Inference of transitions to self-fertilization using haplotype genomic variation

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Stefan Strütt
aus Freiburg i. Br.
Deutschland

Gutachter und Prüfer

1. Prof. Dr. Miltos Tsiantis
2. Prof. Dr. Joachim Krug

Vorsitzender der Prüfungskommission

Prof. Dr. Achim Tresch

Beisitzender

Dr. Stefan Laurent

Zeitraum der letzten mündlichen Prüfung

01.08.2022 – 26.08.2022

Die vorliegende Arbeit wurde am Max-Planck-Institut für Pflanzenzüchtungsforschung in Köln in der Abteilung für vergleichende Entwicklungsgenetik (Direktor: Prof. Dr. Miltos Tsiantis) angefertigt.

Abstract

Abstract

Mating systems play an essential role in the evolution of natural populations. The reproductive mode of a population affects the evolutionary forces and recombination. Shifts in mating systems change major evolutionary traits of natural populations and affect the life-history cycle on many different levels. Among all transitions of mating schemes, a shift from outcrossing to selfing is one of the major shifts in plants. Such shifts have repeatedly occurred on the phylogenetic level. Despite their importance, there were no published tools to estimate such transitions in natural populations using genetic data on a genome-wide level. Existing estimates rely on estimating the loss-of-function mutations of causal loci. However, such estimates rely on the knowledge of the underlying genetic mechanism to induce the shift from outcrossing to selfing. Thus, such estimates are restricted to be conducted on very few species.

In this study, we investigated the genetic consequences of shifts from outcrossing to selfing (Chapter 1). We used extensive simulations of the forward-in-time Wright-Fisher model and the backward-in-time coalescent model. We found the previously described theoretical work on implementing partial selfing in the coalescent to suffice in simulating transitions to selfing. We developed an Approximate Bayesian Computation approach (*tsABC*) to identify and estimate the date of transitions from outcrossing to selfing using a pairwise comparison of genomes (Chapter 2). Finally, in collaboration with Thibaut Sellinger, we introduced the modified *PSMC'* (*teSMC*) to estimate piecewise-constant selfing rates through time jointly with piecewise-constant population sizes for single-population demographies and analyzed its accuracy (Chapter 3). Taken together, we provide not only an approximate Bayesian but also a maximum likelihood approach to identify and estimate transitions to selfing for single populations. We found *tsABC* to be a versatile tool to identify and estimate transitions to selfing. Under carefully made assumptions for the proposed models, transitions to selfing can be detected under a broad range of scenarios. Moreover, the assumed model in the *teSMC* method improved the estimates of demography and detected

transitions to selfing at least as powerful as the *tsABC*. The automized parametrization of *teSMC* allows users with little expertise in scripting to use it.

We used both methods to estimate the transition from outcrossing to selfing for three genetic clusters of *Arabidopsis thaliana*. Our results were consistent with each other and existing estimates from the literature.

With our study, we not only contributed to the understanding of evolutionary processes that formed the genetic diversity of natural populations but also provided two powerful tools to investigate the demographic history of natural populations in the context of transitions to selfing. Recombination provides a molecular clock on a separate time scale compared to mutation that interacts with all the four evolutionary forces at various levels. Eventually, that will contribute to understanding the functions of genes and their relationship and interaction with the bearing individual, the population, and the environment. Taken together, selfing as a breeding scheme or reproductive strategy is a crucial trait that interferes and connects evolutionary forces, adaptive potential, and life-history traits of natural populations.

Zusammenfassung (Abstract in German)

Paarungssysteme spielen eine wichtige Rolle in der Evolution natürlicher Populationen. Die Reproduktionsart einer Population wirkt sich auf die Evolutionskräfte und den Rekombinationsprozess aus. Verschiebungen in Paarungssystemen verändern wichtige evolutionäre Merkmale natürlicher Populationen und beeinflussen den Lebenszyklus auf vielen verschiedenen Ebenen. Unter allen Verschiebungen, die in Paarungssystemen von Pflanzen vorkommen, ist der Übergang vom Auskreuzen zur Selbstbestäubung der am häufigsten vorkommende. Solche Übergänge fanden in der Geschichte der Evolution häufig statt, wie man bei phylogenetischen Analysen sehen kann. Trotz ihrer Bedeutung wurden bisher keine Methoden entwickelt, um solche Übergänge in natürlichen Populationen anhand genetischer Daten auf genomweiter Ebene zeitlich zu bestimmen. Vorhandene Schätzungen beruhen auf der Schätzung des Funktionsverlustes durch Mutationen der beteiligten Loci. Solche Schätzungen beruhen jedoch auf der Kenntnis des zugrunde liegenden genetischen Mechanismus, welcher den Übergang von Auskreuzung zur Selbstbefruchtung bewirkt. Daher ist die Durchführung solcher Schätzungen auf sehr wenige Arten beschränkt.

In dieser Studie untersuchten wir die genetischen Konsequenzen der Verschiebung von Auskreuzung zur Selbstbestäubung (Kapitel 1). Wir simulierten derartige Übergange umfangreich, einerseits mit Wright-Fisher-Modellen, die zeitlich fortlaufend eine gesamte Population inklusive ihrer expliziten Reproduktion simulieren, und andererseits mit Modellen der Koaleszenztheorie, bei der genetische Abstammung in der Zeit rücklaufend simuliert wird. Die Ergebnisse zeigten, dass die zuvor in der Koaleszenstheorie entwickelte theoretische Arbeit zur Implementierung anteiliger Selbstbefruchtung hinreichend sind, um die Übergänge zur Selbstbefruchtung zu simulieren. Wir haben eine näherungsweise bayessche Berechnungs-Methode (*tsABC*) entwickelt, mittels derer wir Übergange vom Auskreuzen zur Selbstbefruchtung, durch das paarweise Vergleichen von Genomen, identifizieren und zeitlich abzuschätzen können (Kapitel 2). Außerdem

erweiterten wir in Zusammenarbeit mit Thibaut Sellinger die PSMC‘-Methode zu *teSMC*, um die stückweise-konstante Selbstbefruchtungsrate zeitgleich mit der stückweise-konstanten Populationsgröße abzuschätzen. Zudem analysierten wir die Genauigkeit dieser Schätzungen (Kapitel 3). Wir entwickelten also nicht nur einen bayesschen, sondern auch eine Maximum-Likelihood-Methode, um Übergänge zur Selbstbefruchtung einzelner Populationen zu identifizieren und zu datieren. Wir stellten fest, dass *tsABC* ein vielseitiges Werkzeug ist, um Übergänge zum Selbsten zu identifizieren und zu datieren. Unter sorgfältig getroffenen Annahmen für die vorgeschlagenen Modelle können Übergänge zur Selbstbefruchtung in einer Vielzahl von Szenarien identifiziert werden. Darüber hinaus verbesserte das angenommene Modell in der *teSMC*-Methode die Schätzungen des zeitlichen Verlaufs der Populationsgröße und identifizierte Übergänge zur Selbstbefruchtung, die mindestens genauso robust waren wie die Schätzungen mit *tsABC*. Die automatisierte Parametrisierung von *teSMC* ermöglicht Benutzern mit wenig Fachwissen die Anwendung.

Wir schätzten den Übergang vom Auskreuzen zur Selbstbefruchtung anhand der Daten von drei genetischen Clustern mit beiden entwickelten Methoden. Die Ergebnise stimmten miteinander und mit bereits veröffentlichten Schätzung überein.

Mit unserer Studie trugen wir nicht nur zum Verständnis evolutionärer Prozesse bei, welche die genetische Vielfalt natürlicher Populationen formten, sondern stellten auch zwei leistungsstarke Methoden zur Verfügung, um die demografische Geschichte natürlicher Populationen im Zusammenhang mit Übergängen zur Selbstbefruchtung zu untersuchen. Rekombination bietet eine molekulare Uhr auf einer anderen Zeitskala als die der Mutation, welche mit allen vier evolutionären Kräften auf verschiedenen Ebenen interagiert. Letztendlich wird dies zum Verständnis der Funktionen von Genen und ihrer Beziehung und Interaktion mit dem jeweiligen Individuum, der Population und der Umwelt beitragen. Zusammenfassend kann die Selbstbestäubung als Züchtungsschema oder Fortpflanzungsstrategie als entscheidendes Merkmal betrachtet werden, welches mit den evolutionären Kräften, dem Adaptionspotenzial und dem Lebenszyklus natürlicher Populationen interagiert.

Table of contents

General introduction 1

The evolution of mating systems from outcrossing to selfing

Keywords 2
Motivational statement and relevance of this project 2
Introduction, general section 2
On the evolution of selfing 3
The selfing syndrome in plants 3
Population dynamics of transitions to selfing 4
Life history of *Arabidopsis thaliana* 5
Coalescent based demographic inference 6
Aim of this project 7

Chapter 1 8

Genetic consequences of transitions from outcrossing to selfing

Keywords 9
Introduction 9
Models and Methods 10
Population genetics models in the context of changing selfing rates 10
Simulation and demographic models in the context of selfing 15
Implementation 17
Classical site-based summarization of genetic diversity 18
Haplotype-based summarization of the genetic diversity 20
Summarizing the clustering property of non-recombining segments 20
Joint summarization of recombination and drift 20

Results 21

Theory predicts the true rescaling of the recombination rate under low mutation rates 21

The rescaled coalescent simulates a single haplotype per individual 23

The coalescent model accurately simulates direct measures of the joint effects of transitions to selfing 25

Summarizing the clustering property of non-recombining segments 27

Observing genetic diversity consequences of selfing on MRCA segments 29

Comparison of the effects of implicit and explicit selfing on the summary statistics 31

Discussion 34

Conclusion 37

Author contributions 37

Chapter 2 *38*

Identifying and estimating transitions from outcrossing to selfing using *tsABC*

Keywords 39

Introduction 39

Existing estimates for *Arabidopsis thaliana* 40

Models and Methods 41

ABC in population genetics 41

Modeling a transition to selfing 41

- Modeling negative linked selection with forward-in-time WF simulations 43
- Coalescent simulations for the ABC 44
- Identifying a transition to selfing 44

Calculation of the summarizing statistics of genetic diversity 44

Dimensionality reduction 45

Parameters and Priors 46

Model choice 47

Parameter inference 48

Simulations and observation of data 48

Simulating negative linked selection and background selection 49

Dating transitions to selfing in *Arabidopsis thaliana* 50
Implementation and software 50
Results **51**
Identifying transitions to selfing 52
Dating transitions to selfing 54
Importance of the ratio between mutation and recombination rate 56
Robustness of *tsABC* to negative linked selection in parameter estimation 57
Application of *tsABC* to *Arabidopsis thaliana* 60
Discussion **62**
Conclusion **65**
Author contributions **66**
Chapter 3 *67*
Inferring demography and piecewise-constant selfing using *teSMC*
Keywords **68**
Introduction **68**
Models and Methods **70**
Hidden Markov models in the coalescent framework 70
SMC-HMM 72
Modifications of *eSMC* to infer changing selfing rates 73
Simulations and inference 74
Results **74**
Theoretical convergence 74
Estimates on simulated data 76
Masking for exonic regions improves *teSMC* inference robustness 82
Application of *tsABC* to *Arabidopsis thaliana* 83
Discussion **84**
Conclusion **87**

Author contributions 87

General discussion 89

Outlook and implications of identifying and estimating transitions from outcrossing to selfing

References 92

Appendix 100

Description of *teSMC* 101

Table of figures 107

Table of tables 116

Abbreviations 117

Acknowledgment 119

Administratives 121

Declaration 122

Publications 123

In preparation 123

Curriculum vitae 124

Personal details 124

Education 124

General introduction

The evolution of mating systems from outcrossing to selfing

The evolution of mating system shifts from outcrossing to selfing

Keywords

Selfing, transition to selfing, self-incompatibility, life-history trait evolution, selfing syndrome, statistical modeling, coalescent theory

Motivational statement and relevance of this project

“Nothing in biology makes sense except in the light of evolution.” (Dobzhansky, 1973, 2013) is probably the most cited statement in lectures and textbooks introducing evolutionary biology. Since genetics gained importance and became centered in biological research, population genetics mark the only field directly accessing, measuring, and modeling evolutionary forces to establish evolutionary hypothesis on experienceable questions and time scales.

I would like to add the importance of modeling further and anticipate methodological criticism of this project by citing George Box: All models are wrong, but some are useful (Box, 1976). Understanding the genetic basis of developmental research on morphological evolution must be embedded into population genetics to provide robustness.

Introduction, general section

Self-fertilization (selfing) is sexual reproduction, i. e. the life-cycle of a species includes gamete formation and reduction of the chromosome set and their fusion to form a complete heterozygote. Selfing describes the mode of sexual reproduction of both gametes descending from the same parental individual. Sexual reproduction is often the mechanism to enforce recombination and, thus, a fundamental mechanism for evolution at the individual, population, and species level (S. P. Otto & Lenormand, 2002). The evolution of sexes is a major topic in biology. The evolution of selfing is an important part of the evolution of sexes. Investigating the rate at which selfing occurs on the phylogenetic level may provide evidence for standing hypotheses based on the theory that has been developed to substantiate basic principles of the theory of evolution (Shimizu & Tsuchimatsu, 2015).

The evolution of sexual reproduction from the ancestral asexual reproduction is associated with compensating for the two-fold cost of sex (J Maynard Smith, 1971; John Maynard Smith & Maynard-Smith, 1978). Asexually reproducing individuals outcompete individuals performing sexual reproduction under otherwise neutral assumptions. Most species, including plants, reproduce sexually (Barton & Charlesworth, 1998). Selfers benefit from the two-fold transmission advantage resulting in automatic selection (Busch & Delph, 2011; Fisher, 1941). However, selfing shuts down recombination. Taken together, it raises the hypothesis that selfing is an evolutionary dead-end (Igic & Busch, 2013; G. L. Stebbins, 1957; G Ledyard Stebbins, 1974; S. I. Wright, Kalisz, & Slotte, 2013). However, data-based evidence is sparse. Thus, a systematic phylogenetic analysis of selfing species is inevitable.

On the evolution of selfing

The selfing syndrome in plants

In plants (and animals), species performing high selfing rates manifest phenotypes distinguishing them from obligatory outcrossing sister species. The so-called ‘selfing syndrome‘ includes small flowers with stigmas and anthers in close proximity, limited pollen, and reduced longevity (Ornduff, 1969; Sicard & Lenhard, 2011). Selfing also correlates with genomic features, established as ‘genomic selfing syndrome’: accumulation of deleterious mutations, smaller genomes and loss of transposable elements, and enhanced structural evolution of chromosomes. Immediate population genetic consequences are a reduced diversity and increased linkage disequilibrium due to reduced heterozygosity (S. C. Barrett, Arunkumar, & Wright, 2014; Glémin & Galtier, 2012; Shimizu & Tsuchimatsu, 2015). The phenotypic and genotypic ‘selfing syndrome’ is hypothesized to be a consequence of selfing; however, the relationship between cause and effect may be hard to state (Cutter, 2019).

Taken together, selfing marks an essential trait with broad effects not only on determining spatial and temporal patterns of genetic diversity but also on morphological and ecological properties, even dispersal and speciation (Cutter, 2019; Epinat & Lenormand, 2009).

Population dynamics of transitions to selfing

Most angiosperms have hermaphroditic flowers providing the general possibility to perform selfing. However, only ~15% are currently known to perform predominant selfing (Goodwillie, Kalisz, & Eckert, 2005). Still, transitions from outcrossing to selfing are the most frequent shifts in mating schemes that have occurred both often and independently throughout the phylogeny (Franklin-Tong, 2008).

There are two main hypotheses about how selfing evolves in a population. The first to state is the automatic selection, i. e. the two-fold transmission advantage of selfing individuals that transmit both and not only one of the haplotypes to the next generation (Fisher, 1941; Goodwillie et al., 2005). That advantage outcompetes outcrossing individuals of the same population under otherwise neutral assumptions. Moreover, selfing is an evolutionary stable strategy (ESS), i. e. no other mating strategy provides an enhanced fitness compared to selfing individuals. Thus transitions to selfing must occur one-directional towards selfing and cannot be reversed (J Maynard Smith, 1971; John Maynard Smith & Maynard-Smith, 1978). The second hypothesis is reproductive assurance which states the advantage of reproduction by selfing in ecological conditions that impede outcrossing (Darwin, 1876). Baker’s law promotes that hypothesis. Furthermore, woody perennial species tend to self at lower rates compared to small herbal plants. Additionally, perennial herbs tend to self on lower rates compared to annuals (Spencer CH Barrett & Harder, 1996). Both are indicative of reproductive assurance. Taken together, selfing provides a fitness advantage and can invade outcrossing populations immediately if there is no disadvantageous effect: Inbreeding depression marks the only short-term negative effect of selfing compared to outcrossing, reviewed by Busch and Delph (2011). However, selfing provides enhanced rates of purging, indicating that the transition phase is a critical phase for the fitness of a population (Arunkumar, Ness, Wright, & Barrett, 2015). However, allotetraploidization has been established to be correlated with transitions to selfing and attenuate the disadvantageous effects of inbreeding (Comai, 2005; Glémin, François, & Galtier, 2019; Sarah P Otto & Whitton, 2000). Long-term wise, populations performing selfing are established to suffer from a reduced adaptation potential which raised the dead-end hypothesis of selfing as a reproductive strategy (Igic & Busch, 2013; G. L. Stebbins, 1957; G Ledyard Stebbins, 1974; S. I. Wright et al., 2013).

Selfing is an ESS; thus, if selfing is prevented, outcrossing is maintained as the reproductive mode in a population. Several self-incompatibility (SI) mechanisms have been discovered (Franklin-Tong, 2008; Nasrallah, 2019). Brassicaceae prove a genetic mechanism consisting of a single locus containing two multiallelic genes. The allelic combination determines an S-haplotype. The recognition of an individual plant to reject self-pollen relies on detecting the S-haplotype. The molecular mechanism is not fully understood however a disruption of the function of one of the genes results in a dominant self-compatibility (SC) that is promoted on the RNA level (de Nettancourt, 1997; Franklin-Tong, 2008; Shimizu & Tsuchimatsu, 2015; Suwabe et al., 2020).

In summary, investigating and dating transitions to selfing throughout the phylogeny will contribute to understanding the evolution of selfing. It may answer open questions (or raise new ones) and help to close some gaps in understanding the evolution of sexes.

Life history of *Arabidopsis thaliana*

The most supported hypothesis of *Arabidopsis thaliana*'s demographic history states that its origin is in Africa and roughly follows the out-of-Africa paradigm of humans (Durvasula et al., 2017). Most sister species of *Arabidopsis thaliana* perform obligate outcrossing. Additionally, the gain of predominant selfing occurred by losing self-incompatibility, the genetic mechanism to enforce outcrossing (Franklin-Tong, 2008; Takayama & Isogai, 2005). Three independent mutations have been identified and were found in every existing population outside of Africa (Durvasula et al., 2017). These findings, taken together, indicate that the transition to selfing occurred before the split and in an ancestral population before migration to Eurasia started (Durvasula et al., 2017).

A high proportion of selfing was observed for independent populations of *Arabidopsis thaliana*. A lower limit of selfing was found to be 0.99715 for seven independent investigated groups, which all performed predominant selfing (Abbott & Gomes, 1989). The flowers of *Arabidopsis thaliana* open late in development after the anthers have elongated, enabling self-pollination during outgrowth (Alvarez-Buylla et al., 2010). Thus, flowers open after fertilization, and only under high pollinator activity, marginal outcrossing rates may be exceeded. Since selfing is an ESS, low and intermediate selfing rates require mechanistic or genetic enforcement. Reduced selfing rate estimates on different *Arabidopsis thaliana* populations recently have not been critically revised and may have arisen under overconfident prior assumptions; e. g. (Sellinger, Abu Awad, Moest, & Tellier, 2020; Tang et al., 2007).

Coalescent based demographic inference

Statistical models provide a systematical approach to jointly sample parameters and statistics under priorly determined theoretical assumptions (D. R. Cox, 1990; D. R. S. E. J. Cox, 1981; McCullagh, 2002). In population genetics, statistical models are used to describe and infer past stochastical processes, e. g., the demographic history (Mark A Beaumont, 2010; M. A. Beaumont & Rannala, 2004; M. A. Beaumont, Zhang, & Balding, 2002). The coalescent model describes the genealogy of a sample from a population backward-in-time. The coalescent theory can be derived from the forward-in-time Wright-Fisher or the Moran model, requiring fewer assumptions than the coalescent model and thus providing a golden standard for simulating diversity. In other words, under certain assumptions, the properties of the Wright-Fisher and the Moran model converge with the properties of the coalescent theory. Indeed, the coalescent theory helped to statistically summarize and understand the effect of evolutionary forces and population genetic properties of populations. In population genetics, opposing to the term definition in descriptive statistics, a population classically is defined as a set of individuals performing random mating. That is a helpful definition for modeling, yet not necessarily true in natural populations. That is why, here, we refer to populations in the modeling context, but to natural populations or genetic groups or clusters if referring to sample sets from natural observations.

Aim of this project

With this study, we aim to develop and provide a method to identify and estimate the time of transitions from outcrossing to predominant selfing using whole-genome haplotype variation. We describe the genetic consequences of transitions to selfing using forward-in-time Wright-Fisher models simulating explicit reproduction under selfing to achieve this goal. Then we compare our findings with simulated genetic data under the coalescent with partial selfing, which approximates implicitly selfing as a reproductive mode, but is computationally more tractable. Based on our findings, we introduce a new summarization statistic that captures the genetic diversity of MRCA segments, which we define as tracts of a pairwise comparison of sequences that descend from the same most recent common ancestor (MRCA) compared to its neighboring tracts. We implement our insights not only in developing an approximate Bayesian computation (ABC) method but also a maximum-likelihood-estimation-method (MLE) based on the Sequential Markovian Coalescent (SMC) to identify and estimate transitions from outcrossing to selfing. We analyze the performance of both the methods and apply them to three distinct genetic clusters of *Arabidopsis thaliana*.

Chapter 1

Genetic consequences of transitions from outcrossing to selfing

Genetic consequences of transitions from outcrossing to selfing

Keywords

Selfing, recombination, mating systems, population genetics, demography

Introduction

The effects of predominant selfing as a reproductive mode on genetic diversity have been described and reviewed in many aspects, e. g, reviewed by Cutter (2019). Selfing occurring via autogamy within a hermaphroditic flower has been described to be associated with the evolution of specific morphological traits of the flower, like reduced petal sizes and pollen reduction (Darwin, 1876; Ornduff, 1969; Wilcock, 1987). Additionally, predominant selfing predicts convergent features of a genomic selfing syndrome: including an increase in the genetic load and more compact genomes (Cutter, 2019). Moreover, theoretical work has been published describing the population genetics aspects of selfing (Glémin, 2021; Hartfield, Bataillon, & Glemin, 2017). However, there is no formal description of the effects of transitions to selfing on a whole-genome level enabling its summarization to be informative about recent and mid-recent transitions to selfing. Multisite statistics, such as linkage, have become only recently of broader interest since next-generation sequencing provides affordable access to such genetic information. Therefore, developing population genetic tools to incorporate such information becomes inevitable.

There are generally two approaches to investigate the effects of selfing on whole-genome genetic diversity. 1) We can describe and investigate genetic patterns of natural populations of which we priorly know that they perform selfing, and 2) we generate a model with which we mimic the effects of selfing and investigate and describe the effects of selfing on genetic diversity. The earlier approach suffers from uncertainties of traits, e. g. life-history traits, and possible confounding factors of the recent evolutionary history of the given populations if unknown. Importantly, we cannot investigate the temporal development of evolving trait processes. Moreover, we must have prior knowledge of the effects of selfing to interpret the genetic diversity in those populations depending on the populations. Here, we aim to investigate the effects of a transition from predominant outcrossing to predominant selfing on whole-genome genetic diversity. Our findings will provide insights into developing a statistical framework to identify and estimate the age of such transitions using whole-genome diversity.

In this study, we simulate and describe genetic diversity using two common model frameworks to investigate how transitions to selfing shape genetic diversity. We use the standard summarizing statistics, the site-frequency spectrum, and linkage-disequilibrium to describe the genetic diversity. Additionally, we develop a novel optimized summarizing statistic to capture the temporal signature of transitions from outcrossing to selfing. Our findings will contribute to the understanding of how to efficiently summarize genetic diversity and, thus, contribute to inferring recent transitions from outcrossing to selfing of populations and species throughout the phylogeny.

Models and Methods

Population genetics models in the context of changing selfing rates

In classical population genetics modeling approaches, we consider partial selfing under a theoretical population of N diploid individuals. N can be any positive integer. If assuming constant population size, we consider the same number of N diploid individuals in the next generation who will be generated from the present generation of individuals through selfing or outcrossing with probability σ and $(1 - \sigma)$, respectively. Both reproductive modes, outcrossing and selfing, are considered sexual reproduction. In other words, individuals of the offspring generation will be generated by the fusion of two gametes produced by the previous generation. During reproduction, gametes are produced through meiosis, a biological process that includes recombining the genetic material via crossovers, e.g., reviewed by Mercier, Mézard, Jenczewski, Macaisne, and Grelon (2015). In our models, the two gametes for the offspring are chosen randomly. However, in the case of outcrossing, the second gamete must be from a different individual compared to the first gamete. In the case of selfing, the same individual must produce the second gamete. In our models, the second gamete can be either the other gamete from the same meiotic process as the first gamete or must have undergone an independent process of meiosis and, thus, recombination. Irrespective of this difference, from a biological perspective, gametes from two different meiotic processes and thus two different recombination events fuse in the case of sexually reproducing organisms. Usually, we do not model the generation of gametes because, from a theoretical perspective and under the assumption of neutral evolution, the generation of the offspring individual will not depend on explicitly simulated meiosis. Thus, each chromosome set can be generated independently. The expected distribution of polymorphic sites in a sample of sequenced individuals is independent of recombination (Fisher, 1958; Kingman, 1977, 1982, 2000; S. Wright, 1931).

Under the assumption of neutral evolution, the distribution of polymorphic sites (single nucleotide polymorphisms, SNPs) in a sample of sequenced individuals from a specified population is determined by the underlying genealogy of the site, which in turn depends on the demographic history of the population. The genealogy of a sample of recombining sequences can be described via a complex graph, the ancestral recombination graph (ARG). The ARG connects the genealogy of each site; the connections of two neighboring trees represent past recombination events between present haplotypes. A tree represents the genealogy of a single site. A tree consists of two properties: 1) Its length and 2) its topology, which, taken together, is the order, number, and timing of the branching process. Each pair of samples traces back its lineages to a node, which represents the most recent common ancestor (MRCA), while we label the node's age as T_{MRCA}. The genealogy of neighboring sites remains unchanged if no recombination event has occurred in the past between them.

Two population parameters determine the distribution and characteristics of genealogies observed in a sample of several recombining genomes: the population mutation rate (θ) and the population recombination rate (ρ). Both variables depend on the effective population size and the per site per generation mutation (μ) or recombination rate (r), respectively. Both μ and r are measures that can be approximated from experimental data. Under outcrossing, the ratio between the population mutation and recombination rates equals the ratio between the per site per generation mutation and recombination rates. The proportion of selfing within a population determines reduced heterozygosity and, thus, a reduced effective population size N_σ. Selfing for several to dozen generations results in the limit of the inbreeding factor F_{IS}. The theoretical expectations of consequences of selfing can be calculated using the mathematical frameworks provided by Fisher, Charlesworth, and further (B. Charlesworth & Charlesworth, 2010; Fisher, 1941): The reduction of the effective population size by a selfing proportion of individuals in the population is simply described being half of the proportion of selfing within the population:

$$N_\sigma = N \cdot \left(1 - \frac{\sigma}{2}\right) \tag{1}$$

or, more generally:

$$N_{F_{IS}} = N \cdot \frac{1}{(1 + F_{IS})} \tag{2}$$

with

$$F_{IS} = \frac{\sigma}{(2 - \sigma)} \tag{3}$$

Again, the inbreeding effect of selfing rescales both θ and ρ through the reduced effective population size. However, the effective recombination rate is further affected by selfing. The reduced heterozygosity silences the effect of recombination. In the most extreme case of complete selfing (we assume no mutation for simplicity here), each individual will have zero heterozygosity.

Recombining two chromosomes will not affect the constitution of the gametes' haplotypes. Thus, the rescaling of the recombination rate is defined as:

$$r_\sigma = r \cdot (1 - F_{IS}) \tag{4}$$

This yields the effective population recombination rate ρ being affected in two distinct ways through selfing:

$$\rho_\sigma = \rho \cdot \frac{(1 - F_{IS})}{(1 + F_{IS})} \tag{5}$$

while θ is rescaled only by the selfing through the rescaled effective population size. Notably, the additional effect on the population recombination rate creates a difference in the ratio between the ratio of θ and ρ, compared to the ratio between the respective per site per generation mutation and recombination rates r and μ:

$$\frac{\vartheta_\sigma}{\rho_\sigma} = \frac{\mu}{r} \cdot \frac{1}{(1 - F_{IS})} \tag{6}$$

The rescaled ration between θ and ρ determines an estimator for the selfing rate (M. Nordborg, 2000). Thus, any summarization of the polymorphism data being informative about both processes potentially is informative about the selfing process and history of a population.

Measures of the effect of the joint rescaling of θ and ρ under partial selfing are descriptive of the reproductive process of a population. This joint rescaling allows modeling a panmictic population with the same levels of drift and recombination in a selfing population. These rescalings are a direct consequence of the reduced heterozygosity. As described above, an offspring reproduced by selfing obtains its haplotypes from either the same haplotype or different haplotypes of the same parent. Thus, on average, each reproduction event via selfing halves the heterozygosity of the offspring compared to its parent. Reduced heterozygosity results in reduced diversity throughout the population, thus explaining the rescaling of the effective population size. The additional rescaling of the population recombination rate ρ again is a direct consequence of the reduced heterozygosity. Under reduced heterozygosity, crossovers tend to recombine homozygous genetic material. Thus, a crossover does not result in a new recombinant. Here, we do not consider the creation of a new haplotype through ongoing mutation. The inbreeding factor F describes the proportion of identity by descent (IBD). In observed data from natural populations, F is estimated from the lack of heterozygosity compared to its expectation under random mating. If we assume t as the number of generations in which the offspring was reproduced by selfing, we find F_t being the the reduced inbreeding factor at time t.

$$F_t = 1 - \left(\frac{1}{2}\right)^t (1 - F_0) \tag{7}$$

with

$$F_0 = 1 - \frac{H_O}{H_E} = 1 - \frac{H_O}{2pq} \tag{8}$$

being any observed lack of heterozygosity at time $t = 0$. In its most extreme case, even after only ten selfing events, the heterozygosity is reduced to a 1024th which we could effectively consider as homozygous. Thus, we conclude that any developed mathematical framework based on the inbreeding factor F holds its limits to selfing.

There are generally two different available models to simulate the genetic diversity of a population: 1) the forward-in-time Wright-Fisher model (WF) and 2) the backward-in-time coalescent model (Fisher, 1958; Kingman, 1982; S. Wright, 1931). The WF simulates all individuals of a population that explicitly undergoes reproduction, including the generation of gametes with meiotic recombination, as a time-forward process. Notably, this model can easily be modified to model complex traits, including life-history traits, e. g. selfing, and violations of the process of neutral evolution. These traits are usually simulated explicitly and, thus, most reliable in mimicking actual biological processes. The coalescent model simulates only the genetic diversity of a sample subset of a population in a time-backward process. The most exciting advantage over the WF is its computational feasibility. Still, the computational demand of many population genetics scenarios is exhausted for computing systems. Second, the coalescent process can be derived from the WF. Both models converge under carefully made assumptions. It is possible to extend the coalescent model to mathematically described life-history traits, e. g. partial selfing using the properties of θ and ρ effectively being rescaled (M. Nordborg, 1997, 2000; Magnus Nordborg, 2001; M. Nordborg & Donnelly, 1997). However, in the framework of the coalescent theory, only a single haplotype per individual is modeled. This sampling scheme is a potential source for breaking the assumptions of the coalescent model with partial selfing. Moreover, it remains to be investigated whether the coalescent model with partial selfing suffices in describing the genetic structure resulting from an unequilibrated population with partial selfing after a transition from outcrossing to selfing.

Simulation and demographic models in the context of selfing

We previously described the properties of the theoretical expectation of the genetic structure arising if a random mating population is partially selfing. θ and ρ will be effectively rescaled via $1/(1+F)$ and $(1-F)/(1+F)$, respectively. When using the coalescent model to simulate transitions to selfing, it is crucial to assure the model's validity. Thus, we propose two different demographic models under three different simulation models. The demographic models include 1) a time-forward transition from predominant outcrossing to predominant selfing under constant population size and 2) a time-forward change from a big population size to a small population size under the rescaling corresponding to the transition from outcrossing to selfing in the first model (**Figure 1**). The simulation models include 1) explicitly selfing under demographic model 1 using time-forward WF, 2) a rescaled time-forward WF under demographic model 1, and 3) a rescaled backward-in-time coalescent under demographic model 1.

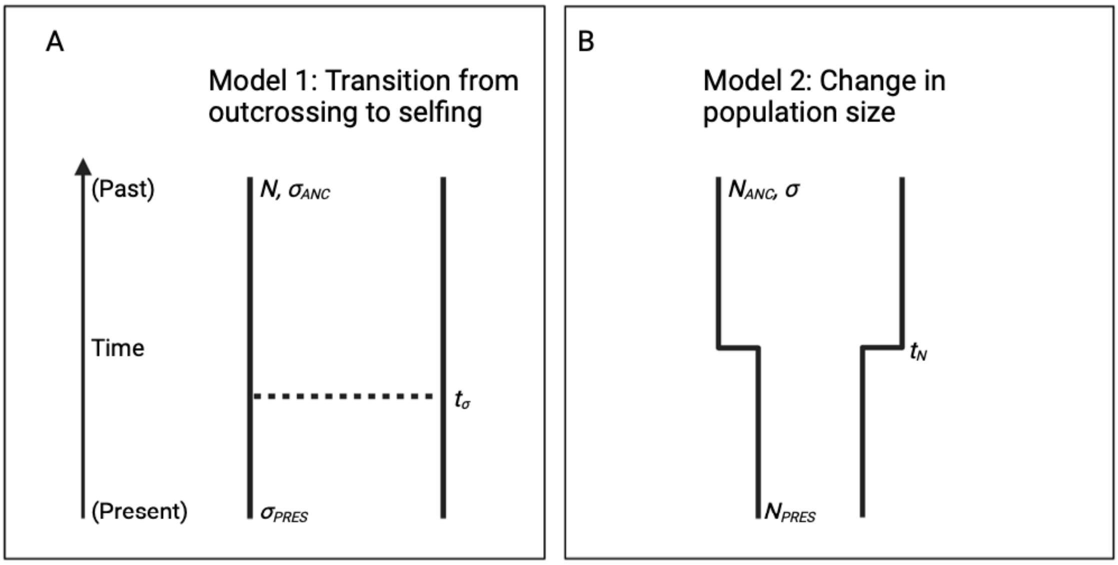

A
Model 1: Transition from outcrossing to selfing
(Past)
Time
(Present)
N, σ_ANC
t_σ
σ_PRES
B
Model 2: Change in population size
N_ANC, σ
t_N
N_PRES

Figure 1. Two population scenarios to investigate the genetic consequences of transitions to selfing: (A) A single population of constant size undergoes a transition from predominant outcrossing to predominant selfing; (B) a single population with constant selfing undergoes a single change of population size according to the rescaled effective population sizes of model 1 in the predominant outcrossing and selfing phase, respectively.

Our theoretical expectations are different for the two demographic models (see above). The second demographic model is a confounding model. It provides an alternative demographic scenario to explain the reduction in diversity, in which no joint change in recombination is considered. Transitions to selfing reduce the effective population size to half in its extreme. Changes in population sizes in the demographic history are common. However, a joint or independent change in the recombination rate in a species is not common. Transitions to selfing rescale diversity and recombination rate, both. Thus, another confounding model could be interesting: A single population undergoes a change in recombination rate but not a change in population size. The third model could capture the effects on the genetic variation through the rescaled ρ, but not ϑ. However, no biological reason is known that the recombination rate changes rapidly in demographic history. Thus, the third model to rescale ρ via a rescaled effective population size is obsolete according to the rescaling of a transition from selfing to outcrossing (backward-in-time). Additionally, it could quickly be ruled out since such a drastic increase in diversity could be detected easily, e. g. approx. 48-fold increase when transitioning from predominant outcrossing ($\sigma = 0.1$) to predominant selfing ($\sigma = 0.99$) backward-in-time.

We successfully identify transitions from outcrossing to selfing if we demonstrate model 1 being more likely under the observed data than model 2. Thus, we need to show how the genetic structure changes specific to a transition to selfing using the two proposed models. Our theoretical expectations are congruent for the three different simulation models (see above). The congruency of the genetic diversity data simulating the consequences of transitions from outcrossing to selfing would enable the use of the coalescent for implementing fast simulations into an inferring method (Chapter 2).

In this chapter, we characterize the specific signals of transitions from predominant outcrossing to selfing on the genetic structure of a population. We provide a summarization of theoretical and measurable statistics. Eventually, we aim to implement our findings into an inference framework to identify and estimate shifts from predominant outcrossing to selfing.

Implementation

To investigate the consequences of a transition from outcrossing to predominant selfing, we simulate genetic variance for a demographic model of a single population undergoing a single-step transition from complete outcrossing ($\sigma = 0$) to predominant selfing ($\sigma = 0.95$, model 1). Furthermore, we compare the genetic variance of this model with the simulated data from a model with a single stepwise change in population size (model 2, **Figure 1**). The change in population size in model 2 rescales the effective population size to the same extent as the transition to selfing does in model 1 (Equation 2).

We simulated a population of 50,000 diploid individuals with chromosomes of $1 \cdot 10^6$ bp length. We set the mutation and recombination rate to $\mu = r = 1 \cdot 10^{-8}$ under neutrality.

To investigate the direct effect on the rescaling of the recombination rate, we counted the effective recombination events before and after a transition to selfing. We defined a recombination event as effective if it created a new haplotype that was not present in the parental gametes, i. e. there is an odd number of recombination events between two SNPs of the parental gamete haplotypes. The rescaling factor f_r was calculated via

$$f_r = \frac{n_{r_{outcrossing}}}{n_{r_{selfing}}} \tag{9}$$

with n_r being the observed number of effective recombination events per generation during the outcrossing or selfing phase of the simulation.

Moreover, we calculated different statistics further to investigate the genetic effects of transitions to selfing. To summarize the genetic data, we sampled 20 haplotypes from different individuals. We calculated the summarizing statistics SFS, LD, $\mathrm{TM_{TRUE}}$, and $\mathrm{TM_{WIN}}$. SFS and LD were calculated on the whole set of 20 sequences. $\mathrm{TM_{TRUE}}$ and $\mathrm{TM_{WIN}}$ (see page 20) were calculated on the pairwise comparison.

We used the *SliM3* suite to simulate a forward-in-time WF population to simulate the explicit transitions to selfing. We explicitly applied the change of the selfing rate at a given time t_σ to the whole population at once. We created the samples at different times after the transition to selfing.

Classical site-based summarization of genetic diversity

In population genetics, classical summary statistics depend on allele frequencies only. Evolutionary forces affect expected frequencies of mutations, but not recombination. We summarized the simulated genetic variation jointly with two classical approaches: 1) The site-frequency spectrum (SFS) and 2) the linkage disequilibrium decay (LD).

The SFS is a two-dimensional statistic summarizing the distribution of allele frequencies of a given sample set. The sampling scheme – whether the samples are taken from a single or multiple populations – and the demographic history, i. e. the magnitude of drift, determines the specific counts of site frequencies and the set of sequences. The expected SFS can be calculated for a single population of constant size since each frequency provides partial information about a specific time frame in the past (Griffiths & Tavaré, 1998). The calculation of site frequencies can be extended to consider a piecewise constant population size function. Reciprocally, estimating a piecewise constant population size can be formalized into ML or approximation by simulation statistical approaches based on a given SFS for different population model assumptions (Boitard, Rodriguez, Jay, Mona, & Austerlitz, 2016; X. Liu & Fu, 2015). However, the recombination rate through time, and thus the selfing rate, does not affect expected site frequencies. The SFS must be extended by a summarizing statistic correlating to recombination rates through time to fulfill the criterion of Bayesian sufficiency for estimating population sizes and recombination rates through time jointly. Thus, we extend the SFS by measuring the linkage disequilibrium (LD).

Linkage disequilibrium relates the dependency of inheritance of site frequencies in a population to an effective recombination rate. The physical distance of sites can be translated to an effective population recombination rate (ρ) of a specific time (Boitard et al., 2016; Hayes, Visscher, McPartlan, & Goddard, 2003):

$$E(r^2) \approx \frac{1}{a + 4Nc} \qquad (10)$$

with $E(r^2)$ being the expected LD measured as r^2, a being a constant dependent on the mutation model, N being the effective population size at the time $1/(2c)$ and c being the recombination rate. Thus, measuring both SFS and LD provides information about the effective population mutation rate through time and the effective population recombination rate through time, which we can parametrize into an effective population size through time and a selfing rate through time.

Haplotype-based summarization of the genetic diversity

Selfing rescales both the effective population size and recombination rate (Equations 2, 3, 4). To obtain a measure sensitive to both of these effects, we chose the following approach: We identified segments and compared two sequences sharing the same most recent common ancestor (MRCA). These MRCA-segments are bounded by the positions at which recombination occurred in the past. Therefore, the distributions of the length of MRCA segments will measure the effective population recombination rate. Thus, characterizing a pairwise comparison of sequences through their MRCA segments by measuring the joint distribution of their T_{MRCA} and length (TL) is sensitive to the effects of selfing.

Note, this summarizing statistic requires knowledge about the boundaries of MRCA segments, i.e., at which position recombination occurred in the past. Thus, we obtain this summarizing statistic only from simulated data.

Summarizing the clustering property of non-recombining segments

In addition to TL distributions, we also calculated the transition matrix of T_{MRCA}s of successive MRCA segments along the genome, which we refer to as TM_{TRUE}. Therefore, we discretized T_{MRCA} values. Then, we counted the frequencies of segment transitions along the simulated sequences for each combination of discrete T_{MRCA} values. We normalized the frequency of transitions from each discretized T_{MRCA} class to obtain the final transition probability matrix.

Note, similar to TL, this summarizing statistic requires knowledge about the boundaries of MRCA segments, i.e., at which position recombination occurred in the past. Thus, we can obtain this summarizing statistic only from simulated data.

Joint summarization of recombination and drift

To measure the combined effects of selfing on the genetic diversity of natural populations, we designed a measure informative about both the recombination and demographic history of a population. We calculated the transition probability matrix of pairwise diversity in non-overlapping windows, which we refer to as

TM_{WIN}. Thus, to summarize the genetic diversity, we do not require any knowledge about the position of recombination events along the sequence.

Results

Theory predicts the true rescaling of the recombination rate under low mutation rates

The rescaling of the effective recombination rate can be predicted using theory (Equation 4). However, in the theoretical model discussed before (REF introduction), we do not consider recombination between haplotypes of the same individual when gametes are formed. However, mutation during gamete formation can result in novel haplotypes if recombination between the haplotypes within an individual is considered. This 'class' of recombination is silent under complete homozygosity. Thus, a particular class of recombination events will not be considered.

We simulated a single population with a given chromosome length to address this issue. We counted the effective recombination events before and after the transition to a given explicitly simulated selfing rate. We considered recombination events as 'effective' if they resulted in novel haplotypes. For comparison, we also simulated the transition to selfing under the rescaling introduced in equations (2) and (4). We counted the number of effective recombination events in each generation using forward-in-time WF simulations under the demographic model 1. We calculated the ratio of the average effective recombination rate before and after the transition to selfing (**Figure 2**, equation 9) to provide a measure for the underlying rescaling parameter (Equation 4).

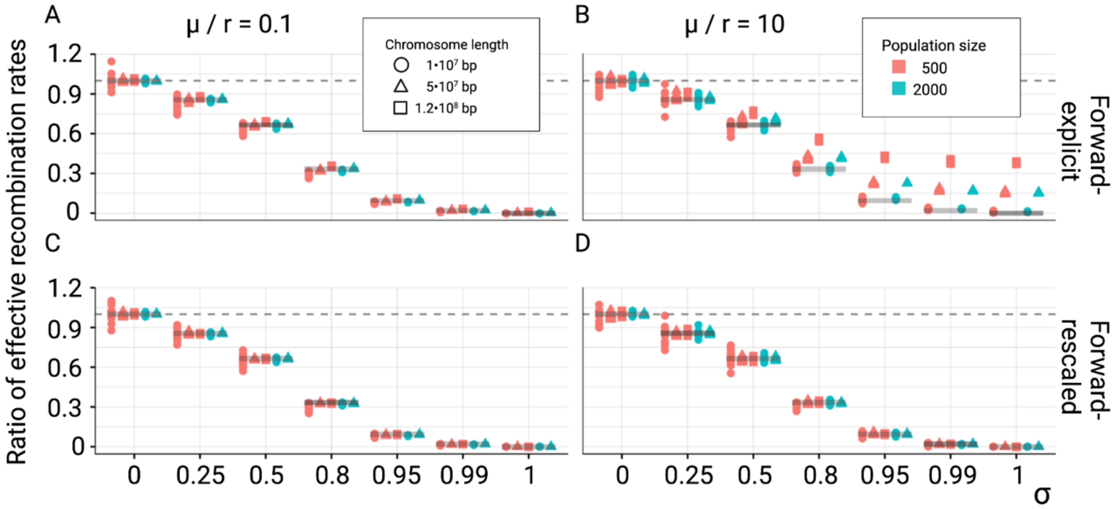


Figure 2. The observed ratio of effective recombination rates of simulated populations under Model 1 in WF simulations: Measured rescaling for $\mu/r = 0.1$ (A, C) and $\mu/r = 10$ (B, D) under the Forward-explicit WF model (A, B) and the Forward-rescaled WF model with partial selfing (C, D). Simulations were performed for two populations of sizes, 500 (red) and 2000 (blue), and three chromosome lengths, 10^7 (circle), $5 \cdot 10^7$ (triangle), $1.2 \cdot 10^8$ bp (square). Results are shown for 12 independent simulations. We counted effective recombination events before and after the transition to selfing and calculated the ratio using equation (9).

For complete outcrossing ($\sigma = 0$), we obtained the ratio of the effective recombination rates $f_r = 1$, under all scenarios. The rescaling factor of the recombination rate varied the most under small chromosome length, and low population sizes around the theoretically expected values. Using the rescaling of the coalescent to mimic transitions to selfing, we obtain a perfect congruence of the simulated rescaling compared to the theoretical expectation. This is different for explicitly simulated selfing rates. With increasing selfing rate and increasing chromosome length, but independent from population size, we obtain an increased rescaling of the effective recombination rate.

Moreover, this effect is even more pronounced for high μ/r ratios, but almost absent for mutation rates being a 10th of the recombination rate. We found this effect was most pronounced for complete selfing, which results in ($\rho = 0$) in the rescaled model, indicating a systematic error in the approximation of the rescaling for partial selfing deriving from a specific class of recombination events. In *Arabidopsis thaliana,* the mutation and recombination rate were estimated to $\mu = 6.95 \cdot 10^{-9}$ and $\mu = 3.6 \cdot 10^{-8}$ resulting in $\mu/r \approx 0.2$ (Ossowski et al., 2010;

P. A. Salomé et al., 2012). Thus, we consider the rescaling proposed in the coalescent with partial selfing being accurate.

The rescaled coalescent simulates a single haplotype per individual

The coalescent with partial selfing provides a model to simulate haplotype genetic diversity for selfing populations. However, the modeled haplotype samples represent distinct individuals. To investigate whether the sampling stage of a selfing population biases the measured genetic diversity in contrast to the rescaled coalescent, we simulated transitions to selfing and calculated F_{IS} under the explicit and the rescaled model for selfing using the forward-in-time WF-model. We compared the individuals' F_{IS} distribution and their average to the expected values (Equation 3). We used different sampling proportions to represent expected different lineage ages (small sample sizes represent old lineages). Except for the complete sampling proportion, we sampled one haplotype per individual and assembled them to represent the sample haploid individuals in the coalescent with partial selfing.

We did not observe differences in the F_{IS} distribution for different sampling proportions. For explicit selfing, the F_{IS} distributed bimodally: A certain proportion of the measured samples is distributed around zero; the other proportion of the measured samples distributes at $F_{IS} = 1$. The average of them exactly meets the theoretical expectation. However, when simulating genetic diversity under the rescaled coalescent with partial selfing, we only obtain a distribution around $F_{IS} = 0$. The observed values distribute with more considerable variance under high selfing rates (**Figure 3**). This effect is observed for any population sampling proportion (0.005 to 1.0). Thus, the coalescent with partial selfing cannot be used to simulate diploid genetic diversity. However, sequencing inbred lines provides an accurate haplotype sample to be analyzed using the coalescent theory.

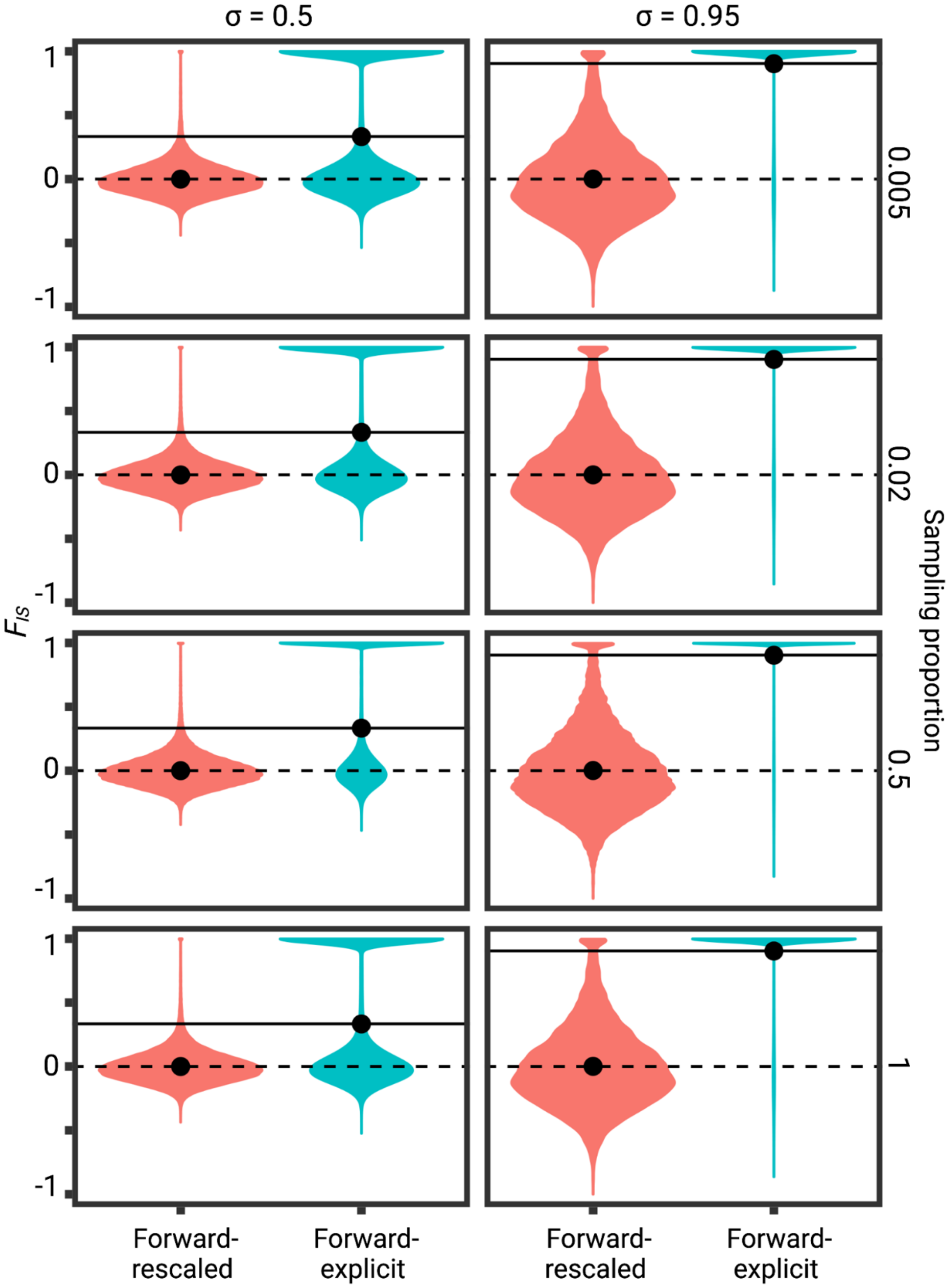


Figure 3. Measured inbreeding factors F_{IS} as one minus the ratio between observed and expected heterozygosity for different sampling proportions from a simulated WF population with constant selfing (sigma = 0.5, left; sigma = 0.95, right). Simulations were done using *SliM3*. Population size was set to $N_e = 10{,}000$, and mutation and recombination rates were set to $\mu = r = 10^{-8}$. After 100,000 generations, samples were taken and repeated every 10,000 generations 30 times. For sample proportion of 0.5 or lower, haplotypes were sampled from different individuals.

The coalescent model accurately simulates direct measures of the joint effects of transitions to selfing

The joint distribution of T_{MRCA} and lengths of MRCA segments (TL) were used to describe the consequences of a transition to selfing at the genomic level and how it differs from a change in population size. In *SliM3* and *msprime*, MRCA segments were analyzed by identifying consequential recombination events in the history of the samples (i.e., events that lead to the inclusion of new MRCA); the genomic position of those recombination events then represent the boundaries of the successive segments, which we refer to as MRCA segment.

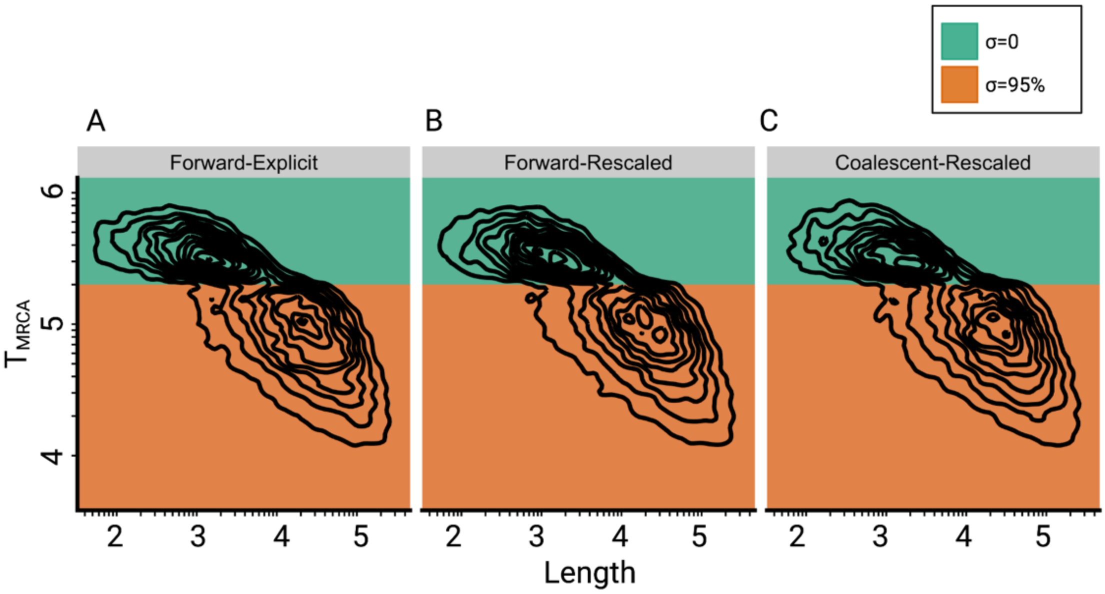


Figure 4. Comparison of the joint distributions of T_{MRCA} and lengths of MRCA segments (TL) under three different simulation approaches. (A) Explicit selfing is implemented in a forward-in-time Wright-Fisher model (*SliM3*). Population size is constant, and the selfing rate changes from outcrossing ($\sigma = 0$, green) to predominant selfing ($\sigma = 0.95$, orange). MRCA segments were defined as contiguous sets of nucleotides sharing the same most recent common ancestor. (B) Shift to selfing is simulated using a forward-in-time Wright-Fisher model (*SliM3*) by rescaling population size and recombination rate at t_σ as suggested by M. Nordborg and Donnelly (1997) (C). Shift to selfing simulated using the coalescent by rescaling population size and recombination rate at t_σ as in panel B. Except for the selfing rates, both axes are scaled in log_{10}.

We found the correlation function between the T_{MRCA} and length of the MRCA segments representing a measure for the selfing rate per time. The T_{MRCA} and the length were simulated and distributed around the expected values, adding a layer of uncertainty when interpreting TL. Pairwise nucleotide diversity can approximate T_{MRCA} when mutation rates are sufficiently large (Ralph, Thornton, & Kelleher, 2020). This is true and indistinguishable using the three proposed simulation models (**Figure 4**).

We simulated TL for a complete-time series of transitions to selfing (**Figure 5**). We found the change in the slope of the underlying correlation between T_{MRCA} and length of MRCA segments to precisely date the time of the transition to selfing and follow strictly theoretical predictions (Equation 6). Furthermore, we found TL to be specific to transitions to selfing (**Figure 6**, panels A and B).

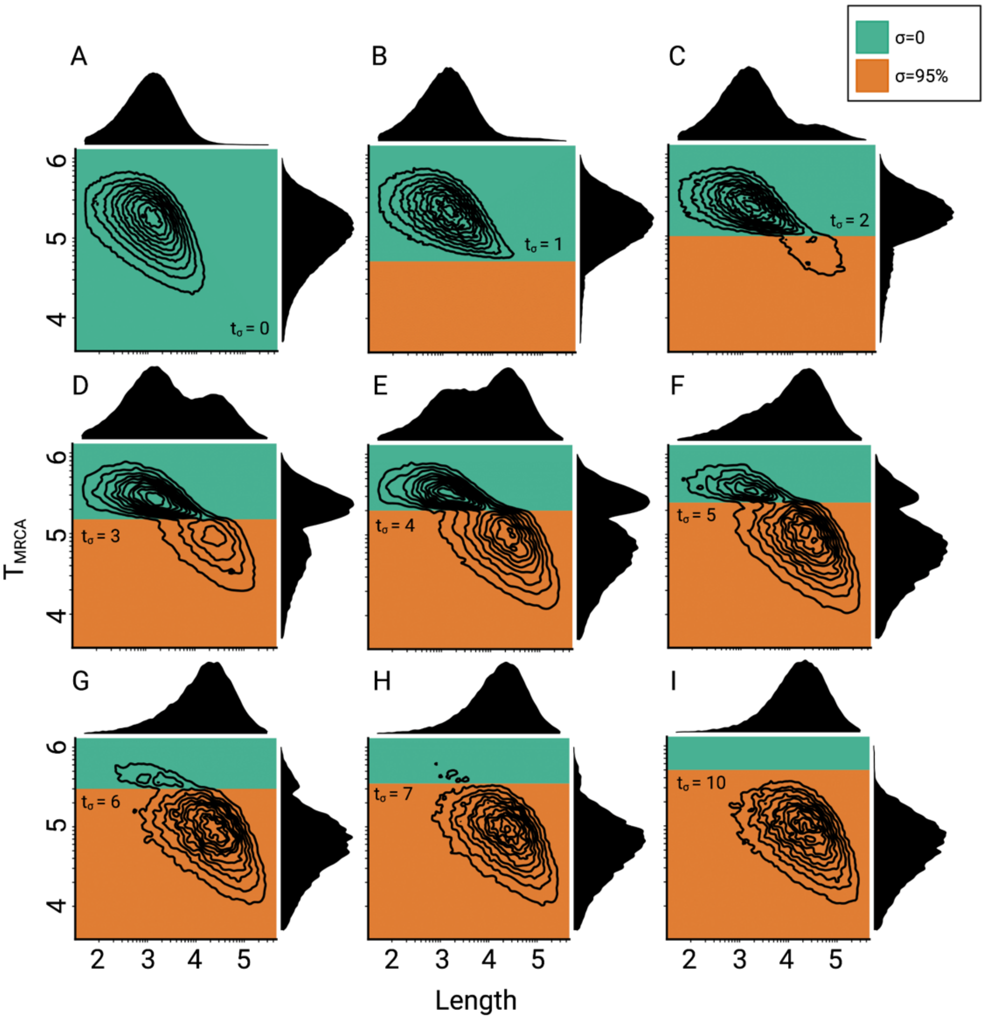


Figure 5. Consequences of a transition to selfing on the genealogies of simulated chromosomes over time. (A-I) Joint and marginal distributions of ages in generations and lengths of MRCA segments (TL) in a population with constant population size and a shift from outcrossing (green) to predominant selfing (orange). MRCA segments were defined as contiguous sets of nucleotides

sharing the same most recent common ancestor. Except for the selfing rates, both axes are scaled in log_{10}.

In summary, we conclude that the coalescent model with partial selfing accurately simulates an effect on genetic diversity being a consequence of transitions to selfing, or more generally, accurately simulates changes in selfing rates through time.

Summarizing the clustering property of non-recombining segments

Recombination is the only event that determines T_{MRCA} of adjacent sites being different. Thus, we measured the probability of T_{MRCA}s of consecutive MRCA segments following each other depending on their T_{MRCA} (TM_{TRUE}). The expected probabilities for MRCA segments following each other were determined by the probabilities of the Poisson process of recombinations. It has been described in the Sequential Markovian Coalescent (SMC) framework for constant recombination rates.

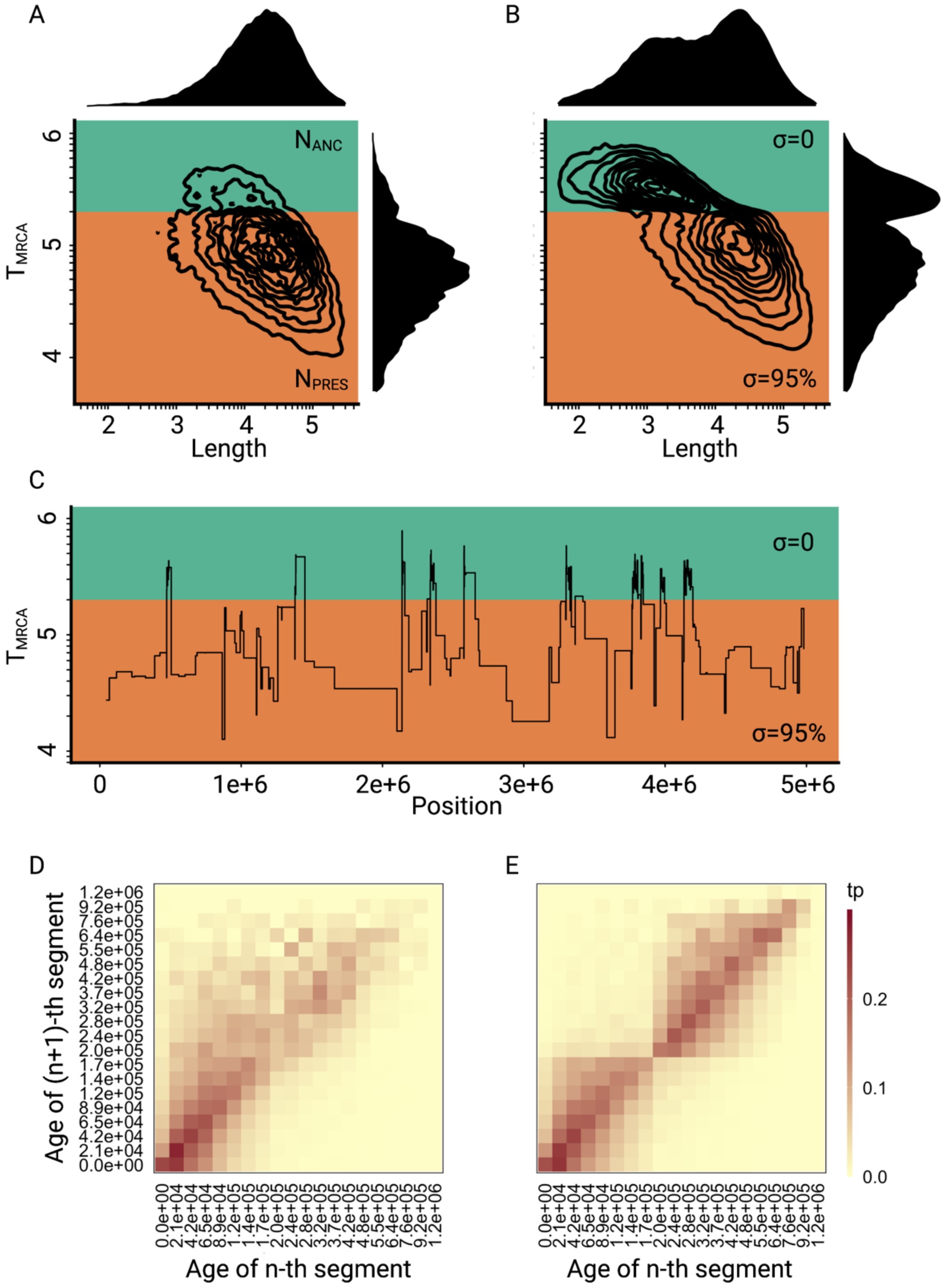


Figure 6. Consequences of a transition to selfing on genealogies of simulated chromosomes. (A) Joint and marginal distributions of ages (T_{MRCA} in generations on a log_{10} scale) and lengths of MRCA segments (in bp on a log_{10} scale) in a selfing population ($\sigma = 0.95$) with a stepwise change from large (green, N_{ANC} = 50,000) to low (orange, $N_{PRES} = 26{,}250$) population size. The population sizes were chosen to correspond to the rescaling of the effective population size by the selfing rates used in panel B. (B) Distribution of ages (T_{MRCA}) and lengths of MRCA segments

(in bp) in a population with a constant population size and a shift from outcrossing (green, $\sigma = 0$) to predominant selfing (orange, $\sigma = 0.95$). (B) (C) Spatial distribution along the genome of a subset of MRCA segments (D) The transition matrix of ages (T_{MRCA}) between adjacent segments along the genome corresponding to the data simulated in panel A. This matrix summarizes the probability that the n^{th} MRCA segment with a given age X is followed by the $(n+1)^{th}$ segment of age Y. The heat colors indicate the transition probabilities (tp). (E) The transition matrix of ages (T_{MRCA}) between adjacent segments along the genome corresponds to the data simulated in panel B. Recombination rate for the simulations was set to $r = 3.6 \cdot 10^{-9}$. T_{MRCA}- and Length-axis are scaled in log_{10}.

Comparing the two demographic models, the transition to selfing and the single change in population size models (**Figure 6**), we found a 2D-bimodal distribution of transition probabilities for transitions to selfing, which is different from a confounding change in population size (**Figure 6**, panel C, D and E). Intuitively, we could explain this finding through the genealogical process of segments that enter a demographic phase with a higher recombination rate (e. g., outcrossing) and, after that, undergo recombination events that only affect the T_{MRCA} of that phase and older. Thus, clusters of MRCA segments appear for each demographic phase, dependent on the recombination rate of that phase.

Observing genetic diversity consequences of selfing on MRCA segments

Pairwise diversity is a direct measure of the T_{MRCA} of two sequences. Recombination does not affect the expected or mean diversity along the sequence. However, recombination affects the variance of the diversity. Thus, similar to TM_{TRUE}, we measured the probability of discretized diversities of consecutive windows along the sequence following each other dependent on their T_{MRCA} (TM_{WIN}). We used a window of size 10,000 bp (**Figure 7**).

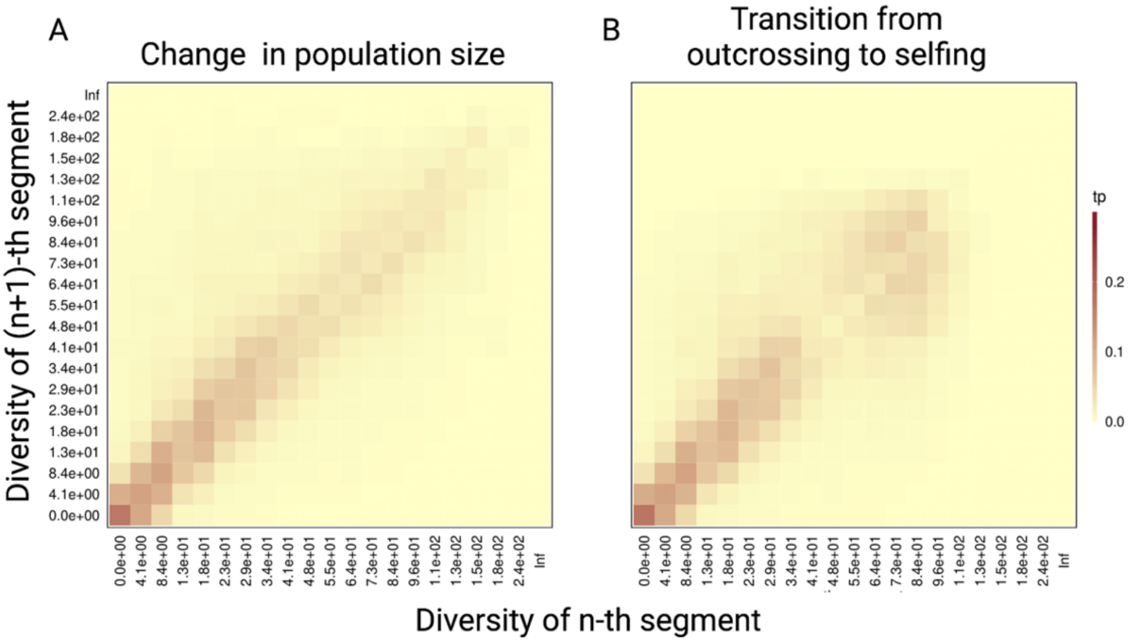


Figure 7. (A) The transition matrix of pairwise diversity ($\mathrm{TM}_{\mathrm{WIN}}$) between adjacent non-overlapping 10 kb windows measured for 1 Mb of data simulated with the population-size-change-model (**Figure 1**) in an outcrossing population with a stepwise change from $\mathrm{N}_{\mathrm{ANC}}$ = 100,000 (green) to low $\mathrm{N}_{\mathrm{PRES}}$=50,500 (orange). The population sizes were chosen to correspond to the rescaling of the effective population size by the selfing rates used in panel B. (B) The same transition matrix of pairwise diversity as in panel A for a constant population ($N = 100{,}000$) but with a transition from outcrossing ($\sigma = 0$) to predominant selfing ($\sigma = 0.99$). The recombination rate was set to $\mu = 1 \cdot 10^{-8}$. These matrices summarize the probabilities that the nth window with a given diversity X is followed by the (n+1)th window of diversity Y. The heat colors indicate the transition probabilities (tp). The demographic model under the simulations for A, a potential confounding model, captures the signal of the rescaled diversity by a transition to selfing, but not the joint rescaling of the recombination rate.

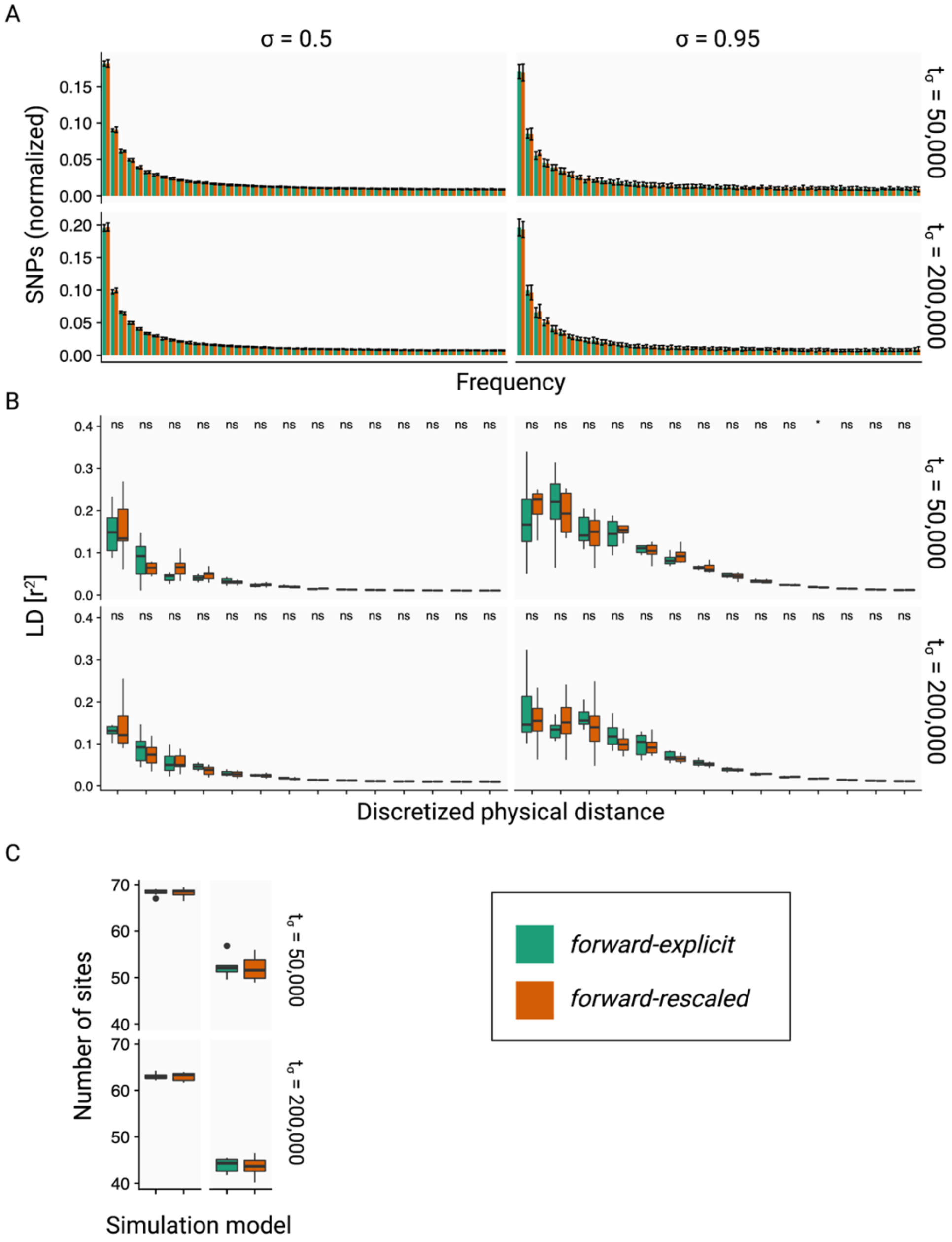
Comparing the two demographic models, the transition to selfing and the single change in population size models (**Figure 1**), we found $\mathrm{TM}_{\mathrm{WIN}}$ to show an increased variance of probability distribution around the diagonal under demographic phases of high recombination rates, e. g. predominant outcrossing, compared to demographic phases of low recombination rates, e. g. predominant selfing, where the transition probability towards a similar diversity is heavily pronounced. This effect is different from the confounding model, where the transition probability towards a similar diversity is also heavily pronounced in the outcrossing phase.

MRCA segments range from a few base pairs to hundreds of thousands of base pairs dependent on their $\mathrm{T}_{\mathrm{MRCA}}$ (**Figure 5**). This range spans around the chosen TM_{WIN} window length. Thus, on the one hand, young MRCA segments containing low numbers of SNPs with a single T_{MRCA} result in an increased probability of transitioning to the same class of low diversity, as the underlying genealogy for multiple each other following windows remains to be the same. That provides a peak transition probability for low diversity to low diversity, depending on the defined discretization boundaries. On the other hand, old but short MRCA segments may contain a high SNP density. However, they are short compared to the TM_{WIN} window length. Thus, a single TM_{WIN} window will span multiple to many MRCA segments resulting in a reduced variance of SNP frequencies for a single window. However, suppose the length of MRCA segments compares well to the TM_{WIN} window length. In that case, we measure the variance of T_{MRCA} dependent on the corresponding temporal recombination rate via its approximate measure of diversity, showing an increased transition probability to other time windows. In other words, the variance of the clustering MRCA segments depends on the effective recombination rate that those segments underwent backward in time, which is high if a recent transition has occurred. Thus, we see a deflection from the diagonal (transition probability to the same discretized T_{MRCA}) in the TM_{WIN} if a transition to selfing has occurred in the past (**Figure** 7).

To conclude, via calculating TM_{WIN}, we measured the combined effects of a rescaled effective population size and recombination rate, both, i. e. specific genetic consequences of transitions to selfing. Thus, TM_{WIN} is a summarizing statistic from which we can obtain the piecewise-constant selfing and population size parameters of the underlying model and identify and date transitions to selfing. Moreover, TM_{WIN} can be used to approximate the calculation of posterior likelihoods of parameter distributions in an Approximate Bayesian Computation (ABC) approach to infer transitions to selfing.

Comparison of the effects of implicit and explicit selfing on the summary statistics

We did not detect differences in simulated data under the explicit selfing or implicit selfing model using SFS, LD (**Figure 8**). The Wilcoxon signed-rank test provided the expected significance proportion of 4.26% under an assumed type 1 error of $\alpha = 0.05$. Given that we performed multiple testing, this met our exact expectation under the assumption of both models simulating the summarizing statistics under the same underlying distribution. Multivariate Kolmogorov-Smirnov testing of TL was oversensitive and detected significance to the same condition. Taken together, under our assumed scenario, we could not find any measurable differences between the two simulation models. In other words, we state the coalescent with partial selfing to be a valid method to simulate and infer transitions to selfing.



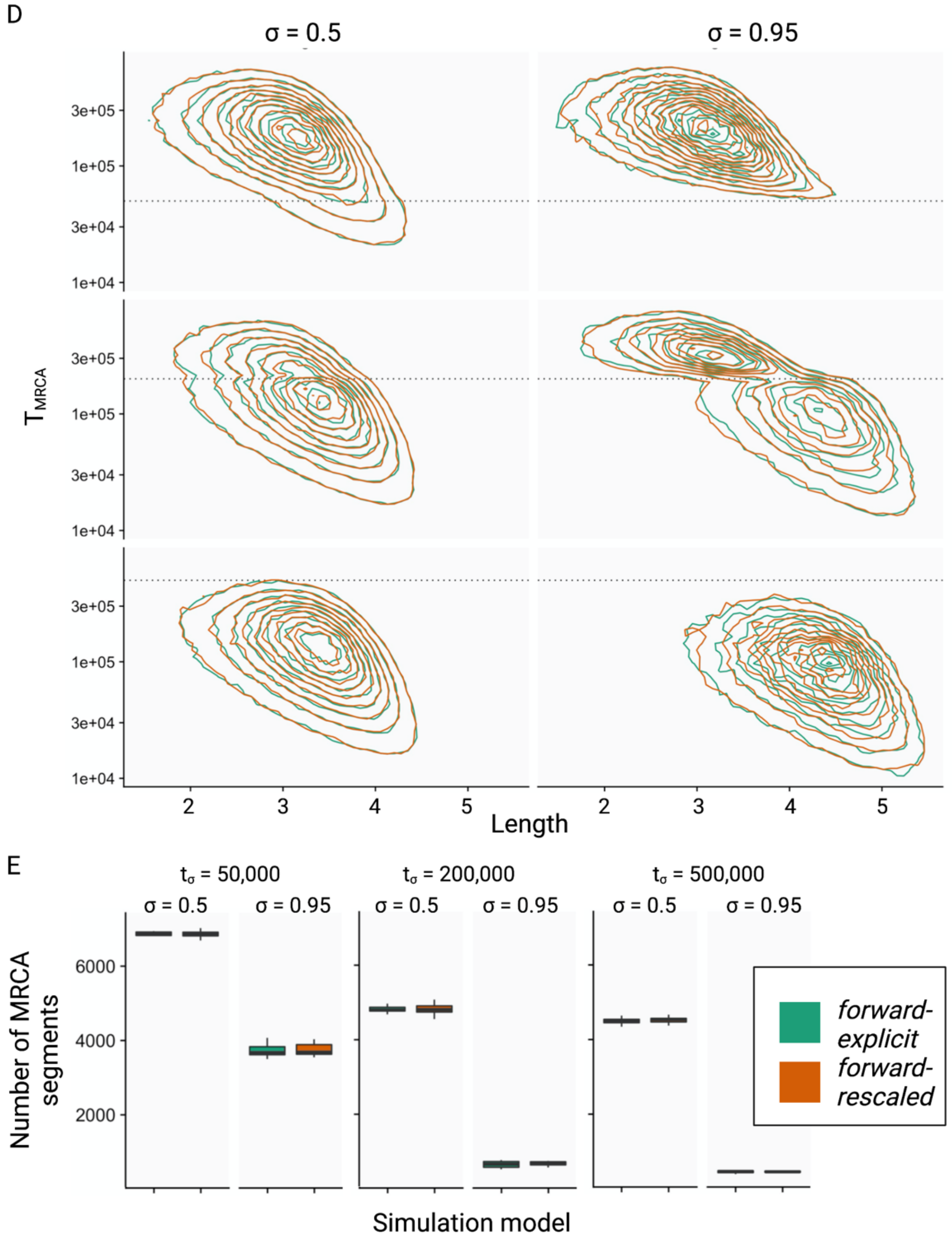


Figure 8. Explicit (green) and implicit (orange) selfing in comparison (A) Site-frequency spectra under two different selfing rates at two time points after a transition to selfing, no significance level ($p > 0.05$) (B) Linkage disequilibrium of discretized distances under two different selfing rates at two time points after a transition to selfing. (C) The number of segregating sites under the same conditions. There was no significance between the models, no significance level ($p > 0.05$). (D) The joint measure of T_{MRCA} (in generations) and length of MRCA segments (TL) for two different selfing rates at three time points after the transition to selfing, the dotted line marks the timing of the transition to selfing (E) Number of according segments (compare to panel D), no significance ($p > 0.05$) was observed. Population size was set to $N_e = 50{,}000$; t_σ was measured in past generations. Significance levels ns ($p > 0.05$), * ($p < 0.05$).

Discussion

This study aimed to summarize genetic data to obtain a specific signal for a transition from outcrossing to selfing. We demonstrated the effects of transitions to selfing and how they relate to the timing of the transition. Additionally, we provided not only a comprehensive summarization of theoretical simulated data but also a new summarization of pairwise genetic variation to measure the effects of transitions to selfing, which can be calculated on observed data from natural populations. Furthermore, we indicated potential limitations to obtaining observable signals of the effects of transitions to selfing on pairwise genetic diversity. Moreover, we have shown the limitations of the coalescent with partial selfing to simulate transitions to selfing.

For the first time, we provide a broad investigation of temporal consequences of transitions to selfing on genome-wide genetic diversities of samples of size two. We demonstrated that a transition to selfing leaves particular marks on the genetic structure in the individuals of a population, depending on the timing of a transition to selfing. Our results hint towards and can be used to develop statistical methods to infer past transitions from outcrossing to selfing (Chapter 2, Chapter 3).

We showed the rescaling introduced in the coalescent framework with partial selfing generally holding when compared to explicit forward-in-time WF simulations. We found the deviation of the coalescent with partial selfing increasing with selfing rates towards one and also with chromosome length, but not with population size. That indicates that a specific type of recombination caused the deviation: The effective recombination concurrently occurs within the meiotic gamete production. This deviation of the effective rescaling compared to the expected from the coalescent with partial selfing indicates potential limits of the coalescent with selfing when considering other life-history traits (e. g., clutch size) and non-neutral scenarios, such as linked or background selection (BGS). In conclusion, we must carefully design coalescent models for partial selfing as it holds for high r/mu ratios and low mutation rates only. However, in most species of interest, e. g., *Arabidopsis thaliana*, we find these prerequisites as given.

Additionally, we provided evidence that the coalescent with partial selfing does not simulate any inbreeding. We conclude that diploidy in partial selfing cannot be modeled under the coalescent. Further, we found a bimodal distribution of observed inbreeding factors for populations with partial selfing. Thus, using an expected inbreeding mean potentially disturbs the accurate simulation of selfing in other model frameworks. However, the diffusion approximation to the coalescent remains robust (Blischak, Barker, & Gutenkunst, 2020).

We used SFS and LD and newly proposed summarizing statistics to describe the effects of transitions to selfing on pairwise genetic diversity: TL, TM_{TRUE}, TM_{WIN}. SFS and LD are classical summarizing statistics used in population genetics; they carry information about temporal changes in population size and recombination rates (Boitard et al., 2016; Tang et al., 2007). The expected SFS can be directly calculated from the demographic history of a single population and vice-versa under the assumption of a constant mutation rate, which is a robust assumption for population genetics time scales (dos Reis, Donoghue, & Yang, 2016; Zuckerkandl & Pauling, 1965). The same is true for the LD-decay under a constant recombination rate assumption. However, we are interested in transitions from predominant outcrossing to predominant selfing, which translates into a severe change in the recombination rate. Thus, the assumption of a constant recombination rate through time is not holding. The possible confounding of LD-decay through a transition to selfing can be ruled out via the joint measure of the SFS.

TL and TM_{TRUE} are theoretical summarizing statistics because they can only be calculated from simulated data. The calculation of these two summarizing statistics depends on the exact knowledge of the position of recombination breakpoints. To obtain our results, we measured the T_{MRCA} but not the diversity. The diversity directly approximates the T_{MRCA}, i. e., molecular clock (dos Reis et al., 2016; Ralph et al., 2020; Zuckerkandl & Pauling, 1965). Under high mutation rates, the diversity converges to the exact T_{MRCA}. Concludingly, relating effective population recombination rates with effective population mutation rates suffers from low $\theta_{\sigma}/\rho_{\sigma}$ ratios.

TMWIN was designed to measure the combined effects of selfing on pairwise genetic diversity along the sequence. This summarization strategy has three advantages: firstly, it captures the distribution of TMRCA (through their positive correlation with pairwise diversity levels); secondly, it captures the effects of the positional clustering in the genome of segments older than t_{σ} (**Figure 6**, panel C) on the pairwise diversity variance; and finally, it also retains some information about the distribution of segment lengths (unlike the matrices in **Figure 6**, panel D, E). The latter is caused by the fact that large segments will contribute more to transition to the same state (i.e., the cells on the diagonal of TMWIN for very recent times), and short segments will be averaged in the sliding window. TMWIN is highly related to the calculated transition matrices in the SMC-HMM frameworks (Chapter 3); those provide the transition probabilities of TMRCAs of a sliding window of size one bp along the sequence. The TMRCAs emit then to a mutated or an unmutated site. The emission depends on the mutation rate only. For a window of size one, the emission can only have two states. However, the underlying hidden states provided a better time resolution, which is implicitly provided by the discretization of diversity, which we use to calculate TMWIN.

In summary, we provided a new way to summarize genetic data to obtain a specific signal for changing selfing rates over time. This is a fundamental base for developing inference methods to estimate changing recombination or selfing rates through time based on whole-genome sequence data. Furthermore, we have studied potential complications of simulation models, e. g. the coalescent with partial selfing, simulating transitions to selfing. Our findings provide an excellent foundation for developing a statistical inference method for estimating breeding shifts from outcrossing to selfing. Exploring the phylogeny of plants of such transitions will contribute to understanding evolutionary processes shaping plant species and populations in the context of the evolution of sexes.

Conclusion

1. Transitions to selfing leave a specific signal in the genetic variation of a population through their combined effect on population mutation and recombination rate.
2. We demonstrated how to summarize the temporal effects of such transitions using both the classical site-based summarizing statistics in population genetics and a novel segment-based summarization of pairwise diversity, which enables measuring transitions to selfing.
3. We investigated the limits of the 'coalescent with selfing' and showed that it can be used for statistical inference under a broad and biologically relevant parameter range.

Author contributions

All methods and data shown in this chapter were developed by the author of this thesis under the supervision of principal investigator Dr. Stefan Laurent and guided by the thesis advisory committee (TAC).

Chapter 2

Identifying and estimating transitions from outcrossing to selfing using *tsABC*

Identifying and estimating transitions from outcrossing to selfing using *tsABC*

Keywords

Selfing, demographic inference, transition/shift mating systems, ABC

Introduction

Technological development progressed in the last decades and provided access to large amounts of genetic data. Phased diversity and haplotype information on entire chromosomes via NGS became accessible. Thus, whole-genome-diversity-based and haplotype-based methods gained importance, and inferring the recombination history became reachable (Marchi, Schlichta, & Excoffier, 2021).

Felsenstein published (Felsenstein, 1988) the Felsenstein equation 1988, which for the first time provided access to the likelihood of genetic data under all possible genealogies. Thus, he introduced the concept of genealogical inference (e. g. demographic history) based on an analytical likelihood function. Providing likelihoods and not only a single expectancy of parameters marked a change in the paradigm of genetic-based demographic inference. Improvements in computational performance further allowed integration over complex likelihood functions, thus, making inference in complex contexts and scenarios tractable. However, for complex models, an analytical likelihood may not be deducible.

Bayesian statistics provide a framework to overcome the urge to rely on a likelihood function. In Bayesian statistics, prior knowledge is implemented into calculating posterior likelihoods using a provided likelihood function. The Approximate Bayesian Computation (*ABC*) framework optimizes a posterior likelihood function using a vector of simulated summarizing statistics if a likelihood function is unknown or intractable or if its calculation is computationally over-demanding. ABC has successfully been introduced into demographic inference in population genetics and used for demographic inference (M. A. Beaumont et al., 2002; Boitard et al., 2016; Wegmann, Leuenberger, Neuenschwander, & Excoffier, 2010).

In this study, we aimed to develop an inference method to identify and estimate transitions to selfing using the ABC framework (*tsABC*). We used a novel summarizing statistic to capture the temporal signature of transitions from outcrossing to selfing (Chapter 1) and compared the improvement in accuracy compared to using the classical summarizing statistics, SFS, and LD. Finally, we infer the transition to selfing of three non-admixed genetic clusters of *Arabidopsis thaliana*.

Existing estimates for *Arabidopsis thaliana*

The timing of the transition to selfing has been estimated in the model species *Arabidopsis thaliana*, reviewed by Tiina M. Mattila, Benjamin Laenen, and Tanja Slotte (2020). In *Arabidopsis thaliana*, Bechsgaard, Castric, Charlesworth, Vekemans, and Schierup (2006) used the S-locus's diversity to infer the timing of the transition to selfing. Under the assumption of selection on a functional S-locus, they inferred a transition to selfing having occurred not later than 413,000 years ago. However, P. Liu, Sherman-Broyles, Nasrallah, and Nasrallah (2007) published PUB8 as an S-locus modifier. Thus, loss of SI in *Arabidopsis thaliana* would be a consequence rather than a cause of a transition to selfing and, thus, potentially more ancient. In a second approach, Tang et al. (2007) concluded from LD patterns that a transition to selfing must have occurred in the magnitude of at least 1 Mio years ago. However, selfing was simulated by a 25-fold increase of recombination rate backward in time under two demographics: constant or exponential growth. These assumptions may not suffice to explain the LD pattern. Little deviation in the exact recombination rates may lead to biased estimates of transitions. Durvasula et al. (2017) concluded the lower boundary of 500,000 years based on the distribution of all known S-haplotypes being present in natural populations in Afrika. Taken together, the estimations of a transition to selfing broadly vary for *Arabidopsis thaliana* depending on the assumptions and used methodology. Mainly, they lack a measure of the direct consequences on the genetic structure of a change in selfing rates.

Models and Methods

ABC in population genetics

Population genetics models often exhibit large nuisance parameter spaces and unknown or prohibited likelihood functions. Parameter inference using Approximate Bayesian Computation (ABC) is a widely used model-based approach in population genetics (Mark A Beaumont, 2010). ABC offers a likelihood-function-free approach for parameter inference (M. A. Beaumont et al., 2002). In the ABC framework, parameters and summarizing statistics of a model are assumed to be random variables that follow a joint probability distribution. Summarizing statistics themselves depend on random nuisance variables that follow a probability distribution. Thus, they result in a Gaussian distribution, resulting in local linearity, which allows for linear regression if the tolerance level of accepted parameters is sufficiently tiny, i. e. the similarity threshold is small. This method's inherent logic provides the mathematical dependence of model parameters given the summary statistics without knowing the likelihood function and provides uncertainty information.

Our developed ABC, *tsABC*, provides a statistical framework for 1) model choice and 2) parameter estimates of a shift-to-selfing model. These two functions of the ABC refer to our biological questions of 1) identifying a transition to selfing and 2) dating the transition to selfing. Furthermore, we provided a complete performance analysis to investigate the accuracy in identifying and dating transitions to selfing in this study. Finally, we apply *tsABC* to data obtained from three natural populations of *Arabidopsis thaliana*.

Modeling a transition to selfing

We considered a single population composed of N diploid individuals to model a transition from outcrossing to predominant selfing. At each generation, each offspring is generated by self-fertilization of a single individual or by outcrossing with probabilities σ and $1 - \sigma$, respectively, where $\sigma = 1$ denotes full selfing and $\sigma = 0$ denotes pure outcrossing. Transitions to predominant selfing were modeled by allowing the selfing rate to change instantaneously from σ_{ANC} to σ_{PRES}

at the time t_σ. The mutation and recombination were set to $1 \cdot 10^{-8}$ events per generation per nucleotide. When needed, the population size was allowed to change instantaneously from N_{ANC} to N_{PRES} at t_N. Thus, we proposed three models in total, two simple models for the performance analysis and one extended model to apply *tsABC* to *Arabidopsis thaliana*: 1) A single population of constant size undergoes a single transition from predominant outcrossing ($\sigma < 0.2$) to predominant selfing ($\sigma > 0.5$), which we refer to as model A; 2) a single population undergoes a single change in population size while keeping the selfing rate constant to predominant selfing ($\sigma > 0.5$), which we refer to as model B; 3) a single population undergoes a transition from outcrossing to selfing at a given time and, additionally, independently a single change in population size, which we refer to as model 3 (**Figure 9**).

We designed model 1 to estimate a transition to selfing in the most concise scenario. We proposed a confounding model 2 that potentially explains the changed diversity through a transition to selfing via a single change in population size while keeping the selfing rate constant to predominant selfing (σ>0.5). Model C combines the earlier models, A and B. Thus, it implicitly allows us to estimate the event contributing to a loss of diversity through time being a population size change instead of a transition to selfing. Consequentially, model C potentially disentangles the transition to selfing from a change in population size.

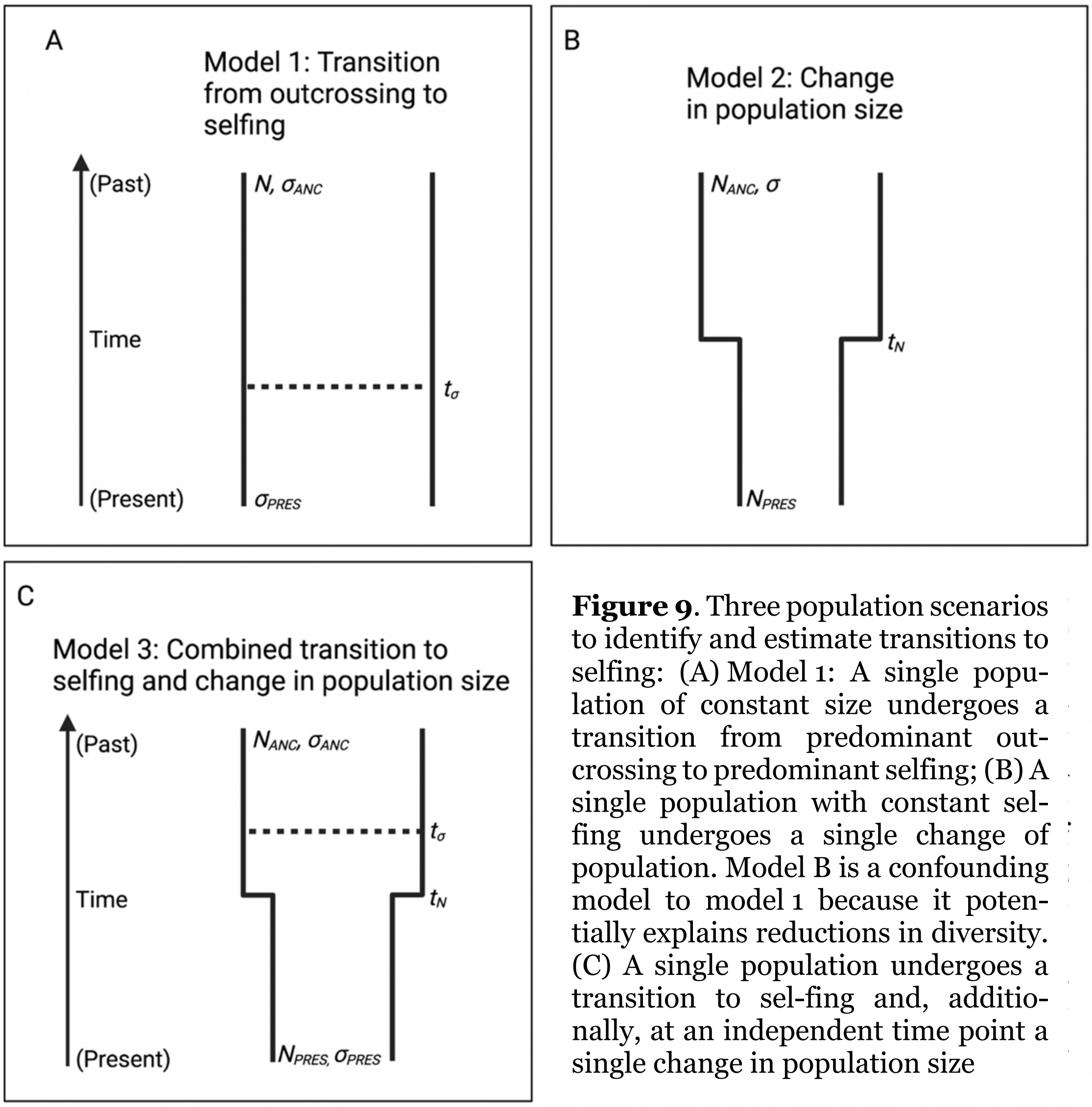

A
Model 1: Transition from outcrossing to selfing
(Past)
Time
(Present)
N, σ_{ANC}
t_σ
σ_{PRES}
B
Model 2: Change in population size
N_{ANC}, σ
t_N
N_{PRES}
C
Model 3: Combined transition to selfing and change in population size
(Past)
Time
(Present)
N_{ANC}, σ_{ANC}
t_σ
t_N
N_{PRES}, σ_{PRES}

Figure 9. Three population scenarios to identify and estimate transitions to selfing: (A) Model 1: A single population of constant size undergoes a transition from predominant outcrossing to predominant selfing; (B) A single population with constant selfing undergoes a single change of population. Model B is a confounding model to model 1 because it potentially explains reductions in diversity. (C) A single population undergoes a transition to sel-fing and, additionally, at an independent time point a single change in population size

Modeling negative linked selection with forward-in-time WF simulations

To extend the performance analysis of *tsABC* to the investigation of its robustness to the negative linked selection, we extended the model to simulate PODs. We simulated the PODs using a forward-in-time WF model. We applied the distribution of fitness effects (DFE) relying on estimates of *Arabidopsis thaliana*. We created a pseudo genome containing exonic and intronic regions with the exact exonic distribution chosen from *Arabidopsis thaliana*. We used the simulation of five independent chromosomes of 1 Mb length mimicking the five chromosomes of *Arabidopsis thaliana*. We used these PODs and repeated the model choice and parameter estimates.

Coalescent simulations for the ABC

Following Nelson, Kelleher, Ragsdale, McVean, and Gravel (2019) who showed that continuous-time coalescent simulation of large sequences causes biases in identity-by-descent and linkage disequilibrium patterns, we implemented a hybrid model in which the first 1,000 generations were simulated using a discrete-time coalescent process. We used this model to generate a genetic variation for a sample of n = 20 haploid genomes, sampled from 20 different individuals and composed of five DNA sequences of one megabase each. The same model was also implemented in a coalescent framework using msprime (Kelleher, Etheridge, & McVean, 2016). The following generations were modeled using the SMC' coalescent algorithm. The coalescent implementation, which runs significantly faster than the forward-WF implementation, was used for the ABC simulations and for generating the PODs for the performance analysis (see below).

Identifying a transition to selfing

In chapter 1, we described two different models to investigate the specific genetic signature of a population that changed selfing rates, e. g. a transition from predominant outcrossing to predominant selfing. Here, we propose the same two demographic models to identify a transition to selfing via an ABC model choice. The demographic models include 1) a time-forward transition from predominant outcrossing to predominant selfing under constant population size and 2) a time-forward change from big population size to a small population size under the rescaling corresponding to the transition from outcrossing to selfing in the first model (**Figure 9**). To analyze the performance of the ABC to estimate transitions to selfing, we use the first model.

Calculation of the summarizing statistics of genetic diversity

In Chapter 1, we introduced summarizing statistics to measure transitions to selfing. However, both TL and TM_{TRUE} are challenging to infer from empirical genetic data. Both require positional knowledge of recombination events. Therefore, we additionally introduced TM_{WIN}, a sliding window approach that captures the specific genetic signal of transitions to selfing. To investigate the performance of *tsABC*, we used the following summarization approaches, which capture the characteristic genetic signal of transitions to selfing: 1) the site frequency spectrum (SFS), which is the distribution of absolute derived allele frequencies in the sample and is known to carry information about past population size changes (Griffiths & Tavaré, 1998); 2) A discretized distribution of linkage disequilibrium (LD) decay inspired from the approach taken by Boitard et al. (2016) who used it jointly with the SFS to estimate past changes in population sizes. Unlike the SFS, which only carries information about N but not the recombination rate (r), LD-decay depends on the product of N and r. Combining both distributions allows capturing the signature of changes in N and r. LD was calculated as r^2 from a subset of 10,000 randomly chosen SNPs and discretized into discrete physical distances with following breakpoints: 6,105; 11,379; 21,209; 39,531; 73,680; 137,328; 255,958; 477,066; 889,175 bp. Unlike in Boitard et al. (2016) the physical distance cannot be generalized to specific parameters in the past demography. That is because the recombination rate is not a fixed parameter and has changed at different times in the past. 3) Window-based transition matrix (TM_{WIN}): While TM_{TRUE} carries a characteristic signal to estimate shifts to selfing, it is not straightforward to calculate it using genetic variation data. This is because the boundaries of MRCA segments are not directly observable and need to be inferred themselves. TM_{WIN} captures some of the information in TM_{TRUE} by substituting the segments by non-overlapping successive genomic windows of 10,000 bp in samples of size two and substituting the T_{MRCA} by the diversity between the two sample sequences.

Dimensionality reduction

The dimensionality of our summary statistic TM_{WIN} is up to 400. To overcome the curse of dimensionality in our ABC, first, we centralized, normalized and Box-Cox transformed each calculated statistic to obtain the orthogonal independent variation. Then, we applied the partial least squares (PLS) analysis to our data as suggested by (Wegmann, Leuenberger, et al. 2010). We chose an appropriate number of PLS components for model selection and parameter estimation according to the chosen combination of summary statistics. SFS and LD are summarizations of lower dimensions. Thus, e. g. folded SFS and discretized LD provided only 17 PLS components. If not stated otherwise, we used the complete 17 PLS to compare with the 20 PLS components of other summarizations.

Parameters and Priors

For the performance analysis, we drew the parameters from uniform or log uniform distributions (**Table 1**). With that, we provide a maximally uninformed prior. Thus, all inference on parameters is model immanent.

Table 1. Parameter priors used for the performance analysis of tsABC

Model	*Parameter*	*lower*	*upper*	*type*
A	N	10,000	200,000.00	logunif
A	SIGMA_PRES	0.5	1.0	uniform
A	SIMGA_ANC	0.0	0.2	uniform
A	T_SIGMA	1,000	500,000	logunif
B	N_ANC	1,000	200,000	logunif
B	N_PRES	1,000	200,000	logunif
B	SIGMA	0.5	1.0	uniform
B	T_N	1,000	500,000	logunif

We aimed to investigate the performance of identifying and estimating times of transition. Thus, the parameters of the PODs were defined for a range of different times of transitions (**Table 2**), but otherwise constant parameters.

Table 2. Parameters for PODs

Model	*Parameter*	*Value*
A	N	40,000
A	SIGMA_PRES	0.99
A	SIMGA_ANC	0.10
A	T_SIGMA	t ϵ(0.05, 10)

Similar to the choice of parameters for models A and B in the performance analysis of the ABC, we chose the priors for the six-parameter model for *Arabidopsis thaliana* (**Table 3**).

Table 3. Parameter priors used to estimate the transition from outcrossing to selfing in Arabidopsis thaliana

Parameter	*lower*	*upper*	*type*
N_PRES	50,000	500,000	logunif
N_ANC	50,000	1,000,000	logunif
T_N	10,000	1,000,000	uniform
SIGMA_PRES	0.5	0.1	uniform
SIMGA_ANC	0.0	0.2	uniform
T_SIGMA	10,000	500,000	logunif

Model choice

The Bayesian framework naturally allows obtaining a posterior density for each proposed model. The Bayes factor is simply defined as the ratio of the marginal densities of two models. In ABC, we approximate those marginal densities by the posterior density. Thus, the Bayes factor

$$B_{AB} = \frac{dens_A(stats_{obs})}{dens_B(stats_{obs})} \qquad (11)$$

provides an approximation to the model support given the observed data (Leuenberger & Wegmann, 2010; Wegmann et al., 2010). We categorize the Bayes factors into negative, barely worth mentioning, substantial, strong, very strong, and decisive (Jeffreys, 1998).

We calculated the Bayes factor for different sets of summarizing statistics. We conducted a model choice using a multinomial regression analysis between proposed models A and B. The calculations were done using the R package abc (Csilléry, François, & Blum, 2012). We accepted 1% of the total number of simulations for the model choice.

Parameter inference

We tested the performance of parameter inference under model A described in the Model selection section to estimate the date of a transition to selfing. We conducted the estimation using the R package *abc* (Csilléry et al., 2012). We accepted the closest 1% of the simulations of the transitioning model. For each corresponding set of summarizing statistics, we estimated the average posterior distribution for the 100 PODs. We show the average quantiles for the following credible intervals: 99%, 95%, 90%, 80%, 50%, 25%, 10%, and the median for the whole time series. We show the performance for each model parameter using the following sets of summary statistics: SFS/LD, TM_{WIN}, and both combined. We used a set of 20 PLS components for each but not for SFS/LD because they consisted of lower dimensionality.

Simulations and observation of data

From 20 individuals, we sampled a single haplotype per individual and five independent regions of 1 MB length mimicking five chromosomes. We used the complete set of 20 haplotypes for the SFS and LD calculation. TM_{WIN} is based on a pairwise comparison of sequences. To calculate TM_{WIN}, we compared all possible 20 *choose* 2 pairs. Thus, the total pairwise length results in 950 MB. We used mutation and recombination rates of $\mu = r = 10^{-8}$.

For the ABC, we created 100,000 datasets for each model. We used corresponding prior parameters for both models (**Table 1**). We tested the performance of the ABC under the assumption of neutrality. The corresponding pseudo-observed datasets (PODs) were created for different transitioning times under the transitioning model and otherwise fixed parameters (**Table 2**). Thus, we obtain a time series for the performance analysis. We tested the performance for following transitioning times in generations: 1,000; 2,000; 3,000; 4,000; 5,000; 6,000; 7,000; 8,000; 9,000; 10,000; 12,000; 16,000; 20,000; 30,000; 40,000; 50,000; 60,000; 70,000; 80,000; 90000; 100,000; 200,000. That translates to coalescent time units ranging from 0.05 to 10. For each condition, we created 100 independent PODs.

Simulating negative linked selection and background selection

Briefly, we simulated the PODs for the background selection (BGS) analysis forward-in-time under Wright-Fisher assumptions. We created 100 independent burn-ins of 10N generations. Starting from those, we 5-times simulated a time series of transitions to explicit selfing under the same parameters as used for the neutral PODs. Each set of five simulations was aggregated and summarized into a single POD.

The simulations with negative linked selection were conducted with *SLiM3*. Deleterious mutations are purged and reduce the genetic variation of a population at linked sites. The reduction of neutral genetic variation due to linkage is referred to as background selection (B. Charlesworth, Morgan, & Charlesworth, 1993; Hudson & Kaplan, 1995). We used the distribution of fitness effects estimated by DFE for *Arabidopsis thaliana* provided by Hämälä and Tiffin (2020). The distribution was used to assign negative selection coefficients to simulated coding non-synonymous genetic variants only (i.e., we did not simulate negative selection on functional non-coding regions). We took care of simulating realistic proportions and spatial distributions of coding sequences by using the positional information of CDS from the annotation of the reference genome of *Arabidopsis thaliana* (Berardini et al., 2015). Except for the DFEs and genetic structure, all other parameters and dataset dimensions were identical to the simulations without negative selection.

Dating transitions to selfing in *Arabidopsis thaliana*

To estimate the transition to selfing for *Arabidopsis thaliana*, we proposed an additional model 3 (**Figure 9**): A single population undergoes a transition from outcrossing to selfing at a given time and, additionally, independently, a single change in population size. Again, this model implicitly estimates the event that contributed to a loss of diversity through time being a population size change instead of a transition to selfing. Thus, we allow disentangling the estimate of a transition to selfing from a change in population size. Furthermore, this model implements the prior knowledge that *Arabidopsis thaliana* as a species has undergone a relatively recent transition to selfing.

Similar to the performance analysis, twelve samples for five independent loci of 1 MB length were simulated to mimic five chromosomes of *Arabidopsis thaliana*. We simulated under a 6-parameter model (**Figure 9**, panel C) that allowed for two independent changes: 1) a transition from predominant selfing ($\sigma > 0.5$) to predominant outcrossing ($\sigma < 0.2$) and 2) a stepwise change of population size. Mutation and recombination rates were set to $6.95 \cdot 10^{-9}$ and $3.6 \cdot 10^{-8}$ per bp per generation, respectively (P. Salomé et al., 2012; Weng et al., 2018).

To obtain the observed summarizing statistics, we masked for exonic regions using the annotation published in TAIR10 (Berardini et al., 2015). The regions were chosen based on homogeneity of recombination rates and diversity (Weng et al., 2018). Our estimates using *tsABC* were based on three independent sample sets; we chose the samples from the CEU, IBnr, and Relicts that belonged at least to 95% to their assigned genetic cluster (Alonso-Blanco et al., 2016).

To estimate the time of the transition to selfing, the model parameters were estimated as described for the ABC performance analysis. We used 20-PLS of the combined summary statistics of SFS/LD and TM_{WIN} (see above).

Implementation and software

Except for BGS, we simulated the genetic data using the coalescent implemented in msprime version 0.7.4 (Kelleher, Etheridge et al. 2016). We simulated the most recent 1,000 generations under the discrete-time-Wright-Fisher model to avoid biases in IBD and the following times under the SMC' model. We used the rescaled coalescent-with-partial-selfing (M. Nordborg & Donnelly, 1997) to simulate transitions to selfing for a recent selfing past. Backward in time, our simulations underwent a transition to outcrossing.

For the BGS-simulations, we wrote a forward-in-time simulator using *SLiM* version 3.6 (Benjamin C. Haller, Galloway, Kelleher, Messer, & Ralph, 2019; Benjamin C Haller & Messer, 2019). We forbid accidental selfing. We used the tree-sequence-recording option. A few lineages have not coalesced after the simulation. Thus, we recapitated the obtained tree-sequences from the pyslim-package version 0.6, which utilizes msprime(Kelleher et al., 2016).

We implemented the entire pipeline into Snakemake 5.13. We have run all simulations on the high-performance cluster HPC of the MPIPZ.

Results

We developed *tsABC*, an approximate Bayesian computation (ABC) method, to estimate population size and selfing rate changes jointly. ABC is a computational approach to estimating posterior probabilities for models and parameters. ABC is well suited for demographic modeling in population genetics because models often have many parameters and no analytically derived or tractable likelihood function (M. A. Beaumont et al., 2002; Csillery, Blum, Gaggiotti, & Francois, 2010). Two advantages of the ABC method allow comparing competing demographic hypotheses based on the Bayes factor. Second, it does not require bootstrapping the data to generate measures of uncertainty for the inferred parameters. A critical aspect of ABC is that it requires a careful summarization of the genomic data into a set of summary statistics that carry information about the parameters of interest (M. A. Beaumont et al., 2002). In the case of a transition to selfing, we require that such summary statistics be informative about coalescence and recombination rates to make changes in selfing rates and population size distinguishable by the ABC model choice (**Figure 10**, panel A, B). Unfortunately, while the lengths of MRCA segments are straightforward to calculate on simulated genealogies, it is more difficult to estimate them based on genomic diversity data alone. We calculated the number of differences for pairs of sampled chromosomes using non-overlapping genomic windows of 10kb. We constructed a transition matrix for pairwise diversity, a summarization we refer to as TM_{WIN} (**Figure 7**). This summarization strategy has three advantages: firstly, it captures the distribution of T_{MRCA} (through their positive correlation with pairwise diversity levels); secondly, it captures the positional clustering in the genome of fragments older than t_{σ}; and finally, it also retains some information about the distribution of segment lengths (**Figure 7**). The latter is because large segments will contribute more to transition to the same state (i.e., the cells on the diagonal of TM_{WIN}. For comparison, we also considered the canonical site-based summarization, the combined site frequency spectrum (SFS), and a discretized distribution of the decay in linkage disequilibrium (LD), as these carry information about temporal changes in population size and selfing rates (see Chapter 1). We, therefore, evaluated the efficiency of three sets of summary statistics: SFS/LD, TM_{WIN}, and SFS/LD/ TM_{WIN} (see methods).

Identifying transitions to selfing

Transitions to selfing result in a reduction of diversity. To test whether *tsABC* identifies transitions to selfing against a model of population census reduction, we conducted a model choice experiment using two competing models (**Figure 10**, panel A, B): We simulated datasets under model 1 and evaluated the ability of *tsABC* to identify the correct model for transitions of varying ages, using different sets of summary statistics to summarize the genetic data.

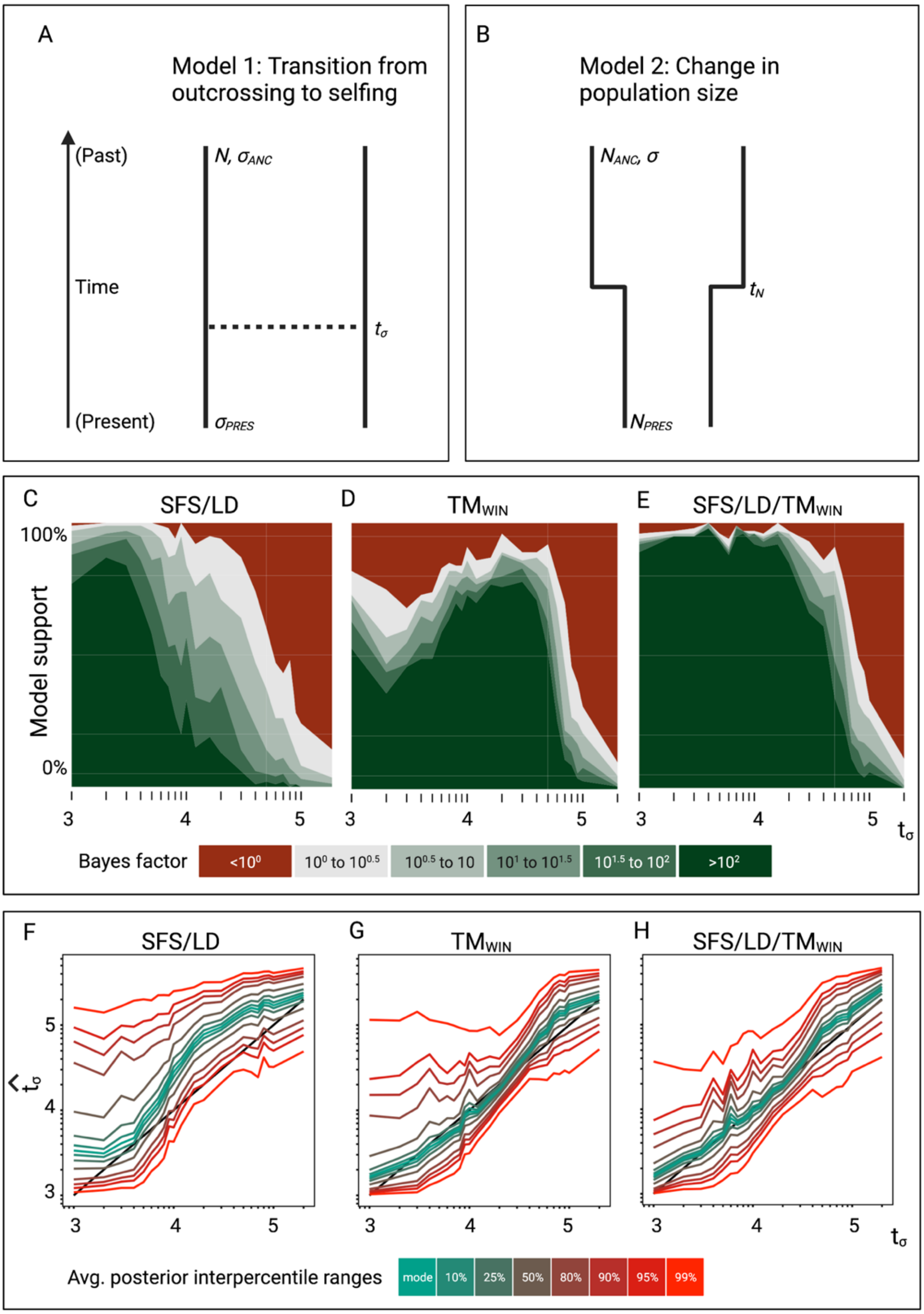


Figure 10. ABC model choice and parameter estimate performance analysis. (A) Demographic model 1 in the model choice analysis: one population with a single transition from predominant selfing to predominant outcrossing (B) Demographic model 2 in the model choice analysis: one

population with constant selfing and a single change in population size. (A, B) The parameters of interest are the population sizes (N_{PRES}, N_{ANC}), the selfing rates (σ_{ANC}, σ_{PRES}), and the time of change in selfing rate and population size (t_σ, t_N). Model 2, a potential confounding model, captures the signal of the rescaled diversity by a transition to selfing, but not the joint rescaling of the recombination rate. (C-E) Performance of the ABC model choice method using three different summarizations of data. C: Combining site frequency spectrum (SFS) and linkage disequilibrium (LD). (D) Window-based transition matrix (TM_{WIN}). E: The combination out of SFS, LD, and TM_{WIN}. The x-axis represents the range of used t_σ values; the y-axis indicates the proportion (out of 100 trials) that the ABC correctly identified the transition-to-selfing model among the two models presented in (A). (F-H): Parameter estimation accuracy for the age of a transition to selfing (100 simulated datasets) under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average interpercentile ranges for 100 posterior distributions corresponding. t_σ-axes are scaled in log_{10}.

We approximated and compared the posterior densities of the transition to selfing model with constant population size to model 2, which implements a population undergoing a change in population size but is constant in selfing. We calculated the Bayes' factors using multinomial logistic regression and the marginal densities of model 1 (transition to selfing) against the marginal densities of model 2 (change in population size). The results depict the proportion of the correct model estimations (**Figure 10**, panel C-E). Depending on our summarizing statistics, our results indicate that we can precisely detect transitions to selfing for times up to 2.5 Ne generations in the past.

Dating transitions to selfing

We evaluated the accuracy of our method for estimating the age of a transition to selfing (t_σ). We simulated 100 datasets under model 1 (**Figure 10**, panel A) with values of tσ ranging from 1,000 to 200,000 generations and used *tsABC* to re-estimate posterior distributions for t_σ and the other model parameters (**Figure 10**, panel F-H, **Figure 11**).

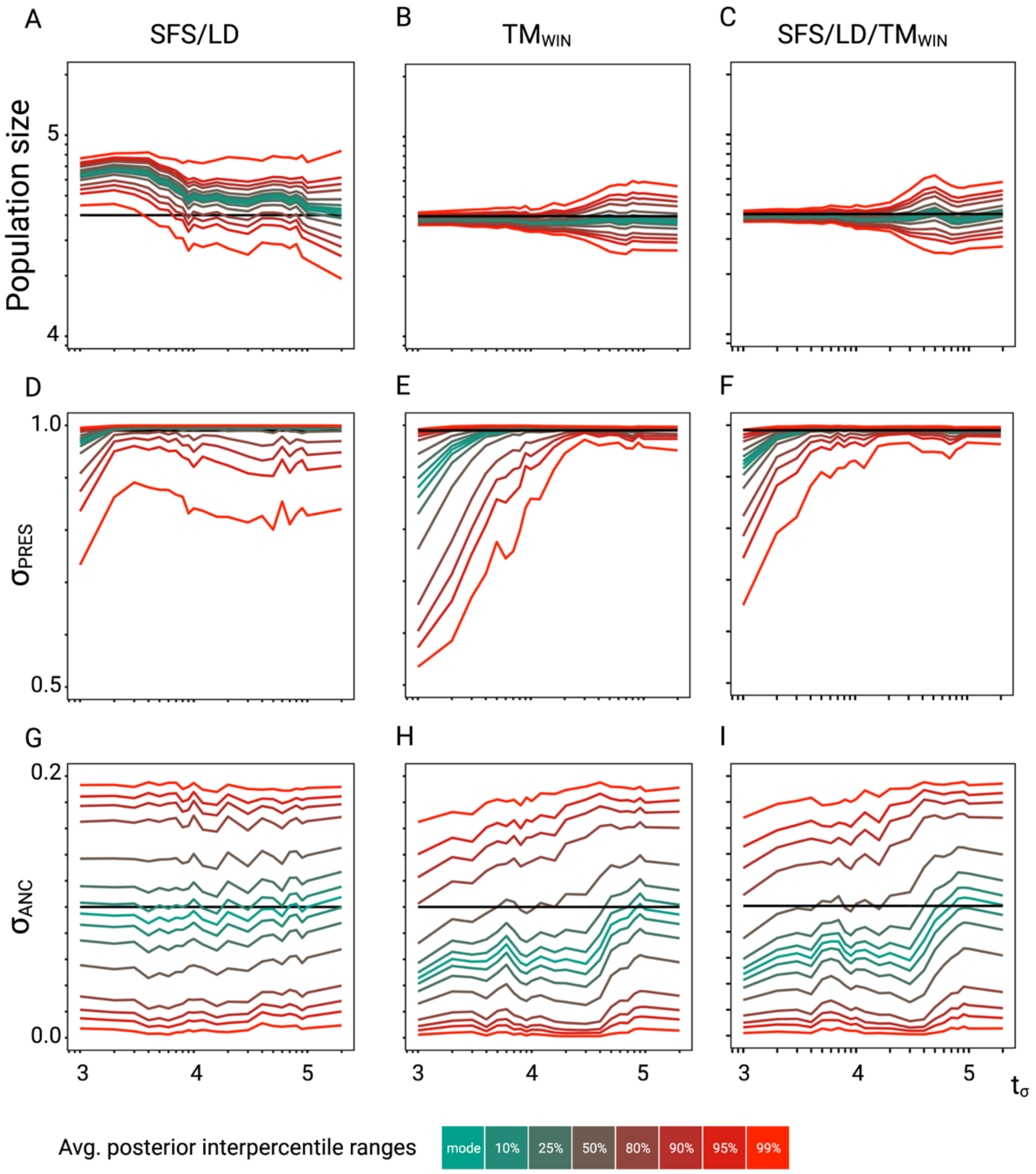


Figure 11. ABC performance analysis: Parameter re-estimation of the three remaining parameters of the model described in **Figure 10**. (A-C) Re-estimation of the population size on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (D-F) Re-estimation of the present selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (G-I) Re-estimation of the ancestral selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average interpercentile ranges for 100 posterior distributions corresponding. Except for the selfing rates, both axes are scaled in log_{10}.

Our estimates were obtained using the same three summarization strategies used for the model choice (SFS/LD, $\mathrm{TM_{WIN}}$, SFS/LD/ $\mathrm{TM_{WIN}}$). The age of a shift to selfing could be well estimated using the $\mathrm{TM_{WIN}}$ approach, while the SFS/LD approach over-estimated t_σ almost over the complete range of values (**Figure 10**, panel F-H). Combining SFS, LD, and $\mathrm{TM_{WIN}}$ does not further improve the accuracy of the estimations (**Figure 10**, panel H). We note that the parameters N and σ_{PRES} (i.e., the population size and the current selfing rate) are better estimated with $\mathrm{TM_{WIN}}$ than with SFS/LD, except for transitions younger than 10^4 generations ago where σ_{PRES} is slightly better estimated with SFS/LD (**Figure 14**, panel D). However, no summary statistics set could estimate the ancestral selfing rate (**Figure 11**, panel G-I).

Importance of the ratio between mutation and recombination rate

The performance of the model choice experiment depends on the specificity of the observed signal, i. e. summarizing statistics, to distinguish between a change in population size and a transition to selfing. The specificity of the signal depends on the detection of MRCA segments (Chapter 1). The SNP density limits the detection of such segments. Thus, to investigate potential information horizon exceedings, we repeated the model choice experiment with an increased r/μ ratio ($r/\mu = 5$).

For $r/\mu = 1$, we maintain consistently high performance in the model choice. However, the power to select the correct model and, thus, the specificity of the observed signal is more uncertain for increased ratios, e. g. here $r/\mu = 5$ (**Figure 12**). Model choice using SFS/LD remains somewhat robust in recent times. $\mathrm{TM_{WIN}}$ based model choice is significantly weaker for the entire investigated time range. However, the model choice experiment performs best when combining SFS/LD and $\mathrm{TM_{WIN}}$. For PODs created with negative linked selection, the performance of model selection is further decreased.

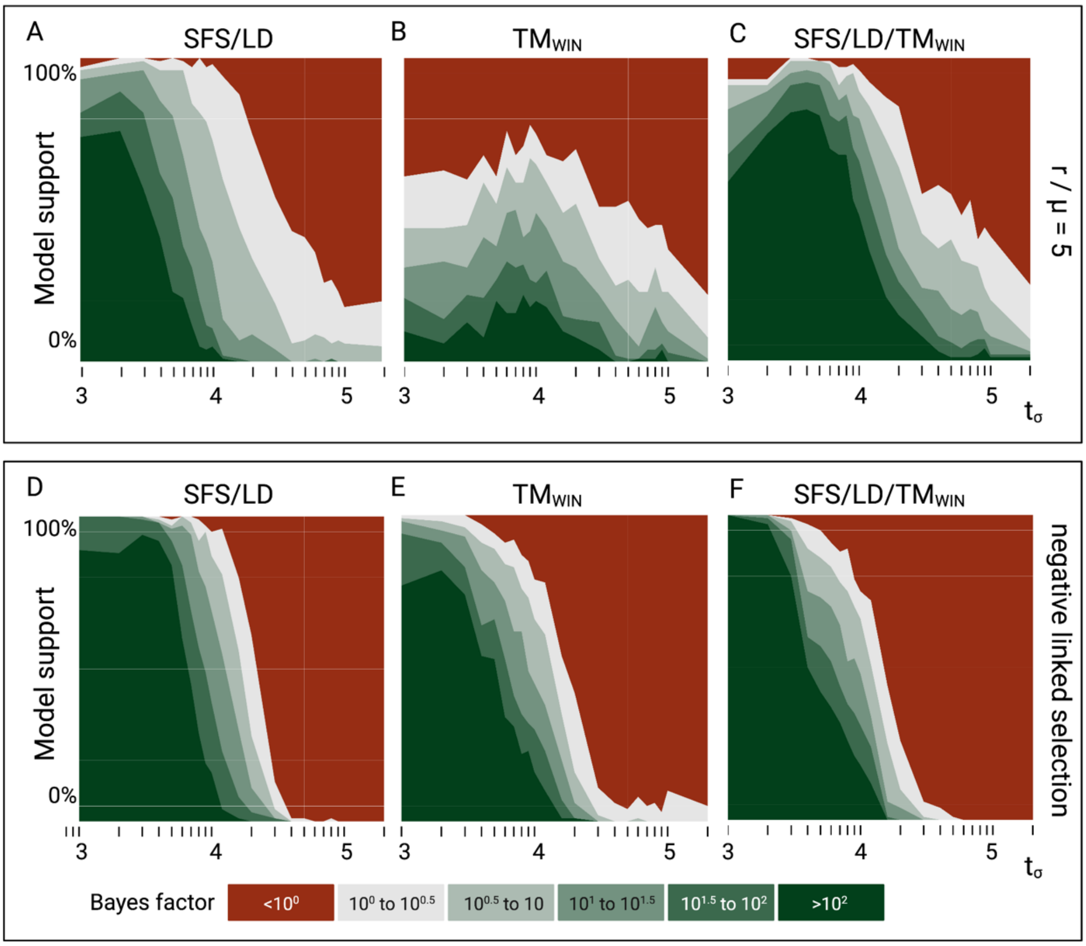


Figure 12. Model choice experiment for different rho-theta-ratios and negative linked selection. The performance analysis for the different conditions was conducted the same way, as shown in **Figure 10**. (A-C) Model choice performance for increased recombination on otherwise same parameters. (D-F) Model choice performance under the same parameters as **Figure 10**, but PODs were simulated under negative linked selection. t_σ-axes are scaled in log_{10}.

Robustness of *tsABC* to negative linked selection in parameter estimation

Background selection (BGS) refers to the effect of deleterious alleles on linked neutral diversity (Charlesworth, et al. 1993; Irwin, et al. 2016). Recently, several studies highlighted that neglecting the effect of BGS in demographic analyses can lead to statistical biases and potential miss-identification of population size changes (Ewing & Jensen, 2016; Johri et al., 2021; Pouyet, Aeschbacher, Thiéry, & Excoffier, 2018; Schrider, Shanku, & Kern, 2016). Because transitions to selfing substantially reduce the recombination rate (up to two orders of magnitude for transitions to predominant selfing), a corresponding increase of linkage between deleterious and neutral alleles is expected. Selfing indeed drastically magnifies the effect of BGS (Kamran-Disfani & Agrawal, 2014). Because *tsABC* ignores the effect of selection, we evaluated the performance when applied to data simulated under a model with both a transition to selfing and background selection. We used *SLiM3* (Benjamin C. Haller et al., 2019; Benjamin C Haller & Messer, 2019) to simulate genomic data with a similar distribution of exonic sequences as in the model species *Arabidopsis thaliana* and modeled negative selection on exonic sequences according to the distribution of fitness effects (DFE) for *Arabidopsis thaliana* published by Hämälä and Tiffin (2020). The *tsABC* estimates remained accurate for estimates using the summarization of unmasked genetic variation (**Figure 13**, **Figure 14**). However, identifying a transition to selfing in the model choice experiment is less accurate (**Figure 12**, panel G-I). These results suggest that parameter estimates of *tsABC* are generally robust to the effect of negative selection on linked neutral sites, even in compact genomes such as the one of *Arabidopsis thaliana*.

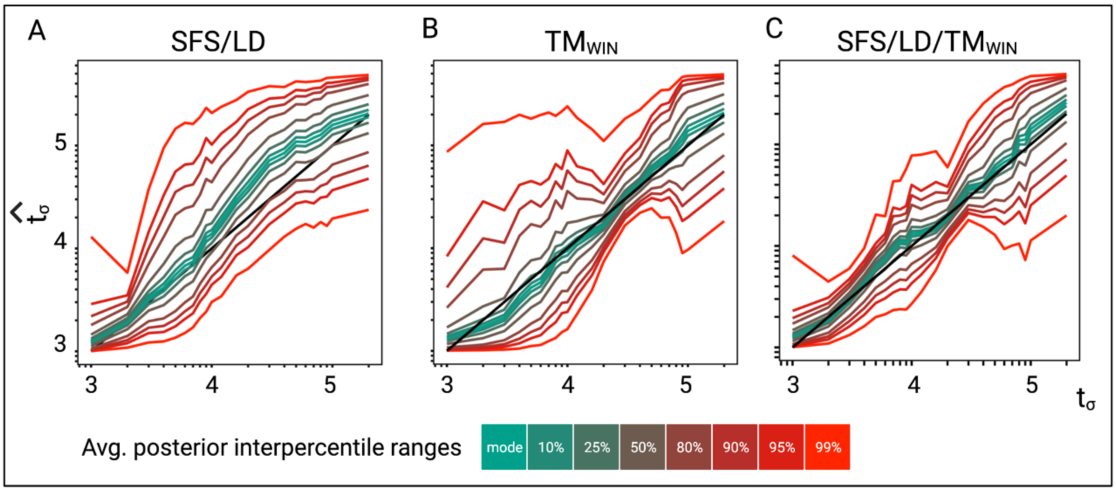


Figure 13. Accuracy of *tsABC* in the presence of background selection (BGS): Inference of times of transition from outcrossing ($\sigma = 0.1$) to predominant-selfing ($\sigma = 0.99$) using *tsABC* using (A) SFS/LD, (B) TM_{WIN} or (C) both. Simulations were done under constant population size and negative selection acting on exonic sequences. The spatial distribution of exonic sequences was fixed and taken from the annotation of *Arabidopsis thaliana*. The negative selection was modeled using a distribution of fitness effects (see methods). The result should be compared with the case without linked negative selection in **Figure 10** (panel F-H). Colored lines represent the interpercentile ranges quantiles for 100 posterior distributions obtained with *tsABC*. Both axes are scaled in log_{10}.

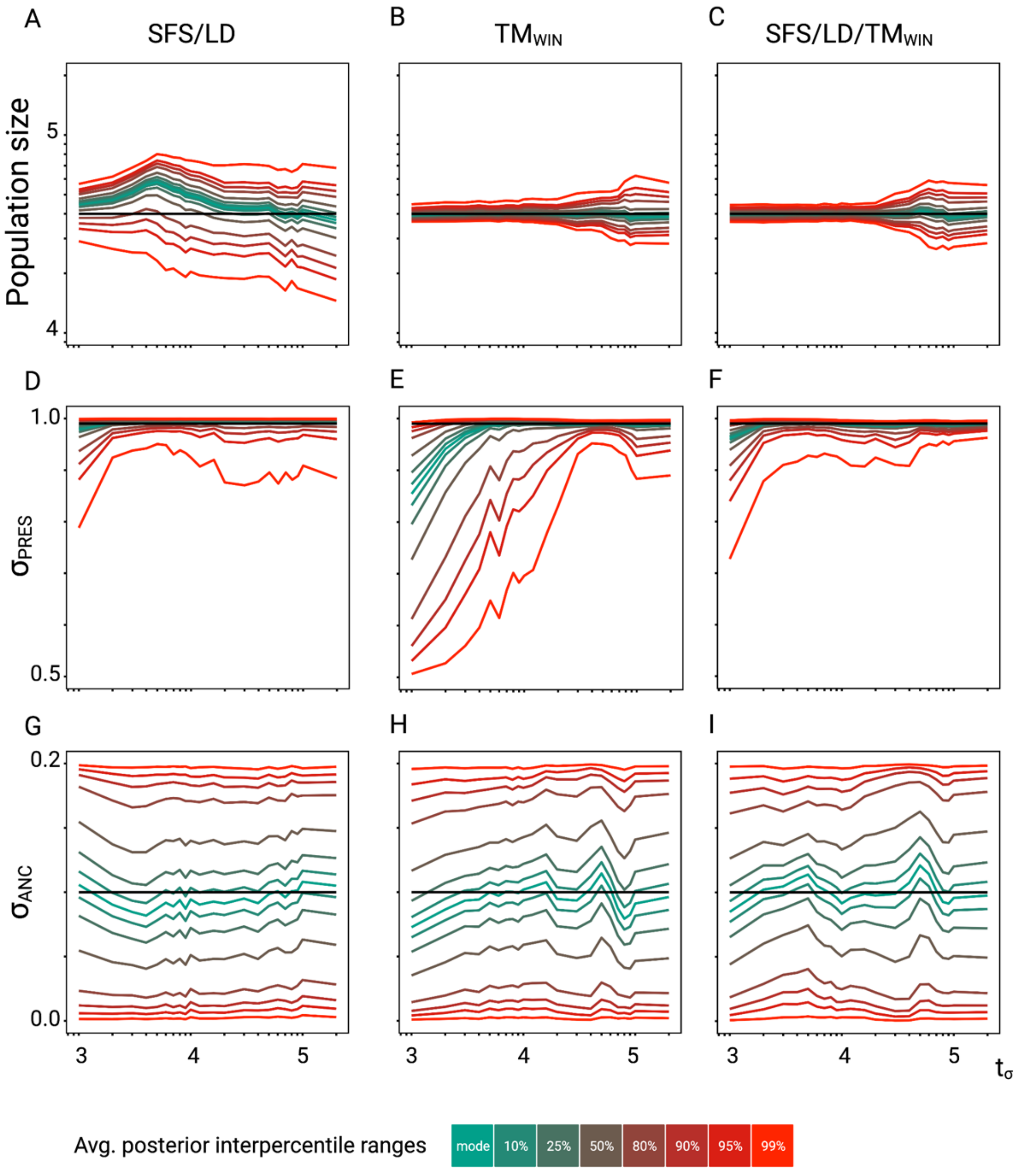


Figure 14. ABC performance analysis under negative linked selection: Parameter re-estimation of the three remaining parameters of the model described in **Figure 10**. (A-C) Re-estimation of the population size on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (D-F) Re-estimation of the present selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (G-I) Re-estimation of the ancestral selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$. Colored lines represent the average interpercentile ranges

for 100 posterior distributions corresponding to the given credible intervals. Except for the selfing rates, both axes are scaled in log_{10}.

Application of *tsABC* to *Arabidopsis thaliana*

To date the transition to selfing in *Arabidopsis thaliana*, we used carefully chosen regions of 1 MB per chromosome to obtain the summarizing statistics. We assured these regions to be in the non-pericentromeric region (Underwood et al., 2018) towards regions of lower diversity to avoid potential biases through associative overdominance in the low recombining regions in the pericentromer (Gilbert, Pouyet, Excoffier, & Peischl, 2020). Further, we calculated the summarizing statistics, SFS, LD, and TM_{WIN}, as described in Chapter 1. We estimated transitions to selfing in *Arabidopsis thaliana* with *tsABC* under the six-parameter model (**Figure 9**, panel C) in the range from 592,321 to 756,976 (**Figure 15**, panel A; **Table 4**).

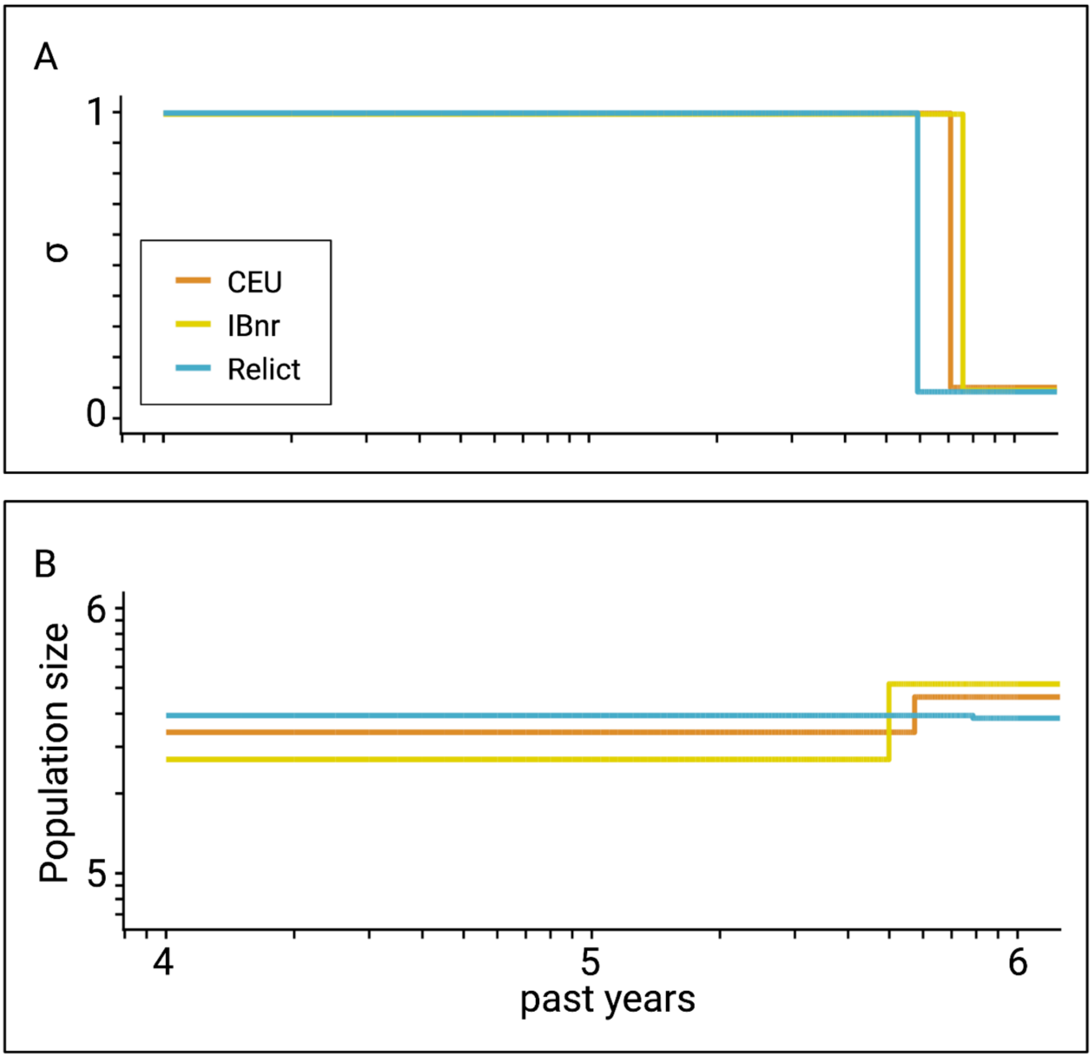


Figure 15. Inference of the time of transition from outcrossing to selfing in *Arabidopsis thaliana* using *tsABC*: A: Inferred transitions from outcrossing to selfing for three independent genetic clusters of *Arabidopsis thaliana* from the 1001 genomes project (CEU, Ibnr, Relicts). (B) Co-estimated population sizes over time with a single population change. Except for the selfing rates, both axes are scaled in log_{10}.

Our estimates are older than the one proposed by Bechsgaard et al. (2006) but younger than the age proposed by Tang et al. (2007). Our estimates of transition to selfing are robust to the geographical origin of the population samples (Iberian non-relicts, Iberian relicts or central European) and range from 592,321 years (CEU) to 756,976 years (IBnr) ago (**Figure 15**, panel A; **Table 4**).

Table 4. Estimated times of transitions from predominant outcrossing to predominant selfing in Arabidopsis thaliana. The demography was estimated for three non-admixed genetic clusters (genetic assignment > 95%) of the 1001 genomes project. *tsABC* estimated the time of a transition from predominant outcrossing to predominant selfing on the entire sample set (n_{CEU}=99, n_{IBnr}=66, $n_{Relicts}$=18). To obtain the summarizing statistics, repeatedly 12 random samples were chosen. The 95% interquartile range CI provides the uncertainty measure on the approximated likelihood function.

Method	*Population*	*Mode*	*95%-IPR*	
tsABC	CEU	707,995	443,485	973,841
tsABC	IBnr	756,976	397,048	988,708
tsABC	Relicts	592,321	386,405	934,499

Discussion

In this study, we developed an inference method to identify and estimate transitions to selfing using the ABC framework (*tsABC*). To date, no method was published to identify and estimate such shifts in reproductive systems using the consequences of such shifts on the genome-wide genetic diversity. Additionally, *tsABC* jointly estimates the demography. We demonstrated how to use a novel summarizing statistic to capture the temporal signature of transitions from outcrossing to selfing. We provided a complete performance analysis to investigate the power of *tsABC* to identify transitions to selfing and show biases and precision in estimating the time of a transition from outcrossing to selfing in a single population. Finally, we estimated the transition to selfing in three independent genetic clusters of *Arabidopsis thaliana*.

Shifts in mating systems are considered a critical evolutionary and ecological process. Timing is a key feature of the transition to selfing (D. Charlesworth & Vekemans, 2005; T. M. Mattila, B. Laenen, & T. Slotte, 2020). Existing hypotheses consider predominant selfing an evolutionary dead-end, reviewed by Igic and Busch (2013). By developing *tsABC*, we provide a method to use genome-wide genetic haplotype variation to date transitions to selfing and, thus, contributing to the evidence for (or against) the existing evolutionary hypothesis of such reproductive shifts.

Using SFS and LD theoretically provides information to relate theta and rho to each other. Thus, their joint measure contains specific information to measure transitions to selfing. However, using *tsABC,* we demonstrated the specificity of that signal being informative only for recent times. Furthermore, although identifying transitions to selfing, SFS/LD alone is not sufficient to infer the correct demographic parameters. Adding the information on past demographic events using $\mathrm{TM}_{\mathrm{WIN}}$ not only provides unbiased parameter inference for recent times but also extends the correct identification to intermediate time ranges (**Figure 10**, panel C-E).

Estimating the correct model parameters exceeds the time range of a robust model choice, i. e., correct identification of a transition to selfing, by more than a magnitude (**Figure 10**, panel F-G). Correctly estimating the model parameters holds for the whole set of parameters. It indicates the potential performance improvement of *tsABC* in the model choice to a similar extent if including prior information and restricting the model parameters to fit ϑ_t, i. e., the temporal population effective diversity. Thus, we enable *tsABC* to implement solely the differences in the genetic structure caused by the rescaled recombination by selfing, but not our assumptions on the demographic history, e. g., the increase of a prior range of a model by factor two potentially halves the marginal probability density of that model.

Classically, population genetics time scales are measured in units of effective population size. The coalescent theory provides mathematical support that the expected (mean) coalescent time of two lineages is 1 scaled in coalescent time; see e. g. (Wakeley, 2009). However, the variance is as significant. Effectively the $\mathrm{T}_{\mathrm{MRCA}}$ converges to 2 for large sample sizes. Thus, the theoretical absolute information horizon is given by the limit of the $\mathrm{T}_{\mathrm{MRCA}} = 2$. Effectively, the information horizon is determined by the SNP density on MRCA segments and their age.

Our study demonstrated that *tsABC* correctly identifies transitions to selfing for times up to 2.5 N_e ($N_e = N_\sigma = 20{,}200$; **Figure 10**, panel E). Dating the transition to selfing is robust for the tested time range from present to 200,000 past generations. However, transitions to selfing reduce diversity; thus, the average of the effective population size exceeds the effective population size at present. Thus, *tsABC* is valid to estimate recent transitions to selfing, but not more ancestral transitions to selfing. The current hypothesis states that selfing is an evolutionary dead-end (Igic & Busch, 2013; G. L. Stebbins, 1957; G Ledyard Stebbins, 1974; S. I. Wright et al., 2013), but selfing is an ESS (J Maynard Smith, 1971; John Maynard Smith & Maynard-Smith, 1978), i. e. a selfer largely outcompetes any outcrosser of the same population throughout the plant phylogeny. Thus, *tsABC* will help to investigate the abundance of recent transitions to selfing throughout the plant phylogeny.

We demonstrated robustly high accuracy in the model choice experiment for low rho/theta ratios, e. g. if $r/\mu = 1$. However, the performance for model choice decreased in more recent times by almost one magnitude (**Figure 12**) for higher r/μ ratios (here, $r/\mu = 5$) . The SNP density per MRCA segment marks a limiting factor in identifying such segments: SNPs approximate the T_{MRCA} of MRCA segments and are the only source to provide information about the underlying genealogy. Thus, the model must be designed carefully under reasonable assumptions. Note, the implementation of prior knowledge, e. g. ecologist providing evidence that a population has undergone a transition to selfing and information on effective population sizes over time would significantly improve identifying transitions to selfing and suggest the better performance of maximum-likelihood methods that co-estimate the demography.

We jointly dated the transition to selfing combined with the demography inference for the first time. The agreement between our estimations for the transition to selfing in *Arabidopsis thaliana* between the two used methods is remarkable. We estimated the transitions to selfing older than previously published estimates (Bechsgaard et al., 2006). However, S-locus-based inferences of shifts to selfing are not only data limitid because of the restricted size of the S-locus, but also to species for which a loss-of-function mutation in the S-locus has caused such transitions. Additionally, the S-locus had to be identified and correctly assembled. This information is only partially available for other species, and the genetic determinism of selfing also varies between genera (Franklin-Tong, 2008). Our only limitation lies in the restrains of the coalescent with partial selfing model (M. Nordborg, 1997, 2000; M. Nordborg & Donnelly, 1997). However, *tsABC* can easily be extended to models which require further modifications to any required complexity, e. g., other reproductive modes, tetraploidization, complex demography, and stepwise transitions to selfing.

In summary, with the developed *tsABC,* we provided a novel method to estimate transitions to selfing to apply to whole-genome diversity observable from natural populations. We analyzed its potential biases and precision under neutral and non-neutral scenarios. We demonstrated the robustness and precision of *tsABC* if the models were designed carefully for recent and intermediate time ranges. We consistently dated the transition to selfing of *Arabidopsis thaliana* using three distinct genetic clusters as proof of principle. Thus, we enable exploring the phylogeny of plants for transitions to selfing to contribute to the understanding of evolutionary processes shaping plant species and populations in the context of the evolution of sexes.

Conclusion

1. *tsABC* identifies transitions to selfing, given that they occurred more recently than 2.5 N_e generations.
2. Our performance analysis indicates potential for improvement in the model choice by including prior information about past effective population sizes and restricting the model parameters to fit ϑ_t (the past pairwise genetic diversity under given past effective population sizes).
3. Estimates on three independent genetic clusters of *Arabidopsis thaliana* are slightly older than previously published estimates and dated the transition to selfing to a range between 592,321 years and 756,976 years ago.
4. The use of *tsABC* will contribute to the phylogenetic exploration and identification of recent transitions to selfing to elaborate on existing

hypotheses on the evolution of mating systems, e. g. the dead-end hypothesis of selfing species.

Author contributions

All methods and data shown in this chapter were developed by the author of this thesis under the supervision of principal investigator Dr. Stefan Laurent and guided by the thesis advisory committee (TAC).

Chapter 3

Inferring demography and piecewise-constant selfing using *teSMC*

Inferring demography and piecewise-constant selfing using *teSMC*

Keywords

SMC, demographic inference, piecewise-constant selfing, population genetics

Introduction

Since the technological development of the last decades provided access to large amounts of genetic data, haplotype information on entire chromosomes via NGS data has become accessible (Marchi et al., 2021). The introduction of the sequentially Markovian coalescent (SMC) provided the statistical framework to analytically describe the genealogy of a pair (or more) samples along the sequence. The genealogy (T_{MRCA} for a sample of size two) emits diversity (e. g. SNPs). Thus, the SMC is a model for relating parameters of genealogy to an observable statistic. Using hidden Markov models (HMM), the likelihood of proposed genealogical models can be optimized (Baum, Petrie, Soules, & Weiss, 1970). Recombination events in between sites may cause differences in the genealogies for the genealogical history older than the recombination event. The SMC describes the current site's genealogy as dependent on the previous site's genealogy and the probability of a recombination event occurring (recombination rate) between these two sites. Thus, the SMC approximates the Ancestral Recombination graph providing information about past demographical processes.

The description of a sequence of states (e. g. genealogies) along a sequence is a well-known paradigm in mathematics. It has been addressed and described as the Markov model. Hidden Markov models (HMM) describe the sequence of observed states (e. g. diversity) depending on underlying hidden states, which emit to the observable states. In the case of a sample of size two, the genealogy is described via the coalescent time (T_{MRCA}). The probability of a mutation having occurred then depends on the T_{MRCA}. Assembling the SMC and the HMM leads to the description of T_{MRCA} along the sequence, which emits an observable diversity.

Baum et al. (1970) introduced an algorithm (Baum-Welch) to optimize the likelihood of a Markov model given a set of observed sequences. Li and Durbin (2011) firstly formalized an SMC-HMM model for pairwise samples to infer an optimized piecewise-constant past population size, known as the pairwise sequentially Markovian coalescent (*PSMC*). Schiffels and Durbin (2014) extended the PSMC framework to multiple samples, the multiple sequentially Markovian coalescent (*MSMC*), using the SMC-prime approximation to the full coalescence model to optimize its composite likelihood given the first coalescent event for the provided sample set.

Moreover, they extended the *PSMC* framework to optimize the likelihood of the model using all possible pairs of a given sample set (*MSMC2*), calling its application to a sample of size two *PSMC'* (pronounce "P – S – M – C – prime"). Recently, the SMC-HMM framework was extended to include other properties of the marginal genealogies of multiple samples, e. g. the T_{MRCA} and the entire length of the genealogy (Upadhya & Steinrücken, 2021).

Previously, Sellinger et al. (2020) extended the *PSMC'* to *eSMC* infer constant seed bank or selfing rates. In this way, they included the estimation of ecological parameters into an SMC-HMM for the first time. Under seed bank or selfing, both, the discrepancy between census and effective population size leads towards the expected $\vartheta/\rho \neq \mu/r$. The underlying model of *eSMC* uses the deviation between the ratios ϑ/ρ and μ/r to infer either seed bank or selfing rates (e. g., equation 6). Their model assumes both rates to be constant and show that the demographic inference is strongly biased if constant seed bank is not considered but less severe biased if constant selfing is not considered.

In this study, we collaborated with Thibaut Sellinger and Aurelien Tellier to introduce and test an inference method to estimate piecewise constant selfing rates and population sizes using the SMC-HMM approach (*teSMC*). Further, we tested the accuracy and performance to estimate transitions from predominant outcrossing to predominant selfing using simulations. Finally, we inferred the transition to selfing of three non-admixed genetic clusters of *Arabidopsis thaliana*.

Models and Methods

Genomes or chromosomes evolve in a temporal process generation by generation, including mutation and recombination during reproduction. Comparing genetic sequences reveals differences, i. e., mutated sites or SNPs. Segments of this comparison that share the same genealogical history are separated by recombination events, leading to differences in the T_{MRCA} or topology of a genealogy. However, for a sample of size two, only the T_{MRCA} determines the genealogy since only a single possible topology for a genealogy exists, i. e., two lineages coalescing at the time T_{MRCA}.

The sequentially Markovian coalescent is a model enabling simulating genealogies and diversity along the sequence. The mutational process is described as a Poisson process overlaid on a simulated genealogical tree in the coalescent framework. Mutations accumulate over time, generally approximated and described as a constant rate (dos Reis et al., 2016; Zuckerkandl & Pauling, 1965) per generation and base pair. Whether a site mutates depends on the stochasticity of the mutational process. Thus, older T_{MRCA} tend to contain more SNPs than segments with younger T_{MRCA}. Consequently, diversity is a statistical measure for T_{MRCA}. Taken together, that enables the development of an HMM using the SMC.

Hidden Markov models in the coalescent framework

A conceptual description of the HMM used in *teSMC* is schematically depicted in **Figure 16**. An HMM itself is a model consisting of five parameters: 1) The set of hidden states, 2) a set of observable states, which sometimes is referred to as a signal, and 3) a transition probability matrix describing the probability of one hidden state to the next, 4) an emission probability matrix describing the probability of observing each signal given each hidden state, and 5) the initial probability, which provides the probability distribution at the first position of the sequence. In the SMC'-HMM for a sample of size two, we translate this into 1) the discretized T_{MRCA}, 2) the number of SNPs at a position, 3) a transition probability matrix describing the transition of any of the discrete T_{MRCA}s to any of the discrete

T_{MRCA}s, 4) an emission probability matrix describing the probability to observe an SNP dependent on the T_{MRCA} and 5) the stationary distribution of the SMC' under the proposed demographic parameters (e. g. constant population size and constant selfing at the beginning before the optimization), respectively.

To discretize the hidden state, we usually use an approximately equidistant log-spaced time distribution to provide an equally distributed information content per hidden state under all constant assumptions. There are only two observable states for a sample size two, as we observe either an SNP or none. The emission probability depends on the T_{MRCA}; it is intuitive to understand that large T_{MRCA}s have a higher probability of emitting a SNP and vice versa. It is straightforward to implement missing data here. To include missing data in our analysis, we expand our set of observables with another element that every hidden state emits equal probability. The initial probability must be defined. We usually use the stationary probability distribution, i. e. the equilibrium distribution, which can be calculated.

The HMM was designed to solve three basic problems: 1) What is the probability that the model generated the observations? 2) What is the most likely sequence of states under the proposed model? 3) How do we need to adjust the model parameters (the initial probabilities, transition probability matrix, and the emission probability matrix) to maximize the likelihood that the HMM produced the sequence of observables, e. g. a sequence of homozygous and heterozygous (SNP) sites. The algorithms to solve these problems are well described. *teSMC* generally uses a modified Baum-Welch algorithm (BW) and optimizes the composite likelihood of the model only based on the transition probabilities. Optimizing the HMM parameters on the full likelihood (answer to problem 1) did not increase the performance (data not shown). The answer to the second problem can be calculated using the Viterbi algorithm. It allows us to infer the positions at which recombination has occurred. However, the likelihood of a state sequence depends highly on the proposed model and may not be informative under constant population size and selfing assumptions to obtain MRCA segments.

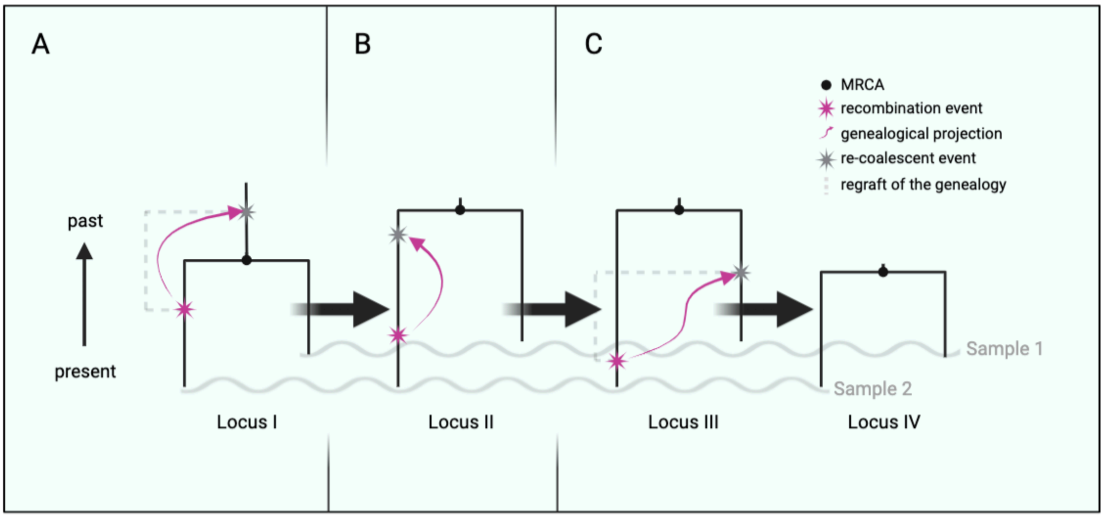


Figure 16. The sequentially Markovian coalescent process for a sample of size two considers recombination events of three different types in the *SMC′* algorithm and the *PSMC'*-based demographic inference methods. Recombination events potentially cause changes to the genealogy of a sample of size two. (A) The T_{MRCA} increases if the recombining lineage re-coalesces at a time older than the T_{MRCA} of the current MRCA, (B) it remains the same if the lineage re-coalesces to itself before the T_{MRCA} of the current MRCA, or (C) it decreases if the lineage re-coalesces to any other lineage but itself before the T_{MRCA} of the current MRCA. The probability of recombining between two loci depends on the integrated recombination rate over time on the genealogy of the current locus. The probability of coalescing follows the assumed coalescent framework.

SMC-HMM

In the SMC-based methods, the waiting time until the next recombination breakpoint along the sequence is dependent on the length of the coalescent tree and the recombination rate. The lineage on which the recombination event occurred will be simulated backward-in-time to coalesce with the remaining lineages, but not itself (McVean & Cardin, 2005) or including itself (Marjoram & Wall, 2006). The probability of a recombination event occurring between two sites depends on the integrated recombination rate over the genealogy, e. g. the coalescent tree of the previous site. Variable recombination rates per time window weighting the total probability of a recombination event enable the implementation of a rescaled SMC, thus selfing. Selfing rescales both the effective recombination and coalescent rates (see Chapter 1). Thus, the parametrization of a population model with piecewise-constant population size and piecewise-constant selfing rates enables the inference of transitions from predominant outcrossing to predominant selfing. An expected piecewise-constant recombination rate affects the transition probabilities of T_{MRCA}s along the sequence. In contrast, the expected piecewise-constant population size will rescale the coalescent times, i. e. the T_{MRCA}, and thus the emission probabilities of SNPs (**Figure 16**).

Modifications of *eSMC* to infer changing selfing rates

Here, we extended *eSMC* (Sellinger et al., 2020) into *teSMC*, allowing the estimation of varying selfing or recombination rates through time, jointly with varying population sizes. To achieve that, *teSMC* no longer assumes the ages of recombination events to follow a uniform distribution along the genealogical branches representing the recombining lineage (Li & Durbin, 2011; Schiffels & Durbin, 2014); but rather let them be a function of the selfing and recombination rate at each piecewise-constant time frame, i. e. hidden state. This enables *teSMC* to jointly infer piecewise constant selfing or recombination rates and population sizes by maximizing the approximated likelihood for the proposed parameter functions, e. g. $\sigma(t)$ or $r(t)$ and $N(t)$ with σ, r, and N being functions of time.

In *teSMC*, the parameter space increased on the magnitude of the number of hidden states compared to *eSMC*. SNPs are a sparse information source; however, they are the only observable source to optimize an inferred demography. Thus, the optimization process potentially benefits from prior knowledge to help the parametrization of the inference model. To account for prior knowledge, two modes are implemented for parameter inference: the free mode, in which each hidden state has its independent selfing rate, and the single-transition mode in which *teSMC* estimates only three parameters: the current and ancestral rates, and the transition time between both rates; this marks a constraint significantly reducing the number of inferred parameters and well suited for the analysis of recent and sudden shifts from outcrossing to predominant self-fertilization. Details about the calculation of the HMM underlying *teSMC* can be obtained in the appendix (Description of *teSMC*).

Simulations and inference

We used the same simulation and population models as in Chapter 1 for the performance analysis of *tsABC*. Briefly, we simulated transitions to selfing under the same parameters as we used for the PODs (**Table 2**). We inferred the population sizes and selfing rates through time on ten pairwise comparisons using *teSMC*.

Results

Briefly, we repeated the performance analysis of chapter 2 but using *teSMC*. We showed the theoretical convergence of *teSMC*, the accuracy of dating transitions to selfing, and its robustness to negative linked selection. Unfortunately, the differences in the theoretical model design and parametrization of the demographic models complicate the comparison of the two methods.

Theoretical convergence

First, to demonstrate the theoretical accuracy of our model and inference method, we analyze its performance when sequences of T_{MRCA} are given as input instead of sequence data. This is termed the best-case convergence of *teSMC* (Sellinger, Abu-Awad, & Tellier, 2021). We simulate data from a population undergoing a substantial bottleneck and simultaneously a transition to selfing or change in recombination rate. We consider such demography complex and difficult to infer. Thus, we obtain a theoretical information horizon of the underlying genealogy independent of the ρ/θ ratio (see Chapter 2) as branch length-based measures of genealogical trees converge to site-based measures for high recombination rates (Ralph et al., 2020). In both cases, the population size and the past selfing/recombination values are recovered with high accuracy (**Figure 17**).

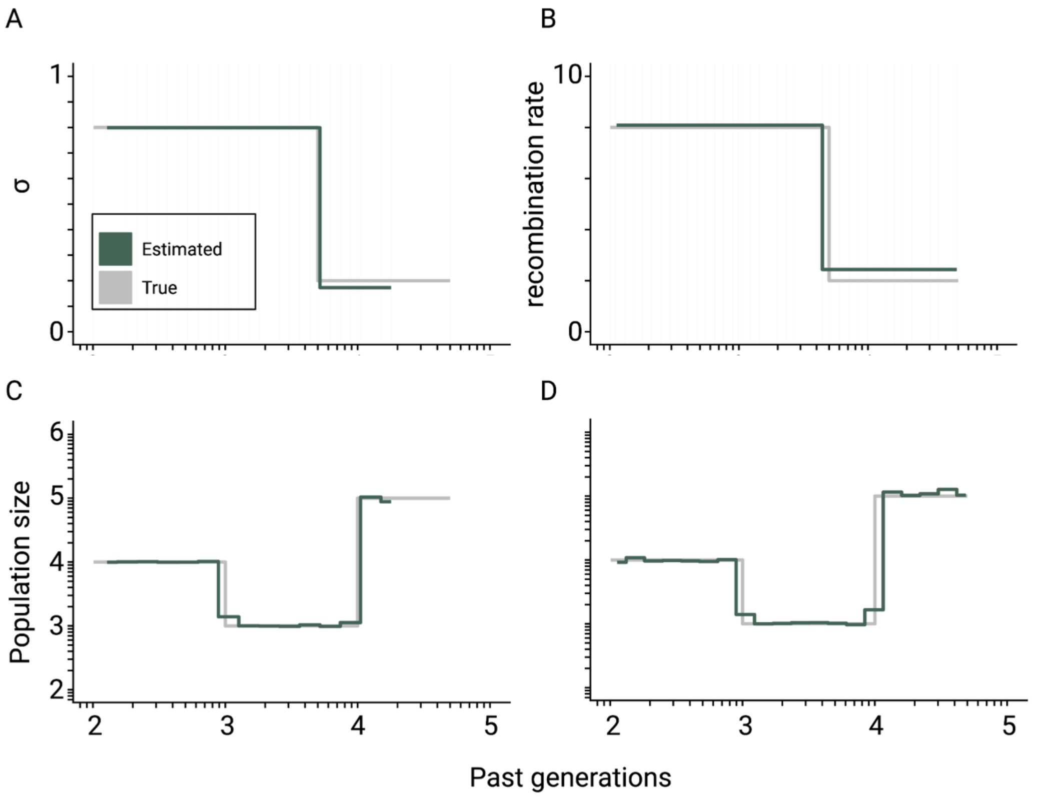


Figure 17. Theoretical convergence of *teSMC* under complex demography. Best-case convergence of *teSMC* using ten sequences (i.e., haploid genomes) of 100 Mb (green) when the population undergoes a bottleneck (true sizes are indicated in black) with either variation of selfing in time (A, C) or variation of recombination rate in time (B, D). The selfing rate through time is represented in A), and the corresponding estimated population size is represented in C), the estimated recombination rate through time (B), and the corresponding population size in D). The recombination rate was set to $r = 1 \cdot 10^{-7}$ and the mutation rate to $\mu = 1 \cdot 10^{-8}$ per generation per bp. Except for the selfing rates, both axes are scaled in log_{10}. Simulations and raw data were provided by Thibaut Sellinger.

Second, to understand the convergence properties of *teSMC*, we analyzed its performance under a simple scenario assuming a constant population size and a constant selfing value of 0.9 given a different amount of data. We compare the *eSMC* method, which estimates a constant selfing rate in time, with *teSMC,* which estimates varying selfing through time. When selfing is known to be constant (*eSMC*), the value of this parameter is recovered with high accuracy and low variance even with the lowest amount of given data (**Figure 18**, panels A and C). However, when it is unknown whether selfing changes through time (*teSMC*), a greater amount of data is required to reduce the variance in the estimation (**Figure 18**, panels B and D).

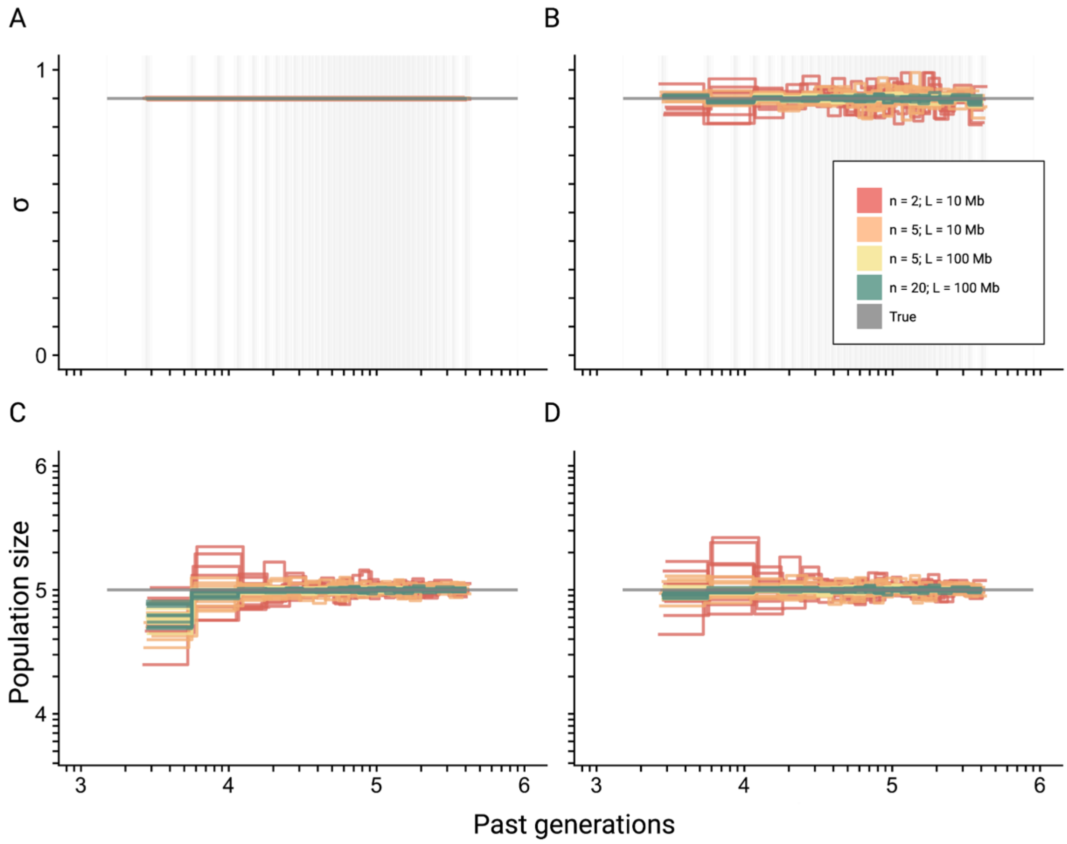


Figure 18. Best-case convergence of *teSMC* for a different amount of data. Best-case convergence of *teSMC* using different combinations of sample sizes ($n = 2$, $n = 5$, or $n = 20$ sequences; i.e., haploid genomes) and sequence lengths ($L = 10\ Mb$ or$L = 100\ Mb$), when population size is constant (N=100,000, black line) with a constant selfing rate of 0.9. The best-case convergence is estimated assuming that selfing is constant (A, C) or varying in time (B, D). The estimated population size assuming constant selfing in time is represented in (C) and the simultaneously estimated selfing rate in (A). The estimated population size assuming varying selfing rate in time is represented in (D) and the simultaneously estimated selfing rate through time in (B). The recombination was set to $r = 1 \cdot 10^{-8}$ per generation per bp. Except for the selfing rates, both axes are scaled in log_{10}. Simulations and raw data were provided by Thibaut Sellinger.

Estimates on simulated data

We now evaluate the statistical accuracy of *teSMC* on neutral polymorphism data from 5 Mb, simulated under a model with constant population size ($N = 40{,}000$) with mutation (μ) and recombination (r) rates of $1 \cdot 10^{-8}$, and with an instantaneous change from outcrossing ($\sigma_{ANC} = 0.1$) to predominant selfing ($\sigma_{ANC} = 0.99$) at the time t_σ (see methods Chapter 2). The single-transition mode estimation procedure performs well over a wide range of t_σ values, although it slightly underestimates the true value for transitions younger than

10,000 generations (corresponding to 0.0625 in units of 4N generations, **Figure 19**). The free mode of *teSMC* performs better before 10,000 generations but slightly overestimated tσ compared to the single-transition mode over the rest of the range. Population sizes estimated under the assumption of a constant selfing rate were consistently larger than the true value in the outcrossing phase and displayed large fluctuation in the selfing phase, which could be mistaken for past population size bottlenecks (**Figure 20**). On the other hand, when *teSMC* is allowed to account for the change in selfing rates, population size estimates (N) remain close to the true values. We note that the increased variance in N in the selfing phase is likely caused by fewer available MRCA segments.

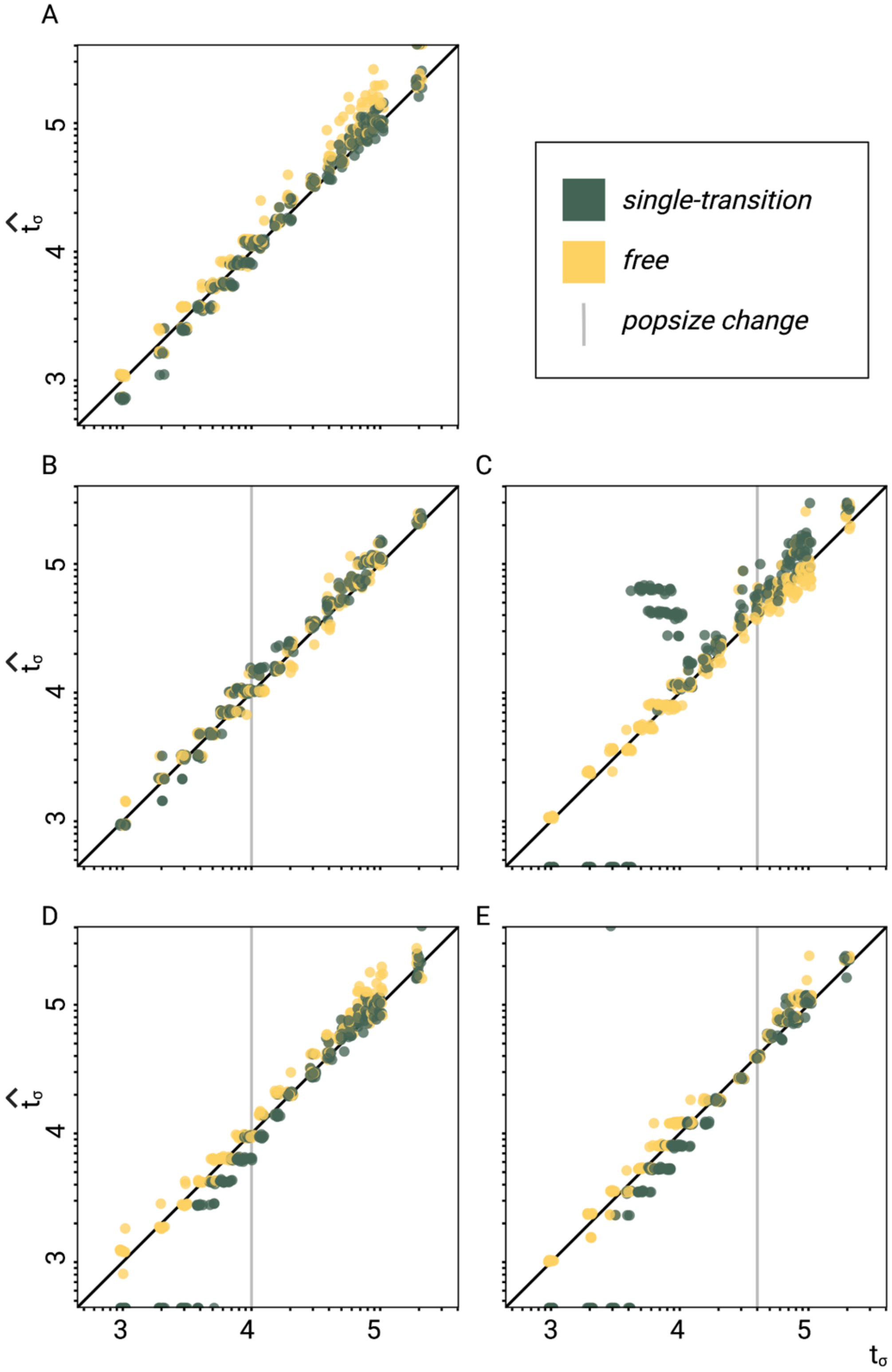


Figure 19. Performance of *teSMC* on simulated polymorphism data. Inference of times of transition from outcrossing (σ = 0.1) to predominantly selfing (σ = 0.99) using neutral simulations. The inference was made using the free mode (yellow) and the one-transition mode

(green) of *teSMC* and ten replicates per time point. (A) Under constant population size. (B-E): simulations were done with an additional change in population size; the vertical grey line indicates the change in population size. (B-C) From NANC = 200,000 to NPRES = 40,000 (population crash) at 10,000 generations (B) or 40,000 generations (C) in the past. D-E) From NANC = 40,000 to NPRES = 200,000 (population expansion) at 10,000 generations (D) or 40,000 generations (E) in the past. Both axes are scaled in log_{10}. Inference was done by Thibaut Selllinger.

Finally, we evaluated the ability of *teSMC* to jointly estimate the age of a transition to predominant selfing and the time of a stepwise change in population size. To achieve that, we used simulated data produced as above, except with the addition of a single stepwise population size reduction (**Figure 19**, panels B and C) or expansion (**Figure 19**, panels D and E). In both cases, our results indicate that *teSMC* can precisely estimate the age of the shift to selfing, regardless of the relative timing of the population size change and the transition to selfing. Also, in most cases, the population sizes inferred by *teSMC* were close to the true simulated values (**Figure 21**). However, when the transition is recent and the present population size is low, this can affect the precision of the estimations of the population sizes (**Figure 21**). We note that, as it is a characteristic of SMC-based methods, *teSMC* failed to recover the population size in very recent times, suggesting a lack of data, i. e. coalescent events (Sellinger et al., 2021). These results demonstrate that transitions to predominant self-fertilization and, more generally, large changes in recombination rate through time can be captured by *teSMC,* and the estimations can be disentangled from changes in population sizes.

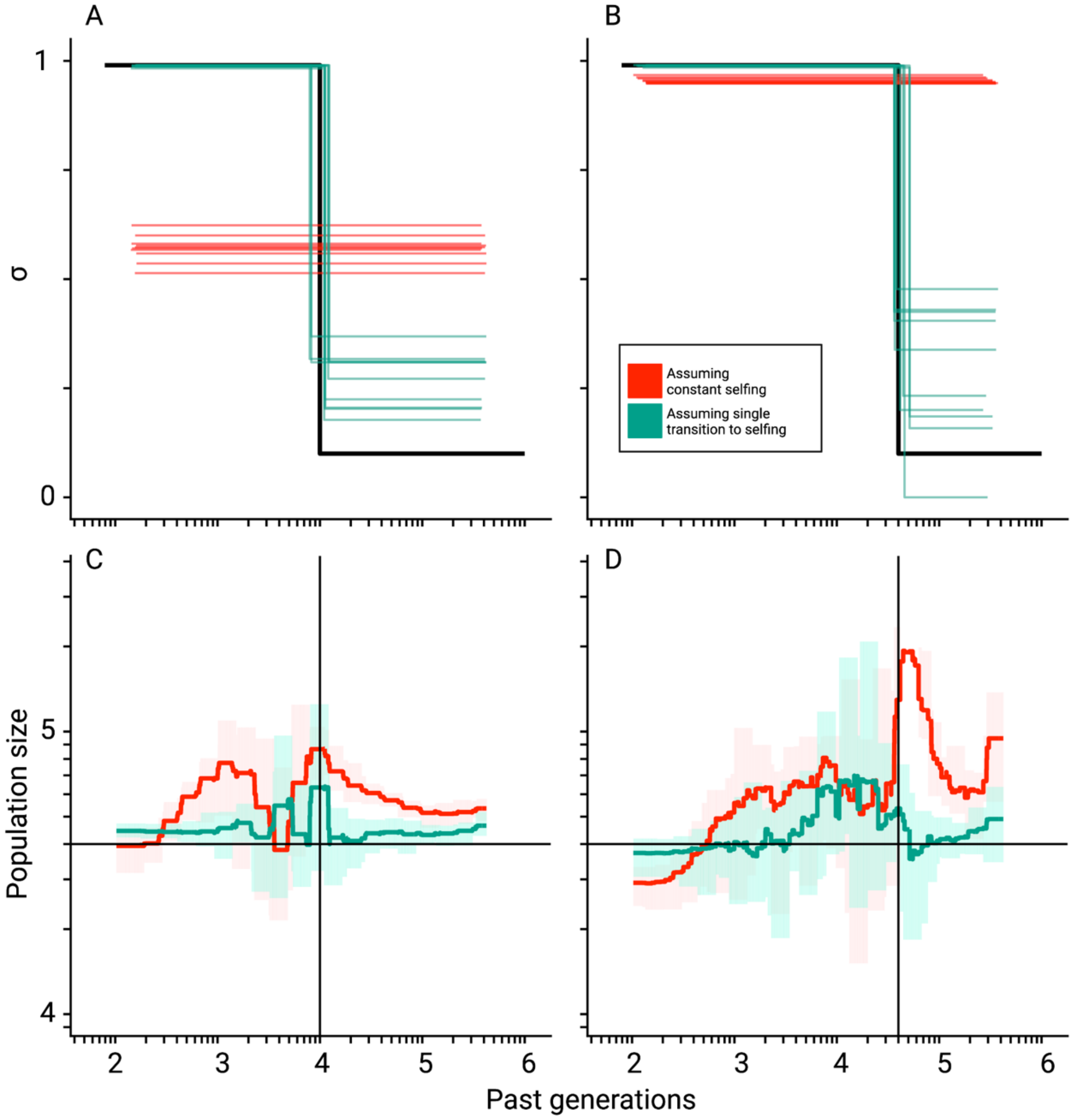


Figure 20. Inference of population sizes when transitions to selfing are not accounted for. Comparisons between true (black lines) and estimated selfing rates and population sizes estimated by *teSMC* for ten replicates. Here simulations were done using a constant population size ($N = 40{,}000$) and a transition to selfing from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$ at $t_\sigma = 10{,}000$ (A,C) and 40,000 (B,D). Five chromosomes of 1 Mb were simulated with mutation and recombination rates set to $\mu = r = 1 \cdot 10^{-8}$ events per generation per bp. Red and green lines indicate results obtained assuming the wrong model (i.e., constant selfing) and the correct model (i.e., single-transition). For the selfing rates (A, B), results for each replicate are indicated with solid lines. For the population sizes, the ten replicates were summarized by the green and red shaded areas, where the width of the shaded area corresponds to the range between the minimum and maximum value observed across replicates. Except for the selfing rates, both axes are scaled in log_{10}.

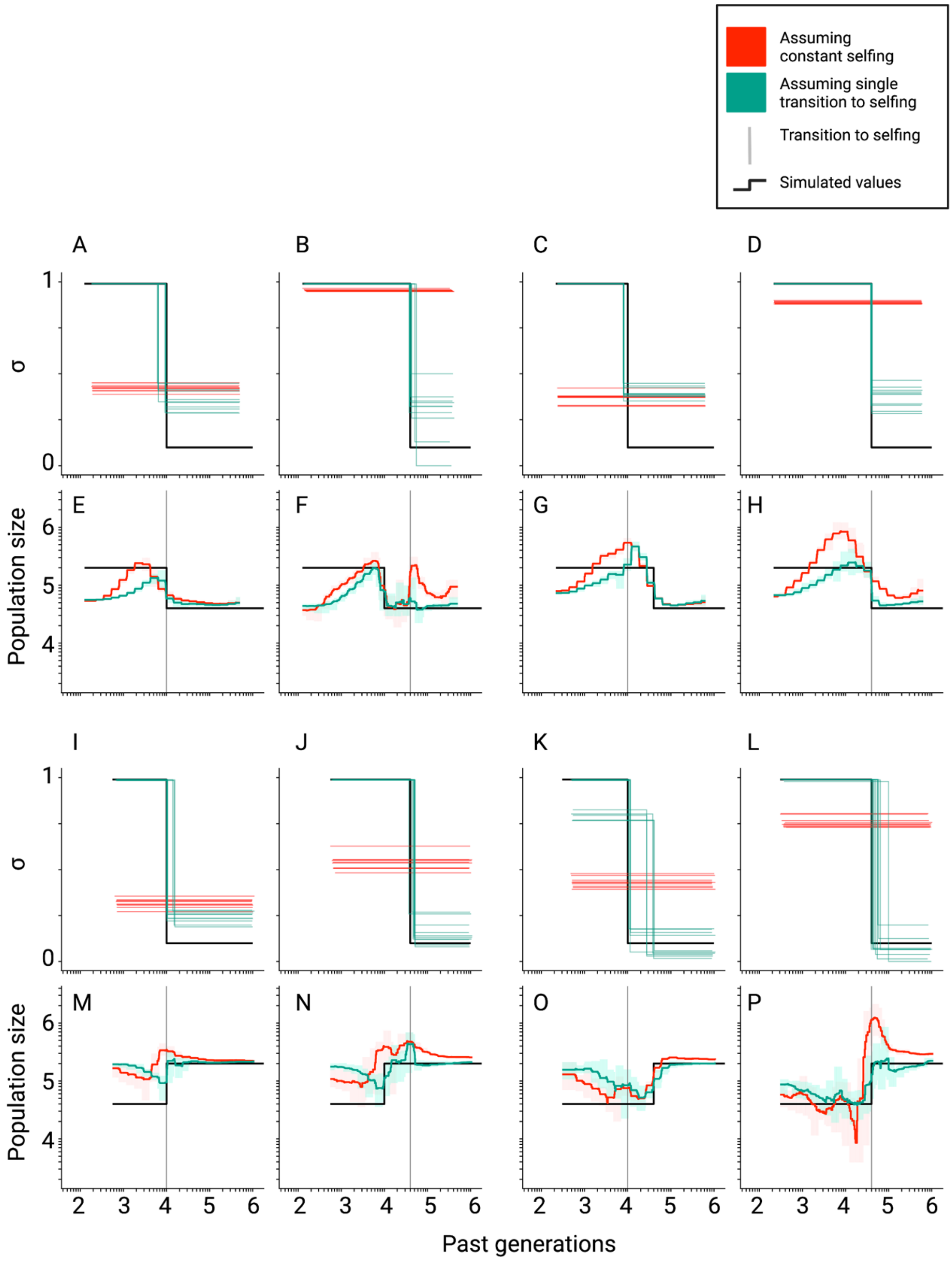


Figure 21. Inference of population sizes and selfing rates estimated by *teSMC* when both parameters change over time. (A-P): Comparisons between true (black lines) and estimated selfing rates and population sizes estimated by *teSMC* for ten replicates. Here simulations were done as in **Figure 20** except for the addition of a single stepwise population size expansion forward-in-time (first and second rows) or contraction (third and fourth row). The transition to selfing occurred from $\sigma_{ANC} = 0.1$ to $\sigma_{PRES} = 0.99$ at $t_\sigma = 10{,}000$ (A, C, E, G, I, K, M, O; first and third column) and $t_\sigma = 40{,}000$ (B, D, F, H, J, L, N, P; second and fourth column). For

the population sizes, the ten replicates were summarized by the green and red shaded areas, where the width of the shaded area corresponds to the range between the minimum and maximum value observed across replicates. Red and green lines indicate results obtained assuming the wrong model (i.e., constant selfing) and the correct model (i.e., single-transition). For the selfing rates, results for each replicate are indicated with solid lines. Except for the selfing rates, both axes are scaled in log_{10}.

Masking for exonic regions improves *teSMC* inference robustness

As described in chapter 2, BGS can lead to statistical biases in the demographic inference if neglected. Transitions to selfing result in a substantial reduction of the recombination rate up to two orders of magnitude. Because *teSMC* ignores the effect of selection, we evaluate its performance when applied to data simulated under a model with both a transition to selfing and background selection. Again, we used *SLiM3* (Benjamin C. Haller et al., 2019; Benjamin C Haller & Messer, 2019) to simulate genomic data with a similar distribution of exonic sequences as in the model species *Arabidopsis thaliana* and modeled negative selection on exonic sequences according to the distribution of fitness effects (DFE) for *Arabidopsis thaliana* published by Hämälä and Tiffin (2020). We found that when exonic sequences are masked, the accuracy of estimating the transition to selfing by *teSMC* improves slightly compared to the unmasked case (**Figure 22**). These results suggest that our approach is somewhat robust to the effect of negative selection on linked neutral sites, even in compact genomes such as *Arabidopsis thaliana*.

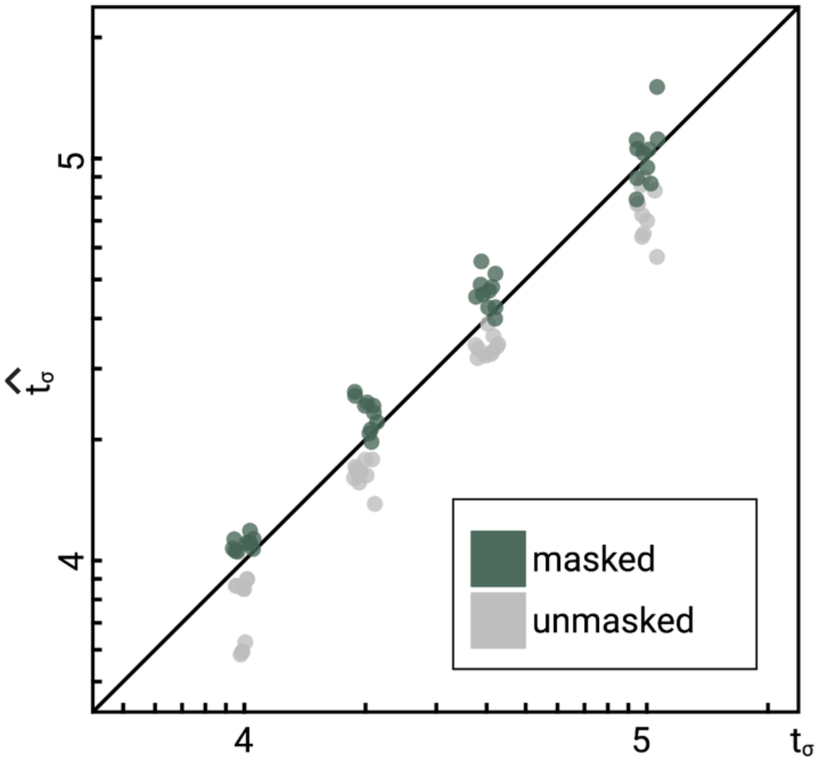


Figure 22. Accuracy of *teSMC* in the presence of background selection (BGS). Inference of times of transition from outcrossing ($\sigma = 0.1$) to predominant-selfing ($\sigma = 0.99$) using *teSMC*. Simulations were done under constant population size and negative selection acting on exonic sequences. The spatial distribution of exonic sequences was fixed and taken from the annotation of *Arabidopsis thaliana*. The negative selection was modeled using a distribution of fitness effects (see Chapter 2). Comparison between simulated values of t_σ and estimates obtained with *teSMC* using the one-transition mode. Estimations were conducted with and without masking exonic sequences subject to negative selection. Both axes are scaled in log_{10}. Inference was done by Thibaut Sellinger.

Application of *tsABC* to *Arabidopsis thaliana*

Similar to inferring transitions to selfing of *Arabidopsis thaliana* using *tsABC*, again, we used *teSMC* to estimate the transition to selfing for three non-admixed genetic clusters. We used the single-transition mode of *teSMC,* parametrizing the model to only allow for a single change in selfing rates. The piecewise constant population size was parametrized to vary during the optimization freely.

Using *teSMC,* we jointly estimated the demography of each *Arabidopsis thaliana* population and the transition to predominant self-fertilization (**Figure 23**). Here, we estimated the transitioning time from predominant outcrossing to selfing ranging from 697,490 to 749,668 years ago, assuming an average of one year per generation (**Table 5**). Furthermore, the estimates of transitions to selfing were robust to the geographical origin of the population sample.

Table 5. Estimated times of transitions from predominant outcrossing to predominant selfing in *Arabidopsis thaliana*. The demography was estimated for three non-admixed genetic clusters (genetic assignment > 95%) of the 1001 genomes project. *teSMC* estimated the time of a transition from outcrossing to selfing on a sample of 20 haplotypes, 1 MB from each of the five chromosomes. Exonic regions were masked. The Relicts sample set consisted of 17 individuals only.

Method	*Population*	*Mode*
teSMC	CEU	697,490
teSMC	IBnr	713,421
teSMC	Relicts	749,668

Discussion

This study introduced and tested an inference method to infer past selfing rates and population sizes of single populations by extending the existing SMC-HMM method *eSMC* to changing selfing rates through time (*teSMC*). We tested the theoretical best-case convergence of *teSMC* on simulated marginal genealogies. Furthermore, we tested the performance of *tsABC* on simulated data under different demographies and for different parametrization modes. Finally, we estimated transitions to selfing consistent with our current estimates using *tsABC* (Chapter 2) and, thus, existing published estimates using different approaches. Together with *tsABC*, we provided two distinct methods to infer transitions to selfing from the genome-wide variance obtainable from natural populations.

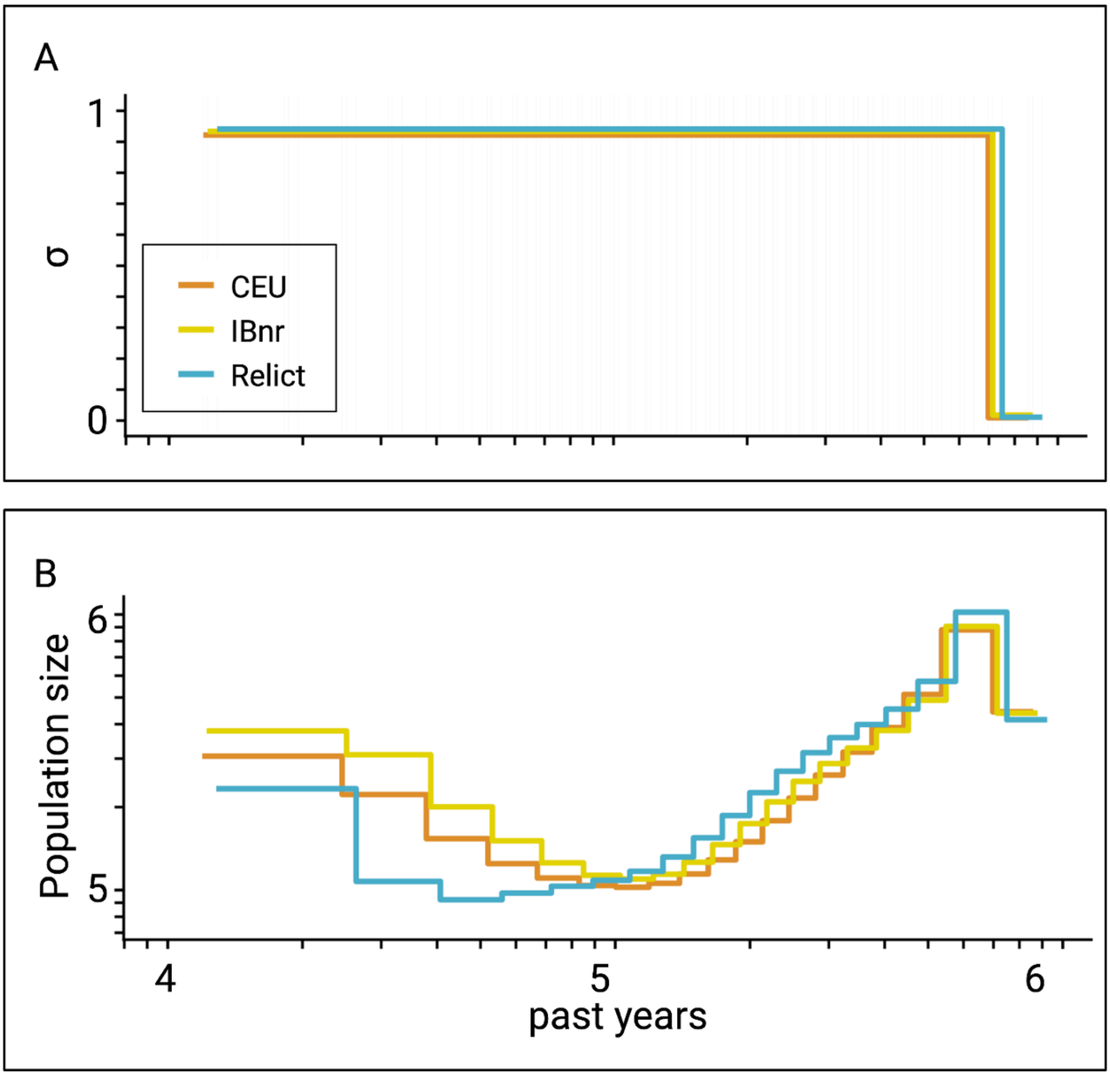


Figure 23. Inference of the time of transition from outcrossing to selfing in Arabidopsis thaliana. (A) Inferred transitions from outcrossing to selfing for three independent genetic clusters of Arabidopsis thaliana from the 1001 genomes project (CEU, IBnr, Relict) using *teSMC* under the one-transition mode. (B) Co-estimated population sizes over time with piecewise constant population size. Except for the selfing rates, both axes are scaled in log_{10}. Inference was done by Thibaut Sellinger.

The correct inference of demographies is critical to understanding evolutionary forces and functions of genes, which is a necessary consequence of the definition of function being the selected effect function (Graur, 2017; Graur, Zheng, & Azevedo, 2015; Graur et al., 2013; P. Brunet & Doolittle, 2014). Thus, the function of DNA, in general, is determined by its interaction with evolutionary forces. Not only are shifts in mating systems considered a critical evolutionary and ecological process, but also the correct inference of inbreeding consistently improves the correct demographic inference (**Figure 20**, **Figure 21**). By introducing *teSMC*, we provided a method to use genome-wide genetic variance to date transitions to selfing on small sample sets.

Our study demonstrated that *teSMC* correctly infers transitions to selfing for the tested time ranges. However, the parametrization of the *teSMC* model complicates the comparison of the likelihoods between the models and different modes. Usually, we obtain better likelihoods for the free mode, inferring transitions to selfing. Different information criteria usually give a penalty to an increased number of parameters, assuming that they are independent. Independency of parameters is not a given in the SMC-HMM method family. However, the estimated selfing history is robust, especially for recent up to intermediate times, making it an excellent method to infer recent shifts in mating systems. Note, assuming constant selfing provides only slightly wrong estimates of recent selfing rates if the transition is not recent (**Figure 20**, panels B and D; **Figure 21**, panels B, F, D, and H). Nevertheless, still, spurious artifacts occur in the inferred demography. We bound the maximum selfing rate to 0.99 because higher values practically do not allow for a change of hidden states when moving along the sequence, potentially biasing the inference of transition to selfing towards older dates.

Chapter 1 demonstrated the consequences of transitions to selfing on the genome-wide genetic structure. We identified the information of transitions to selfing, laying in the correlation of length and diversity of MRCA segments. The introduced *teSMC* is the first member of the SMC-HMM family decoding the position of recombination events under the correct model, i. e., piecewise-constant recombination rates through time, to implicitly identify MRCA segments, providing information about past changes in recombination rates, which we parametrize into transitions to selfing.

With *teSMC*, we jointly dated the transition to selfing combined with the demography inference. The estimates on the same three genetic clusters of *Arabidopsis thaliana* as used in chapter 2 agree with the estimates of *tsABC* (**Table 4**) but are slighly older than previous estimates from literature (see

Chapter 2). Given that two separate approaches agree with each other provides confidence in our estimates. Additionally, the estimated dates being older than any known split in between different *Arabidopsis thaliana* groups and the fact of its transition being older than the migration out of Africa (Durvasula et al., 2017) also raises the hypothesis that selfing contributed to enabling *Arabidopsis thaliana* to migrate commensal with modern humans and early hominins, that migrated out of Africa between 60 kyr to probably more than 400 kyr ago (Malaspinas et al., 2016; Nielsen et al., 2017).

In summary, we introduced *teSMC*, a novel inference method extending existing methods of the SMC-HMM family to estimate changes in recombination rates and selfing. Additionally, we tested the performance of *teSMC* to infer demography and transitions to selfing. Finally, we consistantly dated the transition to selfing of *Arabidopsis thaliana* using three distinct genetic clusters. Thus, we enable exploring the phylogeny of plants for transitions to selfing to contribute to the understanding of evolutionary processes shaping plant species and populations in the context of the evolution of sexes.

Conclusion

1. The new method *teSMC* infers transitions to selfing jointly with the inference of population sizes.
2. Estimates on three genetic clusters of *Arabidopsis thaliana* are slightly older than previously published estimates and dated the transition to selfing from 697,490 to 749,668 years ago.
3. The introduced *teSMC* will facilitate the phylogenetic exploration and identification of recent transitions to selfing to elaborate on existing hypotheses on the evolution of mating systems, e. g. the dead-end hypothesis of selfing species.

Author contributions

Prof. Dr. Sylvain Glémin (TAC member) provided the initial idea of extending the SMC framework to implement selfing. I elaborated the idea under the supervision of Dr. Stefan Laurent, and we initiated a collaboration with Thibaut Sellinger and Auréllien Tellier to develop the method *teSMC* as an extension of their previously published *eSMC* (Sellinger et al., 2020). Thibaut Sellinger formalized and wrote the sequentially Markovian coalescent to infer piecewise-constant selfing rates and piecewise-constant population sizes and implemented the method in the R package *teSMC*. The simulation of transitions to selfing, i. e., changes of recombination rates through time, was implemented using *msprime*, which only provided a usable interface for that in recent versions by using the "from_ts" argument of its simulation function. The performance analysis and application to *Arabidopsis thaliana* were conducted by myself with the help of Thibaut Sellinger.

General discussion

Outlook and implications of identifying and estimating transitions from outcrossing to selfing

Outlook and implications of identifying and estimating transitions from outcrossing to selfing

In this work, we developed and tested methods to identify and estimate transitions to selfing using haplotype genomic variation. We investigated and comprehensively described the consequences of changes in selfing rates on intra-specific genomic variability. We used forward-in-time Wright-Fisher models to explicitly simulate reproduction under selfing and the coalescent with partial selfing, which approximates implicitly selfing. We developed a novel measure of the effects of transitions to selfing on genetic variation based on our theoretical expectation. We provided evidence that the coalescent with partial selfing accurately models transitions to selfing. We developed an ABC and an MLE method to identify and estimate transitions from outcrossing to selfing based on these insights. We provided a complete performance analysis of both methods. Finally, we applied both methods to three distinct genetic clusters of *Arabidopsis thaliana*, consistently providing slightly older estimates than pre-existing estimates of *Arabidopsis thaliana's* transition to selfing.

Based on our insights and investigation of the genetic consequences of transitions to selfing, we introduced a summarization of genetic variance focusing on MRCA segments. Segment-based inference relates to haplotype-based inference methods that recently gained more importance with the rise of high throughput sequencing techniques that provide accessibility to whole genomes of natural populations for many species. Many extensions of our developed methods are possible. For example, *tsABC* could be extended to consider allotetraploid speciation or demographic histories with multiple populations. However, an extension of the *teSMC* to some aspects may seem simple but not necessarily helpful, e. g. *MSMC* with piecewise constant recombination rates. Nevertheless, the implementation of other statistics of the coalescent with partial selfing may contribute to the improvement of estimating shifts in mating systems or the evolution of recombination rates through time and increase the accuracy of estimates also for more ancestral time ranges.

We estimated that the transition to selfing for *Arabidopsis thaliana* occurred between 552,920 and 749,668 years ago. We assumed a generation time of one year per generation (Donohue, 2002). *Arabidopsis thaliana* is an annual plant. Seasonality enforces the assumed generation time. However, an average generation may be shorter in tropical to subtropical regions. We would expect an older estimate for the Relicts. The estimates of *teSMC* are consistent with that assumption, but the estimates of *tsABC* are not in accordance. Furthermore, we provided a single but successful application example in this study. Expanding our research to other populations and species will increase the potential to answer standing biological questions, e. g. the evolutionary relationship of the correlated polyploidy to selfing.

Altogether, this thesis paves the way to expand the research on the evolution of selfing to any sampled population and species. The developed and introduced two statistical methods to identify and estimate transitions to selfing from genome-wide genetic variance will help to explore not only the plant phylogeny on the evolution of mating systems. We expect to confirm and provide evidence for existing hypotheses and raise new ones.

References

Cited scientific literature

Cited scientific literature

Abbott, R. J., & Gomes, M. F. (1989). Population genetic structure and outcrossing rate of Arabidopsis thaliana (L.) Heynh. *Heredity, 62*(3), 411.

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., . . . Zhou, X. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell, 166*(2), 481-491. doi:https://doi.org/10.1016/j.cell.2016.05.063

Alvarez-Buylla, E. R., Benítez, M., Corvera-Poiré, A., Chaos Cador, A., de Folter, S., Gamboa de Buen, A., . . . Sánchez-Corrales, Y. E. (2010). Flower development. *The arabidopsis book, 8*, e0127-e0127. doi:10.1199/tab.0127

Arunkumar, R., Ness, R. W., Wright, S. I., & Barrett, S. C. (2015). The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics, 199*(3), 817-829. doi:10.1534/genetics.114.172809

Barrett, S. C., Arunkumar, R., & Wright, S. I. (2014). The demography and population genomics of evolutionary transitions to self-fertilization in plants. *Philos Trans R Soc Lond B Biol Sci, 369*(1648). doi:10.1098/rstb.2013.0344

Barrett, S. C., & Harder, L. D. (1996). Ecology and evolution of plant mating. *Trends in Ecology & Evolution, 11*(2), 73-79.

Barton, N. H., & Charlesworth, B. (1998). Why Sex and Recombination? *Science, 281*(5385), 1986-1990. doi:doi:10.1126/science.281.5385.1986

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics, 41*(1), 164-171.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics, 41*, 379-406.

Beaumont, M. A., & Rannala, B. (2004). The Bayesian revolution in genetics. *Nat Rev Genet, 5*(4), 251-261. doi:10.1038/nrg1318

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics, 162*(4), 2025-2035.

Bechsgaard, J. S., Castric, V., Charlesworth, D., Vekemans, X., & Schierup, M. H. (2006). The transition to self-compatibility in Arabidopsis thaliana and evolution within S-haplotypes over 10 Myr. *Mol Biol Evol, 23*(9), 1741-1750. doi:10.1093/molbev/msl042

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: making and mining the “gold standard” annotated reference plant genome. *genesis, 53*(8), 474-485.

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2020). Inferring the Demographic History of Inbred Species from Genome-Wide SNP Frequency Data. *Molecular biology and evolution, 37*(7), 2124-2136. doi:10.1093/molbev/msaa042

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach. *PLoS Genet, 12*(3), e1005877. doi:10.1371/journal.pgen.1005877

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association, 71*(356), 791-799. doi:10.1080/01621459.1976.10480949

Busch, J. W., & Delph, L. F. (2011). The relative importance of reproductive assurance and automatic selection as hypotheses for the evolution of self-fertilization. *Annals of Botany, 109*(3), 553-562. doi:10.1093/aob/mcr219

Charlesworth, B., & Charlesworth, D. (2010). Elements of evolutionary genetics.

Charlesworth, B., Morgan, M., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics, 134*(4), 1289-1303.

Charlesworth, D., & Vekemans, X. (2005). How and when did Arabidopsis thaliana become highly self-fertilising. *Bioessays, 27*(5), 472-476. doi:10.1002/bies.20231

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics, 6*(11), 836-846.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science, 5*(2), 169-174.

Cox, D. R. S. E. J. (1981). *Applied statistics : principles and examples*. London; New York: Chapman and Hall.

Csillery, K., Blum, M. G., Gaggiotti, O. E., & Francois, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol, 25*(7), 410-418. doi:10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution, 3*(3), 475-479. doi:10.1111/j.2041-210X.2011.00179.x

Cutter, A. D. (2019). Reproductive transitions in plants and animals: selfing syndrome, sexual selection, and speciation. *New Phytologist, 0*(ja). doi:10.1111/nph.16075

Darwin, C. (1876). The Effects of Cross and Self Fertilization in the Vegetable Kingdom. 1876. *New York, Appleton*.

de Nettancourt, D. (1997). Incompatibility in angiosperms. *Sexual Plant Reproduction, 10*(4), 185-199. doi:10.1007/s004970050087

Dobzhansky, T. (1973). Nothing in Biology Makes Sense except in the Light of Evolution. *The american biology teacher, 35*(3), 125-129. doi:10.2307/4444260

Dobzhansky, T. (2013). Nothing in biology makes sense except in the light of evolution. *The american biology teacher, 75*(2), 87-91.

Donohue, K. (2002). GERMINATION TIMING INFLUENCES NATURAL SELECTION ON LIFE-HISTORY CHARACTERS IN ARABIDOPSIS THALIANA. *Ecology, 83*(4), 1006-1016. doi:https://doi.org/10.1890/0012-9658(2002)083[1006:GTINSO]2.0.CO;2

dos Reis, M., Donoghue, P. C. J., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics, 17*(2), 71-80. doi:10.1038/nrg.2015.8

Durvasula, A., Fulgione, A., Gutaker, R. M., Alacakaptan, S. I., Flood, P. J., Neto, C., . . . Hancock, A. M. (2017). African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A, 114*(20), 5213-5218. doi:10.1073/pnas.1616736114

Epinat, G., & Lenormand, T. (2009). The evolution of assortative mating and selfing with in-and outbreeding depression. *Evolution: International Journal of Organic Evolution, 63*(8), 2047-2060.

Ewing, G. B., & Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Molecular Ecology, 25*(1), 135-141.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual review of genetics, 22*(1), 521-565.

Fisher, R. A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics, 11*(1), 53-63.

Fisher, R. A. (1958). *The genetical theory of natural selection*: Рипол Классик.

Franklin-Tong, V. E. (2008). Self-incompatibility in flowering plants. *Evolution, diversity, and mechanisms, 305.*

Gilbert, K. J., Pouyet, F., Excoffier, L., & Peischl, S. (2020). Transition from background selection to associative overdominance promotes diversity in regions of low recombination. *Current Biology, 30*(1), 101-107. e103.

Glémin, S. (2021). Balancing selection in self-fertilizing populations. *Evolution, n/a*(n/a). doi:https://doi.org/10.1111/evo.14194

Glémin, S., François, C. M., & Galtier, N. (2019). Genome Evolution in Outcrossing vs. Selfing vs. Asexual Species. In M. Anisimova (Ed.), *Evolutionary Genomics: Statistical and Computational Methods* (pp. 331-369). New York, NY: Springer New York.

Glémin, S., & Galtier, N. (2012). Genome Evolution in Outcrossing Versus Selfing Versus Asexual Species. In M. Anisimova (Ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 1* (pp. 311-335). Totowa, NJ: Humana Press.

Goodwillie, C., Kalisz, S., & Eckert, C. G. (2005). The evolutionary enigma of mixed mating systems in plants: occurrence, theoretical explanations, and empirical evidence. *Annu. Rev. Ecol. Evol. Syst., 36*, 47-79.

Graur, D. (2017). An Upper Limit on the Functional Fraction of the Human Genome. *Genome Biology and Evolution, 9*(7), 1880-1885. doi:10.1093/gbe/evx121

Graur, D., Zheng, Y., & Azevedo, R. B. (2015). An evolutionary classification of genomic function. *Genome Biology and Evolution, 7*(3), 642-645.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., & Elhaik, E. (2013). On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution, 5*(3), 578-590. doi:10.1093/gbe/evt028

Griffiths, R. C., & Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Stochastic Models, 14*(1-2), 273-295.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources, 19*(2), 552-566. doi:10.1111/1755-0998.12968

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular biology and evolution, 36*(3), 632-637. doi:10.1093/molbev/msy228

Hämälä, T., & Tiffin, P. (2020). Biased Gene Conversion Constrains Adaptation in &lt;em&gt;Arabidopsis thaliana&lt;/em&gt. *Genetics, 215*(3), 831. doi:10.1534/genetics.120.303335

Hartfield, M., Bataillon, T., & Glemin, S. (2017). The Evolutionary Interplay between Adaptation and Self-Fertilization. *Trends Genet, 33*(6), 420-431. doi:10.1016/j.tig.2017.04.002

Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome research, 13*(4), 635-643.

Hudson, R. R., & Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics, 141*(4), 1605-1617.

Igic, B., & Busch, J. W. (2013). Is self-fertilization an evolutionary dead end? *New Phytologist, 198*(2), 386-397. doi:10.1111/nph.12182

Jeffreys, H. (1998). *The theory of probability*: OUP Oxford.

Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., & Jensen, J. D. (2021). The impact of purifying and background selection on the inference of population history: problems and prospects. *bioRxiv*, 2020.2004.2028.066365. doi:10.1101/2020.04.28.066365

Kamran-Disfani, A., & Agrawal, A. F. (2014). Selfing, adaptation and background selection in finite populations. *J Evol Biol, 27*(7), 1360-1371. doi:10.1111/jeb.12343

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS computational biology, 12*(5), e1004842. doi:10.1371/journal.pcbi.1004842

Kingman, J. F. C. (1977). A note on multidimensional models of neutral mutation. *Theoretical population biology, 11*(3), 285-290. doi:https://doi.org/10.1016/0040-5809(77)90012-0

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications, 13*(3), 235-248. doi:https://doi.org/10.1016/0304-4149(82)90011-4

Kingman, J. F. C. (2000). Origins of the Coalescent: 1974-1982. *Genetics, 156*(4), 1461.

Leuenberger, C., & Wegmann, D. (2010). Bayesian Computation and Model Selection Without Likelihoods. *Genetics, 184*(1), 243. doi:10.1534/genetics.109.109058

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*(7357), 493-496. doi:10.1038/nature10231

Liu, P., Sherman-Broyles, S., Nasrallah, Mikhail E., & Nasrallah, J. B. (2007). A Cryptic Modifier Causing Transient Self-Incompatibility in Arabidopsis thaliana. *Current Biology, 17*(8), 734-740. doi:https://doi.org/10.1016/j.cub.2007.03.022

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics, 47*, 555. doi:10.1038/ng.3254

https://www.nature.com/articles/ng.3254#supplementary-information

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., . . . Crawford, J. E. (2016). A genomic history of Aboriginal Australia. *Nature, 538*(7624), 207-214.

Marchi, N., Schlichta, F., & Excoffier, L. (2021). Demographic inference. *Current Biology, 31*(6), R276-R279.

Marjoram, P., & Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genet, 7*, 16. doi:10.1186/1471-2156-7-16

Mattila, T. M., Laenen, B., & Slotte, T. (2020). Population Genomics of Transitions to Selfing in Brassicaceae Model Systems. *Methods Mol Biol, 2090*, 269-287. doi:10.1007/978-1-0716-0199-0_11

Mattila, T. M., Laenen, B., & Slotte, T. (2020). Population Genomics of Transitions to Selfing in Brassicaceae Model Systems. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (pp. 269-287). New York, NY: Springer US.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics, 30*(5), 1225-1310.

McVean, G. A., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci, 360*(1459), 1387-1393. doi:10.1098/rstb.2005.1673

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., & Grelon, M. (2015). The Molecular Biology of Meiosis in Plants. *Annual Review of Plant Biology, 66*(1), 297-327. doi:10.1146/annurev-arplant-050213-035923

Nasrallah, J. B. (2019). Chapter Sixteen - Self-incompatibility in the Brassicaceae: Regulation and mechanism of self-recognition. In U. Grossniklaus (Ed.), *Current Topics in Developmental Biology* (Vol. 131, pp. 435-452): Academic Press.

Nelson, D., Kelleher, J., Ragsdale, A., McVean, G., & Gravel, S. (2019). Coupling Wright-Fisher and coalescent dynamics for realistic simulation of population-scale datasets. *bioRxiv*, 674440.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature, 541*(7637), 302-310. doi:10.1038/nature21347

Nordborg, M. (1997). Structured coalescent processes on different time scales. *Genetics, 146*(4), 1501-1514.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics, 154*(2), 923-929.

Nordborg, M. (2001). Coalescent theory. *Handbook of statistical genetics*.

Nordborg, M., & Donnelly, P. (1997). The coalescent process with selfing. *Genetics, 146*(3), 1185-1195.

Ornduff, R. (1969). Reproductive biology in relation to systematics. *Taxon, 18*(2), 121-133.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., . . . Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science, 327*(5961), 92-94. doi:10.1126/science.1180677

Otto, S. P., & Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nat Rev Genet, 3*(4), 252-261. doi:10.1038/nrg761

Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics, 34*(1), 401-437.

P. Brunet, T. D., & Doolittle, W. F. (2014). Getting "function" right. *Proceedings of the National Academy of Sciences, 111*(33), E3365-E3365.

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife, 7*, e36317. doi:10.7554/eLife.36317

Ralph, P., Thornton, K., & Kelleher, J. (2020). Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics, 215*(3), 779-797. doi:10.1534/genetics.120.303253

Salomé, P., Bomblies, K., Fitz, J., Laitinen, R., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity, 108*(4), 447-455.

Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity, 108*(4), 447-455. doi:10.1038/hdy.2011.95

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet, 46*(8), 919-925. doi:10.1038/ng.3015

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics, 204*(3), 1207-1223.

Sellinger, T. P. P., Abu Awad, D., Moest, M., & Tellier, A. (2020). Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLoS Genetics, 16*(4), e1008698. doi:10.1371/journal.pgen.1008698

Sellinger, T. P. P., Abu-Awad, D., & Tellier, A. (2021). Limits and convergence properties of the sequentially Markovian coalescent. *Molecular Ecology Resources, 21*(7), 2231-2248.

Shimizu, K. K., & Tsuchimatsu, T. (2015). Evolution of selfing: recurrent patterns in molecular adaptation. *Annual review of ecology, evolution, and systematics, 46*.

Sicard, A., & Lenhard, M. (2011). The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Annals of Botany, 107*(9), 1433-1443.

Smith, J. M. (1971). What use is sex? *Journal of Theoretical Biology, 30*(2), 319-335.

Smith, J. M., & Maynard-Smith, J. (1978). *The evolution of sex* (Vol. 4): Cambridge University Press Cambridge.

Stebbins, G. L. (1957). Self Fertilization and Population Variability in the Higher Plants. *American Naturalist, 91*(861), 337-354. doi:Doi 10.1086/281999

Stebbins, G. L. (1974). *Flowering plants: evolution above the species level.* Retrieved from

Suwabe, K., Nagasaka, K., Windari, E. A., Hoshiai, C., Ota, T., Takada, M., . . . Watanabe, M. (2020). Double-Locking Mechanism of Self-Compatibility in Arabidopsis thaliana: The Synergistic Effect of Transcriptional Depression and Disruption of Coding Region in the Male Specificity Gene. *Frontiers in Plant Science, 11*(1430). doi:10.3389/fpls.2020.576140

Takayama, S., & Isogai, A. (2005). Self-incompatibility in plants. *Annu. Rev. Plant Biol., 56*, 467-489.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y. L., Hu, T. T., . . . Nordborg, M. (2007). The evolution of selfing in Arabidopsis thaliana. *Science, 317*(5841), 1070-1072. doi:10.1126/science.1143153

Underwood, C. J., Choi, K., Lambing, C., Zhao, X., Serra, H., Borges, F., . . . Henderson, I. R. (2018). Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome research, 28*(4), 519-531.

Upadhya, G., & Steinrücken, M. (2021). Inferring Population Size Histories using Coalescent Hidden Markov Models with &lt;em&gt;T&lt;/em&gt;&lt;sub&gt;MRCA&lt;/sub&gt; and Total Branch Length as Hidden States. *bioRxiv*, 2021.2005.2022.445274. doi:10.1101/2021.05.22.445274

Wakeley, J. (2009). Coalescent theory. *Roberts & Company.*

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics, 11*, 116. doi:10.1186/1471-2105-11-116

Weng, M.-L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M. T., Shaw, R. G., . . . Fenster, C. B. (2018). Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in Arabidopsis thaliana. *Genetics, 211*(2), 703-714. doi:10.1534/genetics.118.301721

Wilcock, C. (1987). AJ Richards 1986. Plant breeding systems. George Allen & Unwin, London. 529 pages. ISBN 0-04-581020-6 (hardback), 0-04-581021-4 (paperback). Price:£ 45.00 (hardback),£ 19.95 (paperback). *Journal of Tropical Ecology, 3*(3), 279-280.

Wright, S. (1931). EVOLUTION IN MENDELIAN POPULATIONS. *Genetics, 16*(2), 97-159. doi:10.1093/genetics/16.2.97

Wright, S. I., Kalisz, S., & Slotte, T. (2013). Evolutionary consequences of self-fertilization in plants. *Proc Biol Sci, 280*(1760), 20130133. doi:10.1098/rspb.2013.0133

Zuckerkandl, E., & Pauling, L. (1965). In Evolving Genes and Proteins, ed. by V. Bryson & HJ Vogel. In: New York: Academic Press.

Appendix

Formal description of *tsABC* and *teSMC,* table of figures and tables

Description of *teSMC*

This Appendix was provided by Thibaut Sellinger.

Stefan Strütt, Thibaut Sellinger, Aurélien Tellier, Stefan Laurent

17 mars 2022

1 teSMC

To define our Hidden Markov Model (HMM) we need to define :
- Hidden States
- The signal (observed data)
- A Transition matrix (Probability of jumping from one state to another)
- An Emission matrix (Probability of observing the data given the hidden state)
- An Initial probability (Probability of hidden states at the first position of the sequence)

1.1 Notations and Assumptions

We here define the different notations used and their meaning :
- σ_t : self fertilization rate (between 0 and 1) at time t
- β_t : self fertilization rate (between 0 and 1) at time t
- N_0 : Population size at present time
- r_t : recombination rate per nucleotide per $4N_0$ generation at time t
- μ : Mutation rate per nucleotide per $4N_0$ generation
- μ_b : ratio of mutation rate during the dormant stage over the mutation rate during the active stage per nucleotide per $4N_0$ generation
- u : time at which the recombination occurs (follows a pice-wise uniform distribution)
- L : Sequence length in bp
- N_t : Population size at time t
- χ_t : Scaling factor for the population size at time t ($N_t = \chi_t N_0$)

The model's assumptions are :
- Piecewise constant population size
- Piecewise constant selfing, germination and recombination rate in time
- Constant mutation rate in time
- Constant mutation and recombination rate along the sequence
- Neutrality

1.2 Hidden States

We define our hidden states at one position on the genome as the coalescent time between the two individual at that position. We note that coalescent time t (t>0). A transition from a coalescent time s to time t ($t \neq s$) at the next can only occur if a recombination happened in between the two positions.

1.3 Observations

Our observations, or the signal, is a sequence of 1 and 0. This sequence is build from phasing the DNA sequences of two individual. When going along the sequence, if both nucleotide are similar, then the signal is 0 (no mutation occurred). If both are different, then a mutation occurred, and the signal is 1.

1.4 Transition Matrix

A transition to state t from state s (t $\neq$ s) can only occur if there is a recombination event. Assuming Recombination event on the tree as a Poisson process we have the probability of a recombination :

$$P(rec|s) = (1 - e^{-\int_0^s \frac{2(1-\sigma_k)\beta_k}{2-\sigma_k} 2r_k dk}) \tag{1}$$

We now Assume that a recombination event occurred at time u (<s) where u follows a piecewise uniform distribution (*i.e.* uniform in each hidden state but the density between hidden state is allowed to change) between 0 and s. Then three scenarios are possible. Either the new coalescent time is smaller (t<s),bigger (t>s) or unchanged (t=s).

1.4.1 t<s

The resulting floating branch of the recombination event coalesces at time t < s . This mean it must not coalesce before time t (including itself). In addition we have u<t. The transition probability is therefore :

$$P(t|s,u) = \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv}) \tag{2}$$

1.4.2 t=s

The resulting floating branch of the recombination event self coalesce before time t. We therefore have the transition probability :

$$P(s|s,u) = \int_u^s \frac{2\beta_k^2}{(2-\sigma_k)\chi_k} e^{\int_u^k -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} dk \tag{3}$$

1.4.3 t>s

The resulting floating branch of the recombination event must not coalesce (including itself) before time s. Then no coalescent event must happen before time t. We therefore have the transition probability :

$$P(t|s,u) = \frac{2\beta_t^2}{(2-\sigma_t)\chi_t} e^{\int_u^s -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_s^t -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} \tag{4}$$

1.4.4 Transition probability in continuous time

In the end we have :

$$p(t|s,u) = \begin{cases} (1 - e^{-\int_0^s \frac{2(1-\sigma_k)\beta_k}{2-\sigma_k} 2r_k dk}) \frac{2\beta_t^2}{(2-\sigma_t)\chi_t} (e^{\int_u^t -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv}) & if \quad u < t < s \\ e^{-\int_0^s \frac{2(1-\sigma_k)\beta_k}{2-\sigma_k} 2r_k dk} + (1 - e^{-\int_0^s \frac{2(1-\sigma_k)\beta_k}{2-\sigma_k} 2r_k dk}) \int_u^s \frac{2\beta_k^2}{(2-\sigma_k)\chi_k} e^{\int_u^k -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} dk & if \quad t = s \\ (1 - e^{-\int_0^s \frac{2(1-\sigma_k)\beta_k}{2-\sigma_k} 2r_k dk}) \frac{2\beta_t^2}{(2-\sigma_t)\chi_t} e^{\int_u^s -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_s^t -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} & if \quad t > s \\ 0 & if \quad otherwise \end{cases} \tag{5}$$

Once again, if all $\sigma_k = 0, \beta_k = 1$, we fall back on the probability from PSMC'.
One can find $p(t|s)$ using the total probability formula which is :

$$p(t|s) = \int_0^s p(u)p(t|s,u)du \tag{6}$$

As explained before, the state space must be finite. We therefore discretized time in n intervals. At one point the hidden state is α if $t \in [T_\alpha, T_{\alpha+1}]$, where $\alpha \in [0, (n-1)]$. We define T_α :

$$T_\alpha = -\frac{(2-\sigma_0)}{2\beta_0^2} \ln(1 - \frac{\alpha}{n}) \tag{7}$$

We therefore have :

$$p(\alpha|s) = \int_{T_\alpha}^{T_{\alpha+1}} p(t|s)dt \tag{8}$$

The transition matrix need to be the probability from one state to another. Therefore we need the probability when the coalescent time at the previous position (which is here s) belongs to the state γ. To do this we simply replace s by the average coalescent time t_γ.

1.4.5 Initial Probability

We use the equilibrium probability as initial probability. The equilibrium probability is the probability that the first coalescent happens in each time interval and is thus given by :

$$\begin{aligned}
q_o(\alpha) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha} e^{\int_0^t \frac{-2\beta_v^2}{(2-\sigma_v)\chi_v}dv} dt \\
q_o(\alpha) &= \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha} e^{\int_0^{T_\alpha} \frac{-2\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{T_\alpha}^t \frac{-2\beta_v^2}{(2-\sigma_v)\chi_v}dv} dt \\
q_o(\alpha) &= e^{\int_0^{T_\alpha} \frac{-2\beta_v^2}{(2-\sigma_v)\chi_v}dv} \int_{T_\alpha}^{T_{\alpha+1}} \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha} e^{\frac{-2(t-T_\alpha)\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}} dt \\
q_o(\alpha) &= e^{\sum_{\eta=0}^{\alpha-1} \frac{-2\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}\Delta_\eta} (1 - e^{\frac{-2\Delta_\alpha\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})
\end{aligned} \tag{9}$$

1.4.6 Calculation of t_γ

$$\begin{aligned}
t_\gamma = E[\text{Coalescent time}|\gamma] &= \frac{E[\text{Coalescent time} \cap \gamma]}{P(\gamma)} = \frac{\int_{T_\gamma}^{T_{\gamma+1}} t\Lambda_\gamma e^{-\int_0^t \Lambda_v dv} dt}{q_0(\gamma)} \\
&= \frac{\Lambda_\gamma \int_{T_\gamma}^{T_{\gamma+1}} te^{-\int_0^{T_\gamma} \Lambda_v dv} e^{-\int_{T_\gamma}^t \Lambda_v dv} dt}{q_0(\gamma)} = \frac{\Lambda_\gamma e^{-\int_0^{T_\gamma} \Lambda_v dv} \int_{T_\gamma}^{T_{\gamma+1}} te^{-\int_{T_\gamma}^t \Lambda_v dv} dt}{q_0(\gamma)} \\
&= \frac{\Lambda_\gamma \int_{T_\gamma}^{T_{\gamma+1}} te^{(T_\gamma - t)\Lambda_\gamma} dt}{(1 - e^{-\Delta_\gamma\Lambda_\gamma})} = \frac{T_\gamma - T_{\gamma+1}e^{-\Delta_\gamma\Lambda_\gamma}}{(1 - e^{-\Delta_\gamma\Lambda_\gamma})} + \frac{\int_{T_\gamma}^{T_{\gamma+1}} e^{(T_\gamma - t)\Lambda_\gamma} dt}{(1 - e^{-\Delta_\gamma\Lambda_\gamma})} \\
&= \frac{T_\gamma - T_{\gamma+1}e^{-\Delta_\gamma\Lambda_\gamma}}{(1 - e^{-\Delta_\gamma\Lambda_\gamma})} + \frac{(1 - e^{-\Delta_\gamma\Lambda_\gamma})}{\Lambda_\gamma(1 - e^{-\Delta_\gamma\Lambda_\gamma})} = \frac{T_\gamma - T_{\gamma+1}e^{-\Delta_\gamma\Lambda_\gamma}}{(1 - e^{-\Delta_\gamma\Lambda_\gamma})} + \frac{1}{\Lambda_\gamma}
\end{aligned} \tag{10}$$

Where :

$$\begin{aligned}
\Delta_\gamma &= T_{\gamma+1} - T_\gamma \\
\Lambda_\gamma &= \frac{2\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}
\end{aligned} \tag{11}$$

1.4.7 Calculation of $p(\alpha|\gamma)$

$\alpha < \gamma$ We first need $p(t|t_\gamma)$ when $\alpha < \gamma$, which is obtained as described below :

$$\begin{aligned}
p(t|t_\gamma) &= P_\gamma \int_0^t \frac{\pi_u \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du \\
&= P_\gamma(\sum_{\eta=1}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{\pi_u \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du \\
&\quad + \int_{T_\alpha}^t \frac{\pi_u \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du) \\
&= P_\gamma(\sum_{\eta=1}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{\pi_\eta \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du \\
&\quad + \int_{T_\alpha}^t \frac{\pi_\alpha \frac{2\beta_t^2}{(2-\sigma_t)\chi_t}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du) \\
&= P_\gamma(\sum_{\eta=1}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{\pi_\eta \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(e^{\int_u^{T_{\eta+1}} - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})(e^{\int_{T_{\eta+1}}^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du \\
&\quad + \int_{T_\alpha}^t \frac{\pi_\alpha \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(e^{\int_u^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du) \\
&= P_\gamma(\sum_{\eta=1}^{\alpha-1} \int_{T_\eta}^{T_{\eta+1}} \frac{\pi_\eta \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(e^{-(T_{\eta+1}-u)\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})(e^{\int_{T_{\eta+1}}^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\Pi_\gamma} du \\
&\quad + \int_{T_\alpha}^t \frac{\pi_\alpha \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(e^{-(t-u)\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\Pi_\gamma} du) \\
&= P_\gamma(\sum_{\eta=1}^{\alpha-1} \frac{\pi_\eta \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(1 - e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})(e^{\int_{T_{\eta+1}}^t - \frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}\Pi_\gamma} \\
&\quad + \frac{\pi_\alpha \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(1 - e^{-(t-T_\alpha)\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}\Pi_\gamma}) \\
P_\gamma &= (1 - e^{-(\sum_{\xi=1}^{\gamma-1} \frac{2(1-\sigma_\xi)\beta_\xi}{2-\sigma_\xi} 2r_\xi \Delta_\xi + \frac{(t_\gamma - T_\gamma)2r_\gamma\beta_\gamma 2(1-\sigma_\gamma)}{(2-\sigma_\gamma)})})
\end{aligned} \tag{12}$$

Where :

$$\begin{aligned}
\pi_u &= (\frac{r_u \beta_u 2(1-\sigma_u)}{(2-\sigma_u)}) \\
\Pi_\gamma &= (\sum_{\xi=1}^{\gamma-1} \frac{\Delta_\xi r_\xi \beta_\xi 2(1-\sigma_\xi)}{(2-\sigma_\xi)} + \frac{(t_\gamma - T_\gamma) r_\gamma \beta_\gamma 2(1-\sigma_\gamma)}{(2-\sigma_\gamma)}) \\
&= (\sum_{\xi=1}^{\gamma-1} \Delta_\xi \pi_\xi + (t_\gamma - T_\gamma)\pi_\gamma)
\end{aligned} \tag{13}$$

We can now calculate $p(\alpha|\gamma)$.

$$\begin{aligned}
p(\alpha|\gamma) = \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma 2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(\sum_{\eta=1}^{\alpha-1} \frac{\pi_\eta(1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})(e^{\int_{T_{\eta+1}}^{t} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}\Pi_\gamma} \\
+\frac{\pi_\alpha(1-e^{-(t-T_\alpha)\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}\Pi_\gamma})dt \\
= \frac{P_\gamma 2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}(\sum_{\eta=1}^{\alpha-1} \frac{\pi_\eta(1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})(e^{\int_{T_{\eta+1}}^{T_\alpha} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv})(1-e^{-\Delta_\alpha \frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}\Pi_\gamma} \\
+\frac{\pi_\alpha(\Delta_\alpha - \frac{(1-e^{-\Delta_\alpha \frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})}{\frac{4\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}\Pi_\gamma})
\end{aligned} \tag{14}$$

$\alpha > \gamma$ We first need $p(t|t_\gamma)$ when $\alpha > \gamma$, which is obtained as described below :

$$\begin{aligned}
p(t|t_\gamma) = \int_0^{t_\gamma} \frac{P_\gamma \pi_u \frac{2\beta_t^2}{(2-\sigma_t)\chi_t} e^{\int_u^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv}}{\Pi_\gamma} du \\
= \frac{P_\gamma \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}{\Pi_\gamma}(\sum_{\eta=0}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} \pi_u e^{\int_u^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} du \\
+ \int_{T_\gamma}^{t_\gamma} \pi_u e^{\int_u^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} du) \\
= \frac{P_\gamma \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}{\Pi_\gamma}(\sum_{\eta=0}^{\gamma-1} \int_{T_\eta}^{T_{\eta+1}} \pi_\eta e^{-(T_{\eta+1}-u)\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} du \\
+ \int_{T_\gamma}^{t_\gamma} \pi_\gamma e^{-(t_\gamma-u)\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} du) \\
= \frac{P_\gamma \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}{\Pi_\gamma}(\sum_{\eta=0}^{\gamma-1} \pi_\eta \frac{(1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v}dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv} \\
+\pi_\gamma \frac{(1-e^{-(t_\gamma-T_\gamma)\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}})}{\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v}dv})
\end{aligned} \tag{15}$$

We can now calculate $p(\alpha|\gamma)$.

$$\begin{aligned}
p(\alpha|\gamma) = \int_{T_\alpha}^{T_{\alpha+1}} \frac{P_\gamma \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}{\Pi_\gamma} (\sum_{\eta=0}^{\gamma-1} \pi_\eta \frac{(1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v} dv} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v} dv} \\
+\pi_\gamma \frac{(1-e^{-(t_\gamma - T_\gamma)\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}})}{\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}} e^{\int_{t_\gamma}^{t} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v} dv})dt \\
= \frac{P_\gamma \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}{\Pi_\gamma} (\sum_{\eta=0}^{\gamma-1} \frac{\pi_\eta (1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v} dv} + \frac{\pi_\gamma (1-e^{-(t_\gamma - T_\gamma)\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}})}{\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}}) \\
\int_{T_\alpha}^{T_{\alpha+1}} e^{\int_{t_\gamma}^{T_\alpha} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v} dv} e^{-(t-T_\alpha)\frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}})dt \\
= \frac{P_\gamma}{\Pi_\gamma} (\sum_{\eta=0}^{\gamma-1} \frac{\pi_\eta (1-e^{-\Delta_\eta \frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}})}{\frac{4\beta_\eta^2}{(2-\sigma_\eta)\chi_\eta}} e^{\int_{T_{\eta+1}}^{t_\gamma} -\frac{4\beta_v^2}{(2-\sigma_v)\chi_v} dv} + \frac{\pi_\gamma (1-e^{-(t_\gamma - T_\gamma)\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}})}{\frac{4\beta_\gamma^2}{(2-\sigma_\gamma)\chi_\gamma}}) e^{\int_{t_\gamma}^{T_\alpha} -\frac{2\beta_v^2}{(2-\sigma_v)\chi_v} dv} (1-e^{-\Delta_\alpha \frac{2\beta_\alpha^2}{(2-\sigma_\alpha)\chi_\alpha}}))
\end{aligned} \tag{16}$$

$\alpha = \gamma$ Because probabilities sum up to one. We have the following formula :

$$p(\gamma|\gamma) = 1 - (\sum_{\alpha=0}^{\gamma-1} p(\alpha|\gamma) + \sum_{\alpha=\gamma+1}^{n} p(\alpha|\gamma)) \tag{17}$$

1.5 Emission Matrix

Because of seed banking,the coalescent tree can be very big. In this case the infinite site model hypothesis might be violated,therefore we have the following formula :

$$\begin{aligned}
P(0|\gamma) = e^{-2\mu(((\beta_\gamma + ((1-\beta_\gamma)\mu_b))(t_\gamma - Tc_\gamma)) + \sum_\eta^{\gamma-1}((\beta_\eta + ((1-\beta_\eta)\mu_b))\Delta_\eta))} \\
P(1|\gamma) = 1 - P(0|\gamma)
\end{aligned} \tag{18}$$

Where μ is the mutation rate per nucleotide per N generation, μ_b le ratio of mutation rate during the dormant stage over the one in the active stage, β_γ the germination rate in state γ and $t\gamma$ the average coalescent time in state γ.

Table of figures

Figure 1. Two population scenarios to investigate the genetic consequences of transitions to selfing: (A) A single population of constant size undergoes a transition from predominant outcrossing to predominant selfing; (B) a single population with constant selfing undergoes a single change of population size according to the rescaled effective population sizes of model 1 in the predominant outcrossing and selfing phase, respectively. 16

Figure 2. The observed ratio of effective recombination rates of simulated populations under Model 1 in WF simulations: Measured rescaling for $\mu r = 0.1$ (A, C) and $\mu r = 10$ (B, D) under the Forward-explicit WF model (A, B) and the Forward-rescaled WF model with partial selfing (C, D). Simulations were performed for two populations of sizes, 500 (red) and 2000 (blue), and three chromosome lengths, 107 (circle), $5 \cdot 107$ (triangle), $1.2 \cdot 108$ bp (square). Results are shown for 12 independent simulations. We counted effective recombination events before and after the transition to selfing and calculated the ratio using equation (9)..22

Figure 3. Measured inbreeding factors FIS as one minus the ratio between observed and expected heterozygosity for different sampling proportions from a simulated WF population with constant selfing (sigma = 0.5, left; sigma = 0.95, right). Simulations were done using *SliM3*. Population size was set to $Ne = 10{,}000$, and mutation and recombination rates were set to $\mu = r = 10 - 8$. After 100,000 generations, samples were taken and repeated every 10,000 generations 30 times. For sample proportion of 0.5 or lower, haplotypes were sampled from different individuals. 24

Figure 4. Comparison of the joint distributions of T_{MRCA} and lengths of MRCA segments (TL) under three different simulation approaches. (A) Explicit selfing is implemented in a forward-in-time Wright-Fisher model (*SliM3*). Population size is constant, and the selfing rate changes from outcrossing ($\sigma = 0$, green) to predominant selfing ($\sigma = 0.95$, orange). MRCA segments were defined as contiguous sets of nucleotides sharing the same most recent common ancestor. (B) Shift to selfing is simulated using a forward-in-time Wright-Fisher model

(*SliM3*) by rescaling population size and recombination rate at $t\sigma$ as suggested by M. Nordborg and Donnelly (1997) (C). Shift to selfing simulated using the coalescent by rescaling population size and recombination rate at $t\sigma$ as in panel B. Except for the selfing rates, both axes are scaled in $log10$. 25

Figure 5. Consequences of a transition to selfing on the genealogies of simulated chromosomes over time. (A-I) Joint and marginal distributions of ages in generations and lengths of MRCA segments (TL) in a population with constant population size and a shift from outcrossing (green) to predominant selfing (orange). MRCA segments were defined as contiguous sets of nucleotides sharing the same most recent common ancestor. Except for the selfing rates, both axes are scaled in $log10$. 26

Figure 6. Consequences of a transition to selfing on genealogies of simulated chromosomes. (A) Joint and marginal distributions of ages (T_{MRCA} in generations on a $log10$ scale) and lengths of MRCA segments (in bp on a $log10$ scale) in a selfing population ($\sigma = 0.95$) with a stepwise change from large (green, $N\text{ANC} = 50{,}000$) to low (orange, $N\text{PRES} = 26{,}250$) population size. The population sizes were chosen to correspond to the rescaling of the effective population size by the selfing rates used in panel B. (B) Distribution of ages (T_{MRCA}) and lengths of MRCA segments (in bp) in a population with a constant population size and a shift from outcrossing (green, $\sigma = 0$) to predominant selfing (orange, $\sigma = 0.95$). (B) (C) Spatial distribution along the genome of a subset of MRCA segments (D) The transition matrix of ages (T_{MRCA}) between adjacent segments along the genome corresponding to the data simulated in panel A. This matrix summarizes the probability that the n^th^ MRCA segment with a given age X is followed by the (n+1)^th^ segment of age Y. The heat colors indicate the transition probabilities (tp). (E) The transition matrix of ages (T_{MRCA}) between adjacent segments along the genome corresponds to the data simulated in panel B. Recombination rate for the simulations was set to $r = 3.6 \cdot 10 - 9$. T_{MRCA}- and Length-axis are scaled in $log10$. 28

Figure 7. (A) The transition matrix of pairwise diversity (TM_{WIN}) between adjacent non-overlapping 10 kb windows measured for 1 Mb of data simulated with the population-size-change-model (**Figure 1**) in an outcrossing population with a stepwise change from N_{ANC} = 100,000 (green) to low N_{PRES}=50,500 (orange). The population sizes were chosen to correspond to the rescaling of the effective population size by the selfing rates used in panel B. (B) The same transition matrix of pairwise diversity as in panel A for a constant population ($N = 100{,}000$) but with a transition from outcrossing ($\sigma = 0$) to predominant selfing ($\sigma = 0.99$). The recombination rate was set to $\mu = 1 \cdot 10 - 8$. These matrices summarize the probabilities that the n^th^ window with a given diversity X is followed by the (n+1)^th^ window of diversity Y. The heat colors indicate the transition probabilities (tp). The demographic model under the simulations for A, a potential confounding model, captures the signal of the rescaled diversity by a transition to selfing, but not the joint rescaling of the recombination rate. ... 30

Figure 8. Explicit (green) and implicit (orange) selfing in comparison (A) Site-frequency spectra under two different selfing rates at two time points after a transition to selfing, no significance level ($p > 0.05$) (B) Linkage disequilibrium of discretized distances under two different selfing rates at two time points after a transition to selfing. (C) The number of segregating sites under the same conditions. There was no significance between the models, no significance level ($p > 0.05$). (D) The joint measure of T_{MRCA} (in generations) and length of MRCA segments (TL) for two different selfing rates at three time points after the transition to selfing, the dotted line marks the timing of the transition to selfing (E) Number of according segments (compare to panel D), no significance ($p > 0.05$) was observed. Population size was set to $Ne = 50{,}000$; $t\sigma$ was measured in past generations. Significance levels ns ($p > 0.05$), * ($p < 0.05$). 33

Figure 9. Three population scenarios to identify and estimate transitions to selfing: (A) Model 1: A single popu-lation of constant size undergoes a transition from predominant out-crossing to predominant selfing; (B) A single population with constant sel-fing undergoes a single change of population. Model B is a confounding model to model 1 because it poten-tially explains reductions in diversity. (C) A single population undergoes a transition to sel-fing and, additionally, at an independent time point a single change in population size 43

Figure 10. ABC model choice and parameter estimate performance analysis. (A) Demographic model 1 in the model choice analysis: one population with a single transition from predominant selfing to predominant outcrossing (B) Demographic model 2 in the model choice analysis: one population with constant selfing and a single change in population size. (A, B) The parameters of interest are the population sizes ($NPRES$, $NANC$), the selfing rates (σANC, $\sigma PRES$), and the time of change in selfing rate and population size ($t\sigma$, tN). Model 2, a potential confounding model, captures the signal of the rescaled diversity by a transition to selfing, but not the joint rescaling of the recombination rate. (C-E) Performance of the ABC model choice method using three different summarizations of data. C: Combining site frequency spectrum (SFS) and linkage disequilibrium (LD). (D) Window-based transition matrix (TMWIN). E: The combination out of SFS, LD, and TMWIN. The x-axis represents the range of used $t\sigma$ values; the y-axis indicates the proportion (out of 100 trials) that the ABC correctly identified the transition-to-selfing model among the two models presented in (A). (F-H): Parameter estimation accuracy for the age of a transition to selfing (100 simulated datasets) under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average interpercentile ranges for 100 posterior distributions corresponding. $t\sigma$-axes are scaled in $log10$. 53

Figure 11. ABC performance analysis: Parameter re-estimation of the three remaining parameters of the model described in **Figure 10**. (A-C) Re-estimation of the population size on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (D-F) Re-estimation of the present selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (G-I) Re-estimation of the ancestral selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average interpercentile ranges for 100 posterior distributions corresponding. Except for the selfing rates, both axes are scaled in $log10$. 55

Figure 12. Model choice experiment for different rho-theta-ratios and negative linked selection. The performance analysis for the different conditions was conducted the same way, as shown in **Figure 10**. (A-C) Model choice performance for increased recombination on otherwise same parameters. (D-F) Model choice performance under the same parameters as **Figure 10**, but PODs were simulated under negative linked selection. $t\sigma$-axes are scaled in $log10$. ... 57

Figure 13. Accuracy of *tsABC* in the presence of background selection (BGS): Inference of times of transition from outcrossing ($\sigma = 0.1$) to predominant-selfing ($\sigma = 0.99$) using *tsABC* using (A) SFS/LD, (B) $\mathrm{TM}_{\mathrm{WIN}}$ or (C) both. Simulations were done under constant population size and negative selection acting on exonic sequences. The spatial distribution of exonic sequences was fixed and taken from the annotation of *Arabidopsis thaliana*. The negative selection was modeled using a distribution of fitness effects (see methods). The result should be compared with the case without linked negative selection in **Figure 10** (panel F-H). Colored lines represent the interpercentile ranges quantiles for 100 posterior distributions obtained with *tsABC*. Both axes are scaled in $log10$. 58

Figure 14. ABC performance analysis under negative linked selection: Parameter re-estimation of the three remaining parameters of the model described in **Figure 10**. (A-C) Re-estimation of the population size on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (D-F) Re-estimation of the present selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average quantiles for 100 posterior distributions corresponding to the given credible intervals. (G-I) Re-estimation of the ancestral selfing rate on 100 datasets simulated under a model with constant population size ($N = 40{,}000$) and a change in selfing rate from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$. Colored lines represent the average interpercentile ranges for 100 posterior distributions corresponding to the given credible intervals. Except for the selfing rates, both axes are scaled in $log10$. 59

Figure 15. Inference of the time of transition from outcrossing to selfing in *Arabidopsis thaliana* using *tsABC*: A: Inferred transitions from outcrossing to selfing for three independent genetic clusters of *Arabidopsis thaliana* from the 1001 genomes project (CEU, Ibnr, Relicts). (B) Co-estimated population sizes over time with a single population change. Except for the selfing rates, both axes are scaled in $log10$. 61

Figure 16. The sequentially Markovian coalescent process for a sample of size two considers recombination events of three different types in the *SMC'* algorithm and the *PSMC'*-based demographic inference methods. Recombination events potentially cause changes to the genealogy of a sample of size two. (A) The T_{MRCA} increases if the recombining lineage re-coalesces at a time older than the T_{MRCA} of the current MRCA, (B) it remains the same if the lineage re-coalesces to itself before the T_{MRCA} of the current MRCA, or (C) it decreases if the lineage re-coalesces to any other lineage but itself before the T_{MRCA} of the current MRCA. The probability of recombining between two loci depends on the integrated recombination rate over time on the genealogy of the current locus. The probability of coalescing follows the assumed coalescent framework. 72

Figure 17. Theoretical convergence of *teSMC* under complex demography. Best-case convergence of *teSMC* using ten sequences (i.e., haploid genomes) of 100 Mb (green) when the population undergoes a bottleneck (true sizes are indicated in black) with either variation of selfing in time (A, C) or variation of recombination rate in time (B, D). The selfing rate through time is represented in A), and the corresponding estimated population size is represented in C), the estimated recombination rate through time (B), and the corresponding population size in D). The recombination rate was set to $r = 1 \cdot 10 - 7$ and the mutation rate to $\mu =$

$1 \cdot 10 - 8$ per generation per bp. Except for the selfing rates, both axes are scaled in $log10$. Simulations and raw data were provided by Thibaut Sellinger. 75

Figure 18. Best-case convergence of *teSMC* for a different amount of data. Best-case convergence of *teSMC* using different combinations of sample sizes ($n = 2$, $n = 5$, or $n = 20$ sequences; i.e., haploid genomes) and sequence lengths ($L = 10\ Mb$ or$L = 100\ Mb$), when population size is constant (N=100,000, black line) with a constant selfing rate of 0.9. The best-case convergence is estimated assuming that selfing is constant (A, C) or varying in time (B, D). The estimated population size assuming constant selfing in time is represented in (C) and the simultaneously estimated selfing rate in (A). The estimated population size assuming varying selfing rate in time is represented in (D) and the simultaneously estimated selfing rate through time in (B). The recombination was set to $r = 1 \cdot 10 - 8$ per generation per bp. Except for the selfing rates, both axes are scaled in $log10$. Simulations and raw data were provided by Thibaut Sellinger. 76

Figure 19. Performance of *teSMC* on simulated polymorphism data. Inference of times of transition from outcrossing ($\sigma = 0.1$) to predominantly selfing ($\sigma = 0.99$) using neutral simulations. The inference was made using the free mode (yellow) and the one-transition mode (green) of *teSMC* and ten replicates per time point. (A) Under constant population size. (B-E): simulations were done with an additional change in population size; the vertical grey line indicates the change in population size. (B-C) From NANC = 200,000 to NPRES = 40,000 (population crash) at 10,000 generations (B) or 40,000 generations (C) in the past. D-E) From NANC = 40,000 to NPRES = 200,000 (population expansion) at 10,000 generations (D) or 40,000 generations (E) in the past. Both axes are scaled in $log10$. Inference was done by Thibaut Selllinger. 78

Figure 20. Inference of population sizes when transitions to selfing are not accounted for. Comparisons between true (black lines) and estimated selfing rates and population sizes estimated by *teSMC* for ten replicates. Here simulations were done using a constant population size ($N = 40{,}000$) and a transition to selfing from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$ at $t\sigma = 10{,}000$ (A,C) and 40,000 (B,D). Five chromosomes of 1 Mb were simulated with mutation and recombination rates set to $\mu = r = 1 \cdot 10 - 8$ events per generation per bp. Red and green lines indicate results obtained assuming the wrong model (i.e., constant selfing) and the correct model (i.e., single-transition). For the selfing rates (A, B), results for each replicate are indicated with solid lines. For the population sizes, the ten replicates were summarized by the green and red shaded areas, where the width of the shaded area corresponds to the range between the minimum and maximum value observed across replicates. Except for the selfing rates, both axes are scaled in $log10$.. 80

Figure 21. Inference of population sizes and selfing rates estimated by *teSMC* when both parameters change over time. (A-P): Comparisons between true (black lines) and estimated selfing rates and population sizes estimated by *teSMC* for ten replicates. Here simulations were done as in **Figure 20** except for the addition of a single stepwise population size expansion forward-in-time (first and second rows) or contraction (third and fourth row). The transition to selfing occurred from $\sigma ANC = 0.1$ to $\sigma PRES = 0.99$ at $t\sigma = 10{,}000$ (A, C, E, G, I, K, M, O; first and third column) and $t\sigma = 40{,}000$ (B, D, F, H, J, L, N, P; second and fourth column). For the population sizes, the ten replicates were summarized by the green and red shaded areas, where the width of the shaded area corresponds to the range between the minimum and maximum value observed across replicates. Red and green lines indicate results obtained assuming the wrong model (i.e., constant selfing) and the correct model (i.e., single-transition). For the selfing rates, results for each replicate are indicated with solid lines. Except for the selfing rates, both axes are scaled in $log10$. 81

Figure 22. Accuracy of *teSMC* in the presence of background selection (BGS). Inference of times of transition from outcrossing ($\sigma = 0.1$) to predominant-selfing ($\sigma = 0.99$) using *teSMC*. Simulations were done under constant population size and negative selection acting on exonic sequences. The spatial distribution of exonic sequences was fixed and taken from the annotation of *Arabidopsis thaliana*. The negative selection was modeled using a distribution of fitness effects (see Chapter 2). Comparison between simulated values of $t\sigma$ and estimates obtained with *teSMC* using the one-transition mode. Estimations were conducted with and without masking exonic sequences subject to negative selection. Both axes are scaled in $log10$. Inference was done by Thibaut Sellinger. 83

Figure 23. Inference of the time of transition from outcrossing to selfing in Arabidopsis thaliana. (A) Inferred transitions from outcrossing to selfing for three independent genetic clusters of Arabidopsis thaliana from the 1001 genomes project (CEU, IBnr, Relict) using *teSMC* under the one-transition mode. (B) Co-estimated population sizes over time with piecewise constant population size. Except for the selfing rates, both axes are scaled in $log10$. Inference was done by Thibaut Sellinger.. 85

Table of tables

Table 1. Parameter priors used for the performance analysis of tsABC 46

Table 2. Parameters for PODs 47

Table 3. Parameter priors used to estimate the transition from outcrossing to selfing in Arabidopsis thaliana 47

Table 4. Estimated times of transitions from predominant outcrossing to predominant selfing in Arabidopsis thaliana. The demography was estimated for three non-admixed genetic clusters (genetic assignment > 95%) of the 1001 genomes project. *tsABC* estimated the time of a transition from predominant outcrossing to predominant selfing on the entire sample set (n_{CEU}=99, n_{IBnr}=66, $n_{Relicts}$=18). To obtain the summarizing statistics, repeatedly 12 random samples were chosen. The 95% interquartile range CI provides the uncertainty measure on the approximated likelihood function. 62

Table 5. Estimated times of transitions from predominant outcrossing to predominant selfing in *Arabidopsis thaliana*. The demography was estimated for three non-admixed genetic clusters (genetic assignment > 95%) of the 1001 genomes project. *teSMC* estimated the time of a transition from outcrossing to selfing on a sample of 20 haplotypes, 1 MB from each of the five chromosomes. Exonic regions were masked. The Relicts sample set consisted of 17 individuals only. 84

Abbreviations

ABC	Approximate Bayesian Computation
ANC	Ancestral
BGS	Background selection
bp	Base pair
BW	Baum-Welch
c	Recombination rate scaled in genetic distance
CDS	Coding sequence
CEU	Central European
eSMC	Ecological SMC
ESS	Evolutionary stable strategy
F_{IS}	Inbreeding factor F_{IS}
HMM	Hidden Markov model
HPC	High performance cluster
IBnr	Iberean non-relict
LD	Linkage disequilibrium
MLE	Maximum likelihood estimation
MPG	Max-Planck-Gesellschaft
MPIPZ	Max-Planck-Institut für Pflanzenzüchtungsforschung
MRCA	Most recent common ancestor
MSMC	multiple sequentially Markovian coalescent
μ	Mutation rate per generation per base pair
N	Population size
n	Sample size
PRES	Present
PSMC	Pairwise sequentially Markovian coalescent
PSMC'	Pairwise sequentially Markovian coalescent using the SMC' algorithm
r	Recombination rate per generation per base pair
r2	Unit to measure LD
ρ	Population recombination rate
SFS	Site-frequency spectrum

SI	Self-incompatibility
σ	Selfing rate
SLiM3	Selection on linked mutation 3 (WF-simulator)
SMC	Sequentially Markovian coalescent
TAC	Thesis advisory committee
teSMC	Extended eSMC with piecewise constant selfing
tsABC	ABC to estimate transitions to selfing
ϑ	Population mutation rate; pairwise diversity
TL	The joint distribution of T_{MRCA} and length of MRCA segments
TM_{TRUE}	The probability of T_{MRCA}s of consecutive MRCA segments following each other depending on their T_{MRCA}
TM_{WIN}	The probability of discretized pairwise diversities of consecutive windows along the sequence following each other dependent on their T_{MRCA}
T_{MRCA}	Time to the MRCA
WF	Wright-Fisher

Acknowledgment

Scientific, technical, and private support

Scientific, technical, and private support

I conducted the work of my Ph. D. project, resulting in this thesis at the Max Planck Institute for plant breeding research (MPIPZ) in Cologne, Germany. It would not have been possible without all the great people accompanying me during this journey. I started my time at the MPIPZ with an internship under the outstanding supervision of Lachezar "Luke" Nikolov, learning about comparative developmental genetics in the context of evolutionary biology. I was honored to join Dr. Stefan Laurent in his population genetics group. I would like to thank Dr. Stefan Laurent for his excellent supervision. During my studies, I collaborated with Thibaut Sellinger and Prof. Dr. Aurélien Tellier in the TUM. I would like to thank Thibaut Sellinger and Prof. Dr. Aurélien Tellier. I would like to dedicate special thanks to my thesis advisory committee assuring for quality and progress supervision: 1) Prof. Dr. Silvain Glémin for his expertise in the field and always constructive comments and advice; 2) Dr. Stefan Laurent for his direct supervision and motivational guidance; 3) Prof. Dr. Miltos Tsiantis for his scientific prospicience and hosting in his department since I arrived at the MPIPZ. I would like to thank the graduate school and Stephan Wagner, the Ph. D. coordinator, for his general support. I would like to thank the Ph. D. representatives. I would like to thank Danijel Majic. The figures were adapted with BioRender.com. I would like to thank all members of group Laurent, the natural variation group, the department, and the Institute for providing valuable scientific insights by asking the right question at the right time to guide and evolve this project. I would like to thank the technical support, especially the IT service. I would like to dedicate special thanks to everyone accompanying me through the times of my Ph. D. studies which sometimes may have been seeming stressful. I would like to dedicate special thanks to the people whom I did not mention by name. I would like to thank everyone.

Administratives

Declarations and Publications

Declarations and Publications

Declaration

Erklärung zur Dissertation gemäß der Promotionsordnung vom 12. März 2020

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

Köln, 7. Juni 2022 Stefan Strütt

In preparation

Inference of evolutionary transitions in reproductive modes using whole-genome sequences

Stefan Struett* (1), Thibaut Sellinger* (2), Sylvain Glémin (3), Aurélien Tellier** (2), Stefan Laurent** (1)

(1) Max-Planck Institute for Plant Breeding, Cologne, Germany
(2) Professorship for Population Genetics, Department of Life Science Systems, School of Life Sciences, Technical University of Munich, Freising, Germany
(3) INRAE Rennes, France

* these authors contributed equally,
** corresponding authors: laurent@mpipz.mpg.de and aurelien.tellier@tum.de

Curriculum vitae

Personal details

Name	Stefan Strütt
Geburtsdatum	11.01.1989
Geburtsort	Freiburg im Breisgau, Germany
Nationalität	Deutsch

Education

May 2018 – Ph. D. student at the Max Planck Institute for plant breeding research in Cologne. Thesis: Identifying and estimating shifts from outcrossing to selfing using genome-wide genetic variance

January 2017 to February 2018 – Internship with Luke; following Dr. Lachezar Nikolov for one year performing reverse genetics approach to identify key regulatory genes in leaf shape diversity.

October 2012 to March 2016 – Master's degree in Biological Sciences at the University of Konstanz. Thesis: The non-canonical function of NBS1 in neural development through the Notch pathway (Leibniz institute for aging, Fritz-Lipmann institute, Jena)

October 2008 to September 2012 – Bachelor's degree in Biological Sciences at the University of Konstanz. Thesis: Structural snapshots of ternary complex of Klentaq with modified dNTPs

October 2007 to September 2008 – Mathematics and Physics at the University of Freiburg