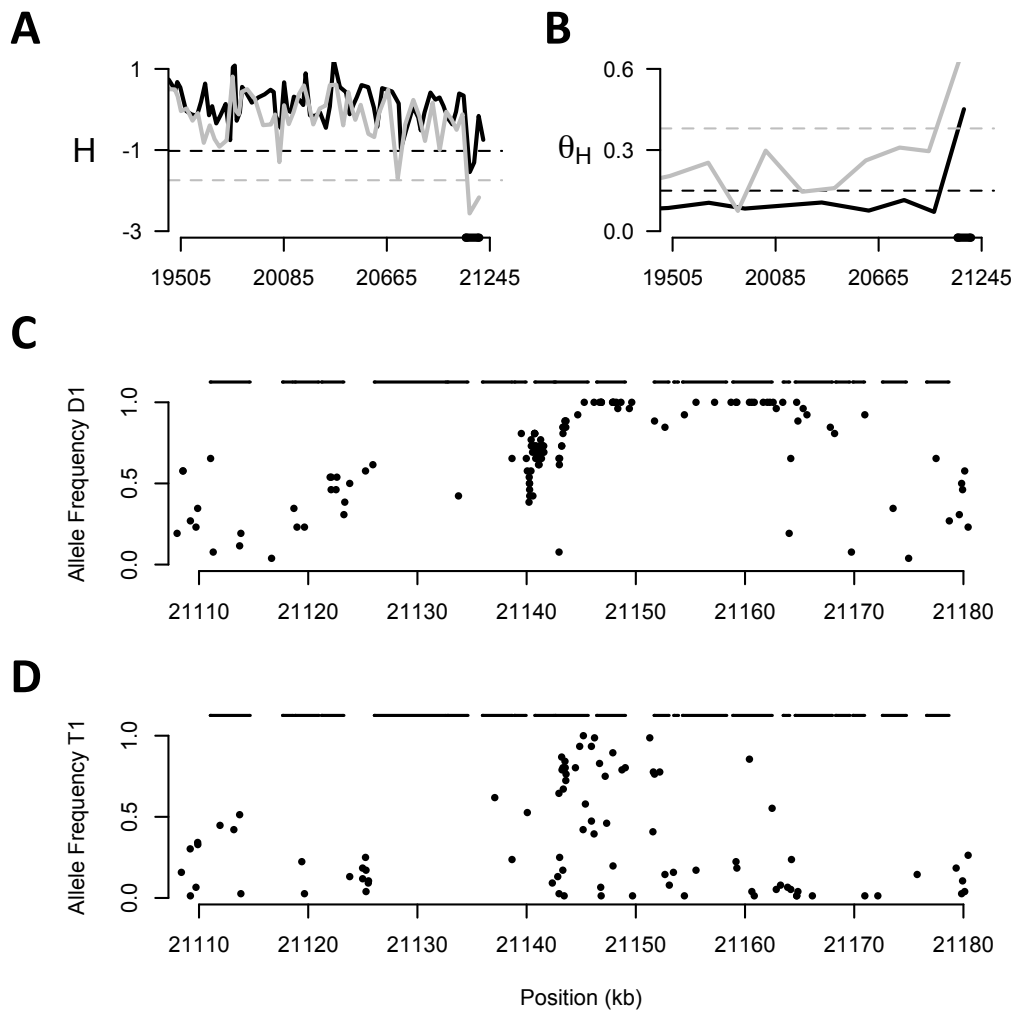


Figure 4.6 Evidence for selection following admixture. **(A)** Fay and Wu's H for 100 SNP windows is below the 5% quantile (dashed line) for both D1 (gray) and T1 (black) near the end of chromosome 5. **(B)** θ_H calculated in 50 SNP windows, using only shared variation unique to D1 and T1. Both D1 (gray) and T1 (black) have an excess of high frequency-shared variation. The dashed lines show the genome-wide 95% quantile for each distribution. **(C,D)** Allele frequency plots for the region spanning the thick black line on the x-axis in **(A)** and **(B)** show a strong enrichment of geographically-unique, high-frequency shared variation for D1 **(C)** and to a lesser degree for T1 **(D)**. In **(C)**, alleles in D1 that are also present in T1 but absent from other Carpathian diploids (D2, D3) are shown. In **(D)**, alleles in T1 are shown if present in D1, but absent from other tetraploids (T5, T7). Many of these variants fall within genic regions (black lines above **C** and **D**).

Second, we found a single genomic region distinct from the one identified in D1 and T1 in which admixture may have been followed by positive selection in populations

D3 and T5 (Figure S4.9). Although selection is not apparent in most other populations at this locus, it may be occurring in a second geographically proximal diploid population D2 (Figure S4.10) that has admixed with T2 (Table 4.1). This region contains a single gene with numerous high-frequency shared SNPs, one nonsynonymous, in D3 and T5. These polymorphisms are absent from other tetraploid populations, but are also present in D2 (Table S4.17). The single gene with high frequency derived polymorphisms in D3 and T5 encodes a protein with pollen allergen domains that is highly homologous to three *A. thaliana* genes in the β -expansin family (AT2G45110, AT1G65680, AT1G65681, Table S4.18). The proteins encoded by these genes are involved in loosening of plant cell walls e.g. during the penetration of pollen tubes through the stigma and style during sexual reproduction (Cosgrove et al. 1997). Expansins may also be important for cell growth in polyploids, and two expansins, including AT1G65680, were found to be under selection in polyploid *A. arenosa* in our previous work (Yant et al. 2013).

4.4 Discussion

Here, we analyze three genomic datasets to assess the demographic history of autotetraploid *A. arenosa* and its diploid relatives. We find that the widespread autotetraploid lineage in *A. arenosa* likely radiated from a single ancestral population ~11,000 - 30,000 generations ago from a diploid lineage closely related to populations found in the Northern Carpathians today. Since *A. arenosa* is generally perennial (Al-Shehbaz and O’Kane 2002), but flowers every year, with railway populations biennial or annual (K Bomblies, P Baduel & B Hunter, unpublished), each generation likely

corresponds to one or two years. These results extend previous work on *A. arenosa*, which suggested the Carpathian Mountains as a center of diversity for the species (Schmickl et al. 2012; Hohmann et al. 2014). Work on other autotetraploids has shown that autotetraploid lineages often arise from more than a single individual (Soltis et al. 1989; Brochmann and Elven 1992; Van Dijk and Bakx-Schotman 1997; Seagraves et al. 1999; Yamane et al. 2003; Yang et al. 2006; Luo et al. 2014), showing that WGD is likely an ongoing mutational process. For *A. arenosa*, the ancestral autotetraploid lineage was likely comprised of multiple individuals (since it is highly diverse and obligately outcrossing), but our analyses strongly suggest that only a single polyploid population gave rise to all our samples of the currently widespread autotetraploid. Unreduced gamete formation from diploids, perhaps elevated in cold stress conditions during periods of glaciation, may have played an important role in the formation of this ancestral gene pool (Ramsey and Schemske 1998).

Our result of a single geographic origin of tetraploid *A. arenosa* is limited to the sampling used in this study. A denser sampling of populations from the Northern Carpathians, or a broader sampling of diploids from the Balkans (Schmickl et al., 2012) and coastal regions of the Baltic Sea (Kolar et al. 2015), as well as tetraploids from these areas, may confirm or change conclusions about tetraploid origins. However, in our population trio analyses above, if a sampled tetraploid population arose from a different, unsampled diploid ancestor, it would have the same time to the most recent common ancestor (TMRCA) to the sampled diploid as the unsampled diploid population from which it arose. However, according to our reconstructions of demographic history,

all sampled tetraploid populations share the same TMRCA with the diploids that we sampled for this study. Thus we can rule out that a more distantly related diploid than the Northern Carpathian diploids gave rise independently to any of the tetraploids within our sample. However, we cannot formally rule out that there were two independent origins from very closely-related unsampled diploids that recently diverged from Northern Carpathian diploids.

Polyploidy is unusual in that it can immediately present a strong gene flow barrier between ploidies even in sympatry, due to the low fertility of progeny from interploidy crosses (e.g. triploids; Ramsey and Schemske 1998). Nevertheless, the autopolyploidization of *A. arenosa* did not create immediate reproductive isolation, as our parametric and nonparametric analyses detect multiple, independent cases of interploidy admixture between geographically proximal populations. Our reconstructions of evolutionary history show this admixture was likely extensive, ancient, and importantly, bidirectional. This is in contrast to a previous report for *A. arenosa*, which suggested that gene flow had occurred only from diploids to tetraploids, not the reverse (Jørgensen et al. 2011). Bidirectional gene flow among ploidies is, however, consistent with findings from other plant species (Thórsson et al. 2001; Ståhlberg 2009).

Gene flow from diploids to tetraploids can occur without the formation of triploids, since diploids produce unreduced gametes at some frequency that can fertilize tetraploids, or neo-tetraploids can arise spontaneously that can also fertilize established tetraploids (Ramsey and Schemske 1998). Our observation of an apparently newly

formed tetraploid in a diploid population supports the possibility that the latter mechanism can occur in *A. arenosa*. Gene flow from tetraploids to diploids, on the other hand, necessitates the formation of triploids, since there is no known mechanism of nondisjunction by which tetraploids can make haploid gametes to regenerate diploids. Though triploids have low fertility, they generally do retain some fertility, allowing gene flow to occur via a so-called “triploid bridge” (Ramsey and Schemske 1998). That this is possible in the *Arabidopsis* genus is supported by the observation that triploids in *A. thaliana* can generate viable aneuploids and populations ultimately resolve to stable diploids and tetraploids over several generations of selfing (Henry et al. 2005). We did not identify any triploids in our sampling of 358 plants, but previous studies have observed rare triploids in *A. arenosa* (Koln k 2007; J rgensen et al. 2011, Kolar et al. 2015), which may suffice to yield substantial gene flow over evolutionary timescales. Since estimates of interploidy gene flow tend to be older than several thousand generations in our models (Table 4.1), it is possible that some degree of interploidy reproductive isolation has evolved and that triploids were once more abundant than they are now.

What the consequences are for diploids of the influx of tetraploid alleles or vice versa is not known. We speculate that interploidy admixture, while generally neutral or likely at times deleterious, could occasionally result in the exchange of beneficial haplotypes. That introgressed regions are beneficial and subsequently experience positive selection seems to be rare, but we do find two cases where admixed populations seem to have experienced selection in introgressed genomic regions (Figure

4.6, S9). In both cases, there is evidence of a strong selective sweep in the admixed diploid and a weaker signature of selection in the corresponding tetraploid. This pattern may be explained by autotetraploids generally having weaker responses to selection (Hill 1971), from selection taking place further in the past in the tetraploid, or both. While these results may also be explained by parallel selection on haplotypes segregating in both ploidies as standing genetic variation, we do not think this is likely as the hitchhiking effect is stronger than expected if the selected SNP(s) persisted as neutral variant(s) as long as the divergence time between D1 and T1 (~35,000 generations). Parallel selection on standing variation this old would likely produce a softer sweep undetectable by Fay and Wu's H (Messer and Petrov 2013). Ultimately, having haplotype information for this region would resolve this uncertainty. It is also possible these loci are not adaptive, but experienced selection upon introgression due to other factors such as meiotic drive (Derome et al. 2004).

After the ancestral tetraploid population arose from its diploid progenitor, it colonized much of Europe via at least four distinct migration routes from its likely origin in the Northern Carpathian Mountains. One lineage is represented in our sample by a single population from the Southern Carpathians that diverged from other tetraploids ~12,000 generations ago (S. Carpathian clade, Figure 4.5). This is the oldest tetraploid divergence time in our sample, and this colonization may have been possible from large ice-free swaths within the Carpathians, even during the last glacial maximum (reviewed in Ronikier 2011). At least two tetraploid lineages then independently colonized southwest Germany and the Alps, diverging from each other ~8,000 generations ago.

The dating of these events strongly depends on the mutation rate we estimated from the data (see Methods) and used in coalescent analyses, but these dates can simply be rescaled if a different mutation rate is discovered.

The migration route of the populations found currently in the Southwestern German Swabian Alb region is unclear and may have occurred along the chains of limestone hills that run across Germany north of the Alps. Finally, a fourth lineage liberated itself from the generally montane niche that other tetraploid lineages are found in and colonized railroad habitats across Central and Northern Europe. This genetically and phenotypically distinct “railroad ecotype” has rapidly traversed large geographic distances such that populations sampled from disparate parts of the range remain very similar, which is not true of the other tetraploid lineages. This suggests a rapid and recent range expansion, likely facilitated by migration along railway networks. To this last point are clearly a few exceptions: we sampled one population from a railway in the Alps (T11) that is genetically primarily an Alpine type. This second colonization of railway habitats may have been facilitated by admixture with the more prevalent railway ecotype found in other parts of Europe (Figure 4.5).

In sum, we show that *A. arenosa* autotetraploids we sampled from 14 widely distributed populations all originated from a single ancestral population that likely arose ~11,000 - 30,000 generations ago in the Northern Carpathian Mountains. This population subsequently split into at least four distinct lineages that colonized the Southern Carpathians, Southwestern Germany, the Alps, and the railways of Central and Northern Europe. We also show evidence that there has been bidirectional interploidy

admixture among geographically proximal diploid and tetraploid populations. Tetraploids that colonized the Alpine region, where no diploids occur, show no evidence of past interploidy admixture. In two instances gene flow between diploids and tetraploids exchanged sets of variants that are associated with selection in both ploidies, suggesting that bidirectional admixture may have functional consequences, though whether the alleles that came under selection are adaptive remains to be tested. Nevertheless, these results suggest that interploidy admixture within multiple-ploidy systems may shape patterns of variation. Our recovery of an apparently newly formed tetraploid individual in a population of diploids suggests that polyploids do arise sporadically within *A. arenosa* diploid populations. The occasional formation of neotetraploids could provide an additional mechanism for gene flow from diploids to tetraploids, and has at least the theoretical potential to generate novel tetraploid lineages, though no independent lineages appear yet to have established themselves widely.

4.5 Methods

Generation of DNA sequence data

We used three DNA sequence datasets in this analysis: Restriction-associated DNA sequencing (RADseq), individual whole-genome sequences (IndSeq), and whole-genome sequencing of population pools (PoolSeq, Table S4.1). The generation of the IndSeq and PoolSeq datasets was described previously (Yant et al. 2013, Wright et al.

2014). We generated the RADseq dataset using a modified a double digest RADseq protocol (Supplementary Methods; Peterson et al. 2012).

DNA sequence alignment and variant calling

For all datasets, we aligned DNA sequences to the *Arabidopsis lyrata* reference genome (Hu et al. 2011) using Stampy v1.0.21 (Lunter and Goodson 2011) with default parameter values. For the IndSeq and PoolSeq datasets, we removed PCR duplicates using Picard (<http://picard.sourceforge.net/>). We locally realigned reads around indels for all datasets using the Genome Analysis Toolkit (GATK v2.7; McKenna et al. 2010). We called sequence variants for the IndSeq and RADseq datasets using the GATK. To maximize variant detection within and between populations for each dataset, we genotyped all individuals irrespective of ploidy simultaneously as diploid or tetraploid, with individual genotypes later extracted from the appropriate file. Potential variants were filtered using the GATK VariantFiltration tool (Supplementary Methods). To call variants in the PoolSeq dataset, we used SNAPE (Raineri et al. 2012, as described in Wright et al. 2014). We also used an additional data filtration step to remove loci that likely contain spuriously mapped sequence reads (Supplementary Methods).

Bioinformatic assessment of ploidy

To determine the ploidy of each sample, we compared non-reference base count distributions of each sample to those expected of a diploid, triploid, and tetraploid. We simulated the expected distribution for diploids as Binomial($n=30$, $p=1/2$), since we only

considered sites with a minimum sequencing depth of 30. To model triploids, we constructed a compound distribution in which non-reference base counts were simulated from either Binomial($n=30$, $p=1/3$) or Binomial($n=30$, $p=2/3$) with probability $2/3$ or $1/3$, respectively, as expected for a neutral mutation frequency spectrum (Fu 1995). For tetraploids, we simulated non-reference base counts from Binomial($n=30$, $p=1/4$), Binomial($n=30$, $p=1/2$), or Binomial($n=30$, $p=3/4$) with probability $6/11$, $3/11$, or $2/11$, respectively. We compared observed non-reference base counts of each sample to these simulated expected distributions using a G statistic in which $G = 2 \sum_i O_i * \ln \left(\frac{O_i}{E_i} \right)$, where O_i are observed frequencies and E_i are expected frequencies. For calculating G , we categorized alternate base count proportions greater than 0.2 and less than 0.8 into twelve bins (increments of 0.05) to avoid sequencing errors that occur at low frequencies. Using G to select the best-fit model for each sample, we show all samples from a collection site were of the same ploidy (Table S4.4) with three exceptions: a putative neotetraploid in an otherwise diploid population (Figure S4.2), and two samples that were likely diploid but not well-modeled by our expected diploid distribution due to greater, unexplained variance in base count frequencies (Figure S4.11).

Principal Component Analysis

We performed principal component analysis (PCA) in R using the package *adegenet* 1.3-5 (Jombart and Ahmed 2011), which accommodates for variable ploidy. All PCAs used SNPs in which all individuals had a sequencing depth of at least 8x. For the

single tetraploid PCA, we allowed up to 30% of individuals to have a sequencing depth of less than 8x and coded that site as missing for those individuals.

Coalescent analyses

We fit demographic models in population trio analyses to observed data using *fastsimcoal2* (Excoffier et al. 2013), a program that uses the coalescent (Kingman 1982) to simulate multi-dimensional AFS and a modified expectation-maximization algorithm to search parameter space and find maximum likelihood estimates (MLEs) for model parameters. After MLEs are obtained, we compared model likelihoods with Akaike information criterion (AIC) to assess which model had a higher probability of being correct given the candidate set of models. To avoid an excess of zeroes in higher-dimensional AFS used for analysis, we used only three populations and sampled only 6 tetraploids and 9-12 diploids for each population. We only considered sites in which all individuals had a sequencing depth of at least 8x, using the common allele of 24 *A. lyrata* genomes as the reference allele.

Since demographic analyses are potentially sensitive to an enrichment of high-frequency alleles due to misspecification of the ancestral allele, we attempted to correct the three-dimensional AFS for misspecified alleles using the following extension of the technique described in Baudry and Depaulis (2003). Although we excluded sites with more than two segregating bases from analyses, multiple mutations may occur at a site and go undetected if the same mutation occurs twice within the species tree. We calculated the probability of a biallelic site experiencing multiple mutations using the

proportion of triallelic sites in the sample, empirically derived estimates of transition and transversion rates, and equation 3 in Baudry and Depaulis (2003). After an uncorrected, unfolded 3D AFS was obtained from the data, we constructed a folded 3D AFS. We multiplied each entry in the folded 3D AFS by the empirically obtained probability of a biallelic site experiencing more than one mutation to obtain frequency-specific proportions of mispolarized alleles. We used these proportions to reorient a proportional number of alleles in the respective unfolded category (Figure S4.12).

To construct 95% parametric bootstrap confidence intervals (CIs), we simulated the same number of sites used in the analysis 100 times, using linkage blocks of 260bp (insert size) for the RADseq dataset or 5kb for the IndSeq and PoolSeq datasets, and a population recombination rate that is roughly twice as large as the population mutation rate. We estimated the mean population recombination rate per bp using *LDhat* (Auton and McVean 2007) on diploid whole-genome sequences, specifically from an 800kb segment on chromosome 2 in four individuals from population D3. We chose this chromosomal segment due to even and high sequencing depths. For each simulated dataset, we ran 50 instances of *fastsimcoal2* to infer the MLEs of parameter values, which were then used to construct confidence intervals.

In order to obtain absolute values for model parameters, a mutation rate must be specified. We calculated the mutation rate for noncoding and 4fold-degenerate sites using a simple isolation-migration (IM) model with populations D2 and D3. We obtained 100 MLEs of all parameter values, including the mutation rate, using 50 *fastsimcoal2* runs each time to obtain parameter estimates. We repeated this analysis separately for

noncoding and 4fold-degenerate sites. The mode of the 100 MLEs for the mutation rate was assumed to be near the true mutation rate, since this held true for 100 datasets simulated with a known mutation rate (Figure S4.13). We thus used a mutation rate of 3.7×10^{-8} and 4.3×10^{-8} for noncoding and 4fold-degenerate sites, respectively.

Comparison of demographic inference among genomic datasets.

In order to evaluate the sensitivity of demographic inference to dataset type, we generated a simple IM model for diploid populations D2 and D3 for all three datasets and calculated the maximum likelihood estimates (MLEs) of model parameters using *fastsimcoal2* (Excoffier et al. 2013). IndSeq and RADseq datasets produced very similar MLEs of parameter values for divergence time and comparable migration rates when only 4-fold degenerate sites (sites which can sustain any mutation without causing an amino acid change) were used (Table S4.5). Thus we used this functional category of sites for coalescent-based reconstructions of history; the use of noncoding sites caused results to differ more significantly between datasets (Table S4.5).

Reconstruction of tetraploid history

We characterized tetraploid population structure using STRUCTURE v2.3.4 (Pritchard et al. 2000), selecting a value of K populations that corresponded to the last largest increase in likelihood before likelihood values approached an asymptote with increasing values of K. We constructed population graphs of the tetraploids with *Treemix* (Pickrell and Pritchard 2012), using population T1 as root and adding migration

edges until residuals did not appreciably decrease (two in our case). We bootstrapped the data by generating 1000 replicates, subsampling every three SNPs each time, and using Newick Utilities (Junier and Zdobnov 2010) to summarize results. Noncoding sites were used for *Treemix* analysis.

CHAPTER 5 – CONCLUSIONS AND FUTURE DIRECTIONS

A major goal of evolutionary biology is to discover the processes that give rise to the stunning organismal diversity on Earth. Natural genetic variation among individuals, which arises from mutations in DNA, creates new phenotypes on which natural selection may act, driving change in populations. Commonly studied mutations are small DNA lesions in which a single nucleotide changes, creating a polymorphism that segregates in a population, or a set of nucleotides are inserted/deleted. However, the process of DNA replication and division may go awry in another important way that gives rise to cells with twice the DNA content. This single mutational event, called whole-genome duplication (WGD), not only creates large-scale phenotypic changes but also gives rise to populations that have different evolutionary dynamics than the population(s) from which they arose (Ramsey and Schemske 2002, Bever and Felber 1992).

I have dedicated my thesis to the study of whole-genome duplication, which is a beautiful error in DNA replication and division that has substantially contributed to organismal diversity in the plant kingdom (Stebbins 1950, Grant 1981, Masterson 1994, Cui et al. 2006). In particular, I have chosen to extend our knowledge about autotetraploids, which arise from WGD events within a diploid species. I began my studies extending an important body of population genetic theory to autotetraploids (Chapter 2), and I applied this theory to revolutionary genomic datasets collected from *A. arenosa*, a species that has diploid and autotetraploid populations (Chapter 4). In the

interim, I took time to question the utility of a popular method I chose to use for sequencing DNA (Chapter 3). The work presented in this thesis goes beyond previous studies of autopolyploids in terms of the amount of data used and the complexities of analyses, and it is the first of its kind to address questions surrounding the evolutionary dynamics of multiple ploidy systems.

However, more work needs to be done both on the *A. arenosa* system and other multiple ploidy systems. For example, although I rigorously demonstrated in Chapter 4 that the autotetraploid race in *A. arenosa* arose from a single ancestral population and subsequently experienced interploidy admixture, I was only able to extend my conclusions to the populations included in our sampling. There are other areas within the *A. arenosa* range in which diploids and tetraploids grow in geographically close localities (see Chapter 4 Discussion), and it will be important to study these regions to see if these tetraploid populations arose from independent ancestral populations or if they arose from the same ancestral population I've discovered. To accomplish this, I am currently collaborating with researchers at the University of Oslo who have extensively sampled *A. arenosa* across its entire range (Kolar et al. 2015). I will help them use the same type of analyses I created in Chapter 4 to directly extend my previous results to the entire species. Knowing how frequently autotetraploids arise from diploids will give us greater insight into how the WGD process contributes to species diversity.

In addition, more studies need to be done on multiple ploidy systems from different species. *A. arenosa* is an obligate-outcrosser, meaning that only gametes from distinct individuals may contribute to progeny in the next generation. However, the

evolutionary dynamics of multiple ploidy systems with plants that are self-compatible, or capable of self-fertilization, may be very different from those observed in *A. arenosa*. For instance, the ability to self-fertilize increases the chance of contributing progeny to the next generation, especially if mates are limited, as is the case with a neopolyploid in a predominately diploid population (Rausch and Morgan 2005). Self-compatible neotetraploids may thus have a much greater chance to give rise to established polyploid populations relative to self-incompatible tetraploids (Rausch and Morgan 2005). Consequently, multiple ploidy systems with self-compatibility may contain numerous, independently derived polyploidy lineages.

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Supplementary Text

Extending Coalescent Theory to Self-Fertilizing Autotetraploids

For a population of autotetraploids that self-fertilize with probability s , the single-generation transition matrix for the ancestral process is given by

$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} + \frac{1}{2}s & \frac{2}{3}(1-s) & \frac{s}{6} \\ \frac{3}{4N} & 1 - \frac{1}{N} & \frac{1}{4N} \\ 0 & 0 & 1 \end{bmatrix}.$$

Using the result of Möhle 1998 and collecting the “fast” events, which occur on the timescale of single generations, into matrix \mathbf{F} and the “slow” events, which occur on the timescale of N generations, into matrix \mathbf{S} , we obtain

$$\mathbf{F} = \begin{bmatrix} \frac{1}{3} + \frac{s}{2} & \frac{2(1-s)}{3} & \frac{s}{6} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} 0 & 0 & 0 \\ \frac{3}{4} & -1 & \frac{1}{4} \\ 0 & 0 & 0 \end{bmatrix}$$

in the limit as N tends to infinity. Here, $\mathbf{F} = \lim_{N \rightarrow \infty} \mathbf{P}$ and $\mathbf{S} = \lim_{N \rightarrow \infty} N(\mathbf{P} - \mathbf{F})$. If the limit $\mathbf{E} = \lim_{t \rightarrow \infty} \mathbf{F}^t$ exists, the continuous-time approximation to this ancestral process with partial selfing is characterized by matrices

$$\mathbf{E} = \begin{bmatrix} 0 & \frac{4-4s}{4-3s} & \frac{s}{4-3s} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{G} = \begin{bmatrix} 0 & -\frac{12s^2-22s+10}{9s^2-24s+16} & \frac{4-4s}{9s^2-24s+16} \\ 0 & -\frac{1}{4-3s} & \frac{1}{4-3s} \\ 0 & 0 & 0 \end{bmatrix}$$

where $\mathbf{G} = \mathbf{E}\mathbf{S}\mathbf{E}$ (Möhle 1998).

Note on extending double reduction to arbitrary n

If four lineages are within an individual and double reduction is possible, three types of events may occur for each pair of lineages in the instantaneous adjustment of the sample: (1) two lineages come from the same gamete and coalesce in the immediately previous generation via double reduction with probability α^2 , (2) two lineages come from the same gamete, do not coalesce in the immediately previous generation, but coalesce in later generations via double reduction with probability $(1-\alpha)\frac{\alpha}{\alpha+2}$, or (3) two lineages get separated into distinct individuals before any coalescence event with probability $(1-\alpha)\left(\frac{2}{\alpha+2}\right)$. Putting these probabilities together, if four lineages are within an individual, a double coalescence event occurs in the

instantaneous adjustment of the sample with probability $\alpha^2 + 2\alpha(1-\alpha)\frac{\alpha}{\alpha+2} + (1-\alpha)^2\left(\frac{\alpha}{\alpha+2}\right)^2$,

a single coalescence event occurs with probability $2\left(\alpha(1-\alpha)\left(\frac{2}{\alpha+2}\right)\right) + 2\left((1-\alpha)^2\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{2}{\alpha+2}\right)\right)$,

and no coalescence events occur with probability $(1-\alpha)^2\left(\frac{2}{\alpha+2}\right)^2$. Thus, the number of

instantaneous coalescence events for m individuals that contain four lineages is

Multinomial with length m and probability vector \mathbf{p} containing the three probabilities described above.

Supplementary Tables

Table S2.1 Plant species with confirmed tetrasomic inheritance

Species	Chromosome associations (metaphase I)	Reference(s)
<i>Acacia nilotica</i>		1
<i>Actinidia chinensis</i>		2
<i>Allium nevilii</i>		3
<i>Arabidopsis lyrata</i>		4
<i>Biscutella laevigata</i>		5, 6
<i>Centaurea jacea</i>	Bivalents	7
<i>Chrysanthemum boreale</i>	Bivalents	8
<i>Dioscorea trifida</i>		9
<i>Epilobium angustifolium</i>	Mostly Bivalents, Some quadrivalents	10, 11
<i>Heuchera grossulariifolia</i>	Mostly Bivalents, Some quadrivalents	12
<i>Heuchera micrantha</i>		13
<i>Lotus corniculatus</i>	Bivalents	14, 15
<i>Maclura pomifera</i>		16
<i>Medicago sativa</i>		17
<i>Paspalum notatum</i> **	Mostly quadrivalents	18
<i>Paspalum simplex</i>	Mostly quadrivalents	19
<i>Prunus spinosa</i>	Bivalents	20
<i>Rorippa amphibia</i>		21
<i>Rorippa sylvestris</i>		21
<i>Rutidosis leptorrhynchoides</i>		22
<i>Thymus praecox</i>		23
<i>Tolmeia menziesii</i>	Bivalents	24
<i>Turnera ulmifolia</i>	Bivalents	25, 26
<i>Vaccinium corymbosum</i>	Bivalents	27, 28

Note: This table lists species that have been demonstrated to be genetically tetrasomic,

meaning that four alleles segregate at single loci. It is meant to show that numerous species have been demonstrated to have tetrasomic inheritance, even if cytologically diploidized, and is not meant to be an exhaustive list.

** *Paspalum notatum* has tetrasomic inheritance at most markers, but in apomictic lines has disomic inheritance around the apospory (apomixis) locus.

References:

1. Mandal, A.K., Ennos, R.A. and Fagg, C.W. (1994) Mating system analysis in a natural population of *Acacia nilotica* subspecies *leiocarpa*. *Theor Appl Genet* 89:931-935.
2. Huang, H., Dane, F., Wang, Z., Jiang, Z., Huang, R. and Wang, S. (1997) Isozyme inheritance and variation in *Actinidia*. *Heredity* 78: 328-336.
3. Rieseberg, L.H. and Doyle, M.F. (1989). Tetrasomic segregation in the naturally occurring autotetraploid *Allium nevillei* (*Alliaceae*). *Hereditas* 111: 31-36.
4. Mable, B.K., Beland, J., and Di Berardo, C. (2004) Inheritance and dominance of self-incompatibility alleles in polyploid *Arabidopsis lyrata*. *Heredity* 93: 476-486.
5. Manton I. (1934) The problem of *Biscutella laevigata* L. *Z. Indukt. Abstammungs-Vererbungslehre* 67: 41-57.
6. Manton I. (1937) The problem of *Biscutella laevigata* L. II. The evidence from meiosis. *Ann. Bot.* 1: 439-462.
7. Hardy, O.J., Vanderhoeven, S., De Loose, M. and Meerts, P. (2000) Ecological, morphological and allozymic differentiation between diploid and tetraploid knapweeds (*Centaurea jacea*) from a contact zone in the Belgian Ardennes. *New Phytol.* 146: 281-29
8. Watanabe, K. (1983) Studies on the control of diploid-like meiosis in polyploidy taxa of *Chrysanthemum*. *Theor Appl Genet* 66:9-14.
9. Bousalem, M., Arnau, G., Hochu, I., Arnolin, R., Viader, V., Santoni, S. and David, J. (2006) Microsatellite segregation analysis and cytogenetic evidence for tetrasomic inheritance in the American yam *Dioscorea trifida* and a new basic chromosome number in the *Dioscoreae*. *Theor Appl Genet* 113: 439-451.
10. Mosquin, T. (1967) Evidence for Autopolyploidy in *Epilobium angustifolium* (Onagraceae). *Evolution* 21: 713-719.
11. Husband, B.C. and Schemske, D.W. (1997) The Effect of Inbreeding in Diploid and Tetraploid Populations of *Epilobium angustifolium* (Onagraceae): Implications for the Genetic Basis of Inbreeding Depression. *Evolution* 51: 737-746.
12. Wolf, P.G., Soltis, P.S., and Soltis, D.E. (1989). Tetrasomic inheritance and chromosome pairing behaviour in the naturally occurring autotetraploid *Heuchera grossulariifolia* (Saxifragaceae). *Genome* 32 : 655 -659.
13. Soltis, D.E. and Soltis, P.S. (1989) Tetrasomic inheritance in *Heuchera micrantha* (Saxifragaceae). *J Hered* 80: 123-126.
14. Dawson, C.D.R. (1941) Tetrasomic inheritance in *Lotus corniculatus* L. *J Genet* 42: 49-73.









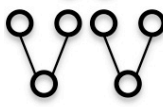


15. Fjellstrom, R.G., Beuselinck, P.R. and Steiner, J.J. (2001) RFLP marker analysis supports tetrasomic inheritance in *Lotus corniculatus* L. *Theor Appl Genet* 102:718-725
16. Laushman, R.H., Schnabel, A., and Hamrick, J.L. (1990). Electrophoretic Evidence for tetrasomic inheritance in the dioecious tree *Maclura pomifera* (Raf.) Schneid. *J Hered*, 87: 469-473.
17. Quiros, C.F. (1982) Tetrasomic segregation for multiple alleles in alfalfa. *Genetics* 101; 117-127.
18. Stein, J., Quarín, C.L., Matínez, E.J., Pessino, S.C., and Ortiz, J.P.A. (2004) Tetraploid races of *Paspalum notatum* show polysomic inheritance and preferential chromosome pairing around the apospory-controlling locus. *Theor Appl Genet* 109:186-191.
19. Pupilli, F., Caceres, M.E., Quarín, C.L. and Arcioni, S. (1997) Segregation analysis of RFLP markers reveals a tetrasomic inheritance in apomictic *Paspalum simplex*. *Genome* 40: 822-828.
20. Leinemann, L. (2000) Inheritance analysis of isozyme phenotypes in tetraploid species using single plant progenies. An example in black thorn (*Prunus spinosa* L.). *Forest Genetics* 7: 205-209.
21. Stift, M., Berenos, C., Kuperus, P. and van Tienderen, P.H. (2008) Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: A general procedure applied to *Rorippa* (Yellow Cress) microsatellite data. *Genetics* 179: 2113-2123.
22. Brown, A.H.D. and Young, A.G. (2000) Genetic diversity in tetraploid populations of the endangered daisy *Rutidosis leptorrhynchoides* and implications for its conservation. *Heredity* 85: 122-129.
23. Landergott, U., Naciri, Y., Schneller, J.J., and Holderegger, R. (2006) Allelic configurations and polysomic inheritance of highly variable microsatellites in tetraploid gynodioecious *Thymus praecox* agg. *Theor Appl Genet* 113: 453-465.
24. Soltis, D.E. and Rieseberg, L.H. Autopolyploidy in *Tolmiea menziesii* (Saxifragaceae): Genetic insights from enzyme electrophoresis. *Am J Bot* 73: 310-318.
25. Tamari, F., Khosravi, D., Hilliker, A.J. and Shore, J.S. (2005) Inheritance of spontaneous mutant homostyles in *Turnera subulata* x *kravovickasii* and in autotetraploid *T. scabra* (Turneraceae). *Heredity* 94: 207-216.
26. Fernández, A. (1987). Estudios cromosómicos en *Turnera* y *Piriqueta* (Turneraceae). *Bonplandia* 6: 1-21.
27. Jelenkovic, G., and Hough, L.F. (1970). Chromosome associations in the first meiotic division in three tetraploid clones of *Vaccinium corymbosum* L. *Can J Genet Cytol* 12: 316–324.
28. Krebs, S.L. and Hancock, J.F. (1989). Tetrasomic inheritance of isoenzyme markers in the highbush blueberry *Vaccinium corymbosum* L. *Heredity* 63: 11–18.






Table S2.2 Patterns of ancestry used to calculate transition probabilities for the ancestral process of 2 lineages sampled from an autotetraploid population. Balls represent autotetraploid individuals, and lines are the transmission of gametes. From left to right, the columns containing probabilities represent the probability of observing the pattern of ancestry and the conditional transition probabilities for the Markov chain.








Pattern	P{Pattern}	p_{11}	p_{12}	p_{13}
	$\frac{1}{N}$	$\frac{5}{6}$		$\frac{1}{6}$
	$\frac{N-1}{N}$	$\frac{1}{3}$	$\frac{2}{3}$	
Pattern	P{Pattern}	p_{21}	p_{22}	p_{23}
	$\frac{1}{N^3}$	$\frac{3}{4}$		$\frac{1}{4}$
	$\frac{N-1}{N^3}$		1	
	$4 \cdot \frac{N-1}{N^3}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$
	$2 \cdot \frac{N-1}{N^3}$	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$
	$2 \cdot \frac{(N-1)(N-2)}{N^3}$		1	
	$4 \cdot \frac{(N-1)(N-2)}{N^3}$	$\frac{3}{16}$	$\frac{3}{4}$	$\frac{1}{16}$
	$\frac{(N-1)(N-2)(N-3)}{N^3}$		1	














Note: For example, for the pattern of ancestry in which two autotetraploids are half-sibs, the probability of coalescence is 1/16 because, conditional on this pattern, there is a 1/4 chance that the two lineages in separate individuals came from the same parent and a 1/4 chance that they originated from the same chromosome. For the case in which two autotetraploids are full-sibs, the probability of coalescence is twice as great since it may occur in each of the two parents.



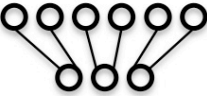
Table S2.3 Patterns of ancestry used to calculate transition probabilities for the ancestral process of 4 lineages sampled from an autotetraploid population. Balls represent autotetraploid individuals, and lines are the transmission of gametes. From left to right, the columns containing probabilities represent the probability of observing the pattern of ancestry and the conditional transition probabilities for the Markov chain.










Pattern	P{Pattern}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}
	$\frac{1}{N}$	$\frac{1}{6}$					$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{N-1}{N}$			1				
Pattern	P{Pattern}	p_{21}	p_{22}	p_{23}	p_{24}	p_{25}	p_{26}	p_{27}
	$\frac{1}{N^3}$	$\frac{1}{8}$					$\frac{5}{8}$	$\frac{1}{4}$
	$\frac{N-1}{N^3}$		$\frac{1}{2}$				$\frac{1}{2}$	
	$4 \cdot \frac{N-1}{N^3}$	$\frac{1}{32}$	$\frac{1}{8}$	$\frac{3}{8}$			$\frac{13}{32}$	$\frac{1}{16}$
	$2 \cdot \frac{N-1}{N^3}$		$\frac{1}{4}$	$\frac{3}{8}$			$\frac{3}{8}$	
	$2 \cdot \frac{(N-1)(N-2)}{N^3}$		$\frac{1}{4}$		$\frac{1}{2}$		$\frac{1}{4}$	
	$4 \cdot \frac{(N-1)(N-2)}{N^3}$		$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{2}$		$\frac{3}{16}$	
	$\frac{(N-1)(N-2)(N-3)}{N^3}$					1		
Pattern	P{Pattern}	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	p_{36}	p_{37}
	$\frac{1}{N^3}$	$\frac{25}{216}$					$\frac{65}{108}$	$\frac{61}{216}$
	$\frac{N-1}{N^3}$			$\frac{25}{36}$			$\frac{5}{18}$	$\frac{1}{36}$














Pattern	P{Pattern}	p_{31}	p_{32}	p_{33}	p_{34}	p_{35}	p_{36}	p_{37}
	$4 \cdot \frac{N-1}{N^3}$	$\frac{5}{216}$	$\frac{5}{18}$	$\frac{5}{36}$			$\frac{107}{216}$	$\frac{7}{108}$
	$2 \cdot \frac{N-1}{N^3}$	$\frac{1}{108}$	$\frac{2}{9}$	$\frac{11}{36}$			$\frac{23}{54}$	$\frac{1}{27}$
	$2 \cdot \frac{(N-1)(N-2)}{N^3}$			$\frac{5}{18}$	$\frac{5}{9}$		$\frac{1}{6}$	
	$4 \cdot \frac{(N-1)(N-2)}{N^3}$	$\frac{1}{216}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{5}{9}$		$\frac{13}{54}$	$\frac{1}{216}$
	$\frac{(N-1)(N-2)(N-3)}{N^3}$			$\frac{1}{9}$	$\frac{4}{9}$	$\frac{4}{9}$		














Pattern	P{Pattern}	p_{41}	p_{42}	p_{43}	p_{44}	p_{45}	p_{46}	p_{47}
	$\frac{1}{N^5}$	$\frac{5}{48}$					$\frac{7}{12}$	$\frac{5}{16}$
	$6 \cdot \frac{N-1}{N^5}$	$\frac{1}{24}$	$\frac{2}{9}$	$\frac{1}{24}$			$\frac{79}{144}$	$\frac{7}{48}$
	$3 \cdot \frac{N-1}{N^5}$		$\frac{5}{18}$	$\frac{5}{24}$			$\frac{17}{36}$	$\frac{1}{24}$
	$12 \cdot \frac{N-1}{N^5}$	$\frac{1}{64}$	$\frac{5}{24}$	$\frac{7}{32}$			$\frac{23}{48}$	$\frac{5}{64}$
	$6 \cdot \frac{N-1}{N^5}$		$\frac{7}{36}$	$\frac{1}{3}$			$\frac{31}{72}$	$\frac{1}{24}$
	$4 \cdot \frac{N-1}{N^5}$	$\frac{1}{96}$	$\frac{5}{24}$	$\frac{1}{4}$			$\frac{15}{32}$	$\frac{1}{16}$
	$3 \cdot \frac{(N-1)(N-2)}{N^5}$		$\frac{5}{18}$	$\frac{1}{12}$	$\frac{1}{6}$		$\frac{4}{9}$	$\frac{1}{36}$














Pattern	$P\{\text{Pattern}\}$	p_{41}	p_{42}	p_{43}	p_{44}	p_{45}	p_{46}	p_{47}
	$12 \cdot \frac{(N-1)(N-2)}{N^5}$	$\frac{1}{64}$	$\frac{13}{72}$	$\frac{1}{24}$	$\frac{7}{24}$		$\frac{59}{144}$	$\frac{35}{576}$
	$24 \cdot \frac{(N-1)(N-2)}{N^5}$		$\frac{1}{8}$	$\frac{5}{32}$	$\frac{3}{8}$		$\frac{47}{144}$	$\frac{5}{288}$
	$12 \cdot \frac{(N-1)(N-2)}{N^5}$		$\frac{7}{72}$	$\frac{5}{48}$	$\frac{1}{2}$		$\frac{41}{144}$	$\frac{1}{72}$
	$24 \cdot \frac{(N-1)(N-2)}{N^5}$	$\frac{1}{192}$	$\frac{19}{144}$	$\frac{11}{96}$	$\frac{3}{8}$		$\frac{199}{576}$	$\frac{1}{36}$
	$6 \cdot \frac{(N-1)(N-2)}{N^5}$		$\frac{1}{18}$	$\frac{5}{48}$	$\frac{7}{12}$		$\frac{1}{4}$	$\frac{1}{144}$
	$1 \cdot \frac{(N-1)(N-2)}{N^5}$				$\frac{5}{6}$		$\frac{1}{6}$	
	$8 \cdot \frac{(N-1)(N-2)}{N^5}$		$\frac{1}{12}$	$\frac{5}{32}$	$\frac{11}{24}$		$\frac{7}{24}$	$\frac{1}{96}$
	$12 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$		$\frac{7}{72}$	$\frac{1}{24}$	$\frac{1}{2}$	$\frac{1}{9}$	$\frac{35}{144}$	$\frac{1}{144}$
	$8 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$	$\frac{1}{192}$	$\frac{5}{48}$	$\frac{1}{32}$	$\frac{5}{12}$	$\frac{1}{6}$	$\frac{49}{192}$	$\frac{1}{48}$
	$3 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$				$\frac{2}{3}$	$\frac{2}{9}$	$\frac{1}{9}$	
	$12 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$		$\frac{1}{36}$	$\frac{5}{96}$	$\frac{13}{24}$	$\frac{2}{9}$	$\frac{11}{72}$	$\frac{1}{288}$
	$6 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$		$\frac{1}{18}$	$\frac{1}{24}$	$\frac{7}{12}$	$\frac{1}{9}$	$\frac{5}{24}$	
	$24 \cdot \frac{(N-1)(N-2)(N-3)}{N^5}$		$\frac{1}{18}$	$\frac{7}{96}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{29}{144}$	$\frac{1}{288}$

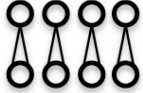
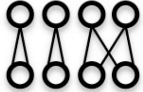
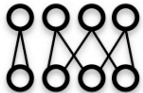
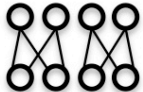









Pattern	P{Pattern}	p_{41}	p_{42}	p_{43}	p_{44}	p_{45}	p_{46}	p_{47}
	$3 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^5}$				$\frac{1}{2}$	$\frac{4}{9}$	$\frac{1}{18}$	
	$12 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^5}$		$\frac{1}{36}$	$\frac{1}{48}$	$\frac{11}{24}$	$\frac{7}{18}$	$\frac{5}{48}$	
	$\frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^5}$				$\frac{1}{3}$	$\frac{2}{3}$		
















Pattern	P{Pattern}	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$\frac{1}{N^7}$	$\frac{3}{32}$					$\frac{9}{16}$	$\frac{11}{32}$
	$8 \cdot \frac{N-1}{N^7}$	$\frac{3}{64}$	$\frac{3}{16}$				$\frac{9}{16}$	$\frac{13}{64}$
	$4 \cdot \frac{N-1}{N^7}$		$\frac{3}{8}$				$\frac{9}{16}$	$\frac{1}{16}$
	$24 \cdot \frac{N-1}{N^7}$	$\frac{3}{128}$	$\frac{3}{16}$	$\frac{9}{64}$			$\frac{33}{64}$	$\frac{17}{128}$
	$24 \cdot \frac{N-1}{N^7}$		$\frac{3}{16}$	$\frac{9}{32}$			$\frac{15}{32}$	$\frac{1}{16}$
	$32 \cdot \frac{N-1}{N^7}$	$\frac{3}{256}$	$\frac{3}{16}$	$\frac{27}{128}$			$\frac{63}{128}$	$\frac{25}{256}$
	$3 \cdot \frac{N-1}{N^7}$			$\frac{9}{16}$			$\frac{3}{8}$	$\frac{1}{16}$
	$24 \cdot \frac{N-1}{N^7}$		$\frac{3}{16}$	$\frac{9}{32}$			$\frac{15}{32}$	$\frac{1}{16}$
	$8 \cdot \frac{N-1}{N^7}$	$\frac{3}{256}$	$\frac{3}{16}$	$\frac{27}{128}$			$\frac{63}{128}$	$\frac{25}{256}$

Pattern	$P\{\text{Pattern}\}$	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$4 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{8}$				$\frac{9}{16}$	$\frac{1}{16}$
	$24 \cdot \frac{(N-1)(N-2)}{N^7}$	$\frac{3}{128}$	$\frac{3}{16}$		$\frac{3}{16}$		$\frac{31}{64}$	$\frac{15}{128}$
	$24 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{16}$		$\frac{3}{8}$		$\frac{13}{32}$	$\frac{1}{32}$
	$48 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{16}$	$\frac{9}{64}$	$\frac{3}{16}$		$\frac{7}{16}$	$\frac{3}{64}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$	$\frac{3}{256}$	$\frac{9}{64}$	$\frac{9}{128}$	$\frac{9}{32}$		$\frac{27}{64}$	$\frac{19}{256}$
	$24 \cdot \frac{(N-1)(N-2)}{N^7}$			$\frac{9}{32}$	$\frac{3}{8}$		$\frac{5}{16}$	$\frac{1}{32}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{32}$	$\frac{9}{64}$	$\frac{3}{8}$		$\frac{23}{64}$	$\frac{1}{32}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{9}{64}$	$\frac{9}{64}$	$\frac{9}{32}$		$\frac{51}{128}$	$\frac{5}{128}$
	$64 \cdot \frac{(N-1)(N-2)}{N^7}$	$\frac{3}{512}$	$\frac{9}{64}$	$\frac{27}{256}$	$\frac{9}{32}$		$\frac{105}{256}$	$\frac{29}{512}$
	$6 \cdot \frac{(N-1)(N-2)}{N^7}$				$\frac{3}{4}$		$\frac{1}{4}$	
	$12 \cdot \frac{(N-1)(N-2)}{N^7}$			$\frac{9}{32}$	$\frac{3}{8}$		$\frac{5}{16}$	$\frac{1}{32}$
	$48 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{32}$		$\frac{9}{16}$		$\frac{21}{64}$	$\frac{1}{64}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{32}$	$\frac{9}{64}$	$\frac{3}{8}$		$\frac{23}{64}$	$\frac{1}{32}$

Pattern	$P\{\text{Pattern}\}$	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$48 \cdot \frac{(N-1)(N-2)}{N^7}$	$\frac{3}{512}$	$\frac{3}{32}$	$\frac{27}{256}$	$\frac{3}{8}$		$\frac{95}{256}$	$\frac{25}{512}$
	$24 \cdot \frac{(N-1)(N-2)}{N^7}$				$\frac{3}{4}$		$\frac{1}{4}$	
	$48 \cdot \frac{(N-1)(N-2)}{N^7}$			$\frac{9}{64}$	$\frac{9}{16}$		$\frac{9}{32}$	$\frac{1}{64}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{64}$	$\frac{9}{64}$	$\frac{15}{32}$		$\frac{41}{128}$	$\frac{3}{128}$
	$16 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{32}$		$\frac{9}{16}$		$\frac{21}{64}$	$\frac{1}{64}$
	$96 \cdot \frac{(N-1)(N-2)}{N^7}$		$\frac{3}{32}$	$\frac{9}{64}$	$\frac{3}{8}$		$\frac{23}{64}$	$\frac{1}{32}$
	$24 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{16}$		$\frac{3}{8}$		$\frac{13}{32}$	$\frac{1}{32}$
	$32 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$	$\frac{3}{256}$	$\frac{9}{64}$		$\frac{9}{32}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{17}{256}$
	$12 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{3}{4}$		$\frac{1}{4}$	
	$24 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{64}$	$\frac{9}{16}$		$\frac{9}{32}$	$\frac{1}{64}$
	$192 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{32}$	$\frac{9}{128}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{3}{128}$
	$48 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{32}$		$\frac{9}{16}$		$\frac{21}{64}$	$\frac{1}{64}$
	$48 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{32}$		$\frac{3}{8}$	$\frac{1}{4}$	$\frac{17}{64}$	$\frac{1}{64}$

Pattern	$P\{\text{Pattern}\}$	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$96 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$	$\frac{3}{512}$	$\frac{3}{32}$	$\frac{9}{256}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{83}{256}$	$\frac{21}{512}$
	$24 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{3}{4}$		$\frac{1}{4}$	
	$48 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{64}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{7}{32}$	$\frac{1}{64}$
	$96 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{64}$	$\frac{9}{128}$	$\frac{15}{32}$	$\frac{1}{8}$	$\frac{35}{128}$	$\frac{1}{64}$
	$16 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{32}$		$\frac{9}{16}$		$\frac{21}{64}$	$\frac{1}{64}$
	$96 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{32}$	$\frac{9}{128}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{5}{16}$	$\frac{3}{128}$
	$24 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$96 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{9}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	
	$48 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{64}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{7}{32}$	$\frac{1}{64}$
	$192 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{128}$	$\frac{9}{16}$	$\frac{1}{8}$	$\frac{15}{64}$	$\frac{1}{128}$
	$96 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{64}$		$\frac{15}{32}$	$\frac{1}{4}$	$\frac{29}{128}$	$\frac{1}{128}$
	$192 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{64}$	$\frac{9}{128}$	$\frac{15}{32}$	$\frac{1}{8}$	$\frac{35}{128}$	$\frac{1}{64}$
	$192 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$		$\frac{3}{64}$	$\frac{9}{128}$	$\frac{15}{32}$	$\frac{1}{8}$	$\frac{35}{128}$	$\frac{1}{64}$

Pattern	P{Pattern}	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$1 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$					1		
	$12 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$32 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$				$\frac{9}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	
	$12 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{64}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{7}{32}$	$\frac{1}{64}$
	$48 \cdot \frac{(N-1)(N-2)(N-3)}{N^7}$			$\frac{9}{128}$	$\frac{9}{16}$	$\frac{1}{8}$	$\frac{15}{64}$	$\frac{1}{128}$
	$6 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{3}{4}$		$\frac{1}{4}$	
	$48 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$		$\frac{3}{32}$		$\frac{3}{8}$	$\frac{1}{4}$	$\frac{17}{64}$	$\frac{1}{64}$
	$16 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$	$\frac{3}{512}$	$\frac{3}{32}$		$\frac{9}{32}$	$\frac{5}{16}$	$\frac{69}{256}$	$\frac{19}{512}$
	$48 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$96 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{9}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	
	$96 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$			$\frac{9}{128}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{11}{64}$	$\frac{1}{128}$
	$96 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$		$\frac{3}{64}$		$\frac{15}{32}$	$\frac{1}{4}$	$\frac{29}{128}$	$\frac{1}{128}$
	$192 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$		$\frac{3}{64}$	$\frac{9}{256}$	$\frac{3}{8}$	$\frac{5}{16}$	$\frac{7}{32}$	$\frac{3}{256}$

Pattern	$P\{\text{Pattern}\}$	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$32 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$		$\frac{3}{64}$		$\frac{9}{32}$	$\frac{1}{2}$	$\frac{21}{128}$	$\frac{1}{128}$
	$4 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$					1		
	$24 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{3}{16}$	$\frac{3}{4}$	$\frac{1}{16}$	
	$24 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$96 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$48 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$			$\frac{9}{128}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{11}{64}$	$\frac{1}{128}$
	$32 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$				$\frac{9}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	
	$192 \cdot \frac{(N-1)(N-2)(N-3)(N-4)}{N^7}$			$\frac{9}{256}$	$\frac{15}{32}$	$\frac{5}{16}$	$\frac{23}{128}$	$\frac{1}{256}$
	$6 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$					1		
	$48 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$				$\frac{3}{16}$	$\frac{3}{4}$	$\frac{1}{16}$	
	$12 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$48 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$96 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$			$\frac{9}{256}$	$\frac{9}{32}$	$\frac{9}{16}$	$\frac{15}{128}$	$\frac{1}{256}$
	$24 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$				$\frac{3}{8}$	$\frac{1}{2}$	$\frac{1}{8}$	
	$32 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)}{N^7}$		$\frac{3}{64}$		$\frac{9}{32}$	$\frac{1}{2}$	$\frac{21}{128}$	$\frac{1}{128}$




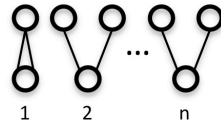
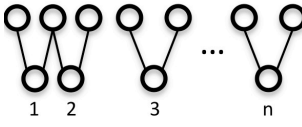
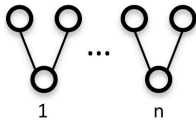
Pattern	P{Pattern}	p_{51}	p_{52}	p_{53}	p_{54}	p_{55}	p_{56}	p_{57}
	$4 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)(N-6)}{N^7}$					1		
	$24 \cdot \frac{(N-1)(N-2)(N-3)(N-4)(N-5)(N-6)}{N^7}$				$\frac{3}{16}$	$\frac{3}{4}$	$\frac{1}{16}$	
	$\frac{(N-1)(N-2)(N-3)(N-4)(N-5)(N-6)(N-7)}{N^7}$					1		

Table S2.4 Patterns of ancestry that are $O(1)$ and $O(1/N)$ used in constructing the coalescent process for an arbitrary sample size n .

Pattern	P{Pattern}	Number of Parents	P{Coalescence Pattern}
	$n \frac{1}{N} + O\left(\frac{1}{N^2}\right)$	$2n - 1$	0
	$\left[\binom{2n}{2} - n \right] \frac{1}{N} + O\left(\frac{1}{N^2}\right)$	$2n - 1$	$\frac{1}{16}$
	$1 - \binom{2n}{2} \frac{1}{N} + O\left(\frac{1}{N^2}\right)$	$2n$	0

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

Supplementary Figures

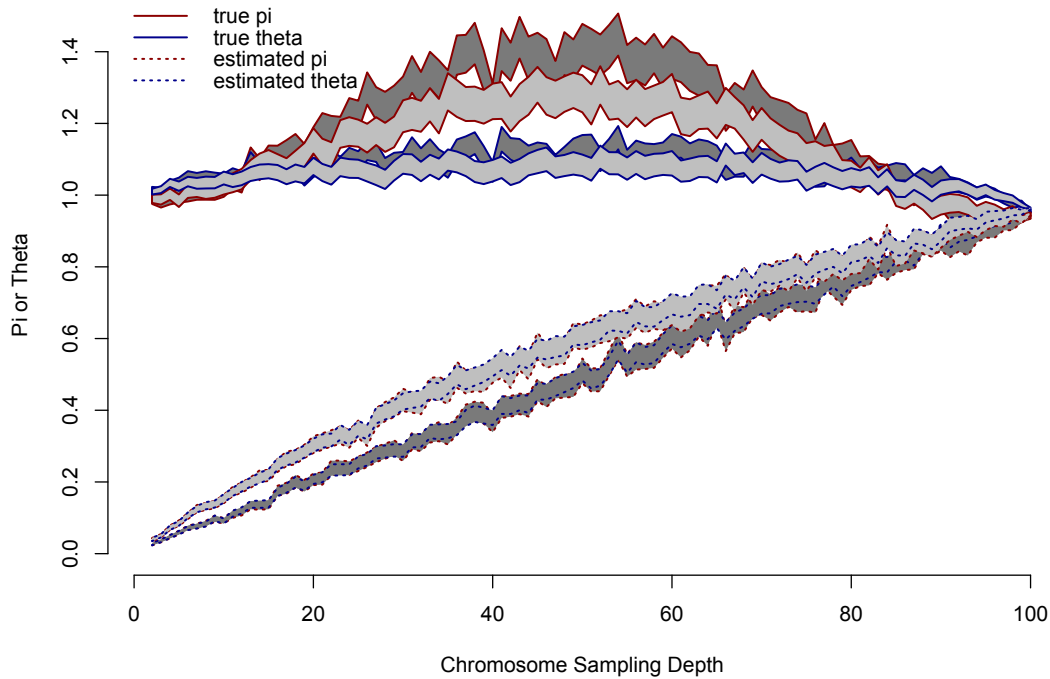


Figure S3.1 True and estimated values of π and θ_w as a function of the chromosome sampling depth for $\theta = \rho = 0.01/\text{bp}$. Light gray regions show the 95% bootstrap percentile confidence intervals (1000 simulations) for simulations with recombination, and dark gray regions are from simulations without recombination. In the absence of recombination, true values of summary statistics vary more as a function of the chromosome sampling depth and estimated values are lower.

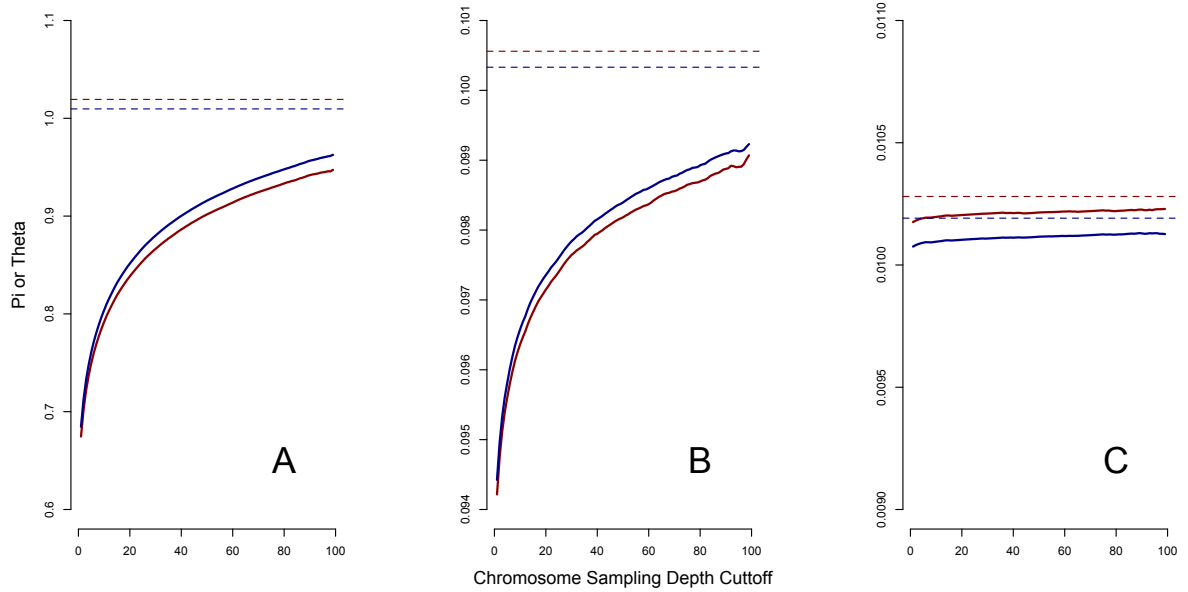


Figure S3.2 Mean values of π_e (solid red) and θ_e (solid blue) for loci with different sampling depth cutoffs (i.e. loci that have at least the specified number of sampled chromosomes with intact recognition sequences). Dashed lines represent the true simulation average for π (red) and θ (blue). **(A)** $\theta=\rho=0.01/\text{bp}$, **(B)** $\theta=\rho=0.001/\text{bp}$, **(C)** $\theta=\rho=0.0001/\text{bp}$. Since 100bp sequences were analyzed, averages on the y-axis are 100 times greater than the per bp parameter values.

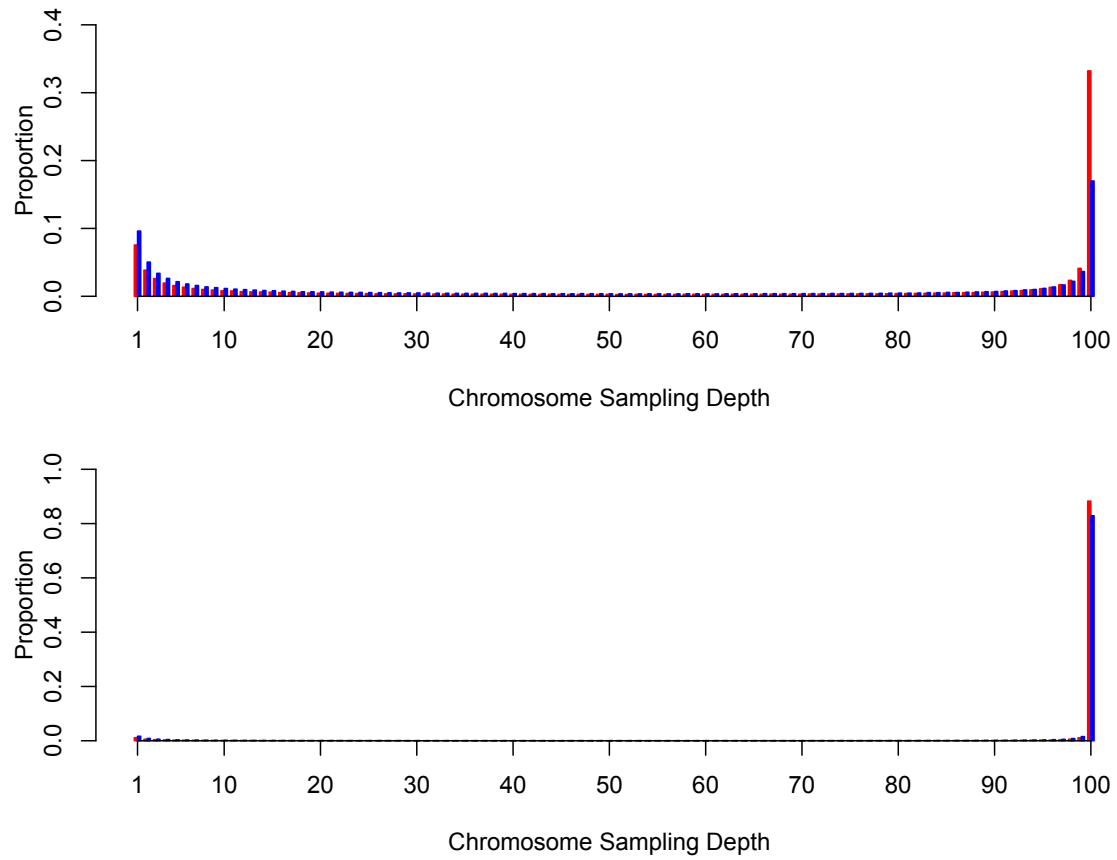


Figure S3.3 Chromosome sampling depth proportions for the standard (red) and double digest (blue) RADseq protocols. The upper graph shows sampling depths for $\theta = \rho = 0.01/\text{bp}$ and the lower graph for $\theta = \rho = 0.001/\text{bp}$.

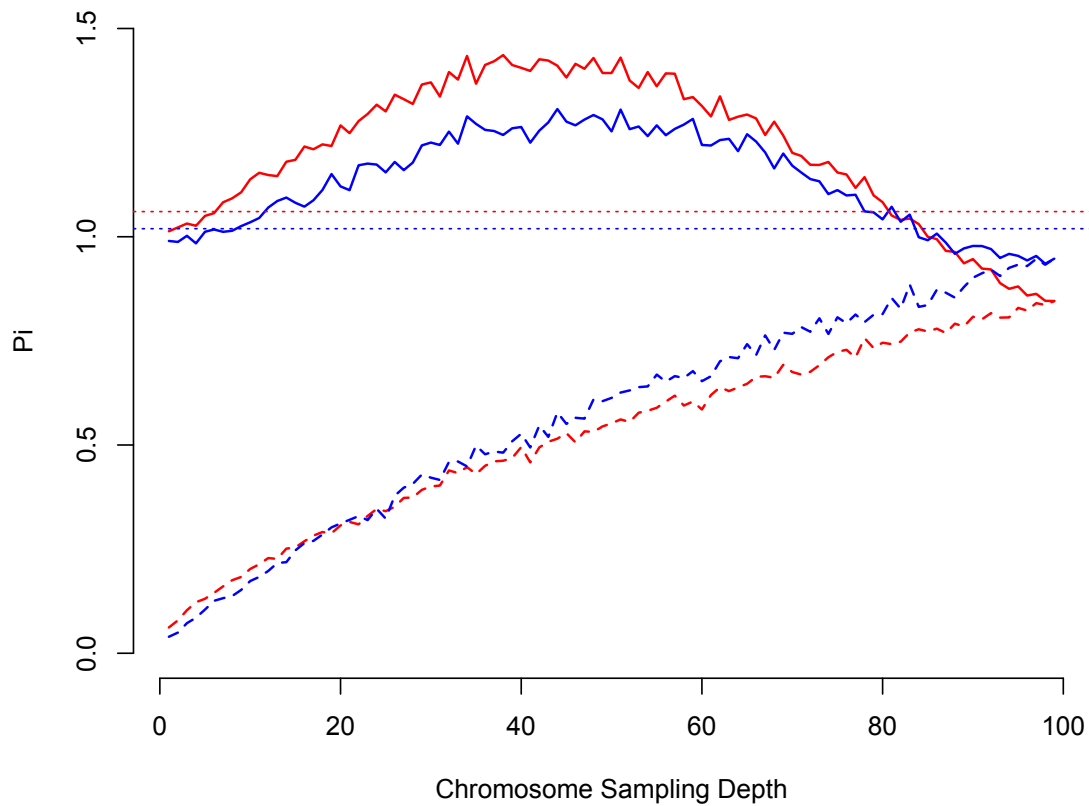


Figure S3.4 True π (solid lines) and estimated π (dashed lines) vary as a function of chromosome sampling depth for the standard (blue) and double-digest (red) RADseq protocols. Dotted lines represent the true simulation averages of π . Loci with higher chromosome sampling depths (*i.e.* near 100) have true and estimated values of π that are below the true simulation average, especially for the double-digest RADseq protocol.

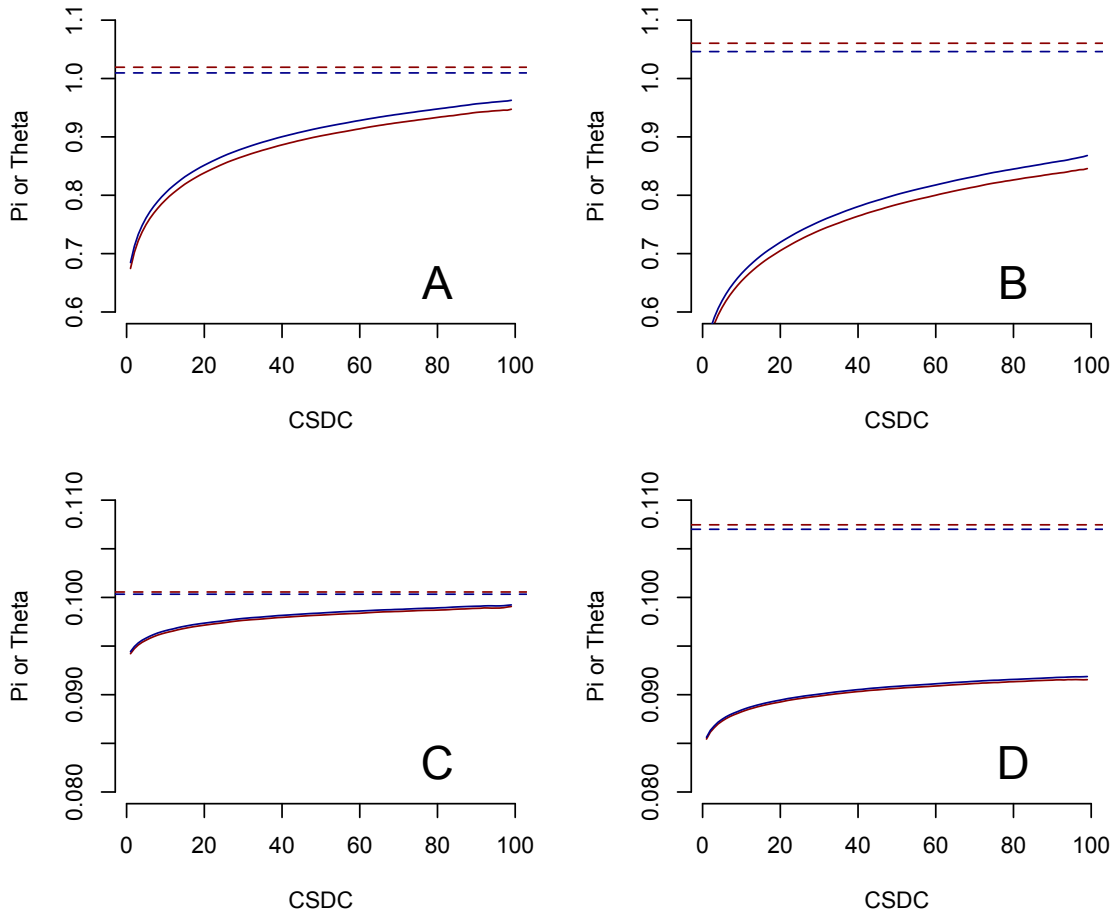


Figure S3.5 Mean values of π_e (solid red) and θ_e (solid blue) for loci with different sampling depth cutoffs (CSDC, *i.e.* loci that have at least the specified number of sampled chromosomes with intact recognition sequences). Dashed lines represent the true simulation average for π (red) and θ (blue). **(A)** Standard RADseq with $\theta=\rho=0.01/\text{bp}$, **(B)** double digest RADseq with $\theta=\rho=0.01/\text{bp}$, **(C)** standard RADseq with $\theta=\rho=0.001/\text{bp}$, **(D)** double digest with $\theta=\rho=0.001/\text{bp}$. Since 100bp sequences were analyzed, averages on the y-axis are 100 times greater than the per bp parameter values.

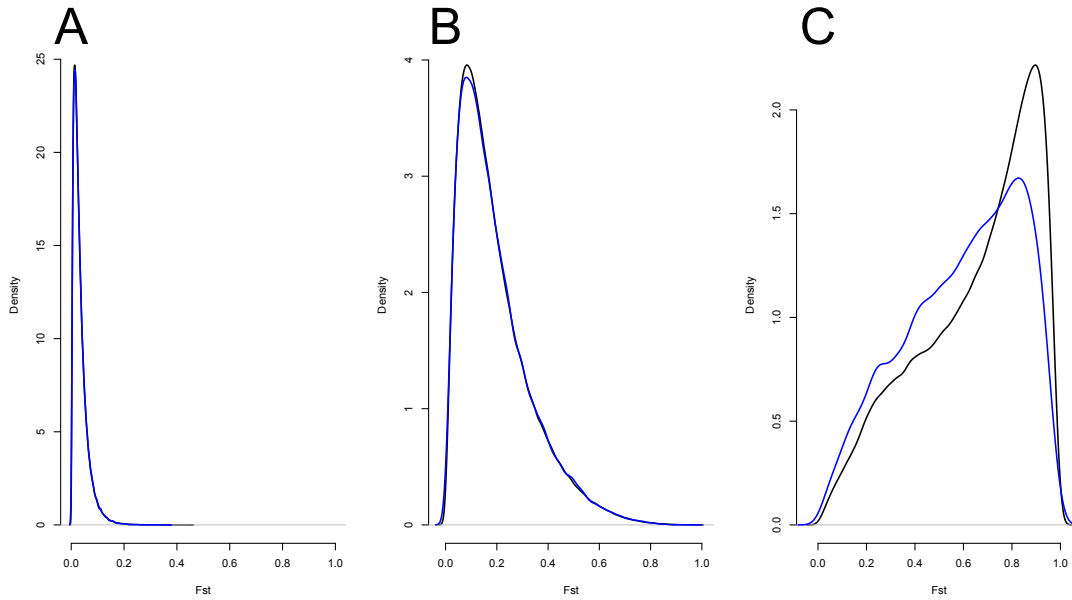


Figure S3.6 Distribution of estimated F_{st} when all haplotypes are sampled (blue) versus the true distribution (black), for $N_m=10$ (A), $N_m=1$ (B), and $N_m=0.1$ (C).

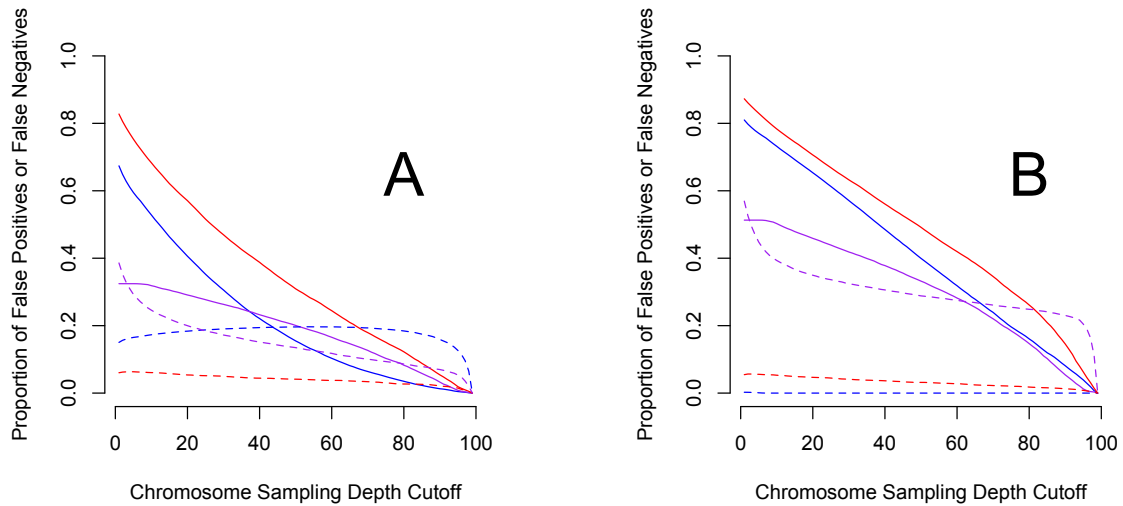


Figure S3.7 Proportion of estimated π (red), θ_w (blue), or D (purple) 5% outlier loci that are false positives (solid lines) or false negatives (dashed lines) relative to the true distribution for different chromosome sampling depth cutoffs. (A) Standard RADseq protocol. (B) Double digest protocol.

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Supplementary Text

Generation of double-digest RADseq dataset

We used only one methylation-insensitive restriction enzyme (HpyCH4V) to generate blunt-end DNA fragments. We constructed adapters to have T overhangs and custom barcodes as described in Peterson *et al.* (1) and combined the A-tailing step with the adapter-ligation step. We sequenced DNA libraries on an Illumina HiSeq 2500 at the Harvard University Center for Systems Biology core facility. We used custom Perl scripts to sort sequence reads according to adapter barcodes.

GATK filter expression for VariantFiltration

RADseq data

```
"QD < 2.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSumTest < -12.5" --filterName "TET.3.21.14"
```

Individual whole-genome sequences

```
"QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MappingQualityRankSumTest < -12.5 || ReadPosRankSum < -8.0" --filterName "TET.2.10.14"
```

Filtering loci with spuriously mapped sequence reads

Allele frequency spectra for all populations, both diploid and tetraploid, exhibited an enrichment of alleles at a frequency of 50% in the sample, which can arise from reads mapped to duplicate loci, or in the case of tetraploids, from chromosome pairing preferences. In the former case, we expect read coverage to be higher at these

sites. Consequently, we created an additional data filtration step to exclude loci at which reads mapped from potentially paralogous loci as follows: Using 8 diploid whole-genome sequences for populations D2 and D3, we targeted loci with erroneously mapped reads by identifying annotated genes or intergenic segments (no longer than 2kb) in which all 8 diploids were heterozygous at more than 2 sites within the defined region. We excluded these loci from all downstream analyses, as it is unlikely for all 8 diploids from two distinct populations to be heterozygous at 3 or more sites within a gene or intergenic segment according to Hardy-Weinberg equilibrium. We used only diploid genomes for this analysis in order to prevent ascertainment bias against potentially diploidized loci in the autotetraploids.

Concordance between datasets

Despite the different DNA library preparations and variant calling methods, allele frequency estimates from the RADseq and PoolSeq datasets are highly correlated with the same polymorphic sites identified in the IndSeq dataset (Pearson's $r \sim 80\%$, Table S4.2). However, estimates of genetic diversity differ (Table S4.3). As expected, RADseq underestimates diversity relative to IndSeq (2), while PoolSeq estimates are only slightly elevated. The difference between $\theta\pi$ and θ_w (as calculated in (3) and (4), respectively) is larger for RADseq and PoolSeq than IndSeq, which may be attributable to a variety of differences between the datasets (i.e. sample sizes) or the way they are bioinformatically processed. For example, large-sample SNP calling with GATK may cause a slight excess of intermediate-frequency alleles (5) while pooled data creates

challenges for calling low-frequency SNPs since sequencing errors occur at similar frequencies (6).

References

1. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7. doi:10.1371
2. Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22, 3179–3190.
3. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123, 585–595.
4. Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7, 256–276.
5. Nevado, B., Ramos-Onsins, S. E., & Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology*, 23, 1764–1779.
6. Cutler, D. J., & Jensen, J. D. (2010). To pool, or not to pool? *Genetics*, 186(1), 41–3.

Supplementary Figures

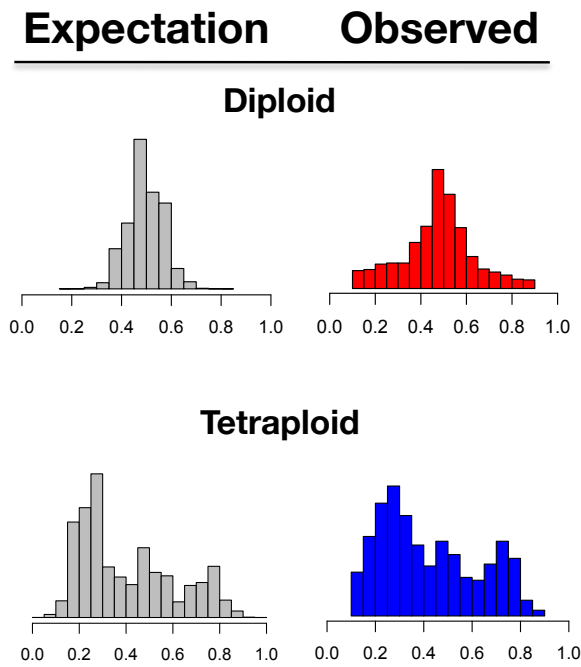


Figure S4.1 Ploidy assessment of individuals. Expected (gray) and observed (colored) histograms of raw, non-reference base counts for diploids (red) and tetraploids (blue) at all polymorphic sites with at least 50X sequencing depth. The x-axis of each histogram is the frequency of a non-reference base within an individual. See main text for further details.

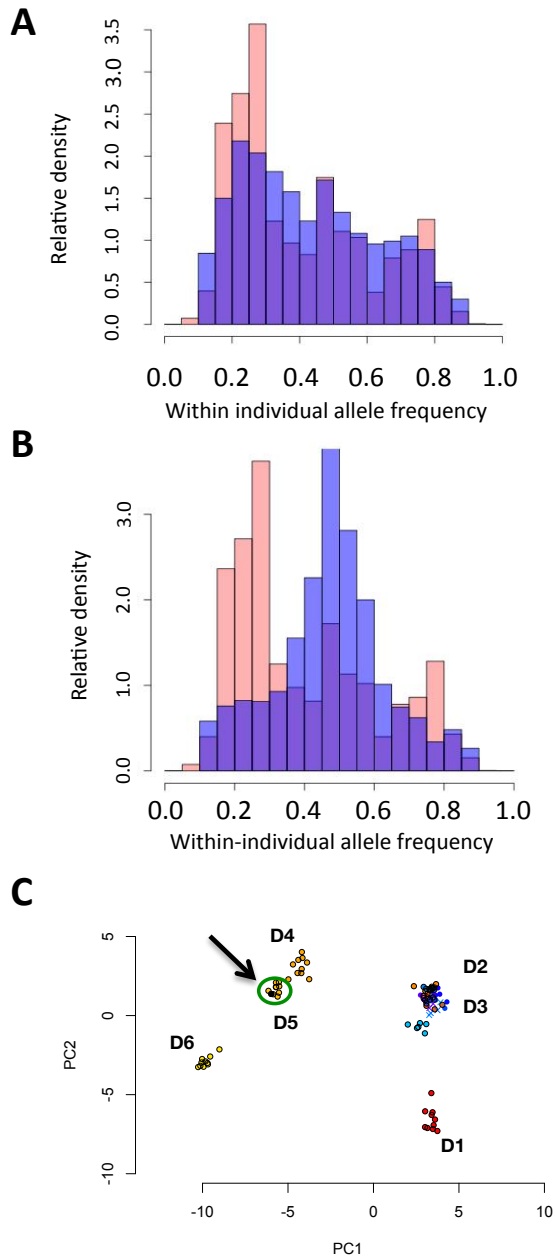
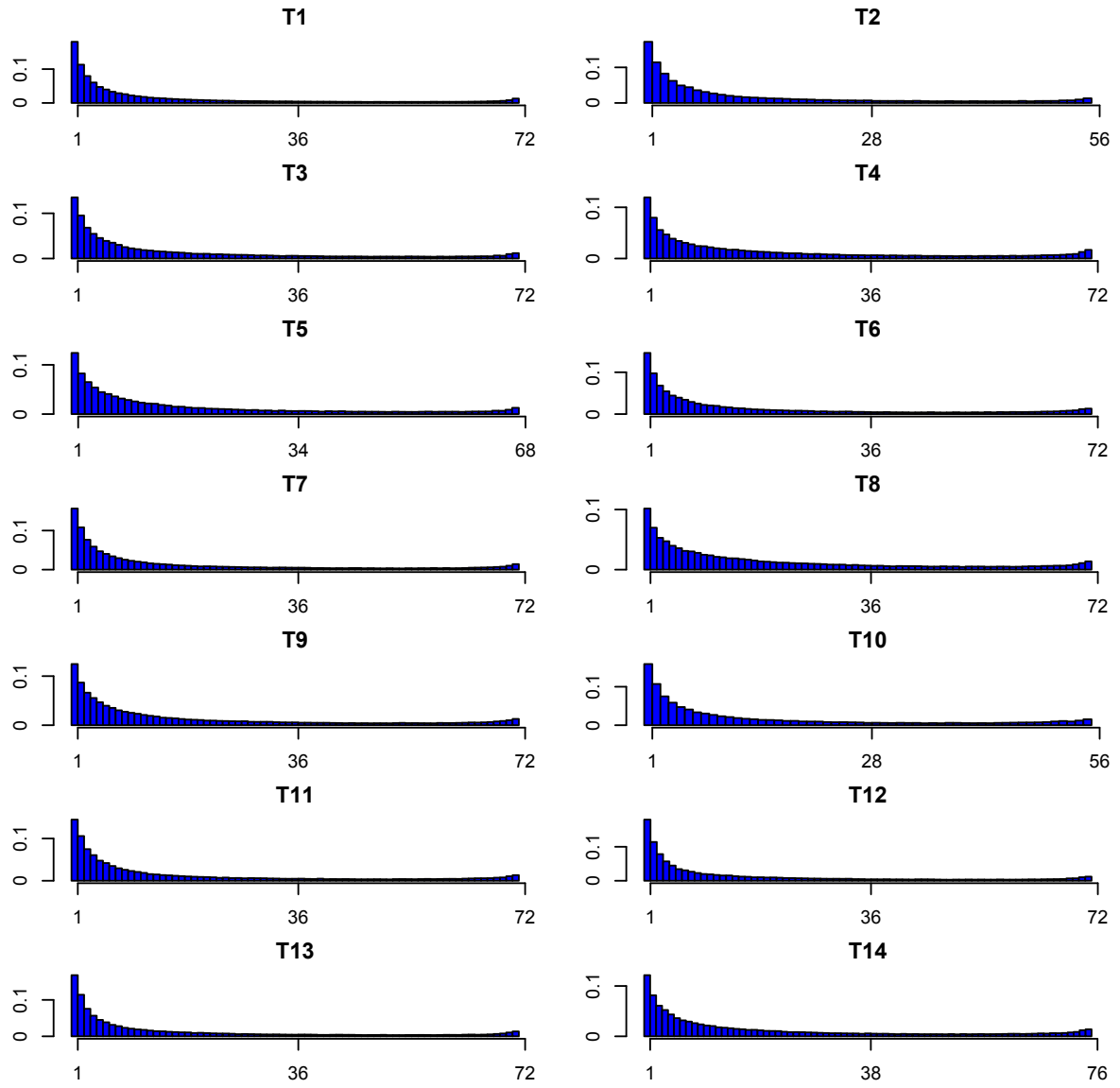


Figure S4.2 Candidate neoautotetraploid in population D5. An individual collected from population D5 may be a spontaneously arisen autotetraploid. **(A)** Histogram of unfiltered, non-reference base counts for individual D5-21 (blue) superimposed on the expected distribution for an autotetraploid individual (red). **(B)** The same set of histograms for another individual from population D5, which show the sample is diploid. All other individuals from population D5 resembled this pattern with the exception of D5-21. **(C)** Principal Component Analysis shows D5-21 (black dot near arrow) is closely related to other diploids from population D5 (circled in green).

A



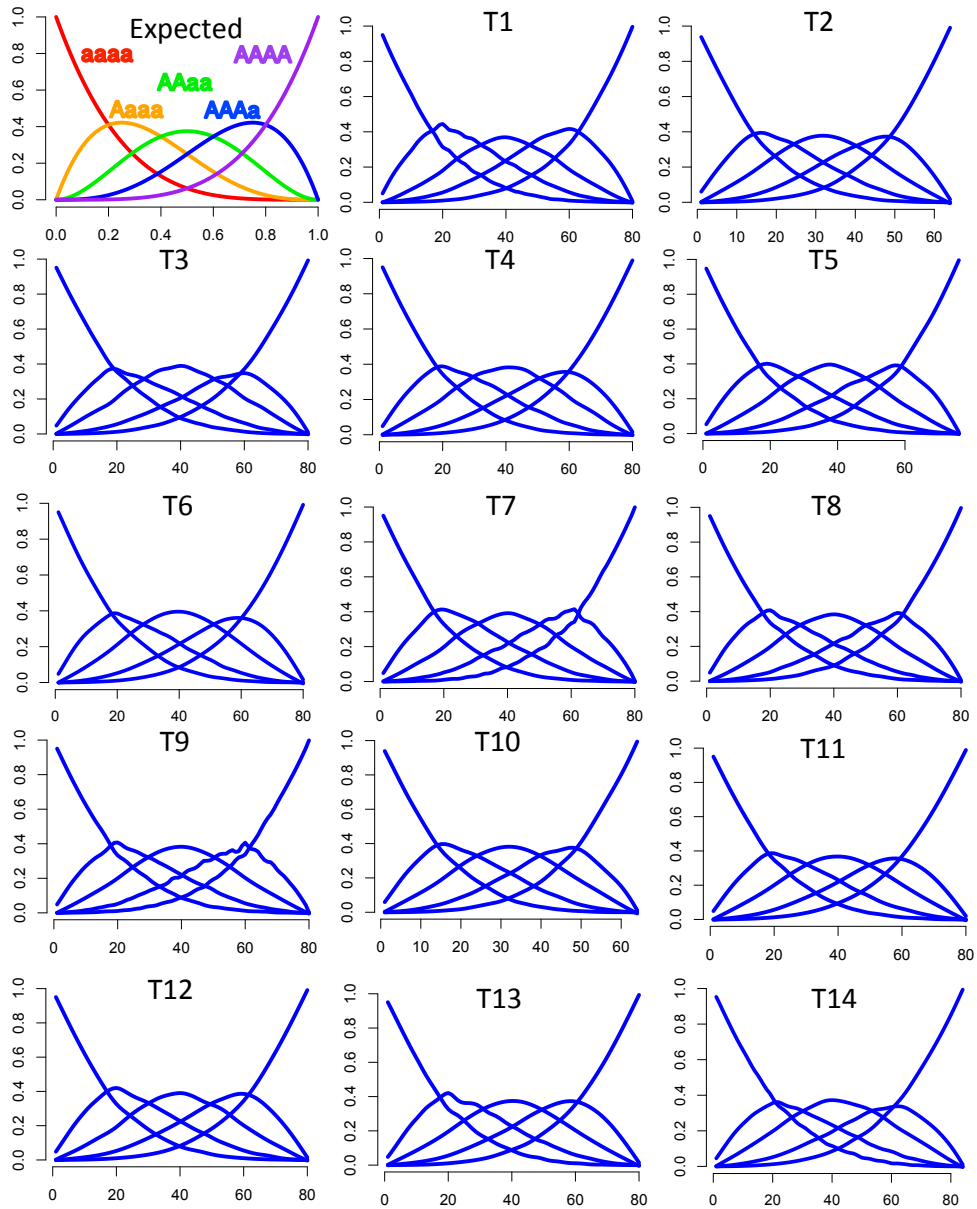
B

Figure S4.3 Population genetic data of autotetraploids consistent with tetrasomic inheritance. **(A)** Allele frequency spectra from autotetraploid populations do not display an enrichment of intermediate-frequency alleles at 50% for RADseq datasets. **(B)** Genotype proportions (y-axis) as a function of allele frequency (x-axis) for each tetraploid population along with expected values under tetrasomic inheritance. These data are consistent with autotetraploidy and tetrasomic inheritance in *A. arena*.

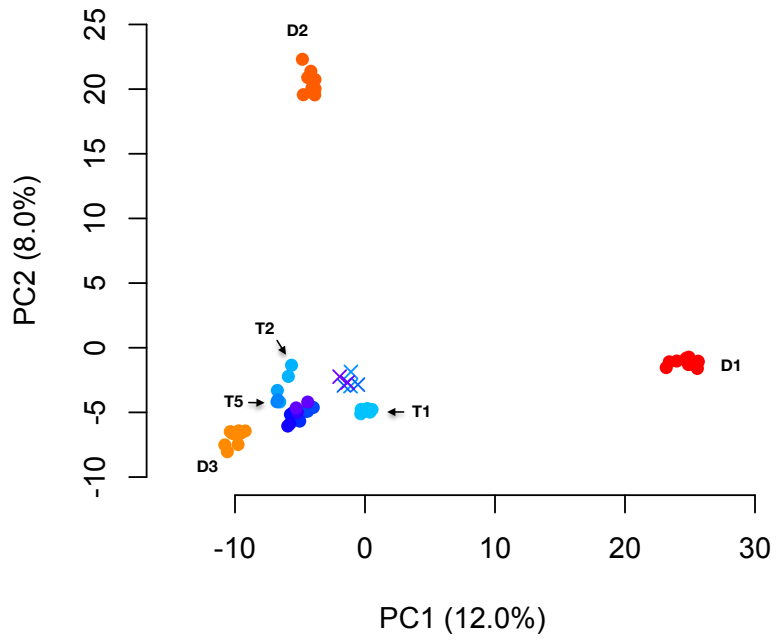


Figure S4.4 PCA of D1,D2, and D3 with subsample of tetraploids. PCA of 30 Carpathain diploids (D1-D3) and the same number of tetraploids from across the range.

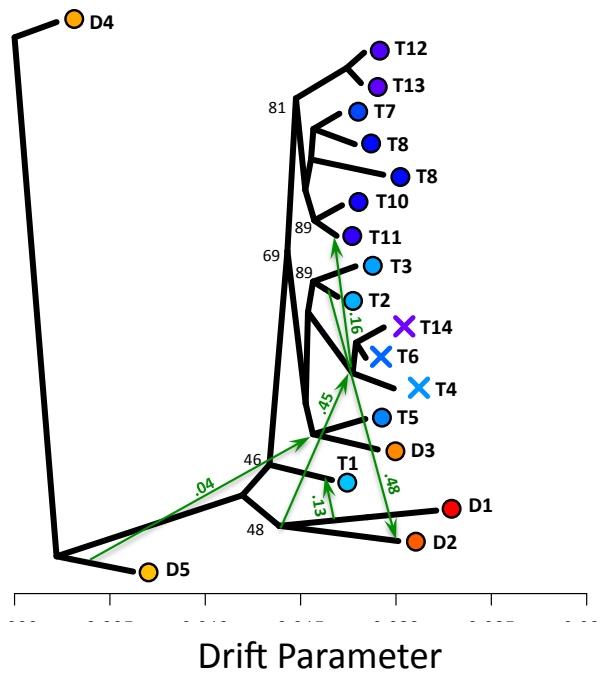


Figure S4.5 *Treemix*¹ population graph for the RADseq dataset. Only bootstrap values below 90% are shown. Adding 5 migration edges (green arrows) allowed the RADseq population graph to fit the data best. This graph suggests a single origin of the tetraploid race, but bootstrap support for important nodes are low. However, there appears to be extensive interploidy admixture between diploid and tetraploid populations that are geographically close.

1. Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, 8. doi:10.1371.

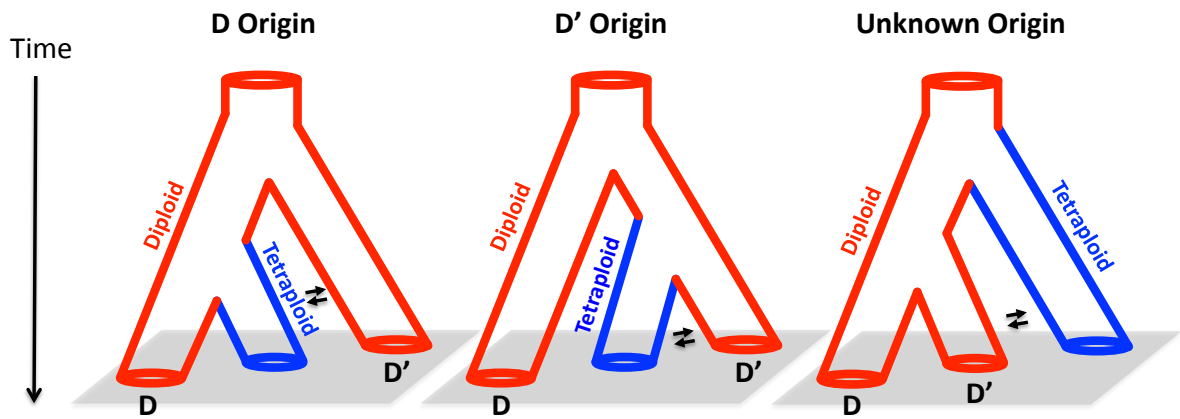
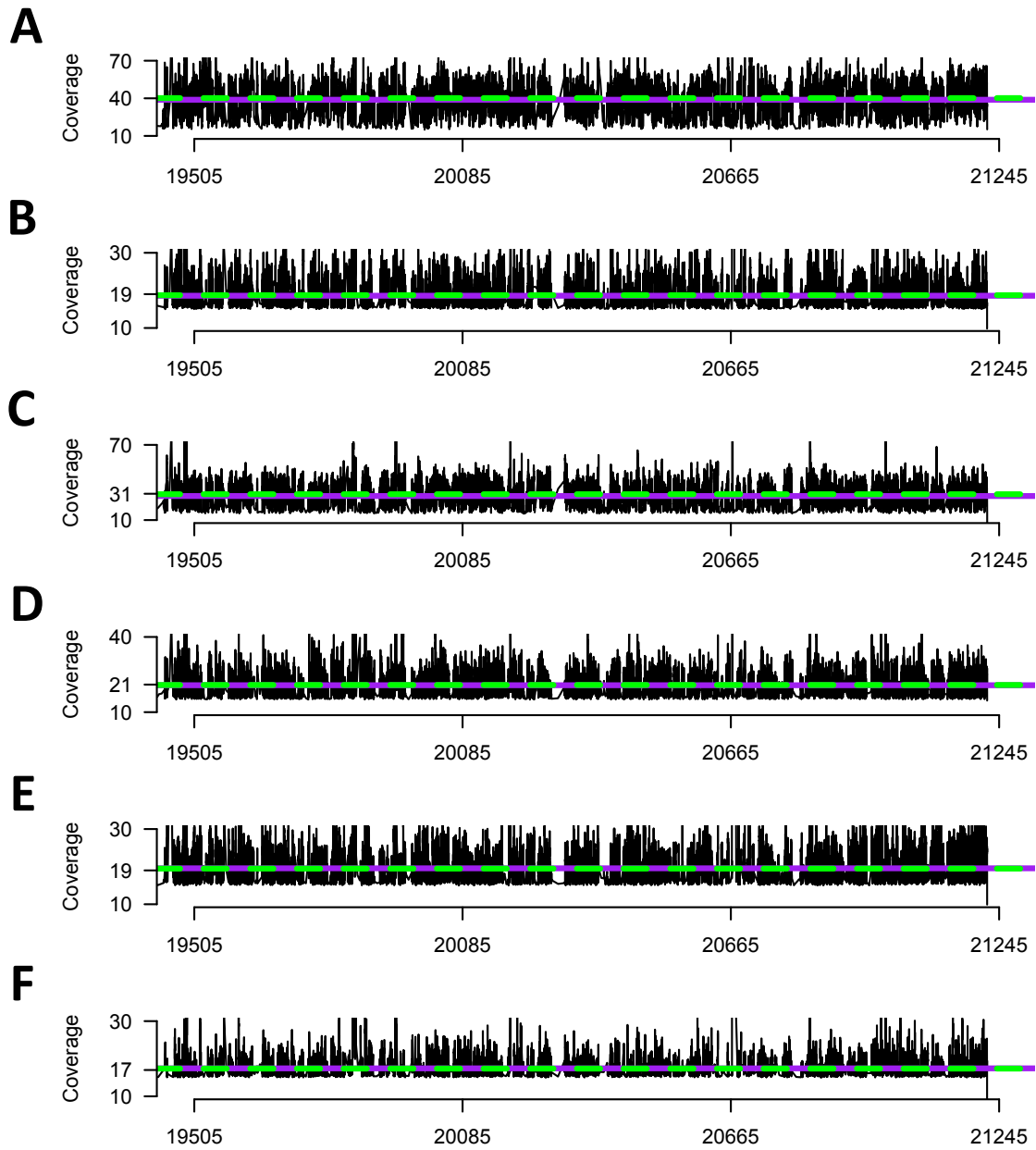


Figure S4.6 Population trio analysis of two diploids and one tetraploid. The geographic origin of the tetraploid race may be inferred if certain models explain the data better and suggest tetraploids are more closely related to particular diploid gene pools (here, either D or D'). However, the tetraploid may be more distantly related to the diploids (far right), which arose from an ancestral population with an unknown geographic location.



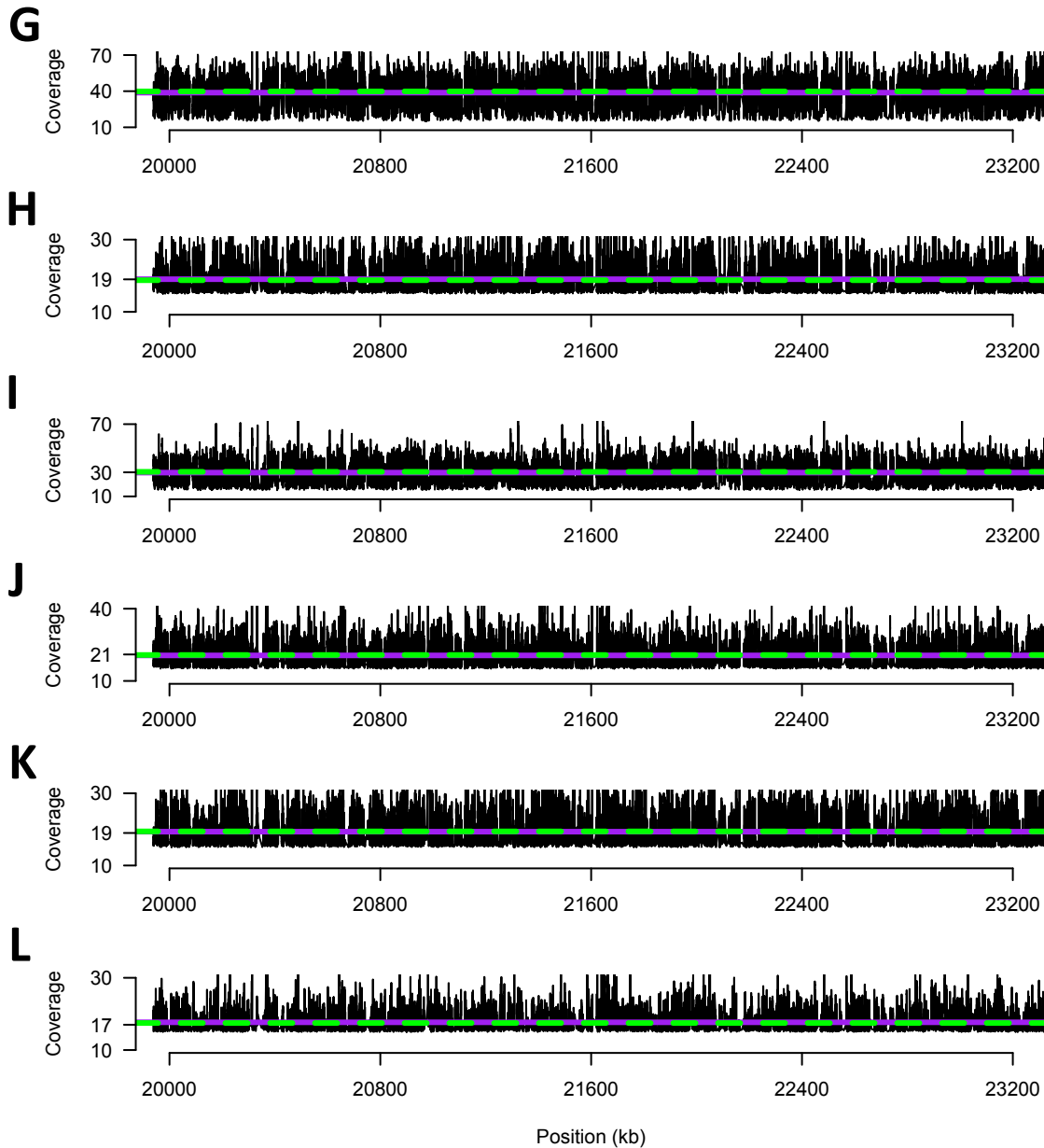


Figure S4.7 Local sequencing depths for regions potentially under selection. Sequencing depths for D1 (A), D2 (B), D3 (C), T1 (D), T5 (E), and T7 (F) within the region shown in Figure 6. Each point represents the average sequencing depth of a 50bp window using only sites with a depth of at least 15x (the sequencing depth cutoff used for analyses). The solid purple line is the median genome-wide depth for 50bp windows, while the dashed green line is the median for the region shown in Figure 6. Likewise, (G-L) are sequencing depths for the same populations, in the same order (D1 (G), D2 (H), D3 (I), T1 (J), T5 (K), and T7 (L)), for the region shown in Figure S9.

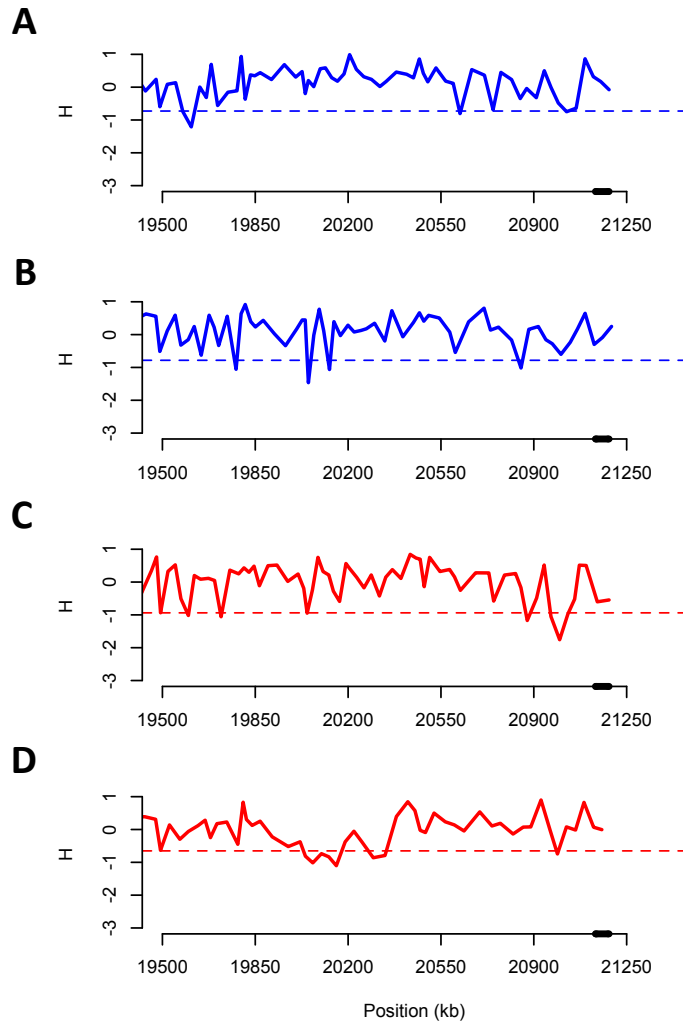


Figure S4.8 Fay and Wu's H near end of chromosome 5 for other populations. Fay and Wu's H calculated for 100 SNP windows near the end of chromosome 5 for T7 (**A**), T5 (**B**), D3 (**C**), and D2 (**D**). The black line on the x-axis corresponds to the same region highlighted in Figure 6, and the dashed line represents the 5% quantile for each distribution.

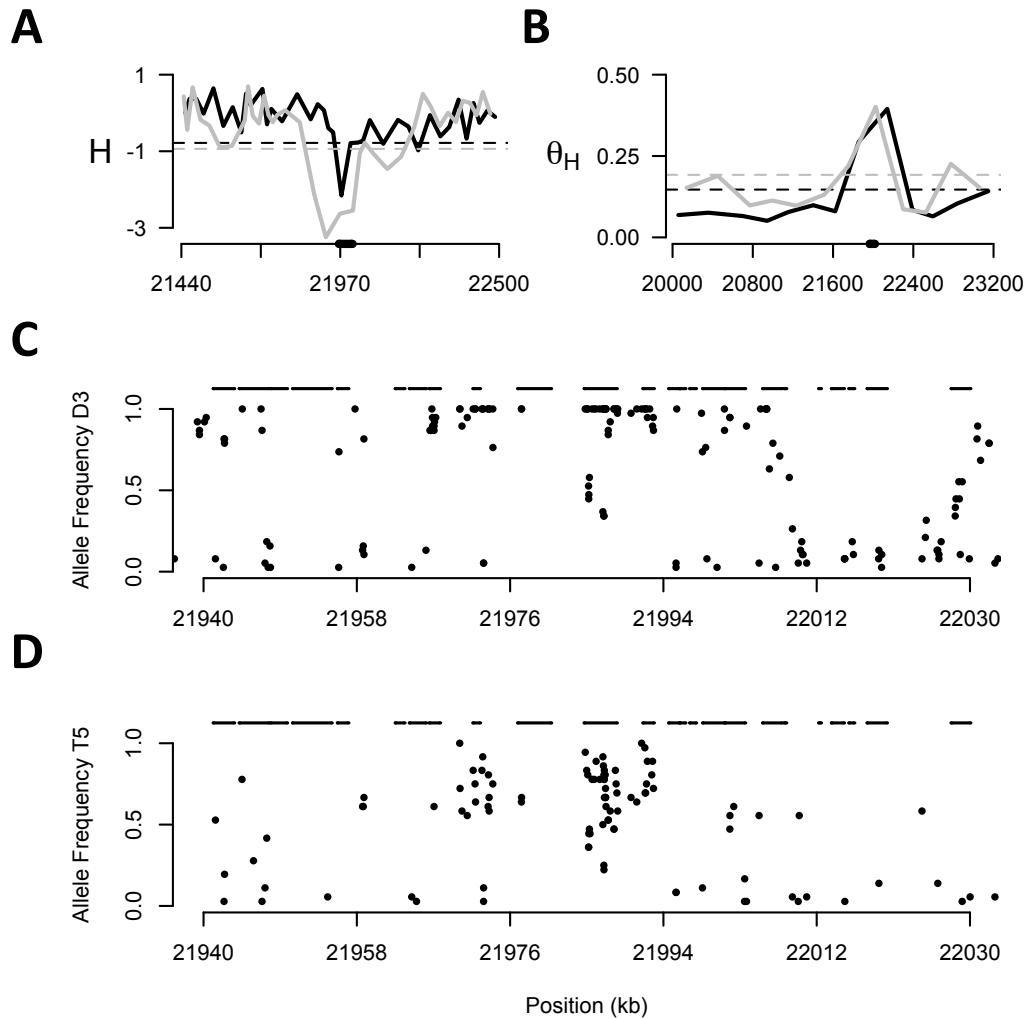


Figure S4.9 Evidence of selection after admixture between D3 and T5. **(A)** Fay and Wu's H for 100 SNP windows is below the 5% quantile (dashed line) for both D3 (gray) and T5 (black) on chromosome 4. **(B)** θ_H calculated in 50 SNP windows, using only shared variation unique to D1 and T1. Both D1 (gray) and T1 (black) have an excess of high frequency-shared variation. Allele frequency plots for the region spanning the black line on the x-axis show a strong enrichment of geographically-unique, high-frequency shared variation for D1 **(C)** and T1 **(D)**. In panel **(C)**, alleles in D3 that are also present in T5 but absent from another Carpathian diploid (D3) are highlighted. These alleles were allowed to be present in D2 due to its close relation to D3. Likewise, for panel **(D)**, alleles in T5 are highlighted if present in D3 but absent from other tetraploids (T1, T7). Thus, many of the high-frequency alleles in these admixed populations are shared and not present in other populations or the same ploidy. Many of these variants fall within genic regions (black lines above **C** and **D**).

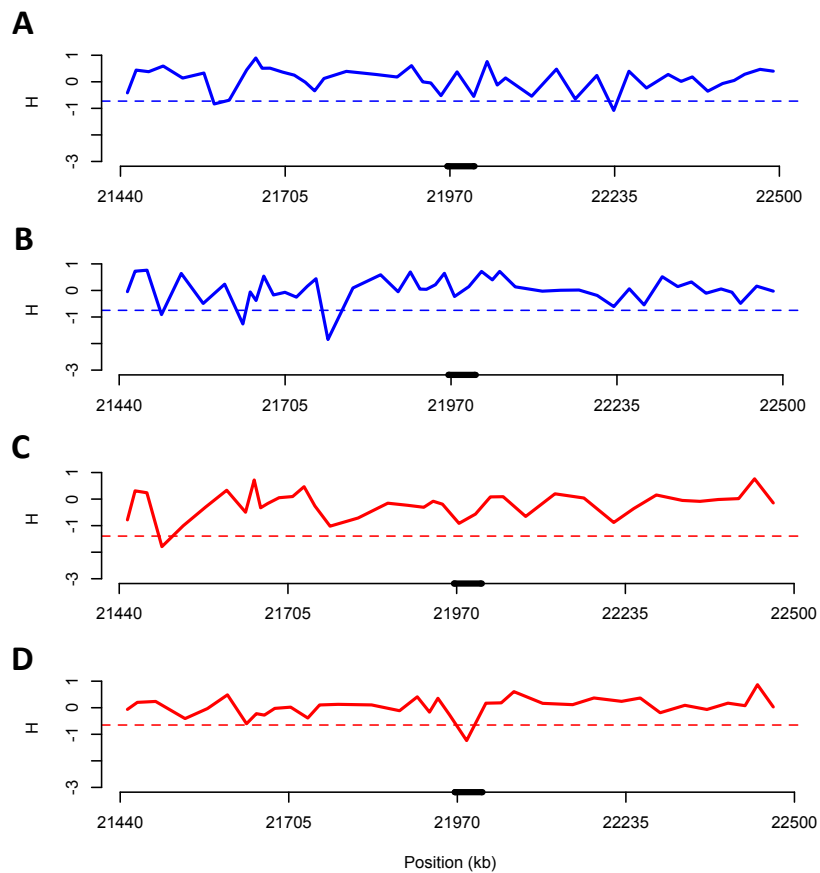


Figure S4.10 Fay and Wu's H on chromosome 4 for other populations. Fay and Wu's H calculated for 100 SNP windows on chromosome 4 for T7 (**A**), T1 (**B**), D1 (**C**), and D2 (**D**). The black line on the x-axis corresponds to the same region highlighted in Figure S9, and the dashed line represents the 5% quantile for each distribution. There appears to be no evidence of selection at this locus in other populations except for D2 (**D**), a diploid population geographically close to D3 and T5.

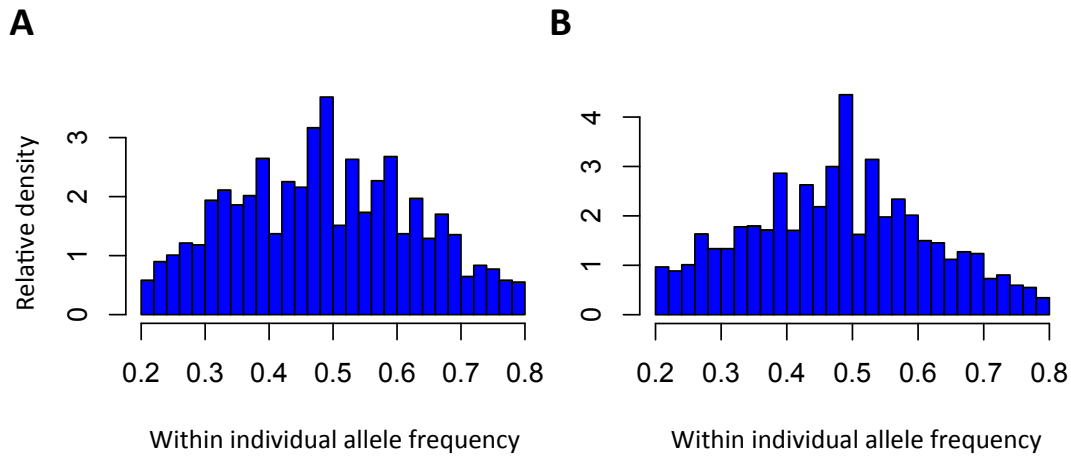


Figure S4.11 Other potential polyploid samples within diploid population D6. Although the alternate base count frequencies for individual D6-12 (**A**) and D6-26 (**B**) were best modeled by the triploid and tetraploid models, respectively, the raw data suggest they are diploid with greater, unexplained variance in base count frequencies around an expected mean (for diploids) of 0.5.

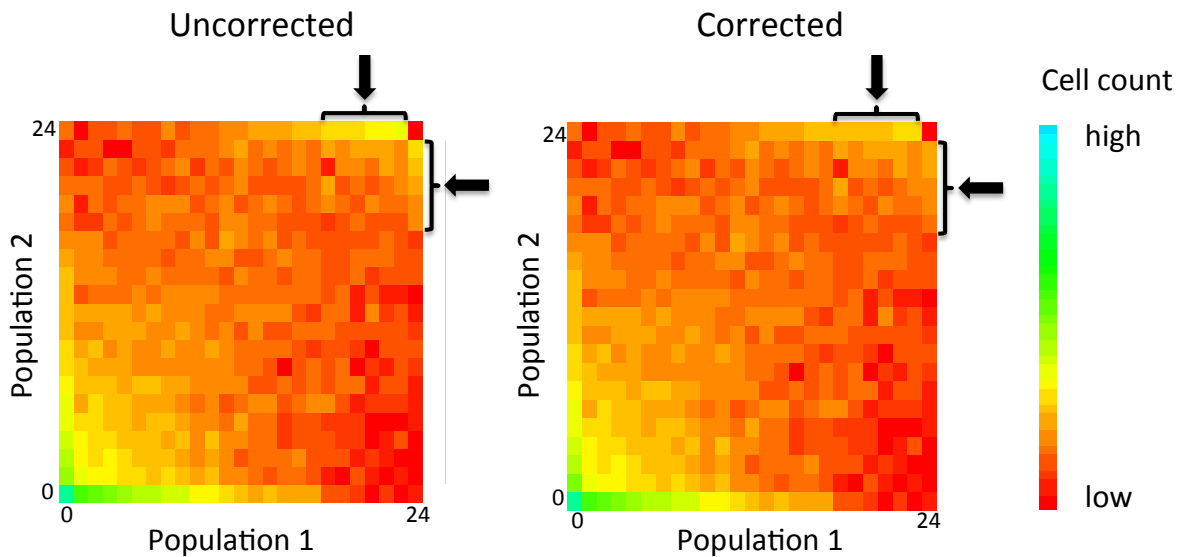


Figure S4.12 Correction of 3D-AFS for demographic inference. Allele frequency spectra were corrected for ancestral state misspecification, which creates an excess of high-frequency derived alleles. Shown here is an application of the technique in Baudry and Depaulis 2003 to a two-dimensional frequency spectrum of two tetraploid samples each containing six individuals (or 24 haplotypes). The black arrows point to the high-frequency allele categories in the uncorrected (left) and corrected (right) spectra; note the number of variants fixed in one sample and high frequency in the other decrease after correction. The excess of high-frequency variants in the uncorrected spectrum is

likely due to ancestral state misspecification from multiple mutations occurring between *A. arenosa* and the *A. lyrata* reference panel.

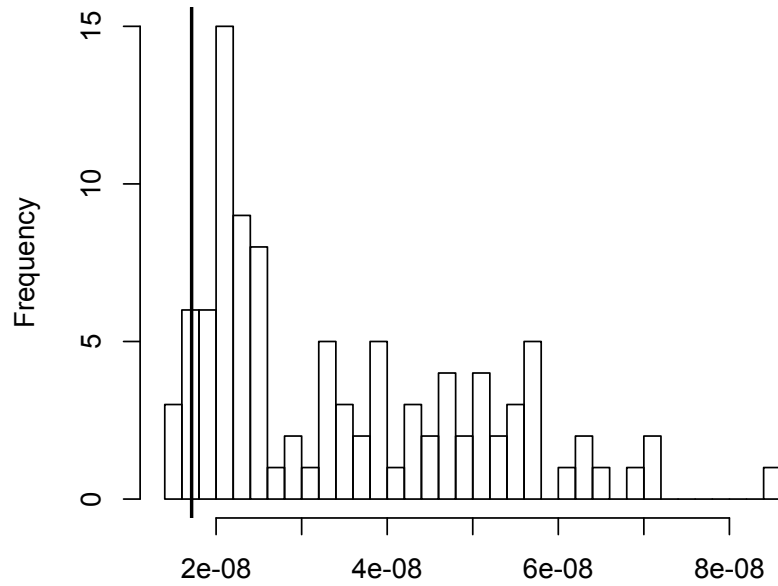


Figure S4.13 Estimation of mutation rate from simulated data. Histogram of the maximum likelihood estimates of mutation rates from 100 data sets simulated with fixed parameter values. The vertical line shows the input mutation rate value used to simulate the data, which is very close to the mode of the distribution.

Supplementary Tables

Table S4.1 Population and sample size composition of each data set.

RADseq		POOLseq		INDseq	
Population	sample size	Population	sample size	Population	sample size
T1	20	T1	19	T11	3
T2	16	T5	9	T12	3
T3	20	T7	10	T14	2
T4	20	D1	13	D2	4
T5	19	D2	9	D3	4
T6	20	D3	19	Total	16
T7	20	D4	10		
T8	20	Total	89		
T9	20				
T10	16				
T11	20				
T12	20				
T13	20				
T14	21				
D1	12				
D2	11				
D3	21				
D4	11				
D5	15				
D6	16				
Total	358				

Note: D's represent diploid populations, and T's represent tetraploid populations according to Figure 1.

Table S4.2 Pearson correlation of allele frequencies in either the RADseq dataset (**A**) or the POOLseq dataset (**B**) with the same alleles discovered in the INDseq data.

A	RADseq population	sample size	correlation
	D2	18	0.75
	D3	38	0.85
	T8	72	0.81
	T11	72	0.87
	T12	72	0.87

B	POOLseq population	sample size	correlation
	D2	18	0.77
	D3	38	0.80

Table S4.3 Estimates of the population mutation rate θ using Watterson's estimator (θ_w) and the average number of pair wise differences (θ_π) for the three datasets: INDseq (A), RADseq (B), and POOLseq (C).

INDseq		
Population	θ_w	θ_π
D2	0.022	0.023
D3	0.024	0.024
T8	0.039	0.039
T11	0.038	0.038
T12	0.039	0.039

RADseq		
Population	θ_w	θ_π
D2	0.017	0.019
D3	0.019	0.020
T8	0.020	0.027
T11	0.026	0.029
T12	0.025	0.028

POOLseq		
Population	θ_w	θ_π
D2	0.026	0.031
D3	0.024	0.031

Table S4.4 Best-fit model of non-reference base counts for each individual.

Individual	Best Model	Individual	Best Model	Individual	Best Model	Individual	Best Model	Individual	Best Model	Individual	Best Model	Individual	Best Model	Individual	Best Model
D1-10	2	D3-7	2	D6-3	2	T2-9	4	T4-9	4	T7-1	4	T9-1	4	T11-13	4
D1-13	2	D3-8	2	D6-4	2	T3-1	4	T5-1	4	T7-10	4	T9-10	4	T11-14	4
D1-15	2	D3-9	2	D6-5	2	T3-10	4	T5-10	4	T7-11	4	T9-11	4	T11-15	4
D1-16	2	D4-1	2	D6-7	2	T3-11	4	T5-11	4	T7-12	4	T9-12	4	T11-16	4
D1-17	2	D4-10	2	D6-8	2	T3-12	4	T5-12	4	T7-13	4	T9-13	4	T11-17	4
D1-18	2	D4-11	2	T1-12	4	T3-13	4	T5-13	4	T7-14	4	T9-14	4	T11-18	4
D1-19	2	D4-12	2	T1-13	4	T3-14	4	T5-14	4	T7-15	4	T9-15	4	T11-19	4
D1-24	2	D4-14	2	T1-14	4	T3-15	4	T5-15	4	T7-16	4	T9-16	4	T11-2	4
D1-25	2	D4-18	2	T1-15	4	T3-16	4	T5-16	4	T7-17	4	T9-17	4	T11-20	4
D1-5	2	D4-19	2	T1-16	4	T3-17	4	T5-17	4	T7-18	4	T9-18	4	T11-3	4
D1-6	2	D4-2	2	T1-17	4	T3-18	4	T5-18	4	T7-19	4	T9-19	4	T11-4	4
D1-8	2	D4-20	2	T1-18	4	T3-19	4	T5-19	4	T7-2	4	T9-2	4	T11-5	4
D2-1	2	D4-3	2	T1-2	4	T3-2	4	T5-2	4	T7-20	4	T9-20	4	T11-6	4
D2-10	2	D4-5	2	T1-21	4	T3-20	4	T5-3	4	T7-3	4	T9-3	4	T11-7	4
D2-11	2	D5-1	2	T1-22	4	T3-3	4	T5-4	4	T7-4	4	T9-4	4	T11-8	4
D2-2	2	D5-12	2	T1-23	4	T3-4	4	T5-5	4	T7-5	4	T9-5	4	T11-9	4
D2-3	2	D5-13	2	T1-26	4	T3-5	4	T5-6	4	T7-6	4	T9-6	4	T12-1	4
D2-5	2	D5-16	2	T1-27	4	T3-6	4	T5-7	4	T7-7	4	T9-7	4	T12-10	4
D2-6	2	D5-17	2	T1-3	4	T3-7	4	T5-8	4	T7-8	4	T9-8	4	T12-11	4
D2-7	2	D5-18	2	T1-34	4	T3-8	4	T5-9	4	T7-9	4	T9-9	4	T12-12	4
D2-8	2	D5-2	2	T1-4	4	T3-9	4	T6-1	4	T8-1	4	T10-1	4	T12-13	4
D2-9	2	D5-21	4	T1-5	4	T4-1	4	T6-10	4	T8-10	4	T10-10	4	T12-14	4
D3-1	2	D5-22	2	T1-6	4	T4-10	4	T6-11	4	T8-11	4	T10-11	4	T12-15	4
D3-10	2	D5-23	2	T1-7	4	T4-11	4	T6-12	4	T8-12	4	T10-12	4	T12-16	4
D3-11	2	D5-28	2	T1-8	4	T4-12	4	T6-13	4	T8-13	4	T10-13	4	T12-17	4
D3-12	2	D5-4	2	T2-1	4	T4-13	4	T6-14	4	T8-14	4	T10-14	4	T12-18	4
D3-13	2	D5-5	2	T2-10	4	T4-14	4	T6-15	4	T8-15	4	T10-15	4	T12-19	4
D3-14	2	D5-6	2	T2-11	4	T4-15	4	T6-16	4	T8-16	4	T10-16	4	T12-2	4
D3-15	2	D5-9	2	T2-12	4	T4-16	4	T6-17	4	T8-17	4	T10-2	4	T12-20	4
D3-16	2	D6-11	2	T2-13	4	T4-17	4	T6-18	4	T8-18	4	T10-3	4	T12-3	4
D3-17	2	D6-12	3	T2-14	4	T4-18	4	T6-19	4	T8-19	4	T10-4	4	T12-4	4
D3-18	2	D6-14	2	T2-15	4	T4-19	4	T6-2	4	T8-2	4	T10-5	4	T12-5	4
D3-19	2	D6-17	2	T2-16	4	T4-2	4	T6-20	4	T8-20	4	T10-6	4	T12-6	4
D3-2	2	D6-18	2	T2-2	4	T4-20	4	T6-3	4	T8-3	4	T10-7	4	T12-7	4
D3-20	2	D6-2	2	T2-3	4	T4-3	4	T6-4	4	T8-4	4	T10-8	4	T12-8	4
D3-21	2	D6-20	2	T2-4	4	T4-4	4	T6-5	4	T8-5	4	T10-9	4	T12-9	4
D3-3	2	D6-25	2	T2-5	4	T4-5	4	T6-6	4	T8-6	4	T11-1	4	T13-1	4
D3-4	2	D6-26	4	T2-6	4	T4-6	4	T6-7	4	T8-7	4	T11-10	4	T13-10	4
D3-5	2	D6-27	2	T2-7	4	T4-7	4	T6-8	4	T8-8	4	T11-11	4	T13-11	4
D3-6	2	D6-29	2	T2-8	4	T4-8	4	T6-9	4	T8-9	4	T11-12	4	T13-12	4

Note: Three models of non-reference base count distributions were tested for each individual: diploid (2), triploid (3), and tetraploid (4). Shown here is the model that fit the observed data best using a *G* statistic (see Materials and Methods). Three samples, highlighted in red, have non-reference base count distributions that best fit a polyploid model despite being collected from an otherwise diploid population.

Table S4.5 Contrasting the maximum likelihood estimates (MLEs) of a simple isolation-migration (IM) model of two diploid populations (D2, D3) sequenced in all three datasets.

Parameter	4-fold degenerate sites			Parameter	Noncoding sites		
	INDseq $n_{D2}=8, n_{D3}=8$ 156,760 sites	RADseq $n_{D2}=18, n_{D3}=24$ 47,113 sites	POOLseq $n_{D2}=18, n_{D3}=38$ 739,728 sites		INDseq $n_{D2}=8, n_{D3}=8$ 11,325,439 sites	RADseq $n_{D2}=18, n_{D3}=24$ 290,326 sites	POOLseq $n_{D2}=18, n_{D3}=38$ 8,639,105 sites
Divergence	24615 (11514, 37480)	21908 (10691, 29631)	11866 (5880, 16953)	Divergence	63674 (58262, 71474)	31351 (26339, 39205)	33366 (31037, 48595)
N_{D2}	155286 (79724, 181797)	69113 (46014, 89697)	60658 (29934, 74827)	N_{D2}	116702 (97389, 130965)	73389 (65452, 79065)	133504 (121756, 182340)
N_{D3}	105791 (58728, 128833)	111881 (73994, 130336)	53458 (27164, 66627)	N_{D3}	165075 (142798, 172954)	106324 (99061, 114846)	148982 (136730, 197768)
N_{Anc}	228276 (159393, 302083)	448073 (350585, 539454)	404866 (384179, 458626)	N_{Anc}	221904 (166477, 275901)	192970 (174886, 236687)	308738 (300853, 332773)
$N_{D2}m_{D3D2}$	0.33 (0.012, 1.16)	0.14 (0.006, 0.42)	0.01 (0.005, 0.21)	$N_{D2}m_{D3D2}$	0.49 (0.29, 0.90)	0.37 (0.23, 0.53)	0.01 (0.01, 0.10)
$N_{D3}m_{D2D3}$	0.039 (0.007, 0.789)	0.72 (0.021, 0.82)	0.01 (0.004, 0.29)	$N_{D3}m_{D2D3}$	1.12 (0.66, 1.80)	0.71 (0.54, 1.14)	0.03 (0.03, 0.19)

Note: The analysis was repeated for both 4fold-degenerate and noncoding sites. Beneath each dataset label are the sample sizes used for each diploid population. Shown are the MLEs of divergence time (generations), population sizes (N_{D2} , N_{D3} , and $N_{ancestral}$), and population migration rates (where m_{D2D3} is the migration rate from D2 to D3).

Table S4.6 Comparison of model likelihoods using AIC to infer number of autotetraploid origins with the RADseq dataset.

	D1,T1,T7		D2,T2,T7		D1,T4,T7		D3,T5,T7	
	Model A	Model B	Model A	Model B	Model A	Model B	Model A	Model B
Max $\log_{10}(Lhood_i)$	-57193.012	-57244.078	-41399.139	-41424.2	-47571.272	-47629.888	-35150.671	-35183.609
AIC_i	263419.6	263654.7	190686.1	190801.5	219109.8	219379.7	161910.8	162062.5
Δ_i	0	235.1	0	115.4	0	269.9	0	151.7
w_i	~1	8.9*10 ⁻⁵²	~1	8.7*10 ⁻²⁶	~1	2.5*10 ⁻⁵⁹	~1	1.1*10 ⁻³³

	D1,T1,T13		D2,T2,T13		D1,T4,T13		D3,T5,T13	
	Model A	Model B	Model A	Model B	Model A	Model B	Model A	Model B
Max $\log_{10}(Lhood_i)$	-54442.147	-54522.276	-40612.069	-40652.252	-44756.02	-44827.987	-33922.588	-33955.301
AIC_i	250751.4	251120.4	187061.5	187246.5	206145.1	206476.5	156255.3	156405.9
Δ_i	0	369	0	185	0	331.4	0	150.6
w_i	~1	7.5*10 ⁻⁸¹	~1	6.7*10 ⁻⁴¹	~1	1.1*10 ⁻⁷²	~1	2.0*10 ⁻³³

Notes: Using the RADseq data, the maximum likelihood of model A and model B was computed for each population trio, with each trio consisting of a geographically proximal diploid and tetraploid as well as a tetraploid outgroup, either T7 (A) or T13 (B). The highlighted rows contain the Akaike weights, or the relative likelihood of each

model given the candidate set of models. For all population trios examined, a single origin of the tetraploid, model A, is unambiguously supported.

$$AIC_i = 2(d) - 2\ln(Lhood_i)$$

Where d is the number of parameters (18)

$$\Delta_i = AIC_i - \min(AIC)$$

$$w_i = \frac{e^{-0.5\Delta_i}}{\sum_r e^{-0.5\Delta_r}}$$

Table S4.7 Comparison of model likelihoods using AIC to infer number of autotetraploid origins with the PoolSeq dataset.

	D1,T1,T7		D3,T5,T7	
	Model A	Model B	Model A	Model B
Max log₁₀(Lhood_i)	-151899.163	-151959.533	-114186.854	-137029.986
AIC_i	699557.5	699835.5	525885.9	631082.4
Δ_i	0	278	0	105196.5
w_i	~1	4.3*10 ⁻⁶¹	~1	~0

Note: Using the POOLseq data, the maximum likelihood of model A and model B was computed for each population trio, with each trio consisting of a geographically proximal diploid and tetraploid as well as a tetraploid outgroup T7. The highlighted rows contain the Akaike weights, or the relative likelihood of each model given the candidate set of models. The calculation of Akaike weights is described in Table S4.6.

Table S4.8 MLEs of population trio models using noncoding sites.

Parameter	Population Trio			
	D3,T5,T7	D1,T1,T7	D2,T2,T7	D1,T4,T7
Adm_{DT}	0.53 (0.39, 0.59)	0.43 (0.35, 0.53)	0.50 (0.37, 0.56)	0.30 (0.22, 0.42)
Adm_{TD}	0.26 (0.16, 0.40)	0.17 (0.12, 0.28)	0.20 (0.15, 0.35)	0.12 (0.07, 0.21)
T_{Adm}	4638 (3186, 7215)	12484 (10297, 15017)	4748 (3171, 6576)	9752 (7671, 13843)
D_1	6169 (3885, 9808)	17361 (12428, 19641)	6667 (4104, 10298)	11099 (8650, 15990)
D_2	59295 (41573, 75965)	110066 (95595, 133400)	53875 (32289, 49758)	80652 (61511, 95734)
N_D	43721 (31102, 61832)	52198 (45740, 59630)	27091 (17898, 35600)	50824 (42982, 61011)
N_T	36039 (24061, 53546)	128686 (100619, 142437)	53307 (34522, 73677)	56704 (45616, 74825)
$N_{T'}$	69019 (43314, 111716)	118502 (87292, 129380)	71323 (43394, 100805)	86845 (67725, 116131)

Notes: These values were obtained using noncoding sites in the RADseq dataset. Shown are the admixture proportions from diploids to tetraploids (Adm_{DT}) and tetraploids to diploids (Adm_{TD}) going backwards in time, the time of admixture (T_{Adm}), the divergence time between tetraploids (D_1) and the ancestral tetraploid and diploid (D_2), and the population sizes of the diploid (N_D), admixed tetraploid, and the outgroup tetraploid ($N_{T'}$). Divergence times are expressed in terms of generations, and population sizes are in haploid number of chromosomes.

Table S4.9 MLEs of model parameters using population trios and higher sequencing depth cutoffs (12X) for tetraploids.

Parameter	Population Trio			
	D3,T5,T7	D1,T1,T7	D2,T2,T7	D1,T4,T7
Adm_{DT}	0.32	0.34	0.25	0.13
Adm_{TD}	0.42	0.11	0.22	0.12
T_{Adm}	6478	9681	5196	5923
D_1	9019	10806	7391	7167
D_2	30312	69589	26397	40518
N_D	67156	48689	44285	58470
N_T	46813	96627	60189	38969
$N_{T'}$	120788	84998	90689	54572

Notes: These values were obtained using fourfold degenerate sites in the RADseq dataset. Shown are the admixture proportions from diploids to tetraploids (Adm_{DT}) and tetraploids to diploids (Adm_{TD}) going backwards in time, the time of admixture (T_{Adm}),

the divergence time between tetraploids (D_1) and the ancestral tetraploid and diploid (D_2), and the population sizes of the diploid (N_D), admixed tetraploid, and the outgroup tetraploid (N_T). Divergence times are expressed in terms of generations, and population sizes are in haploid number of chromosomes.

Table S4.10 Relative likelihoods of models involving bidirectional, unidirectional, or no admixture.

A

		D1,T1,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max $\log_{10}(Lhood_i)$		-57178.569	-57198.122	-57191.146
AIC_i		263353	263441.1	263405
Δ_i		0	88.1	52
w_i		~1	7.4×10^{-20}	5.1×10^{-12}

		D2,T2,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max $\log_{10}(Lhood_i)$		-41399.016	-41417.748	-41452.850
AIC_i		190685.5	190769.8	190927.4
Δ_i		0	84.3	241.9
w_i		~1	4.9×10^{-19}	3.0×10^{-53}

		D1,T4,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max $\log_{10}(Lhood_i)$		-47563.871	-47576.324	-47582.698
AIC_i		219075.7	219131.1	219156.4
Δ_i		0	55.4	80.7
w_i		~1	9.3×10^{-13}	3.0×10^{-18}

		D3,T5,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max $\log_{10}(Lhood_i)$		-35149.814	-35165.802	-35227.818
AIC_i		161906.9	161978.5	162260.1
Δ_i		0	71.6	353.2
w_i		~1	2.8×10^{-16}	2.0×10^{-77}

B

		D4,T12,T13	
Model		bidirectional	No admixture
# parameters		18	15
Max $\log_{10}(Lhood_i)$		-38643.499	-38627.862
AIC_i		177995.9	177917.9
Δ_i		78	0
w_i		1.2×10^{-17}	~1

C

		D1,T1,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max log₁₀(Lhood_i)		-151900.68	-151905.32	-151939.09
AIC_i		699564.5	699583.9	699735.4
Δ_i		0	19.4	170.9
w_i		~1	6.1*10 ⁻⁵	7.8*10 ⁻³⁸

		D3,T5,T7		
Model		bidirectional	unidirectional	No admixture
# parameters		18	17	15
Max log₁₀(Lhood_i)		-114133.72	-114152.50	-114140.05
AIC_i		525641.2	525725.7	525664.4
Δ_i		0	84.5	23.2
w_i		~1	4.5*10 ⁻¹⁹	9.2*10 ⁻⁶

Notes: (A) Allowing gene flow from tetraploids to diploids (forward in time) confers a significantly better fit of the model to the data. (B) As a negative control, models were made with tetraploid populations that have not experienced admixture, and the model with no admixture fits the data significantly better than one that allows for bidirectional interploidy admixture. (C) These results are validated with the populations present in the PoolSeq dataset.

Table S4.11 MLEs of model parameters using a population trio consisting of a diploid and two tetraploids not suspected of having experienced any interploidy admixture.

Parameter	Trio
	D4,T12,T13
Adm_{DT}	0.011 (0.007, 0.028)
Adm_{TD}	0.010 (0.004, 0.027)
T_{Adm}	2703 (1024, 5087)
D₁	3612 (2260, 6366)
D₂	60109 (36683, 61758)
N_D	100266 (79327, 112840)
N_T	43082 (24703, 68137)
N_{T'}	60322 (34134, 100754)

Notes: These values were obtained using four-fold fegenerate sites in the RADseq dataset. Parameter labels correspond to those listed in Table 4.1.

Table S4.12 Results of population trio analyses with two diploids and one tetraploid.

Diploid Populations	Tetraploid Population				
	T1	T2	T4	T5	T7
D1, D2	(D1,D2)	D2	D2	-	D2
D1, D3	D3	-	D3	D3	D3
D2, D3	-	D3	-	D3	(D2,D3)

Notes: The two diploid populations used are labeled in the leftmost column and the tetraploid labeled along the top. Listed in the table is the diploid population that is sister to the tetraploid, as that topology had the highest likelihood according to AIC analyses. Since tetraploid populations T1, T2, T4, and T5 are admixed, trio analyses that do not include the geographically proximal diploid population with which they have exchanged genes were not performed. The AIC analyses to obtain these results are in Table S4.13.

Table S4.13 Akaike weights for population trio analyses of two diploids and one tetraploid (Table S4.12).

Model*	D1,D2,T1			D1,D3,T1		
	D1	D2	(D1,D2)	D1	D3	(D1,D3)
Max log ₁₀ (Lhood _i)	-27146.995	-27153.65	-27140.506	-31005.042	-30994.352	-31001.975
AIC _i	125046.5	125077.2	125016.6	142813.5	142764.3	142799.4
Δ _i	29.9	60.6	0	49.2	0	35.1
w _i	3.2*10 ⁻⁷	6.9*10 ⁻¹⁴	~1	2.1*10 ⁻¹¹	~1	2.4*10 ⁻⁸

Model*	D1,D2,T2			D2,D3,T2		
	D1	D2	(D1,D2)	D2	D3	(D2,D3)
Max log ₁₀ (Lhood _i)	-24928.226	-24927.594	-24929.801	-19155.662	-19138.698	-19152.866
AIC _i	114828.7	114825.8	114836	88245.08	88166.96	88232.21
Δ _i	2.9	0	10.2	78.12	0	65.25
w _i	0.19	4.9*10 ⁻³	0.81	1.1*10 ⁻¹⁷	~1	6.8*10 ⁻¹⁵

Model*	D1,D2,T4			D1,D3,T4		
	D1	D2	(D1,D2)	D2	D3	(D2,D3)
Max log ₁₀ (Lhood _i)	-25200.907	-25195.947	-25205.529	-26415.776	-26396.409	-26417.042
AIC _i	116084.5	116061.6	116105.8	121679.1	121590	121685
Δ _i	22.9	0	44.2	89.1	0	95
w _i	1.1*10 ⁻⁵	~1	2.5*10 ⁻¹⁰	4.5*10 ⁻²⁰	~1	2.3*10 ⁻²¹

Model*	D1,D3,T5			D2,D3,T5		
	D1	D3	(D1,D3)	D2	D3	(D2,D3)
Max log ₁₀ (Lhood _i)	-27599.888	-27591.275	-27607.842	-19006.428	-18994.784	-18998.471
AIC _i	127132.2	127092.5	127168.8	87557.84	87504.21	87521.19
Δ _i	39.7	0	76.3	53.63	0	16.98
w _i	2.4*10 ⁻⁹	~1	2.7*10 ⁻¹⁷	2.3*10 ⁻¹²	~1	2.1*10 ⁻⁴

Model*	D1,D2,T7			D1,D3,T7			D2,D3,T7		
	D1	D2	(D1,D2)	D1	D3	(D1,D3)	D2	D3	(D2,D3)
Max log ₁₀ (Lhood _i)	-26706.522	-26696.616	-26704.645	-29445.383	-29426.321	-29448.386	-20531.748	-20525.234	-20519.103
AIC _i	123018.1	122972.5	123009.4	135631	135543.2	135644.8	94582.19	94552.2	94523.96
Δ _i	45.6	0	36.9	87.8	0	101.6	58.23	28.24	0
w _i	1.3*10 ⁻¹⁰	~1	9.7*10 ⁻⁹	8.6*10 ⁻²⁰	~1	8.7*10 ⁻²³	2.3*10 ⁻¹³	7.4*10 ⁻⁷	~1

Notes: The calculation of Akaike weights is described in Table S4.6.

*the diploid population to which the tetraploid is most closely related, with (D,D') representing the tetraploid being an outgroup to the two diploids used.

Table S4.14 Likelihood analyses of tetraploid tree topologies using coalescent simulations.

Tree topology	Max log ₁₀ (Lhood _i)	AIC _i	Δ _i	w _i
(T1,(T3,(T7,T13)))	-473386.177	2180043.771	534.0062907	1.1015E-116
(T1,(T7,(T3,T13)))	-473309.549	2179690.886	181.121332	4.67735E-40
(T1,(T13,(T3,T7)))*	-473270.219	2179509.765	0	1
(T3,(T1,(T7,T13)))	-473781.074	2181862.339	2352.574067	0
(T3,(T7,(T1,T13)))	-474299.099	2184247.932	4738.167202	0
(T3,(T13,(T1,T7)))	-474351.843	2184490.827	4981.062283	0
(T7,(T1,(T3,T13)))	-474038.117	2183046.066	3536.300752	0
(T7,(T3,(T1,T13)))	-474240.093	2183976.199	4466.434547	0
(T7,(T13,(T1,T3)))	-474178.843	2183694.133	4184.367891	0
(T13,(T1,(T3,T7)))	-473830.382	2182089.411	2579.645784	0
(T13,(T3,(T1,T7)))	-474242.791	2183988.624	4478.859295	0
(T13,(T7,(T1,T3)))	-474145.681	2183541.416	4031.651247	0
((T1,T3),(T7,T13)) #2nd node earlier	-473918.543	2182495.407	2985.642167	0
((T1,T7),(T3,T13)) #2nd node earlier	-473993.274	2182839.556	3329.791119	0
((T1,T13),(T3,T7)) #2nd node earlier	-474029.421	2183006.019	3496.254195	0
((T1,T3),(T7,T13)) #1st node earlier	-474068.016	2183183.756	3673.990727	0
((T1,T7),(T3,T13)) #1st node earlier	-474274.129	2184132.941	4623.17611	0
((T1,T13),(T3,T7)) #1st node earlier	-474065.13	2183170.465	3660.700207	0

Notes: A simple divergence model (no migration) was used with 4 tetraploid populations and 10 parameters in total. Calculation of Akaike weights are described in Table S4.6.

*Most likely topology

Table S4.15 Allele frequencies for each population for within the region under selection as described in Figure 4.6.

mRNA PAC ID	scaff	position	D1	T1	Functional Category	D3	D2	T5	T7	
16045957	5	21136047	1.00	1.00	SYNONYMOUS	1.00	1.00	NA	1.00	
	5	21136879	1.00	1.00	INTRON	NA	1.00	1.00	1.00	
	5	21138079	1.00	1.00	INTRON	0.71	0.67	0.81	1.00	
	5	21138160	1.00	1.00	SYNONYMOUS	0.74	0.33	0.39	NA	
	5	21138231	1.00	1.00	INTRON	0.71	0.39	0.36	0.75	
	5	21138276	1.00	1.00	INTRON	0.39	0.50	0.64	0.65	
	5	21138294	1.00	1.00	INTRON	0.82	0.50	1.00	1.00	
	5	21138456	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00	
5	21138506	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00		
16036121	5	21139646	1.00	1.00	SYNONYMOUS	0.74	NA	0.61	0.55	
16040231	5	21141284	0.77	0.82	SYNONYMOUS	0.00	0.00	NA	0.15	
	5	21141926	1.00	1.00	NON_SYNONYMOUS	0.79	1.00	1.00	0.60	
	5	21142084	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00	
	5	21142422	1.00	1.00	NON_SYNONYMOUS	NA	NA	NA	NA	
	5	21142467	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	0.72	NA	
16061301	5	21143198	1.00	1.00	INTRON	1.00	1.00	0.86	1.00	
	5	21143330	0.81	0.80	INTRON	0.00	0.00	0.00	0.00	
	5	21143412	1.00	1.00	SYNONYMOUS	1.00	0.83	0.92	0.80	
	5	21143418	0.85	0.79	SYNONYMOUS	0.00	0.00	0.00	0.00	
	5	21143512	0.88	0.84	INTRON	0.00	0.00	0.00	0.00	
	5	21143534	0.88	0.80	INTRON	0.00	0.00	0.00	0.00	
	5	21143592	0.85	0.91	INTRON	0.66	0.22	0.61	0.38	
	5	21143610	0.88	0.76	INTRON	0.00	0.00	0.00	0.00	
	5	21143624	1.00	1.00	NON_SYNONYMOUS	0.68	1.00	0.64	0.60	
	5	21144022	0.96	0.79	SYNONYMOUS	0.00	0.11	0.00	0.18	
	5	21144308	1.00	0.78	INTRON	0.00	0.17	NA	0.08	
	5	21144380	1.00	1.00	INTRON	1.00	1.00	1.00	1.00	
	5	21144468	1.00	0.80	SYNONYMOUS	0.00	0.06	0.00	0.00	
	5	21144675	0.92	0.80	INTRON	0.00	0.06	0.00	0.20	
	5	21144859	1.00	0.93	SYNONYMOUS	0.00	0.11	0.00	0.00	
	5	21144901	1.00	0.87	SYNONYMOUS	0.00	0.17	0.00	0.18	
	5	21145084	1.00	1.00	NON_SYNONYMOUS	0.00	0.28	0.00	0.08	
	5	21145197	1.00	1.00	INTRON	0.00	0.06	0.00	0.00	
	5	21145298	1.00	0.87	INTRON	0.00	0.17	0.00	0.13	
	16065113	5	21146677	1.00	0.83	INTRON	0.00	0.00	0.00	0.00
5		21146840	1.00	1.00	INTRON	0.00	0.00	NA	0.15	
5		21147477	0.96	0.88	SYNONYMOUS	0.11	0.17	0.00	NA	
5		21147565	0.96	0.88	NON_SYNONYMOUS	0.00	0.06	0.00	NA	
5		21147869	1.00	0.89	SYNONYMOUS	0.00	0.00	0.00	0.15	
5		21147895	1.00	0.89	NON_SYNONYMOUS (A218T)	0.00	0.00	0.00	0.00	
5		21147942	1.00	0.82	INTRON	0.00	0.06	0.00	0.05	
5		21148017	1.00	0.78	SYNONYMOUS	0.00	0.00	0.00	0.08	
5		21148101	1.00	0.88	SYNONYMOUS	1.00	1.00	1.00	NA	
5		21148132	1.00	0.89	INTRON	0.00	NA	NA	NA	
5		21148208	1.00	0.87	INTRON	0.00	0.17	0.00	0.23	
5		21148212	1.00	0.87	INTRON	0.00	0.00	0.00	0.23	
5		21148901	1.00	0.84	NON_SYNONYMOUS	0.00	0.17	0.06	0.18	
5		21148912	1.00	0.84	SYNONYMOUS	0.03	0.11	0.06	0.18	
16036656		5	21151713	0.88	0.76	SYNONYMOUS	0.00	0.00	0.00	0.00
		5	21152192	0.88	0.78	INTRON	0.00	0.56	0.00	0.00
	5	21152858	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00	
	5	21152958	1.00	1.00	NON_SYNONYMOUS	NA	NA	NA	NA	
16062538	5	21154294	1.00	1.00	NON_SYNONYMOUS	NA	1.00	1.00	NA	
	5	21154311	1.00	0.76	SYNONYMOUS	1.00	1.00	0.89	NA	
	5	21154445	1.00	1.00	NON_SYNONYMOUS	0.92	0.72	0.81	0.63	
	5	21154459	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00	
	5	21154491	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00	
	5	21154536	1.00	1.00	SYNONYMOUS	0.76	NA	NA	1.00	
	5	21154579	1.00	0.79	NON_SYNONYMOUS	1.00	1.00	1.00	1.00	
	5	21154642	0.81	0.80	INTRON	0.74	0.78	0.31	0.00	
	5	21154675	0.81	0.82	INTRON	0.63	0.78	0.78	1.00	
	5	21154785	1.00	0.88	SYNONYMOUS	1.00	1.00	0.75	NA	
	5	21154826	1.00	0.78	NON_SYNONYMOUS	1.00	1.00	1.00	NA	
	5	21154927	1.00	0.93	NON_SYNONYMOUS	1.00	1.00	1.00	NA	
	5	21155017	1.00	0.91	NON_SYNONYMOUS	NA	0.72	0.72	1.00	
	5	21155033	1.00	1.00	NON_SYNONYMOUS	NA	1.00	0.72	1.00	
	5	21155106	1.00	1.00	SYNONYMOUS	0.63	0.61	0.67	0.58	
	5	21155127	1.00	1.00	SYNONYMOUS	1.00	1.00	0.83	1.00	
	5	21155406	1.00	1.00	SYNONYMOUS	NA	1.00	0.72	1.00	
	5	21155737	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	0.92	1.00	

	5	21157829	0.96	1.00	NON_SYNONYMOUS	NA	1.00	0.83	1.00
16045888	5	21159411	1.00	0.83	SYNONYMOUS	0.00	0.11	0.31	0.00
	5	21159999	1.00	0.80	SYNONYMOUS	NA	0.00	0.25	0.00
	5	21160203	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21160398	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21160404	1.00	0.86	SYNONYMOUS	0.00	0.00	0.00	0.00
	5	21160570	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	NA
	5	21160839	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21161550	1.00	0.79	SYNONYMOUS	0.29	0.00	0.33	0.28
	5	21161883	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21162041	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21162045	1.00	0.83	SYNONYMOUS	0.00	0.00	0.50	0.00
	5	21162066	0.96	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21162191	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
16053029	5	21163545	1.00	1.00	NON_SYNONYMOUS	0.00	NA	0.17	NA
	5	21163553	1.00	1.00	NON_SYNONYMOUS	1.00	NA	1.00	NA
	5	21163620	1.00	1.00	NON_SYNONYMOUS	1.00	NA	NA	NA
	5	21163684	1.00	1.00	SYNONYMOUS	1.00	NA	1.00	NA
	5	21163745	1.00	0.99	NON_SYNONYMOUS	1.00	NA	1.00	0.45
	5	21163843	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	NA
16065728	5	21164710	1.00	1.00	INTRON	1.00	NA	1.00	1.00
	5	21164759	1.00	0.91	NON_SYNONYMOUS	1.00	1.00	0.75	0.68
	5	21164827	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21164862	0.81	0.76	INTRON	NA	0.72	0.61	0.60
	5	21164884	0.77	1.00	INTRON	NA	1.00	1.00	1.00
	5	21164885	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21164946	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	NA	1.00
	5	21165171	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21165216	1.00	1.00	INTRON	0.97	1.00	0.78	0.98
	5	21165218	0.77	1.00	INTRON	1.00	1.00	NA	NA
	5	21165291	1.00	1.00	INTRON	NA	1.00	1.00	1.00
	5	21165343	1.00	1.00	INTRON	1.00	1.00	1.00	NA
	5	21165355	1.00	1.00	INTRON	1.00	1.00	1.00	NA
	5	21165416	1.00	1.00	INTRON	0.89	1.00	1.00	1.00
	5	21165458	1.00	0.86	INTRON	0.89	0.83	0.92	0.70
	5	21165558	1.00	1.00	INTRON	0.82	NA	0.89	0.75
	5	21165581	1.00	1.00	INTRON	NA	NA	1.00	NA
	5	21165619	1.00	1.00	INTRON	0.87	0.89	0.89	0.80
	5	21165627	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21165666	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	NA
	5	21165689	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21165788	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21165794	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21165916	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21165964	1.00	0.91	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166046	1.00	0.83	INTRON	1.00	1.00	1.00	1.00
	5	21166080	1.00	0.91	INTRON	1.00	0.83	1.00	1.00
	5	21166164	1.00	1.00	NON_SYNONYMOUS	NA	NA	1.00	1.00
	5	21166195	1.00	1.00	NON_SYNONYMOUS	1.00	0.78	1.00	1.00
	5	21166198	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166242	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166284	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166333	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166414	1.00	1.00	INTRON	NA	NA	1.00	1.00
	5	21166509	1.00	1.00	SYNONYMOUS	NA	NA	1.00	NA
	5	21166612	1.00	1.00	INTRON	1.00	NA	1.00	NA
	5	21166615	1.00	1.00	INTRON	1.00	NA	1.00	NA
	5	21166680	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21166735	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21166767	1.00	1.00	INTRON	NA	NA	1.00	1.00
	5	21166780	1.00	1.00	INTRON	1.00	NA	1.00	1.00
	5	21166851	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166873	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21166908	1.00	1.00	INTRON	1.00	NA	1.00	1.00
	5	21166923	1.00	1.00	INTRON	1.00	0.83	1.00	1.00
	5	21166969	1.00	1.00	INTRON	NA	NA	1.00	1.00
	5	21166981	1.00	1.00	INTRON	1.00	NA	1.00	1.00
	5	21167023	1.00	1.00	NON_SYNONYMOUS	1.00	NA	1.00	1.00
	5	21167039	1.00	1.00	SYNONYMOUS	1.00	NA	1.00	1.00
	5	21167056	1.00	1.00	NON_SYNONYMOUS	1.00	NA	1.00	1.00
	5	21167098	1.00	1.00	NON_SYNONYMOUS	1.00	NA	1.00	1.00
	5	21167225	1.00	1.00	SYNONYMOUS	0.92	NA	1.00	1.00
	5	21167234	1.00	1.00	SYNONYMOUS	1.00	NA	1.00	1.00
	5	21167290	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	NA	1.00
	5	21167331	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21167342	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21167350	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
	5	21167382	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
	5	21167404	1.00	1.00	NON_SYNONYMOUS	1.00	0.72	1.00	1.00
	5	21167488	1.00	1.00	INTRON	1.00	1.00	1.00	1.00

5	21167509	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
5	21167593	1.00	1.00	INTRON	NA	1.00	1.00	1.00
5	21167596	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
5	21167603	1.00	1.00	INTRON	1.00	0.78	1.00	1.00
5	21167645	1.00	1.00	INTRON	1.00	1.00	1.00	1.00
5	21167738	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	1.00
5	21167833	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00
5	21167901	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	1.00

Table Notes: Allele frequencies are listed in the columns labeled by population. Highlighted rows are high-frequency shared variants between D1 and T1 but absent from other populations. NA represents sites with insufficient coverage (>10x).

Table S4.16 Annotated genes within the genomic region shown in Figure 4.6 C,D

mRNA PAC ID	Scaffold	Lower Coordinate	Upper coordinate	<i>A. thaliana</i> ortholog	PANTHER
16047203	5	21117650	21118572		
16063875	5	21118773	21120916	AT3G63250.1	homocysteine s-methyltransferase
16045618	5	21121211	21123181	AT3G63260.1	Serine/threonine-protein kinase
16049082	5	21126050	21132592	AT3G63280.1	Serine/threonine-protein kinase NEK
16060429	5	21132744	21134546	AT3G63290.1	Oxidoreductase
16045957	5	21135947	21138641	AT3G63300.1	
16036121	5	21138867	21139874	AT3G63310.1	
16040231	5	21140754	21142483	AT3G63320.1	
16061301	5	21142600	21145560	AT3G63330.1	Protein Phosphatase 2C
16065113	5	21146408	21148971	AT3G63340.1	Protein Phosphatase 2C
16036656	5	21151672	21152962	AT3G63350.1	Heat shock transcription factor
16055279	5	21153475	21153827		
16062538	5	21154273	21158242	AT3G63370.1	
16045888	5	21158874	21162410	AT3G63380.1	
16053029	5	21163489	21164010		
16065728	5	21164587	21167936	AT3G63400	Peptidyl-prolyl cis-trans isomerase

Table notes: Highlighted genes display many sites that are high frequency in D1 and T1 and absent in other populations (See Figure 4.6, Table S4.15).

Table S4.17 Allele frequencies for each population for within region under selection as described in Figure S4.9.

mRNA PAC ID	scaff	position	D3	T5	Functional Category	D2	D1	T1	T7
16045866	4	21965095	1.00	1.00	INTRON	0.61	0.88	0.45	0.40
	4	21965325	1.00	1.00	NON_SYNONYMOUS	0.78	NA	1.00	0.80
	4	21965498	1.00	1.00	SYNONYMOUS	0.72	NA	1.00	NA
16048949	4	21966694	0.87	1.00	SYNONYMOUS	0.06	0.00	0.21	0.48
	4	21966700	0.87	1.00	SYNONYMOUS	0.06	0.00	0.21	0.45
	4	21966706	1.00	1.00	SYNONYMOUS	0.72	1.00	1.00	1.00
	4	21966811	1.00	0.81	SYNONYMOUS	0.00	0.00	0.00	0.25
	4	21966878	0.95	0.78	INTRON	0.00	0.00	0.03	0.40
	4	21966916	0.89	0.83	INTRON	0.22	0.69	0.37	0.38
	4	21966988	1.00	1.00	INTRON	1.00	1.00	1.00	NA
	4	21967195	1.00	1.00	INTRON	1.00	1.00	NA	1.00
	4	21967205	1.00	1.00	INTRON	1.00	1.00	NA	1.00
	4	21967240	0.95	0.81	INTRON	0.11	0.08	0.00	0.08
	4	21967243	1.00	1.00	INTRON	NA	1.00	0.99	NA
	4	21967267	0.95	0.81	INTRON	0.06	0.00	0.01	0.05
	4	21967301	1.00	0.97	INTRON	0.94	1.00	NA	NA
	4	21967305	1.00	0.94	INTRON	0.61	1.00	NA	1.00
	4	21967315	1.00	1.00	INTRON	1.00	0.81	0.76	NA
4	21967333	1.00	1.00	INTRON	0.50	0.42	0.37	0.48	
16059281	4	21971658	1.00	0.83	NON_SYNONYMOUS	NA	0.00	0.00	0.00
	4	21971684	1.00	0.89	SYNONYMOUS	NA	NA	0.11	0.85
	4	21972086	1.00	1.00	INTRON	1.00	NA	NA	NA
16056840	4	21977109	1.00	1.00	INTRON	1.00	1.00	0.99	NA
	4	21977131	1.00	0.81	INTRON	1.00	0.00	0.00	NA

	4	21977184	1.00	0.89	SYNONYMOUS	0.72	0.00	0.00	NA
	4	21977403	1.00	0.78	INTRON	0.83	1.00	NA	NA
16063626	4	21984705	1.00	1.00	SYNONYMOUS	0.83	0.42	0.32	0.35
	4	21984756	1.00	1.00	SYNONYMOUS	1.00	NA	1.00	NA
	4	21984820	1.00	0.94	NON_SYNONYMOUS (S45A)	0.67	0.00	0.00	0.00
	4	21984834	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	NA
	4	21984991	1.00	0.83	SYNONYMOUS	0.72	0.00	0.00	0.00
	4	21985024	1.00	1.00	SYNONYMOUS	1.00	1.00	0.97	0.85
	4	21985057	1.00	0.86	SYNONYMOUS	0.89	0.00	0.01	0.30
	4	21985118	1.00	0.81	INTRON	0.83	0.00	0.00	0.00
	4	21985470	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	NA	1.00
	4	21985600	1.00	1.00	NON_SYNONYMOUS	1.00	NA	1.00	NA
	4	21985615	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	NA
	4	21985622	1.00	0.78	SYNONYMOUS	0.72	0.00	0.00	0.00
	4	21985652	1.00	0.92	SYNONYMOUS	1.00	1.00	0.86	0.70
	4	21985722	1.00	0.83	NON_SYNONYMOUS	0.83	0.15	0.29	0.00
	4	21985882	1.00	0.78	INTRON	0.78	0.00	0.00	0.00
	4	21985922	1.00	1.00	INTRON	1.00	NA	NA	1.00
	4	21985998	1.00	0.86	INTRON	0.72	0.08	0.00	0.00
	4	21986013	1.00	0.86	INTRON	0.72	NA	0.46	0.30
	4	21986074	1.00	1.00	INTRON	1.00	0.88	0.84	1.00
	4	21986099	1.00	0.89	INTRON	0.72	0.00	0.00	0.00
	4	21986112	1.00	0.89	INTRON	0.67	NA	0.00	0.00
	4	21986115	1.00	1.00	INTRON	1.00	0.96	0.91	1.00
	4	21986230	1.00	1.00	INTRON	1.00	NA	1.00	1.00
	4	21986404	1.00	0.83	INTRON	1.00	0.00	NA	0.00
	4	21986479	1.00	1.00	INTRON	NA	0.73	0.89	1.00
	4	21986487	1.00	1.00	INTRON	1.00	0.69	0.86	1.00
	4	21986524	1.00	0.78	INTRON	0.67	0.00	0.00	0.00
	4	21986735	1.00	0.89	INTRON	0.78	0.00	NA	0.00
	4	21986753	1.00	0.89	INTRON	0.72	0.00	NA	0.00
	4	21986845	1.00	0.92	INTRON	0.83	0.23	0.11	NA
	4	21986898	1.00	0.92	INTRON	0.89	0.00	0.00	0.00
	4	21986940	1.00	1.00	INTRON	NA	0.46	1.00	1.00
4	21986941	1.00	1.00	INTRON	NA	0.38	0.25	NA	
4	21986944	1.00	1.00	INTRON	1.00	0.46	1.00	1.00	
4	21986955	1.00	0.86	INTRON	NA	0.00	0.00	0.00	
4	21987008	1.00	0.81	INTRON	0.72	0.00	0.00	NA	
4	21987011	1.00	0.81	INTRON	NA	0.00	0.00	0.00	
4	21987039	1.00	0.78	INTRON	0.78	0.00	0.00	0.00	
4	21987046	1.00	0.78	INTRON	0.78	0.00	0.00	0.00	
4	21987063	1.00	0.83	INTRON	0.72	0.00	0.00	0.00	
4	21987164	1.00	0.81	SYNONYMOUS	0.56	0.00	0.00	0.00	
4	21987624	1.00	1.00	SYNONYMOUS	1.00	1.00	1.00	NA	
4	21987688	1.00	1.00	NON_SYNONYMOUS	1.00	1.00	1.00	NA	
4	21988281	1.00	1.00	NON_SYNONYMOUS	NA	1.00	NA	1.00	
4	21988358	1.00	0.83	SYNONYMOUS	0.61	0.00	0.00	0.00	

Table Notes: Allele frequencies are listed in the columns labeled by population. Highlighted rows are high-frequency shared variants between D3 and T5 but absent from other populations. NA represents sites with insufficient coverage (>10x).

Table S4.18 Genes within the genomic region shown in Figure S4.9 C,D

mRNA PAC ID	Scaffold	Lower Coordinate	Upper coordinate	<i>A. thaliana</i> ortholog	PANTHER	PFAM
16048372	4	21955708	21956963	AT2G45040.1	Matrix metalloproteinase	Matrixin, Putative peptidoglycan binding domain
16048972	4	21962501	21963548	AT2G45040.1	Transcription factor GATA, zinc finger (GATA type)	GATA zinc finger
16045866	4	21964161	21966093			
16048949	4	21966487	21967726	AT2G45070.4	Sec61beta-ProV protein	Sec61beta family
16059281	4	21971631	21972418	AT2G45080.1		Cyclin
16056840	4	21976885	21980773	AT2G45100.1	Transcription initiation factor IIB-related	Brf1-like TBP-binding domain
16063626	4	21984688	21988484	AT2G45110.1, AT1G65680.1, AT1G65681.1		Pollen allergen, Rare lipoprotein A (RlpA)-like double-psi beta-barrel

Table notes: Highlighted genes display many sites that are high frequency in D3 and T5 and absent in other populations (See Figure S4.9, Table S4.17).

BIBLIOGRAPHY

- Al-Shehbaz IA, O’Kane SL. 2002. Taxonomy and Phylogeny of Arabidopsis (Brassicaceae). The Arabidopsis Book.
- Andolfatto P, Davison D, Erezyilmaz D, Hu T, Mast J, Sunayama-Morita T, Stern D. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research* 21: 610-617.
- Arnold B, Bomblies K, Wakeley J. 2012. Extending coalescent theory to autotetraploids. *Genetics*, 192: 195-204.
- Arnold B, Corbett-Detig R, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol Ecol* 22: 3179-3190.
- Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. *Genome Res* 17: 1219-1227.
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3: E3376.
- Barringer B. 2007. Polyploidy and self-fertilization in flowering plants. *American Journal of Botany* 94: 1527–1533.
- Baudry E, Depaulis F. 2003. Effect of misoriented sites on neutrality tests with outgroup. *Genetics* 165: 1619-1622.
- Bever J, and Felber F. 1992. The theoretical population genetics of autopolyploids. *Oxford Surveys in Evolutionary Biology* 8: 185-217.
- Bomblies K, Madlung A. 2014. Polyploidy in the Arabidopsis genus. *Chromosome Res* 22: 117-34.
- Botstein D, White R, Skolnick M, Davis R. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32: 314-331.
- Brochmann C, Elven R. 1992. Ecological and genetic consequences of polyploidy in arctic *Draba* (Brassicaceae). *Evol Trends Plants* 6: 111-124.

- Butruille D, and Boiteux L. 2000. Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proc. Natl. Acad. Sci. USA* 97: 6608-6613.
- Cannings C. 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* 6: 260-290.
- Charlesworth B, Morgan T, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134: 1289-1303.
- Ciu L, Wall P, Leebens-Mack J, Lindsay B, Soltis D, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* 16: 738–749.
- Clark L, Stewart J, Nishiwaki A, Toma Y, Kjeldsen J, Jorgensen U, et al. 2015. Genetic structure of *Miscanthus sinensis* and *Miscanthus sacchariflorus* in Japan indicates a gradient of bidirectional but asymmetric introgression. *Journal of Experimental Botany*. doi:10.1093.
- Comai L. 2005. The advantages and disadvantages of being polyploidy. *Nat. Rev. Genet.* 6: 836-846.
- Cooper D, Antonarakis S, Krawczak M. 1995. The nature and mechanisms of human gene mutation. In: Scriver CR, Beaudet AL, Sly WS, Valle D (Eds.), *The metabolic and molecular bases of inherited disease* 7th ed McGraw-Hill, New York, 259–291.
- Corbett-Detig R, and Hartl D. 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genetics* 9: doi:10.1371.
- Cosgrove D, Bedinger P, Durachko D. 1997. Group I allergens of grass pollen as cell wall-loosening agents. *Proc Natl Acad Sci USA* 94: 6559-6564.
- Crow J. 1954. Random mating with linkage in polysomics. *Amer. Nat.* 88:431-434.
- Cui L, Wall P, Leebens-Mack J, Lindsay B, Soltis D, Doyle J, Soltis P, Carlson J, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16: 738-749.
- Cutler D, Jensen J. 2010. To pool, or not to pool? *Genetics* 186: 41-43.

- Davey J, Hohenlohe P, Etter P, Boone J, Catchen J, Blaxter M. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* 12: 499-510.
- Dehal P, Boore J. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* 3: doi:10.1371.
- Derome N, Métayer K, Montchamp-Moreau C, Veuille M. 2004. Signature of Selective Sweep Associated with the Evolution of sex-ratio Drive in *Drosophila simulans*. *Genetics* 166: 1357-1366.
- Doyle J, Flagel L, Paterson A, Rapp R, Soltis D, Soltis P, Wendel J. 2008. Evolutionary genetics of genome merger and doubling in plants. *Annual Review of Genetics* 42: 443-461.
- Emerson K, Merz C, Catchen J, Hohenlohe P, Cresko W, Bradshaw W, Holzapfel C. 2010. Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci* 107: 16196-200.
- Ewing G, Hermisson J. 2010. MSMS: A coalescent simulation program including recombination, demographic structure, and selection at a single locus. *Bioinformatics* 26: 2064-2065.
- Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa V, Foll M. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics* 9: doi:10.1371.
- Fay J, Wu C. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155: 1405-1413.
- Felber F, Bever J. 1997. Effect of triploid fitness on the coexistence of diploids and tetraploids. *Biol. J. Linn. Soc.* 60: 95-106.
- Fu Y. 1995. Statistical properties of segregating sites. *Theor Popul Biol* 48: 172-197.
- Gallardo M, Bickham J, Honeycutt R, Ojeda R, Köhler N. 1999. Discovery of tetraploidy in a mammal. *Nature* 401: 341.
- Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J, Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol Ecol* 22: 3165-3178.
- Gompert Z, Forister ML, Fordyce JA, Nice CC, Williamson RJ, Buerkle C. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* 19: 2455-2473.

- Gompert Z, Lucas L, Nice C, Fordyce J, Forister M, Buerkle C. 2012. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly. *Evolution* 66-7: 2167-2181.
- Grant V. 1981. *Plant speciation*, 2nd edition. New York, Columbia University Press. p. 563.
- Gregory T, Mable B. 2005. Polyploidy in animals, pp427-517 in *The Evolution of the Genome*, edited by T. R. Gregory. Elsevier, San Diego.
- Grinstead C, Snell J. 1997. *Introduction to Probability*, 2nd edition. American Mathematical Society. Providence, Rhode Island, USA.
- Haldane J. 1930. Theoretical genetics of autopolyploids. *J. Genet.* 22:359-372.
- Henry I, Dilkes B, Young K, Watson B, Wu H, Comai L. 2005. Aneuploidy and genetic variation in the *Arabidopsis thaliana* triploid response. *Genetics* 170: 1979-1988.
- Hill R. 1971. Selection in Autotetraploids. *Theor Appl Genet* 41: 181-186.
- Hohenlohe P, Amish S, Catchen J, Allendorf F, Luikart G. 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology* 11: 117-122.
- Hohenlohe P, Bassham S, Currey M, Cresko W. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Phil. Trans. R. Soc. B* 367: 395–408.
- Hohenlohe P, Bassham S, Etter P, Stiffler N, Johnson E, Cresko W. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genetics* 6: e1000862.
- Hohmann N, Schmickl R, Chiang T, Lucanova M, Kolar F, Marhold K, Koch M. 2014. Taming the wild: resolving the gene pools of non-model *Arabidopsis* lineages. *BMC Evol Biol* 14: e224.
- Hollister J, Arnold B, Svedin E, Xue K, Dilkes B, Bomblies K. 2012. Genetic Adaptation Associated with Genome-Doubling in Autotetraploid *Arabidopsis arenosa*. *PLoS Genetics* 8: doi:10.1371.
- Hu T, Pattyn P, Bakker E, Cao J, Cheng J, Clark R, Fahlgren N, Fawcett J, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43, 476-481.

- Hudson R. 2002. Generating samples under a Wright– Fisher neutral model. *Bioinformatics* 18: 337–338.
- Husband B. 2004. The role of triploid hybrids in the evolutionary dynamics of mixed-ploidy populations. *Biol. J. Linn. Soc.* 82: 537-546.
- Husband B, Sabara H. 2011. Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, *Chamerion angustifolium* (Onagraceae). *New Phytologist* 161: 703–713. doi:10.1046.
- Jiao Y, Wickett N, Ayyampalayam S, Chanderbali A, Landherr L, Ralph P, Tomsho L, Hu Y, Liang H, Soltis P. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97-100.
- Jombart T, Ahmed I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27: 3070–3071.
- Jørgensen M, Ehrich D, Schmickl R, Koch M, Brysting A. 2011. Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae). *BMC Evol Biol* 11: 346.
- Junier T, Zdobnov E. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* 26, 1669-1670.
- Kellis M, Birren B, Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617-624.
- Kingman J. 1982a. The coalescent. *Stochastic process. Appl.* 13: 235-248.
- Kingman J. 1982b. On the Genealogy of Large Populations. *J Appl Probab* 19: 27-43.
- Kolar F, Lucanova M, Zaveska E, Fuxova G, Mandakova T, Spaniel S, Senko D, Svitok M, Kolnik M, Gudzinskas Z, et al. 2015. Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the *Arabidopsis arenosa* group (Brassicaceae). *Biol J Linn Soc*: in press.
- Kolnik M. 2007. *Arabidopsis*. In: Marhold K, Martonfi P, Mereda P, Mraz P, editors. *Chromosome number survey of the ferns and flowering plants of Slovakia*. Bratislava: VEDA, pp 94-102.
- Langley C, Stevens K, Cardeno C, Lee Y, Schrider D, Pool J, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192: 533-598.

- Leggatt R, Iwama G. 2003. Occurrence of polyploidy in the fishes. *Rev. Fish Biol. Fish.* 13: 237–246.
- Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 000: 1-3.
- Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21: 936-939.
- Luo J, Gao Y, Ma W, Bi X, Wang S, Wang J, Wang Y, Chai J, Du R, Wu S, et al. 2014. Tempo and mode of recurrent polyploidization in the *Carassius auratus* species complex (Cypriniformes, Cyprinidae). *Heredity* 112: 415-427.
- Luo Z, Zhang Z., Zhang R, Pandey M, Gailing O, et al. 2006. Modeling population genetic data in autotetraploid species. *Genetics* 172: 639-646.
- Mable B. 2004. Polyploidy and self-compatibility: Is there an association? *New Phytologist* 162: 803–811.
- Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264: 421-424.
- Mather K. 1935. Reductional and equational separation of the chromosomes in bivalents and multivalents. *J. Genet.* 30: 53-78.
- Mather K. 1936. Segregation and linkage in autotetraploids. *J. Genet.* 32: 287-314.
- Maynard-Smith J, and Haigh J. 1974. The hitchhiking effect of a favorable gene. *Genet Res* 23: 23-35.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-1303.
- Messer P, Petrov D. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* 28: 659-669.
- Miyashita N, Langley C. 1988. Molecular and phenotypic variation of the white locus region in *Drosophila melanogaster*. *Genetics* 120: 199–212.
- Möhle M. 1998a. A convergence theorem for markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.* 30: 493-512.

- Möhle M. 1998b. Coalescent results for two-sex population models. *Adv. Appl. Prob.* 30: 513-520.
- Möhle M. 1998c. Robustness results for the coalescent. *J. Appl. Prob.* 35: 438-447.
- Möhle M, Sagitov S. 2001. A classification of coalescent processes for haploid exchangeable population models. *Ann. Appl. Probab.* 29: 1547-1562.
- Moody M, Mueller L, Soltis D. 1993. Genetic variation and random drift in autotetraploid populations. *Genetics* 134: 649-657.
- Nordborg M, Donnelly P. 1997. The coalescent process with selfing. *Genetics* 146: 1185-1195.
- Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 40: 646-649.
- Oberle B, Montgomery R, Beck J, Esselman E. 2012. A morphologically intergrading population facilitates plastid introgression from diploid to tetraploid *Dodecatheon* (Primulaceae). *Botanical Journal of the Linnean Society* 168: 91–100.
- Otto S, Whitton J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* 2000. 34:401-37.
- Pandey K. 1977. Origin of complementary incompatibility systems in flowering plants. *Theoretical and Applied Genetics*, 49: 101–109.
- Parchman T, Gompert Z, Mudge J, Schilkey F, Benkman C, Buerkle C. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* 21: 2991–3005.
- Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytol* 186: 5-17.
- Peterson B, Weber J, Kay E, Fisher H, Hoekstra H. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7: doi:10.1371.
- Petit C, Bretagnolle F, Felber F. 1999. Evolutionary consequences of diploid-polyploid hybrid zones in wild species. *Trends Ecol. Evol.* 14: 306-311.
- Pfender W, Saha M, Johnson E, Slabaugh M. 2011. Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theor Appl Genet* 122: 1467–1480.

- Pickrell J, Pritchard J. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* 8: doi:10.1371.
- Pool J, Corbett-Detig R, Sugino R, Stevens K, Cardeno C, et al. 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. Accepted. *PLoS Genetics*.
- Pool J, Hellmann I, Jensen J, Nielsen R. 2010. Population genetic inference from genomic sequence variation. *Genome Research* 20: 291–300.
- Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
- Ptacek M, Gerhardt H, Sage R. 1994. Speciation by polyploidy in treefrogs: multiple origins of the tetraploid, *Hyla versicolor*. *Evolution* 48: 898-908.
- Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. 2012. SNP calling by sequencing pooled samples. *BMC Bioinformatics* 13: 239.
- Ramsey J, Schemske D. 1998. Pathways, mechanisms, and rates of polyploidy formation in flowering plants. *Annu Rev Ecol Syst* 29: 467-501.
- Rausch J, Morgan M. 2005. The effect of self-fertilization, inbreeding depression, and population size on autopolyploid establishment. *Evol*: 59:1867-1875.
- Rodriguez D. 1996. A model for the establishment of polyploidy in plants. *The Am. Nat.* 147: 33-46.
- Rokas A, Abbot P. 2009. Harnessing genomics for evolutionary insights. *Trends in Ecology and Evolution* 24: 192-200.
- Ronfort J. 1999. The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. *Genetical Research* 74: 31–42.
- Ronfort J, Jenczewski E, Bataillon T, Rousset F. 1998. Analysis of population structure in autotetraploid species. *Genetics* 150: 921-930.
- Ronikier M. 2011. Biogeography of high-mountain plants in the Carpathians: An emerging phylogeographical perspective. *Taxon* 60: 373-389.

- Schmickl R, Paule J, Klein J, Marhold K, Koch M. 2012. The evolutionary history of the *Arabidopsis arenosa* complex: Diverse tetraploids mask the Western Carpathian center of species and genetic diversity. *PLoS ONE* 7: doi:10.1371.
- Segraves K, Thompson J, Soltis P, Soltis D. 1999. Multiple origins of polyploidy and the geographic structure of *Heuchera grossulariifolia*. *Mol Ecol* 8: 253-262.
- Sexton O. 1980. Polyploidy in animal evolution: summary, pp 379-381 in *Polyploidy, biological relevance*, edited by W. H. Lewis. Plenum Press, New York.
- Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics* 169: 1061–1070.
- Soltis D, Buggs R, Doyle J, Soltis. 2010. What we still don't know about polyploidy. *Taxon* 59: 1387-1403.
- Soltis D, Soltis P, Ness B. 1989. Chloroplast-DNA Variation and Multiple Origins of Autopolyploidy in *Heuchera micrantha*. *Evolution* 43: 650–656.
- Soltis D, Soltis P, Schemske D, Hancock J, Thompson J, et al. 2007. Autopolyploidy in angiosperms: Have we grossly underestimated the number of species? *Taxon* 56: 13–30.
- Sonnleitner M, Weis B, Flatscher R, García P, Suda J, Krejčíková J, et al. 2013. Parental ploidy strongly affects offspring fitness in heteroploid crosses among three cytotypes of autopolyploid *Jacobaea carniolica* (Asteraceae). *PLoS ONE* 8: doi:10.1371.
- Ståhlberg D. 2009. Habitat differentiation, hybridization and gene flow patterns in mixed populations of diploid and autotetraploid *Dactylorhiza maculata* s.l. (Orchidaceae). *Evol Ecol* 23, 295-328.
- Stebbins G. 1947. Types of polyploids: their classification and significance. *Adv. Genet.* 1: 403-429.
- Stebbins G. 1950. *Variation and Evolution in Plants*. New York: Columbia University Press.
- Stift M, Berenos C, Kuperus P, van Tienderen P. 2008. Segregation Models for Disomic, Tetrasomic and Intermediate Inheritance in Tetraploids: A General Procedure Applied to *Rorippa* (Yellow Cress) Microsatellite Data. *Genetics* 179: 2113-2123.

- St. Onge, K, Foxe J, Li J, Li H, Holm K, et al. 2012. Coalescent-based analysis distinguishes between allo- and autopolyploid origin in Sheperd's Purse (*Capsella bursa-pastoris*). *Mol. Biol. Evol.* In press.
- Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437-460.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.
- The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796–815.
- Thompson J, Lumaret R. 1992. The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends in Ecology & Evolution (Personal Edition)*, 7: 302–307.
- Thornton K. 2009. Automating approximate Bayesian computation by local linear regression. *BMC Genet.* 10: 35.
- Thórsson T, Salmela E, Anamthawat-Jónsson K. 2001. Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch. *J Hered* 92: 404-8.
- Tiffin P, Gaut B. 2001. Sequence Diversity in the Tetraploid *Zea perennis* and the Closely Related Diploid *Z. diploperennis*: Insights From Four Nuclear Loci. *Genetics* 158: 401-412.
- Van Dijk P, Bakx-Schotman T. 1997. Chloroplast DNA phylogeography and cytotype geography in autopolyploid *Plantago media*. *Mol Ecol* 6: 345-352.
- Van Tassell C, Smith T, Matukumalli L, Taylor J, Schnabel R, Lawley C, Haudenschild C, Moore S, Warren W, Sonstegard T. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* 5: 247-252.
- Wakeley J. 1999. Non-equilibrium migration in human history. *Genetics* 153: 1863-1871.
- Wakeley J. 2008. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.
- Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics* 159: 893-905.

- Watterson G. 1975. On the number of segregating sites in genetical models without recombination. *Theoret Pop Biol* 7: 256-276.
- Weir B, Cockerham C. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358-1370.
- Wolf P, Soltis D, Soltis P. 1990. Chloroplast-DNA and allozymic variation in diploid and autotetraploid *Heuchera grossulariifolia* (Saxifragaceae). *Am. J. Bot.* 77: 232-244.
- Wood T, Takebayashi N, Barker M, Mayrose I, Greenspoon P, Rieseberg L. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci USA* 106: 13875–13879.
- Wright S. 1938. The Distribution of gene frequencies in population of polyploids. *Proc. Nat. Acad. Sci.* 24: 372-377.
- Wright K, Arnold B, Xue K, Surinova M, O'Connell J, Bomblies K. 2014. Habitat and cytotype associated selection on meiosis proteins in *Arabidopsis arenosa*. *Mol Biol Evol*, in press.
- Yamane K, Yasui Y, Ohnishi O. 2003. Intraspecific cpDNA variations of diploid and tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). *Am J Bot* 90: 339-346.
- Yang W, Glover B, Rao G, Yang J. 2006. Molecular evidence for multiple polyploidization and lineage recombination in the *Chrysanthemum indicum* polyploid complex (Asteraceae). *New Phytol* 171: 875-886.
- Yant L, Hollister J, Wright K, Arnold B, Higgins J, Franklin C, Bomblies K. 2013. Meiotic Adaptation to Genome Duplication in *Arabidopsis arenosa*. *Curr Biol* doi:10.1016.