HAL
open science

Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes

Katja Reichel

To cite this version:

Katja Reichel. Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes. Agronomie. Agrocampus Ouest, 2015. Français. NNT : 2015NSARC123 . tel-01493824

HAL Id: tel-01493824

https://theses.hal.science/tel-01493824v1

Submitted on 22 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RÉSUMÉ

Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes

Les systèmes reproducteurs déterminent comment le matériel génétique est transmis d'une génération à la suivante [...]. Les espèces qui combinent de la reproduction sexuée et asexuée/clonale sont très répandues [... mais] l'effet de leur système reproducteur sur leur évolution reste énigmatique et discuté.
L'objectif de cette thèse est de modéliser la dynamique des fréquences génotypiques d'une population avec une combinaison de reproduction sexuée et/ou clonale dans des cycles de vie principalement diploïdes [. ... Un] modèle du type chaine de Markov avec temps et états discrets sert de base mathématique pour décrire [leurs] changements [...] au cours du temps.
Les résultats montrent que la reproduction partiellement asexuée peut en effet modifier la dynamique de la diversité génomique par rapport à une reproduction strictement sexuée ou strictement asexuée. [...] L'histoire démographique a un rôle important pour les organismes partiellement clonaux et doit être prise en compte dans toute analyse [...].
Cette thèse fait des recommandations pour la collecte des données et une hypothèse de base pour l'interprétation des données de génétique/génomique [...]. Ces résultats ont des retombées dans plusieurs domaines, allant de la recherche fondamentale [...] à des applications en agriculture [...], pêche [...] et protection de la nature [...].

Mots-clés : évolution, biologie théorique, diversité génétique, microsatellites, polymorphismes nucléotidiques, génomique, apomixie, agamospermie, multiplication végétative, parthénogenèse, mutation, dérive génétique, démographie, hybridation

ABSTRACT

Effects of partial asexuality on the dynamics of genotype frequencies in dominantly diploid populations

Reproductive systems determine how genetic material is passed from one generation to the next, making them an important factor for evolution. Organisms that combine sexual and asexual/clonal reproduction are very widespread [... yet] the effects of their reproductive system on their evolution are still controversial and poorly understood.
The aim of this thesis was to model the dynamics of genotype frequencies under combined sexual/clonal reproduction in dominantly diploid life cycles [. ... A] state and time discrete Markov chain model served as the mathematical basis to describe [their] changes [...] through time.
The results demonstrate that partial clonality may indeed change the dynamics of genomic diversity compared to either exclusively sexual or exclusively clonal populations. [...] Time has a crucial role in partially clonal populations and needs to be taken into account in any analysis of their genomic diversity.
This thesis provides recommendations for data collection and a null hypothesis for the interpretation of population genetic/genomic data [...]. Moreover, it includes new methods for the analysis of genotype-based population genetic Markov chain models. These results have a high potential relevance in several areas, ranging from basic research [...] to applications in agriculture [...], fisheries [...] and nature conservation [...].

Keywords: *evolution, theoretical biology, genetic diversity, microsatellites, single nucleotide polymorphisms, genomics, apomixis, agamospermy, vegetative multiplication, parthenogenesis, mutation, genetic drift, demography, hybridization*

Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes

AGROCAMPUS OUEST • Institut supérieur des sciences agronomiques, agroalimentaires, horticoles et du paysage
65 rue de Saint-Brieuc – CS84215 – F-35042 Rennes Cedex
Tél. : 02 23 48 50 00
www.agrocampus-ouest.fr

Thèse **C-123 — 2015-29** • REICHEL Katja

AGRO CAMPUS OUEST

UNIVERSITÉ EUROPÉENNE DE BRETAGNE

Katja REICHEL • **10 décembre 2015**

Thèse AGROCAMPUS OUEST
sous le label de l'Université Européenne de Bretagne
pour obtenir le grade de
DOCTEUR D'AGROCAMPUS OUEST
Spécialité Biologie et Agronomie

ÉCOLE DOCTORALE • Vie-Agro-Santé (VAS)

LABORATOIRE D'ACCUEIL • UMR INRA - AGROCAMPUS OUEST - Université de Rennes 1 : Institut de génétique, environnement et protection des plantes (IGEPP)

Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes

David CAUSEUR
Professeur, AGROCAMPUS OUEST, Laboratoire de mathématiques appliquées / *président*

Sylvain GLEMIN
Directeur de recherche, CNRS Montpellier, Uppsala Universitet / *rapporteur*

Christoph HAAG
Chargé de recherche, CNRS Montpellier / *rapporteur*

Emmanuelle PORCHER
Professeur, MNHN Paris / *examinatrice*

Denis ROZE
Chargé de recherche, CNRS Roscoff / *examinateur*

Timothy SHARBEL
Professeur, IPK Gatersleben, University of Saskatchewan / *examinateur*

Jean-Christophe SIMON
Directeur de recherche, UMR INRA-AO-UR1 IGEPP / *directeur de thèse*

Solenn STOECKEL
Chargé de recherche, UMR INRA-AO-UR1 IGEPP / *co-directeur de thèse*

EFFECTS OF PARTIAL ASEXUALITY ON THE DYNAMICS OF GENOTYPE FREQUENCIES IN DOMINANTLY DIPLOID POPULATIONS

EFFETS DE LA REPRODUCTION PARTIELLEMENT ASEXUEE SUR LA DYNAMIQUE DES FREQUENCES GENOTYPIQUES EN POPULATIONS MAJORITAIREMENT DIPLOIDES

Thesis

submitted by

Katja Reichel

to obtain a doctoral degree in biology from the

Institut Supérieur des Sciences Agronomiques, Agro-Alimentaires, Horticoles et du Paysage (Agrocampus Ouest)

Defended in Rennes, 10th December 2015

Jury members

Sylvain GLEMIN	CNRS Montpellier, Uppsala Universitet	rapporteur
Christoph HAAG	CNRS Montpellier	rapporteur
Emmanuelle PORCHER	MNHN Paris	examiner
Denis ROZE	CNRS Roscoff	examiner
Timothy SHARBEL	IPK Gatersleben, University of Saskatchewan	examiner
David CAUSEUR	Agrocampus Ouest Rennes	examiner
Solenn STOECKEL	INRA Le Rheu	adviser
Jean-Christophe SIMON	INRA Le Rheu	senior adviser

Seht ihr den Mond dort stehen?
Er ist nur halb zu sehen,
Und ist doch rund und schön!
So sind wohl manche Sachen,
Die wir getrost belachen,
Weil unsre Augen sie nicht sehn.

Der Mond ist aufgegangen
Matthias Claudius (1778)

A Foreword

The work presented in this thesis was carried out during three years, from November 2012 to October 2015, at the Institute for Genetics, Environment and Plant Protection (IGEPP), a part of the French National Institute of Agricultural Research (INRA) situated in Le Rheu near Rennes, Brittany. It was co-financed to equal parts by the Doctoral Studies Allowance Programme (ARED) of the Brittany Region, and the Department of Plant Health and Environment (SPE) of the INRA. The subject of the thesis was embedded in the collaborative research project CLONIX (ANR-11-BSV7-0007), financed by the French National Agency for Research (ANR).

This document is structured as a thesis by publication based on four manuscripts of research articles. According to the French Government Resolution of 6th January 2005 (NOR: MENS0402905A) and the internal standards of the Life Agronomy Health (VAS) doctoral school, it includes a summary of at least ten percent of the text (by number of pages) in French. In addition, the abstract is provided in German.

The progress of this thesis was reported twice to a thesis advisory committee, whose members were: Sophie ARNAUD-HAOND (Ifremer Sète), Stéphane DE MITA (INRA Nancy), Fabien HALKETT (INRA Nancy), Florent MALRIEU (François Rabelais University Tours), Jean-Pierre MASSON (INRA Le Rheu), Denis ROZE (CNRS Roscoff), Jean-Christophe SIMON (INRA Le Rheu, senior advisor), Solenn STOECKEL (INRA Le Rheu, advisor) and Christian WALTER (Agrocampus Ouest Rennes, tutor).

Official logo of the CLONIX project, designed by the author.

B Abstract – Résumé – Kurzfassung

Abstract Effects of partial asexuality on the dynamics of genotype frequencies in dominantly diploid populations

Reproductive systems determine how genetic material is passed from one generation to the next, making them an important factor for evolution. Organisms that combine sexual and asexual/clonal reproduction are very widespread, both throughout the earth's biomes and on the eukaryotic tree of life. However, the effects of their reproductive system on their evolution are still controversial and poorly understood.

The aim of this thesis was to model the dynamics of genotype frequencies under combined sexual/clonal reproduction in dominantly diploid life cycles, with the view of establishing a reference for future field studies. This involves two subtypes of partially clonal reproduction: either both reproductive modes co-occur ("acyclic partial clonality"), or they alternate ("cyclic clonality"). For both, a state and time discrete Markov chain model served as the mathematical basis to describe changes of the genotype frequencies through time.

The results demonstrate that partial clonality may indeed change the dynamics of genomic diversity compared to either exclusively sexual or exclusively clonal populations. Moreover, both subtypes have different effects under selectively neutral conditions: while acyclic partial clonality leads to increased variation in the frequency of heterozygous genotypes within the population, the patterns observed under cyclic clonality depend on the sampling time (before or after sexual reproduction) and show a stronger effect on allele frequencies. The dynamics of population heterozygosity were also slowed down under acyclic partial clonality, yet this effect did not generally lead to slower adaptation under selection. Time has a crucial role in partially clonal populations and needs to be taken into account in any analysis of their genomic diversity.

This thesis provides recommendations for data collection and a null hypothesis for the interpretation of population genetic/genomic data from dominantly diploid partially clonal organisms. Moreover, it includes new methods for the analysis of genotype-based population genetic Markov chain models. These results have a high potential relevance in several areas, ranging from basic research, e.g. on the evolution of sex and speciation, to applications in agriculture (e.g. partially clonal crops, pests and pathogens), fisheries (e.g. primary producers and plankton) and nature conservation (e.g. threatened or invasive species).

Keywords: evolution, theoretical biology, genetic diversity, microsatellites, single nucleotide polymorphisms, genomics, apomixis, agamospermy, vegetative multiplication, parthenogenesis, mutation, genetic drift, demography, hybridization

Résumé Effets de la reproduction partiellement asexuée sur la dynamique des fréquences génotypiques en populations majoritairement diploïdes

Les systèmes reproducteurs déterminent comment le matériel génétique est transmis d'une génération à la suivante, de ce fait ils sont un facteur important pour l'évolution des organismes. Les espèces qui combinent de la reproduction sexuée et asexuée/clonale sont très répandues, pas seulement dans les biomes mondiaux mais également sur l'arbre de vie des eucaryotes. Pourtant l'effet de leur système reproducteur sur leur évolution reste énigmatique et discuté.

L'objectif de cette thèse est de modéliser la dynamique des fréquences génotypiques d'une population avec une combinaison de reproduction sexuée et/ou clonale dans des cycles de vie principalement diploïdes, dans la perspective d'établir une référence pour des études de terrain. Deux formes de reproduction partiellement asexuée sont considérées : soit les deux modes de reproduction se produisent en parallèle (« asexualité partielle acyclique »), soit ils arrivent en alternance (« asexualité cyclique »). Dans les deux cas, un modèle du type chaine de Markov avec temps et états discrets sert de base mathématique pour décrire les changements des fréquences génotypiques au cours du temps.

Les résultats montrent que la reproduction partiellement asexuée peut en effet modifier la dynamique de la diversité génomique par rapport à une reproduction strictement sexuée ou strictement asexuée. De plus, les deux formes ont des effets différents dans des conditions sélectivement neutres. L'asexualité partielle acyclique produit plus de variabilité dans la fréquence des génotypes hétérozygotes dans la population, tandis que les configurations observées sous asexualité cyclique dépendent du moment d'échantillonnage (avant ou après la reproduction sexuée) et on y observe des effets plus importants sur les fréquences alléliques. Par ailleurs, l'évolution de l'hétérozygotie au niveau de la population est ralentie avec l'asexualité partielle acyclique, même si cet effet ne mène pas pour autant à une adaptation généralement plus lente sous sélection. L'histoire démographique a un rôle important pour les organismes partiellement clonaux et doit être prise en compte dans toute analyse de leur diversité génomique.

Cette thèse fait des recommandations pour la collecte des données et une hypothèse de base pour l'interprétation des données de génétique/génomique des populations chez les organismes principalement diploïdes avec une reproduction partiellement asexuée. Au-delà, elle contient des nouvelles méthodes pour l'analyse des modèles du type chaine de Markov basés sur les fréquences génotypiques en génétique des populations. Ces résultats ont des retombées dans plusieurs domaines, allant de la recherche fondamentale (par exemple sur l'évolution de la sexualité et la spéciation) à des applications en agriculture (par exemple plantes cultivées partiellement clonaux, pathogènes et ravageurs des cultures), pêche (par exemple producteurs primaires et plancton) et protection de la nature (par exemple espèces menacées ou invasives).

mots clés : évolution, biologie théorique, diversité génétique, microsatellites, polymorphismes nucléotidiques, génomique, apomixie, agamospermie, multiplication végétative, parthénogenèse, mutation, dérive génétique, démographie, hybridation

Kurzfassung Die Effekte partiell asexueller Reproduktion auf die Häufigkeitsdynamik von Genotypen in dominant diploiden Populationen

Fortpflanzungssysteme bestimmen darüber, wie genetische Information von einer Generation zur nächsten weitergegeben wird. Sie sind somit ein wichtiger Evolutionsfaktor. Organismen, die sexuelle und asexuelle/klonale Fortpflanzung miteinander kombinieren können, sind sehr weit verbreitet, sowohl in den verschiedenen Biomen der Erde als auch im Stammbaum der Eukaryoten. Die evolutionären Auswirkungen ihres Fortpflanzungssystems sind jedoch noch immer umstritten und rätselhaft.

Ziel dieser Doktorarbeit war es, die zeitliche Entwicklung der relativen Häufigkeiten von Genotypen in einer Population bei kombinierter sexueller und klonaler Vermehrung in dominant diploiden Lebenszyklen zu modellieren, um eine Referenz für zukünftige Feldstudien zu erhalten. Dies beinhaltet zwei Formen von partiell asexueller Reproduktion: entweder treten beide Fortpflanzungsmodi parallel auf (azyklische partielle Asexualität), oder in alternierenden Phasen (zyklische Asexualität). Bei beiden Formen dienten diskrete, endliche Markov'sche Ketten als Basis für die mathematische Beschreibung der Häufigkeitsdynamik verschiedener Genotypen über die Zeit.

Die Ergebnisse zeigen, dass partiell asexuelle Fortpflanzung im Vergleich mit ausschließlich sexueller oder ausschließlich klonaler Vermehrung in der Tat die zeitliche Entwicklung genomischer Diversität beeinflussen kann. Darüber hinaus haben beide Formen partieller Asexualität unter selektiv neutralen Bedingungen unterschiedliche Auswirkungen: Während die azyklische Form zu größerer Variabilität in der relativen Häufigkeit heterozygoter Genotypen in der Population führt, hängen bei zyklischer Asexualität die Beobachtungen vom Zeitpunkt der Probennahme ab (vor oder nach der sexuellen Phase) und zeigen stärkere Folgen für die Häufigkeiten von Allelen. Auch verändert sich die Häufigkeit heterozygoter Individuen unter azyklischer partieller Asexualität langsamer, was jedoch unter Selektion nicht zu einer generell verlangsamten Anpassung führt. Die Zeit spielt in partiell klonalen Populationen eine Schlüsselrolle und sollte bei allen Analysen ihrer genomischen Diversität in Betracht gezogen werden.

Diese Doktorarbeit beinhaltet Empfehlungen zur Probennahme bei populationsgenetischen/-genomischen Studien an partiell klonalen, dominant diploiden Organismen, sowie eine Referenz für deren Auswertung. Darüber hinaus umfasst sie neue Analysemethoden für Markov'sche Ketten, welche die Entwicklung der Frequenzen von Genotypen beschreiben. Diese Ergebnisse haben ein hohes Anwendungspotential in mehreren Gebieten, von der Grundlagenforschung (z.B. Fragen nach der Evolution der Sexualität und Artbildung) bis hin zu Landwirtschaft (z.B. partiell asexuelle Kulturpflanzen, Schädlinge und Pathogene), Fischerei (z.B. Primärproduzenten und Plankton) und Naturschutz (z.B. bedrohte oder invasive Arten).

Schlüsselworte: Evolution, Theoretische Biologie, genetische Diversität, Mikrosatelliten, Einzelnukleotid-Polymorphismen, Genomik, Apomixis, Agamospermie, vegetative Vermehrung, Parthenogenese, Mutation, Genetische Drift, Demografie, Hybridisierung

C Thanks

"Der Zufall kann große Dinge tun", great things can be done by chance, was the inscription some former students left on a sculpture in the central courtyard of my school. Who would have thought that, one day, I would come to live in a small town in Brittany and work as a doctoral student for the French National Institute for Agricultural Research, on a topic that sounds like a sequel to my high school research project about asexual reproduction in higher plants! Chance indeed – or was it? Looking back, I feel that I have been extremely fortunate, and I would like to thank all people who – knowingly or unknowingly, willingly or unwillingly – have helped "chance" along, or at least did not slam the door in its face.

First of all, I would like to thank my supervisors Jean-Christophe Simon and Solenn Stoeckel, for their willingness to take on the "adventure" of having a German PhD student and their good-humored support throughout these past three years. Especially without Solenn, this thesis project would never have been possible – not just because I would not have managed to rent a flat without the help of a native speaker. It is a comfort to know that, even more than 1 200 km from home, one can still find people who share the same crazy interest and enthusiasm for science, evolution and the role of asexual reproduction within it all. Come to think of it, it might even be easier to find them here than elsewhere.

Being a doctoral student in France has some peculiarities, among them the "comité de these" (PhD committee) – I have felt extremely honoured, though also rather intimidated, to see six busy scientists from all over France leave their work just to come and talk with me about mine. I would like to thank Sophie Arnaud-Haond, Stéphane De Mita, Fabien Halkett, Florent Malrieu, Jean-Pierre Masson and Denis Roze for their interest, their patience, their helpful criticism and their ideas that have helped me to advance. Moreover, it appears that the thesis defense will follow the same principle, only with different people from a yet wider geographical area – I would like to thank Sylvain Glémin, Christoph Haag, Emanuelle Porcher, Denis Roze, Timothy Sharbel and David Causeur for having agreed to participate in my PhD jury.

In addition to my PhD committee meetings, I have profited from numerous discussions during the annual meetings of the CLONIX project group, at the five international and two national (French) conferences where I was given the chance to present my work, with colleagues, former colleagues, former colleague's colleagues and others. In alphabetical order, I would like to thank in particular: Jurgen Angst, Yoann Bourhis, Magda Castel, Judith Fehrer, Julie Jaqiéry, Melodie Kuenegel, Judith Lichtenzveig, Pierre Nouhaud, Nicolas Parisey, Sylvain Poggi, Christiane Ritz, Romuald Rouger, Myriam Valero and Karsten Wesche. Ingo Uhlemann inspired my first steps towards studying plant reproduction, Frank Richter may be held (partially) responsible for the profusion of ternary diagrams within this thesis, while the Ginko in the introduction appears in fond memory of Harald Walther.

During my time in Le Rheu, I also had the chance to acquire some teaching experience: by co-supervising the Master thesis of Cédric Midoux and the research internships of Valentin Bahier, François Timon and Clément Barthélémy. I would like to thank Malika Aïnouche for giving me the opportunity to try my hand at a one-year teaching assistantship at the

Université Rennes 1, my colleagues Abdelkader Aïnouche, Julien Boutte, Abdelhak El Amrani, Morgane Gicquel, Helène Rousseau, Armel Salmon, Agnès Schermann, Cécile Sulmon and Michèle Tarayre-Renouard for their friendliness and support, and all students for their patience and resistance to my errors. Teaching lab courses in French about plant life cycles, morphology and ecology would not have been possible without first learning some French myself, and I would like to thank Xavier Ségalen, Annick Le Gall and Laetitia Burmalo for helping me to do so. Annick Le Gall has also earned my warmest thanks for proofreading the French translations within my thesis.

Finally, I would like to thank those people who have made my life easier and the distance from home more bearable during the last three years: my tutor Christian Walter, whose advice I appreciated; Patricia Nadan, Pascale Leneve, Anne-Sophie Grenier and Géraldine Blondel, who helped me organize and find things; Akiko Sugio and Alexandre Robert-Seilaniantz, for mental support; my colleagues at the IGEPP, for the comforting murmur on the corridor; my teachers and fellow students at the "La Flume" music school; and my friends from the local bike club "Le Rheu à vélo".

The last place is the place of honour – which goes, as always, to my family.

D Table of contents

A	Foreword		5
B	Abstract – Résumé – Kurzfassung		7
C	Thanks		10
D	Table of contents		12
Part I	**Introduction**		**15**
1	Task		17
2	Context		
	2.1	Genetic diversity – a key to success?	19
	2.2	Reproduction and inheritance – an ancient quest	20
	2.3	The evolution of (a)sex – an ongoing debate	21
	2.4	Partial asexuality – a pervasive phenomenon	22
3	Current knowledge		
	3.1	Reproductive systems – what is what	25
	3.2	Previous studies	27
		3.2.1. Selectively neutral diversity	27
		3.2.2. Diversity under selection	30
4	Approach		32
Part II	**Methods**		**37**
5	Mathematical model		
	5.1	Model choice	39
	5.2	Model assumptions and structure	41
	5.3	Model equations	43
	5.4	Model analysis	48
Article I		*Interpretation and approximation tools for big, dense Markov chain transition matrices in population genetics*	51
	5.5	Digression: Extending the triangle	77

Part III Results **81**

6 Selectively neutral diversity

6.1 Neutral diversity under acyclic partial asexuality 83

Article II *Rare sex or out of reach equilibrium? The dynamics of* F_{IS} *in partially clonal organisms* 87

6.2 Multilocus simulation of a demographic bottleneck 133

6.3 Neutral diversity under cyclical parthenogenesis 135

Article III *Effects of complex life-cycle on genetic diversity: The case of cyclical parthenogenesis* 141

7 Diversity under selection

7.1 Effects of selection under acyclic partial asexuality 166

Article IV *Partial asexuality and the speed of adaptation* 171

7.2 Linkage 187

Part IV Discussion and conclusion **191**

8 Synthesis

8.1 Main findings 193

8.1.1. General remarks 193

8.1.2. Genotype frequency dynamics under partial asexuality 194

8.1.3. Meselson effect 197

8.2 Practical implications 198

8.3 Contribution to evolution of sex debate 199

8.4 Perspectives 202

Part V Bibliography **205**

E Author information 217

Part I Introduction

1 Task

Organisms who may reproduce both sexually and asexually, called “partially asexual”, are extremely common on earth. However, most population genetic theory accounts only for either exclusively sexual or exclusively asexual reproduction. Though there are many empirical studies on partially asexual species, putting observations into context can therefore be difficult. Developing population genetic theory that accounts for both sexual and asexual reproduction may therefore help to understand the evolution of such organisms, and eventually the evolutionary history of their reproductive system itself.

This thesis presents results from a mathematical model for the population genetics of partially asexual species. Its central question is,

“How do the patterns and dynamics of genetic diversity change in partially asexual species, compared to their exclusively sexual/asexual counterparts?”

Special attention is given to comparatively small populations, as there is both a special need (conservation genetics) and a special scarcity of appropriate theory for this case. Consequently, the model (modeled quantities, input parameters) is oriented towards the methods currently in use for such population genetic studies: it describes the dynamics of genotype frequencies at individual single-nucleotide polymorphism (SNP)/microsatellite loci, with and without selection. Though asexual reproduction may occur during both the haploid and diploid phase of a (sexual) life cycle, and though there are many instances of partially asexual polyploid species, the scope of this thesis is limited to populations of dominantly diploid (i.e. diplontic) organisms, as a basis for future development.

In this thesis, modeling is used as a deductive tool, i.e. based on theory in contrast to empirical-based inductive models. The model results are thus intended to serve as a null hypothesis, to test if field data conform to preconceived ideas about the biology of the studied populations. Beside this direct application, the results are also discussed in the broader context of the evolution of reproductive systems.

Mission

Les organismes également capables de reproduction sexuée et asexuée, appelés « partiellement asexuée », sont très répandus sur la Terre. Cependant, la plupart des théories en génétique des populations ne prend seulement en compte soit la reproduction exclusivement sexuée, soit exclusivement asexuée. Bien qu'il y ait beaucoup d'études empiriques sur les espèces partiellement asexuées, indiquer les résultats dans un contexte peut être difficile. Le développement de la théorie en génétique des populations qui inclut aussi bien la reproduction sexuée qu'asexuée pourrait en conséquence aider à comprendre l'évolution de tels organismes, et éventuellement l'histoire évolutive de ce système reproducteur en lui-même.

Cette thèse de doctorat présente des résultats d'un modèle mathématique pour la génétique des populations des espèces partiellement asexuées. Sa question centrale est,

« Comment les structures et les dynamiques de la diversité génétique changent-elles chez les espèces partiellement asexuées, par rapport à leurs pendants exclusivement sexués/asexués ? »

Les égards sont focalisés sur des populations comparativement petites, comme il y a également un besoin particulier (méthodes génétiques pour la protection des espèces) et une rareté particulière de modèles pour ce cas spécifique. Le modèle (quantités représentées, paramètres d'entrée) est ainsi orienté vers des méthodes actuellement en cours d'utilisation pour certaines études de génétique des populations : il décrit les fréquences génotypiques à des loci uniques, tels que des polymorphismes nucléotidiques (SNP) ou microsatellites, avec et sans sélection. Même si la reproduction asexuée peut se dérouler également pendant la phase haploïde et diploïde d'un cycle de vie (sexué), et même s'il existe de nombreux exemples d'espèces partiellement asexuées polyploïdes, le cadre de cette thèse s'est limité aux populations principalement diploïdes (i.e. diplophasiques et haplodiplophasiques avec dominance de la phase diploïde) comme base à des développements futurs.

Dans cette thèse de doctorat, la modélisation est utilisée comme outil de déduction, soit basé sur la théorie contrairement aux modèles inductifs basés sur des données empiriques. Les résultats du modèle sont ainsi destinés à servir comme hypothèse de base, afin de tester si des données du champ sont conformes à des idées préformées sur la biologie des populations étudiées. Au-delà de cette application directe, les résultats sont également discutés dans le contexte plus vaste de l'évolution des systèmes reproducteurs.

2 Context

2.1 Genetic diversity – a key to success?

Looking back through geological time, one has to acknowledge that most life forms known from fossils do not exist anymore. Taking seed plants as an example, the first fossils that can be assigned to current genera based on their morphology appear during the Tertiary (with few notable exceptions, e.g. *Ginko* which is known since the Jurassic; Taylor et al. 2008).

Would extinct taxa have had more chance to survive if they had been more genetically diverse? Perhaps not – conditions have changed, evolution has moved on, and the early land plants of the Rhynie chert would today surely be overgrown and outcompeted by physiologically more intricate angiosperms (Channing & Edwards 2013). Or perhaps yes – though they may look somewhat different, modern angiosperms certainly draw on their genetic heritage from the early days of land plant evolution. We do not know how many of the extinct taxa really died out, and how many just changed beyond recognition.

Genetic diversity may be an important asset for species' survival by allowing the offspring to differ heritably from its less fit parents. However, this includes the risk that the offspring may be even less fit, turning each mechanism that increases offspring diversity into a double-edged sword. The diversity of reproductive systems among biota may, at least in part, be a direct result of this dilemma. In connection with information on the origin and prevalence of different reproductive systems, understanding their impact on genetic diversity (Duminil et al. 2007, 2009) may therefore tell us a great deal about how evolution works under different conditions.

This thesis contributes to the debate on the importance of genetic diversity for the potential to evolve, by mathematically describing the dynamics of genetic diversity in populations with partially asexual reproduction. The effect of this reproductive system is controversial: asexual reproduction is sometimes seen as an evolutionary dead-end and hindrance to adaptation, yet it could also be a mechanism to escape extinction and even an ecological asset (e.g. Honnay & Bossuyt 2005, Silvertown 2008, Hörandl 2009, Van Drunen et al. 2015). Extending the theoretical basis of our understanding for evolution under partial asexuality and providing a reference for field studies may help to put this discussion on firmer ground.

Today, the impact of a single species – ours – on the entire biosphere produces worrying results (Vitousek et al. 1997). Though our efforts to understand, predict and direct evolution may be first and foremost directed towards safeguarding the future of our own descendants, this goal cannot be reached without preserving a functioning environment for them to live in, including "friendly neighbors (and their kids)" from other species.

2.2 Reproduction and inheritance – an ancient quest

Reproduction, the ability of organisms to give rise to something similar, though not identical to themselves, is one of the defining features of life (Trifonov 2011). In evolutionary biology, reproduction is especially important since it defines how genetic diversity is passed on through generations – to use a mathematical analogy, reproduction is the "recurrence equation" of the "sequence" of organisms through time.

Human interest in reproduction and inheritance goes back at least until the earliest remaining records of Western natural philosophy and science. In the fourth book of his "On the generation of animals" (4th century BCE), the Greek philosopher Aristotle asked: "Again, for what reason is a child generally like its ancestors, even the more remote?" (Aristotle 2002). This question has accompanied natural historians, researchers in medicine and biologists for generations, and it is only since the 19th century that answers are beginning to take shape (Jahn 2004).

Different forms of reproduction are already distinguished in the works of Aristotle: according to the number of parents involved, reproduction may be biparental (two parents, typically dimorphic as a male and a female), uniparental (one parent) or abiogenetic (no parent). Though later research brought a lot of adjustments to Aristotle's views – most notably about abiogenesis, which is now assumed to have happened only in the ancient past (Pasteur 1864, Woese 1987)– this basic system is still often used (and it leaves us with the intriguing question why there are rarely more than two parents, but see Bonen et al. 2007). A more detailed review of reproductive systems from today's perspective is given in chapter 3.1.

The "laws" of inheritance were also always especially interesting for breeders of domesticated plants and animals. This context originally motivated Gregor Mendel's famous experiments on garden peas and other plants (Mendel 1865). Though he never knew it, Mendel also inadvertently became the first "geneticist" to study inheritance under partial asexuality (Mendel 1869, Koltunow et al. 2011): to complement his experiments on garden peas and other plants, he tried to achieve artificial crosses with different species of hawkweeds (genera *Hieracium* and *Pilosella*, Asteraceae). Despite great efforts, which may even have contributed to Mendel's deteriorating eyesight (Dostál 2015), he obtained only very few "hybrid" seeds due to extensive asexual seed production in his study species. He eventually died before bringing his experiments to a statistically satisfying result.

Population genetics, the study of genetic diversity at the population level, was a direct result of the rediscovery of Mendel's work at the beginning of the 20th century. One of its most important foundations, the Hardy-Weinberg equilibrium (Hardy 1908, Weinberg 1908), is the direct extension of Mendel's "laws" of inheritance to whole groups of organisms. The theoretical framework of population genetics, which draws heavily on mathematics and statistics, first permitted to discuss the quantitative outcomes of different reproductive systems in an evolutionary context and compare them with each other, as shall be done in this thesis.

Sadly, Mendel's incomplete hawkweed results did not inspire a similar theoretical development as did his pea experiments, and the first mathematical models for the population genetics of partially asexual species only appeared in the 1970ies (Asher 1970, Marshall & Weir 1979; see chapter 3.2). Still today, the scarcity of methods that are adapted to partially asexual species make studying their population genetics challenging (Halkett et al. 2005, Arnaud-Haond et al. 2007). Methods in population genetics continue to evolve, especially as the sequencing of whole genomes steadily becomes cheaper and a wealth of new data becomes available (e.g. Brookes 1999, Vitti et al. 2013). Yet before theories that were developed based on exclusively sexual populations can be safely used also for partially asexual species, the underlying assumption that both systems are identical, at least in those respects on which the theory relies, has to be verified. Our results may prove highly useful for pinpointing such issues and finding a solution.

Recently, the focus of research on reproduction and inheritance has somewhat shifted from the mechanism itself to understanding its environmental and genetic regulation and its evolutionary significance. Studies working on the regulation of different modes of reproduction showed that, though reproductive modes with identical outcomes for offspring diversity may have arisen multiple times in different evolutionary lineages, the mechanisms involved can sometimes be similar across large phylogenetic distances (e.g. Wang et al. 2004, Le Trionnaire et al. 2008, Sharbel et al. 2009, Hand & Koltunow 2014). These results make the question of the role of different reproductive systems and genetic diversity for the evolution of species more complex on the one hand, since different evolutionary backgrounds have to be taken into account, but on the other hand give more hope for resolving it by looking for similarities between the highly different cases. By developing a very basic and general model, which is not parameterized to fit only one species, we hope that our results will be widely applicable and allow such far-reaching comparisons.

2.3 The evolution of (a)sex – an ongoing debate

Why sexual and asexual reproduction co-exist is a long-standing question in evolutionary biology (e.g. Darwin 1860). Asexual reproduction does not need "costly" meiosis and no potentially dangerous search for a mating partner, does not disturb a functioning genotype by recombination, increases the chances of parental "selfish genes" to be transmitted and, compared to dioecy, does not produce a fraction of offspring that cannot give birth by itself (Maynard Smith 1978, Otto 2009). With all these advantages on its side, why does it not take over? Or, if sexual reproduction possesses some advantage that makes up for its "cost", why is asexual reproduction still going on? These questions are especially striking with respect to partially asexual species, where both modes of reproduction not just exist in closely related taxa, but within the same individual.

Both meiosis and mitosis are shared traits (synapomorphies) of the whole eukaryote clade and probably originated very early on in its history (Cavalier-Smith 2002, Bogdanov 2003, Wilkins & Holliday 2008). In theory, all eukaryotic organisms could therefore reproduce both sexually and asexually, thus making use of the relative advantages of either reproductive

system just as the situation calls for (compare Raven et al. 2004). However, not all of them do. Among those who have given up one or the other mode of reproduction at some point in their evolution, most appear to have opted for exclusive sexuality. Exclusively asexual reproduction is usually considered a recent and possibly short-lived "experiment" (but compare Neiman et al. 2009); whether or not "ancient asexual scandals" exist (Judson & Normark 1996, Signorovitch et al. 2015) can currently not be said with certainty, though it may be safe to assume that sexual reproduction is extremely rare in some species.

Partially asexual species may be especially interesting in the "evolution of sex" debate, as they constitute a "natural laboratory" to study the relative merits of both reproductive modes. Studies that manipulate the frequency of asexual vs. sexual reproduction to see how it affects the populations, or that, conversely, change the environmental conditions to see how it affects reproduction, are still rare (Becks & Agrawal 2012). However, though both modes of reproduction may directly compete for the same resources within a single individual, this is not always the case (e.g. Yu et al. 2001, Van Drunen et al. 2015). Potential advantages of partial asexuality in itself, rather than just as a way to combine the advantages of sexual and asexual reproduction, make the situation more complex.

By studying partial asexuality in its own right, we hope to draw attention to possible peculiarities of this reproductive system that might contribute to the "evolution of sex" debate. Partially asexual species have often been indiscriminately counted towards the "sexual camp" (e.g. Hartfield et al. 2012), even though previous studies already hinted otherwise (Marshall & Weir 1979, Berg & Lascoux 2000). We particularly concentrated on models of finite populations, including stochastic effects, and on providing not just static means and equilibria, but also a dynamic perspective. These two conditions have been suggested as important for models of the evolution of reproductive systems (Otto 2009), and even though we shall not directly address this question, our results may provide a basis for future development. The main goal of this thesis is to provide a reference and aid to the interpretation for future field studies – including experimental evolution – that may give new ideas and impulses for the ongoing debate.

2.4 Partial asexuality – a pervasive phenomenon

Partially asexual reproduction is extremely widespread in nature, both among unicellular and multicellular eukaryotes: it is common in protists (Speijer et al. 2015), fungi (Taylor et al. 1999, 2015) and the different clades of "algae" (Collado-Vides 2001). Plants are particularly notorious for it (Fryxell 1957, Grant 1976, Durka 2002, Richards 2003, Hojsgaard et al. 2014): well over half of all angiosperm species in Central Europe possess at least one way of clonal reproduction (Klimeš et al. 1997, Klimeš & Klimešová 1999), and the situation for other plant groups may be similar (e.g. Shaw & Goffinet 2000). Partially asexual reproduction in animals is usually considered to be somewhat rarer – nevertheless, several animal clades such as rotifers, platyhelminths or arthropods (Normark 2003) include partially asexual species, and there appear to be even some among vertebrates (Neaves & Baumann 2011, Avise 2015). Evaluating the prevalence of (partial) asexuality in animals may be more challenging than in

other organisms, as it often occurs by mechanisms which are easily confounded with different forms of selfing/automixis (see chapter 3.2).

Partially clonal species are directly important for human economy. In agriculture, one can encounter them in all capacities except as farm animals: there are partially clonal crop and fodder plants such as strawberries, potatoes and many species of grass (McKey et al. 2010); partially clonal pests, such as aphids; partially clonal parasites such as rust fungi; and partially clonal pathogens, such as oomycetes of the genus *Phytophthora*. Partially clonal trees such as willows or wild cherry are used in horti- and silviculture. In fishery, partially clonal species appear e.g. as sea grasses (e.g. used as building insulation material), different algae (e.g. for food) and zooplankton species such as the members of the genus *Daphnia* (nutrition for planktivorous fish). Yet the perhaps most direct impact on human life have several partially clonal pathogens (e.g. *Leishmania, Plasmodium, Trypanosoma, Toxoplasma* – Tibayrenc et al. 1990).

In most of these examples, the partially asexual reproductive system is directly important, either for the epidemiology of the "unwanted" or the propagation and harvesting of the "wanted". As an example, though strawberries are grown for their berries (sexual reproduction), they are usually propagated from runners (asexual reproduction) rather than seeds. For potatoes, the organs of asexual reproduction (tubers) are the actual crop, though sexual reproduction is important for breeding new varieties. Clonal propagation has the advantage of maintaining the properties of hybrid cultivars, which makes it a highly desirable trait for practically all cultivated species. The development of techniques for "artificial" clonal propagation (e.g. by tissue culture in orchids, Philip & Nainar 1986, or nuclear transfer in livestock, Wilmut et al. 1997) has therefore recently been complemented by attempts to genetically modify exclusively sexual crop plants so that they become capable of asexual seed production (van Dijk & van Damme 2000).

On a larger scale, partially clonal organisms dominate several of the earth's biomes (Klimeš et al. 1997, Baird et al. 2009), including grasslands/seagrass meadows, coral reefs, mangroves and tundra. Though they may appear in many different ecological roles, partially clonal species are most often associated with ecosystem engineers and invasive species (Silvertown 2008). In fact, both roles may be linked, and by the reproductive system: as an example, the European beachgrass *Ammophila arenaria* L. is famous for its ability to colonize and stabilize sand dunes with its roots and rhizomes. Because of this "useful" property, the plant was actively introduced outside its native range, where it now proliferates (mainly clonally) and threatens autochthonous ecosystems (Hilton 2006). However, partially clonal species may also be rare or threatened themselves: in some cases such as *Spiraea* (Brzyski & Culley 2011, Dajdok et al. 2011) or *Opuntia* (Reyes-Agüero et al. 2006, Helsen et al. 2009), one partially clonal genus may contain both threatened and invasive species. A reference for the population genetics of small partially clonal populations is particularly interesting for conservation, as it can be applied to rare and threatened species as well as to newly introduced and potentially invasive clonal organisms.

Several partially asexual species appear to be of (natural) hybrid origin, and/or are polyploid. It has been suggested that "accidental" inter-specific hybrids which show some

fitness advantage over their parents, but have a disturbed meiosis due to genomic incompatibilities, may "use" clonal reproduction to persist until eventually re-acquiring functional sexuality (Grant 1976, Chapman et al. 2003). Potentially improved colonizing abilities of partially or exclusively clonal (especially agamospermous) plants compared to their exclusively sexual relatives could lead to different distributional patterns ("geographical parthenogenesis" hypothesis, Hörandl 2006): As an example, the distributional range of clonal and partially clonal groups within *Taraxacum* and *Rubus*, compared to their exclusively sexual congeners, extends further north in Europe, which could be due to faster range expansion after the last glaciation. The results of this thesis may be partially applicable to such species as well, though polyploidy is not yet included in our model.

3 Current knowledge

3.1 Reproductive systems – what is what

Life on earth employs a kaleidoscopic diversity of reproductive systems, and research about them has produced an almost equal diversity of terms and, sometimes incongruent or even incompatible, definitions (compare e.g. Asher 1970, Grant 1976, Mogie 1986, de Meeûs et al. 2007, Vallejo-Marín et al. 2010, Nougué et al. 2015). To avoid confusion, and without any guarantee for its completeness or usefulness in other contexts, we will give a short overview of the framework used in this thesis.

The system in figure 3.1 is based on the effect of different reproductive modes on the genetic diversity of the resulting offspring, and secondarily on the developmental/ putative regulatory mechanisms involved. Per definition, asexual (synonymous to clonal) reproduction results in the offspring being identical to its parent except for mutations. In consequence, all reproductive modes that may involve some form of recombination, i.e. reassortment of chromosomes (typically, but not exclusively, by chromosome segregation during meiosis and subsequent fusion of gametes) and/or crossing-over (exchange of genetic material between chromosomes, genetic recombination in the narrow sense), are sexual. For some reproductive modes, e.g. central fusion (fusion of the products of meiosis I to reconstitute a diploid zygote-like cell), it may be difficult to ascertain whether or not crossing-over is possible; for others, such as post-meiotic duplication of the chromosomes, sexual reproduction may not involve an actual fusion of gametes (syngamy).

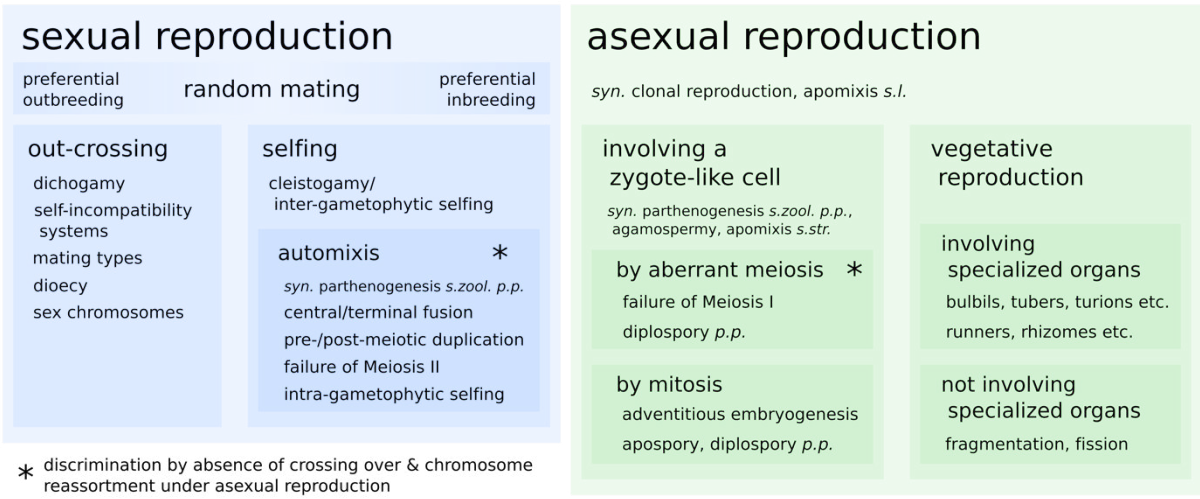

Figure 3.1 A system of reproductive modes. *syn.* synonymous to; *s.l.* = sensu lato, in the broad sense; *s.str.* = sensu stricto, in the narrow sense; *s.zool.* = sensu zoologico, in the zoological sense; *p.p.* = pro parte, in part

Asexual reproduction is sometimes treated as a form of growth, which especially suggests itself in cases of vegetative reproduction that occurs without any specialized structures.

However, in contrast to growth asexual reproduction may involve cell divisions other than "regular" mitosis and always results in (factually or potentially) physiologically independent copies of the whole parent organism, called ramets. Following common practice e.g. in human monozygotic twins, individuals are defined as ramets and not based on having unique genotypes (genets). A population is a group of conspecific individuals co-existing in space and time.

Having thus clarified what is meant by sexual and asexual reproduction, respectively, partial asexuality (synonymous to partial clonality) requires that these two be combined in the individual's life cycle. Similar to parallel and series connections in electric circuits, this may be achieved in two ways (compare figure 3.2): either the two modes of reproduction occur in parallel, or they succeed each other periodically. The second case, which typically involves several rounds of asexual reproduction that alternate with a single round of sexual reproduction, is generally referred to as cyclical parthenogenesis for historical reasons (first description in aphids: Bonnet 1745, Owen 1849). We will follow this nomenclature, though "cyclical asexuality" or "cyclical clonality" would be less ambivalent ("parthenogenesis" is used differently in a zoological vs. botanical context) and more appropriate according to our system. In contrast, the first case will be called acyclic partial asexuality, or just partial asexuality if the exclusion of the cyclical case is clear from the context.

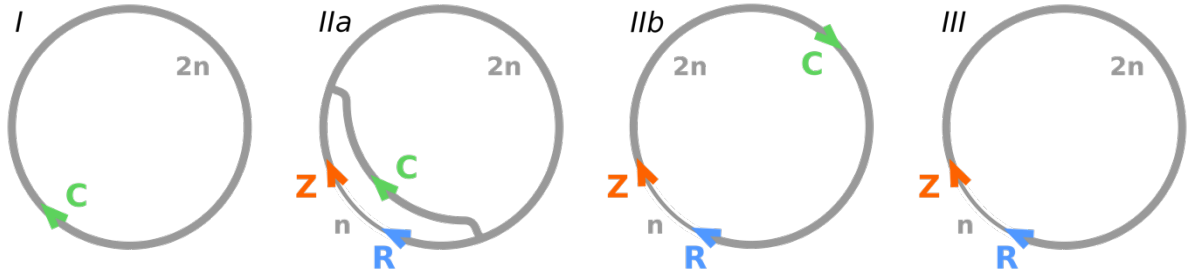

Figure 3.2 Asexual and sexual life cycles. I – exclusive asexuality; II – partial asexuality in the wide sense: IIa – acyclic partial asexuality or partial asexuality in the narrow sense, IIb – cyclical parthenogenesis; III – exclusive sexuality. Sexual reproduction: R – reduction (meiosis), Z – fertilization (syngamy). Asexual reproduction: C, which may represent several rounds of asexual reproduction under cyclical parthenogenesis.

Les systèmes reproducteurs – quelques repères

La vie sur terre utilise une diversité kaléidoscopique de systèmes de reproduction, et la recherche à leur sujet a produit une diversité presque égale des termes et des définitions parfois incongrues, voire incompatibles (voir par exemple Asher 1970, Grant 1976 Mogie 1986, de Meeûs et al. 2007, Vallejo-Marín et al. 2010, Nougué et al. 2015). Pour éviter toute confusion, et sans aucune garantie quant à l'exhaustivité ou l'utilité dans d'autres contextes, nous allons donner un bref aperçu du cadre utilisé dans cette thèse.

Le système de la figure 3.1 est basé sur l'effet des modes de reproduction différents selon la diversité génétique de la descendance, et secondairement selon les mécanismes impliqués du développement / de la régulation putative. Par définition, la reproduction asexuée (synonyme de clonale) mène à une progéniture identique à son parent à l'exception

des mutations. En conséquence, tous les modes de reproduction qui peuvent impliquer une certaine forme de recombinaison, que ce soit par réassortiment des chromosomes (généralement, mais pas exclusivement, par la ségrégation des chromosomes lors de la méiose et la fusion subséquente de gamètes) et / ou enjambement (échange de matériel génétique entre les chromosomes, recombinaison génétique au sens strict), sont sexués. Pour certains modes de reproduction, par exemple la fusion centrale (fusion des produits de la méiose I pour reconstituer une cellule diploïde pareille à un zygote), il peut être difficile de déterminer si l'enjambement est possible ; pour d'autres, comme la duplication post-méiotique des chromosomes, il est possible que la reproduction sexuée n'implique pas une fusion réelle des gamètes (syngamie).

La reproduction asexuée est parfois considérée comme une forme de croissance, ce qui est le plus évident dans les cas de la reproduction végétative qui se produit sans aucune structure spécialisée. Cependant, contrairement à la croissance, la reproduction asexuée peut entraîner des divisions cellulaires autres que la mitose « régulière » et se traduit toujours en copies (de fait ou potentiellement) physiologiquement indépendantes de l'organisme parent, appelées des « ramets ». Conformément à la pratique courante comme par exemple chez les jumeaux monozygotes humains, les individus sont définis comme ramets et non basés sur le fait d'avoir des génotypes uniques (« genets »). Une population est un groupe d'individus de la même espèce coexistant dans l'espace et le temps.

Ayant ainsi clarifié ce que l'on entend, respectivement, par la reproduction sexuée et asexuée, l'asexualité partielle (synonyme de clonalité partielle) signifie que les deux sont combinés dans le cycle de vie de l'individu. Par analogie avec des circuits électriques en parallèle et en série, cela peut être réalisé de deux façons (à comparer avec la figure 3.2) : soit les deux modes de reproduction se produisent en parallèle, soit ils se succèdent périodiquement. Dans le deuxième cas, cela implique généralement plusieurs cycles de reproduction asexuée qui alternent avec un seul tour de reproduction sexuée, et est généralement appelé « parthénogenèse cyclique » pour des raisons historiques (première description chez les pucerons : Bonnet 1745, Owen 1849). Nous suivrons cette nomenclature, même si « asexualité cyclique » ou « clonalité cyclique » seraient moins ambivalents (« parthénogenèse » est utilisé différemment dans un contexte botanique ou zoologique) et plus appropriés en fonction de notre système. En revanche, le premier cas sera appelé asexualité partielle acyclique, ou seulement asexualité partielle si l'exclusion du cas cyclique est évidente à cause du contexte.

3.2 Previous studies

3.2.1. Selectively neutral diversity

The first population genetic models for acyclic partially asexual species were primarily interested in its effect on population heterozygosity: Asher (1970) compared different forms of parthenogenesis in animals, but focused mainly on automixis and did not yet include combined sexual/asexual reproduction in their models. This was first done by Marshall & Weir (1979), who modeled the population genetic effect of combined agamospermy, selfing

and random mating in a population otherwise conforming to the Hardy-Weinberg assumptions, i.e. in the absence of mutation and genetic drift. They concluded that additional asexual reproduction has absolutely no effect on the equilibrium heterozygosity, which only depends on the relative rates of random mating and selfing, but could, however, considerably slow down the approach to this equilibrium. Probably because of this apparent indifference, no further models of partial asexuality were published for almost 15 years. Building directly on the results of Marshall & Weir (1979), Overath & Asmussen (2000a, b) wrote a model for the co-inheritance of "cytonuclear factors" (e.g. mitochondrial or plastid genes) under partial asexuality (even including tetraploidy), a line of research that apparently has not been further developed since.

With the new technique of coalescence models (see chapter 3.1) and by integrating demography and population genetics, Orive (1993) analyzed populations with complex life cycles including clonality. Based on sample life cycles parameterized from field data, she showed that partial clonality can lower the inbreeding effective population size (i.e. individuals are on average more closely related) in partially clonal species compared to their exclusively sexual counterparts. Also using a coalescence model, Bengtsson (2003) postulated that partially clonal reproduction should only change the patterns of genetic diversity under very high rates of clonality. If very old, such populations would have highly divergent allelic copies within the same individual, even if sexual reproduction occurred at low rates. In a second model for shorter time spans, ignoring mutation, he also found a "memory effect" for past genotypic diversity in partially asexual populations. Ceplitis (2003) challenged the correctness of Bengtsson's model, but nevertheless came to similar long-term results (only high rates of clonality have an effect, which is similar to exclusive clonality) and did not look at the short-term model again. Yonezawa et al. (2004) further developed effective population sizes under partial clonality.

The effect of partial clonality in finite-sized, subdivided populations with mutation on the population genetic parameters F_{IS} and F_{ST} was first studied by Balloux et al. (2003), the first in a series of related articles (including de Meeûs & Balloux 2005, de Meeûs et al. 2006). In contrast to the first results of Marshall & Weir (1979), they concluded that the equilibrium heterozygosity (and more specifically, F_{IS}) is affected by partial clonality, but – similar to the findings of Bengtsson (2003) and Ceplitis (2003) – only at nearly exclusive clonality. The results of research on expected patterns of genetic diversity under partial asexuality, and their potential for the estimation of rates of asexual reproduction from population genetic data, were subsequently summarized in Halkett et al. (2005).

Models for the population genetic effect of cyclical parthenogenesis are even rarer, and typically very much tailored towards the situation in particular species. In consequence, they may already involve additional parameters such as spatial substructure (i.e. several subpopulations connected by migration) or selfing, making comparison with other cases more difficult: Berg & Lascoux (2000) modeled sub-divided populations of daphnia in order to explain their differentiation from each other. Similarly, Prugnolle et al. (2005b, c) looked at cyclically parthenogenetic parasites (platyhelminths) but interpreted their results primarily in terms of selfing, migration and reproductive success (selection). The model for daphnia by Vanoverbeke & De Meester (2010) is somewhat different, focusing only on

genotype dynamics during a single asexual phase, but with (selective) environmental constraints. It demonstrated a pattern of “clonal erosion”, i.e. a successive loss of clonal lineages/genotypes.

La diversité sélectivement neutre

Les premiers modèles en génétique des populations pour des espèces partiellement asexuées acycliques s’intéressaient principalement à son effet sur l’hétérozygotie des populations : Asher (1970) a comparé différentes formes de la parthénogenèse chez les animaux, mais il s’est principalement occupé de l’autofécondation et n'a pas encore inclut la combinaison entre reproduction sexuée / asexuée dans son modèle. Cela a été fait en premier par Marshall & Weir (1979), qui ont modélisé l'effet en génétique de la population de la combinaison entre l’agamospermie, l’autofécondation et l'accouplement aléatoire pour une population qui était par ailleurs conforme aux hypothèses de Hardy-Weinberg, c’est-à-dire en l’absence de la mutation et de la dérive génétique. Ils ont conclu que la reproduction asexuée supplémentaire n'avait absolument aucun effet sur l'hétérozygotie d'équilibre, qui ne dépend que des taux relatifs d'accouplement aléatoire et d’autofécondation, mais qu’il pourrait toutefois ralentir considérablement l'approche de cet équilibre. Probablement à cause de cette indifférence apparente, aucun autre modèle de l'asexualité partielle n’a été publié pendant presque 15 ans. En se basant directement sur les résultats de Marshall & Weir (1979), Overath & Asmussen (2000a, b) ont écrit un modèle sur le co-héritage des « facteurs cyto-nucléaires » (par exemple des gènes mitochondriaux ou plastidiques) sous asexualité partielle (y compris même de la tétraploïdie), une ligne de recherche qui, apparemment, n'a pas encore été développée depuis.

Avec la nouvelle technique de modèles de coalescence (voir chapitre 3.1) et en liant la démographie à la génétique des populations, Orive (1993) a analysé les populations avec des cycles de vie complexes, y compris la clonalité. Basée sur les cycles de vie d’exemples paramétrés à partir de données de terrain, elle a montré que la clonalité partielle peut réduire la taille effective de consanguinité de la population (c’est-à-dire que les individus sont en moyenne plus étroitement liés) en espèces partiellement clonales par rapport à leurs homologues exclusivement sexués. Également à l'aide d'un modèle de coalescence, Bengtsson (2003) a affirmé que la reproduction partiellement clonale ne devait pas modifier les schémas de diversité génétique excepté avec des taux de clonalité très élevés. Si elles sont très vieilles, ces populations auraient des copies alléliques très divergentes dans le même individu, même si la reproduction sexuée se produit à des taux faibles. Dans un deuxième modèle, pour des périodes plus courtes, en ignorant la mutation, il a également trouvé un « effet mémoire » pour la diversité génotypique précédente dans des populations partiellement asexuées. Ceplitis (2003) a contesté l’exactitude du modèle de Bengtsson, mais néanmoins il est arrivé à des résultats similaires à long terme (seuls les taux élevés de clonalité ont un effet, qui est similaire à la clonalité exclusive) et il n’a pas regardé le modèle à court terme de nouveau. Yonezawa et al. (2004) ont développé des tailles de population efficaces sous clonalité partielle.

L'effet de la clonalité partielle en population de taille déterminée, subdivisée et avec de la mutation, sur les paramètres de génétique des populations F_{IS} et F_{ST} a d'abord été étudié

par Balloux et al. (2003), le premier d'une série d'articles connexes (y compris de Meeûs & Balloux 2005, de Meeûs et al. 2006). Contrairement aux premiers résultats de Marshall & Weir (1979), ils ont conclu que l'hétérozygotie d'équilibre (et plus précisément, F_{IS}) est affectée par une clonalité partielle, mais - similaire aux conclusions de Bengtsson (2003) et Ceplitis (2003) - seulement à clonalité presque exclusive. Les résultats de la recherche sur les tendances attendues de la diversité génétique sous asexualité partielle, et leur potentiel pour l'estimation des taux de reproduction asexuée à partir de données génétiques de la population, ont ensuite été résumés dans Halkett et al. (2005).

Les modèles pour l'effet sur la génétique des populations de la parthénogenèse cyclique sont encore plus rares, et typiquement très spécialisés à la situation des espèces particulières. En conséquence, ils peuvent déjà contenir des paramètres supplémentaires, tels que sous-structure spatiale (c'est à dire plusieurs sous-populations liées par migration) ou l'autofécondation, rendant plus difficile la comparaison avec d'autres cas : Berg & Lascoux (2000) ont modélisés des populations des daphnies sous-divisées pour expliquer leur la différenciation. De même, Prugnolle et al. (2005b, c) ont regardé des parasites cycliquement parthénogénétiques (de plathelminthes) mais ont interprété leurs résultats principalement en termes d'autofécondation, migration et succès de la reproduction (sélection). Le modèle pour les daphnies par Vanoverbeke & De Meester (2010) est quelque peu différent, en se concentrant uniquement sur la dynamique des génotypes pendant une phase asexuée simple, mais avec des contraintes environnementales (sélectifs). Il a montré un modèle de « l'érosion clonale », c'est-à-dire une perte successive de lignées clonales / génotypes.

3.2.2. Diversity under selection

Selection in partially asexual species has received some attention in the context of the evolution of sex, although most models for this question only compared exclusive sexuality and exclusive asexuality directly. Modifier models, where the rate of asexual reproduction is not constant, but in fact controlled by each individual's genotype at a single locus, play a great role in this respect (Marshall & Brown 1981, Roze 2009, Roze & Michod 2010). Another group of models looked at the fitness of multilocus genotypes assuming infinite population size and number of loci (no back mutation), or sexual reproduction partially or exclusively by selfing (Muirhead & Lande 1997, their model also includes cyclic parthenogenesis; Masel & Lyttle 2011, Marriage & Orive 2012). Models for single loci typically ignore mutation (but see Lokki 1976), look only at a specific selection scenario (but see Overath & Asmussen 1998) and are also focused on infinite populations (Marshall & Weir 1979). Ryndin et al. (2001) proposed a model for two (partially) linked loci in an infinite population without mutation under dynamically changing selection. However, a "simple" single locus model, describing the expected genotype frequency patterns (and their dynamics) under various selective scenarios for a finite population with mutation still seems to be missing.

La diversité sous sélection

La sélection chez les espèces partiellement asexuées a reçu une certaine attention dans le contexte de l'évolution du sexe, bien que la plupart des modèles de cette question ont

seulement comparé la sexualité exclusive à l'asexualité exclusive en direct. Des modèles avec des modificateurs, où le taux de reproduction asexuée n'est pas constant, mais en fait contrôlé par le génotype de chaque individu à un seul locus, jouent un grand rôle à cet égard (Marshall & Brown 1981, Roze 2009, Roze & Michod 2010). Un autre groupe de modèles a observé la valeur sélective des génotypes à multiples loci en supposant une taille infinie de la population et un nombre de loci infinie (pas de mutation de retour), ou la reproduction sexuée se fait partiellement ou exclusivement par autofécondation (Muirhead & Lande 1997, leur modèle comprend également la parthénogenèse cyclique; Masel & Lyttle 2011, Mariage & Orive 2012). Des modèles pour des loci uniques ignorent généralement la mutation (mais voir Lokki 1976), ne regardent seulement qu'un scénario de sélection spécifique (mais voir Overath & Asmussen 1998) et ne mettent également l'accent que sur les populations infinies (Marshall & Weir 1979). Ryndin et al. (2001) ont proposé un modèle pour deux loci (partiellement) liées dans une population infinie sans mutation sous un régime de sélection qui se change dynamiquement. Cependant, un modèle « simple » d'un seul locus, décrivant les résultats attendus pour les fréquences des génotypes (et leur dynamique) selon divers scénarios sélectifs pour une population finie avec de la mutation semble être encore à faire.

4 Approach

The first step towards the description of the patterns and dynamics of genotype frequencies under partial asexuality is the development of a suitable mathematical model. The basis of the model we will use here was already established before (Stoeckel & Masson 2014); in the course of this thesis, the model equations were rearranged and considerably extended (cyclical parthenogenesis, multiple alleles, selection). The model is discussed in detail in the next chapter.

Our model provides a lot of detail, but also produces a lot of data; to cope with this situation, we developed new techniques for the analysis and visualization of the model results, which are included in the first article:

Article I: ***Interpretation and approximation tools for big, dense Markov chain transition matrices in population genetics***

The core of our population genetic model is a large, square matrix that contains information on both the static (final, equilibrium) and the dynamic (short-term) behavior of the modeled genotype frequencies. Based on a specialized ternary (i.e. triangular) plot for population genetic data, called *de Finetti* diagram after its inventor, we developed concise and comprehensive methods to display the different aspects of the model results. In addition, we developed a method by which very big and memory intensive matrices can be reduced in size independently of their content, using an approximation algorithm that keeps their mathematical properties.

The first article is followed by a short subchapter, which includes additional information about *de Finetti* diagrams and their potential as a teaching tool.

The following part of the thesis presents the results from the analysis of our mathematical model. We first focused on evolution under neutral conditions, as including selection would be an extension of this case and best studied in comparison with it. Our findings constitute the second article:

Article II: ***Rare sex or out of reach equilibrium? The dynamics of F_{IS} in partially clonal organisms***

We modeled the neutral dynamics of genotype frequencies at a single locus over time. Compared to the exclusively sexual case, partial asexuality mainly affected the dynamics of heterozygosity within the population, which we described using the population genetic parameter $\mathrm{F_{IS}}$. We found that $\mathrm{F_{IS}}$ increases its variation and needs longer to return to its equilibrium mean value under partial asexuality.

The second article is followed by a subchapter that presents first results for multilocus simulations of a bottleneck effect in partially asexual species.

In the course of preparing article II, we realized that its results would not be directly applicable to populations undergoing cyclical parthenogenesis. However, cyclical parthenogenesis is of special interest to the work group (studying aphids) in which this

thesis was prepared. Therefore the next step was to compare neutral genotype/ F_{IS} dynamics under exclusively asexual, exclusively sexual and both kinds of partially asexual reproduction; the results form the next article:

Article III: ***Effects of complex life-cycles on genetic diversity: The case of cyclical parthenogenesis***

We compared the expected equilibrium distributions of F_{IS} for four different reproductive systems, i.e. exclusive sexuality, exclusive clonality, acyclic partial clonality and cyclical parthenogenesis. For the latter case, we also took into account different numbers clonal generations and different sampling strategies (before sexual phase, after sexual phase). The results show that each case is distinct.

The second part of the results deals with the genetic diversity at single loci under selection; again, acyclic partial asexuality is assumed to obtain the results presented in the next article:

Article IV: ***Partial clonality and the speed of adaptation***

We modeled genotype frequencies in partially clonal populations at loci under selection according to four different selection scenarios. The results show that partial clonality is hardly ever optimal for adaptation based on single loci, both in terms of speed and final mean population fitness.

This manuscript is followed by a final subchapter that includes first steps towards an extension of the model to two loci.

The thesis ends with a global discussion and conclusion.

Approche

La première étape vers la description des caractéristiques et de la dynamique des fréquences génotypiques sous asexualité partielle est l'élaboration d'un modèle mathématique approprié. La base du modèle que nous utiliserons ici a déjà été établie précédemment (Stoeckel & Masson 2014); dans le cadre de cette thèse, les équations du modèle ont été réorganisées et considérablement étendues (parthénogenèse cyclique, allèles multiples, la sélection). Le modèle est présenté en détail dans le chapitre suivant.

Notre modèle fournit beaucoup de détails, mais produit également un grand nombre de données ; pour faire face à cette situation, nous avons développé de nouvelles techniques pour l'analyse et la visualisation des résultats des modèles, qui sont inclus dans le premier article :

Article I : ***Outils pour l'interprétation et l'approximation des matrices de transition grandes et denses des chaînes de Markov en génétique des populations***

Le cœur de notre modèle de génétique des populations est une grande matrice carrée, qui contient des informations à la fois sur le comportement statique (finale, d'équilibrium) et dynamique (à court terme) des fréquences génotypiques modélisées. À partir d'un diagramme ternaire spécialisé (c'est-à-dire triangulaire) adapté à des données de génétique des populations, appelé diagrammes de *de Finetti* d'après son inventeur, nous avons

développé des méthodes concises et copieuses pour exposer les différents aspects des résultats du modèle. En outre, nous avons développé une méthode par laquelle des matrices très grandes qui occupent beaucoup de mémoire peuvent être réduites en taille indépendamment de leur contenu, en utilisant un algorithme d'approximation qui maintient leurs propriétés mathématiques.

Le premier article est suivi d'une courte section de chapitre, qui comprend des informations supplémentaires sur les diagrammes de *de Finetti* et leur potentiel en tant qu'outil d'enseignement.

La partie suivante de la thèse présente les résultats de l'analyse de notre modèle mathématique. Nous avons d'abord mis l'accent sur l'évolution dans des conditions neutres, comme la sélection serait une extension de ce cas et étudié le meilleur en comparaison avec elle. Nos résultats constituent le deuxième article :

Article II : ***Sexe rare ou hors portée de l'équilibre ? La dynamique d'F_{IS} chez les organismes partiellement clonaux***

Nous avons modélisé la dynamique neutre des fréquences génotypiques à un seul locus au fil du temps. Par rapport au cas exclusivement sexué, l'asexualité partielle a principalement affecté la dynamique de l'hétérozygotie de la population, ce que nous avons décrits en utilisant le paramètre de génétique des populations F_{IS}. Nous avons constaté que le F_{IS} augmente sa variation et prend plus de temps pour revenir à sa valeur moyenne d'équilibre sous asexualité partielle.

Le deuxième article est suivi d'un sous-chapitre qui présente les premiers résultats des simulations à multiples loci avec un effet de goulot d'étranglement chez les espèces partiellement asexuées.

Au cours de la préparation de l'article II, nous avons réalisé que ces résultats ne seraient pas directement applicables à des populations subissant une parthénogenèse cyclique. Cependant, la parthénogenèse cyclique est d'un intérêt particulier pour le groupe de travail (qui fait des études sur les pucerons) dans lequel cette thèse a été préparée. Par conséquent, la prochaine étape était de comparer la dynamique neutre des génotypes / F_{IS} sous asexualité exclusive, sexualité exclusive et les deux types de reproduction partiellement asexuée ; les résultats forment le prochain article :

Article III : ***Effets des cycles de vie complexes sur la diversité génétique : le cas de la parthénogenèse cyclique***

Nous avons comparé les distributions d'équilibre attendus du F_{IS} pour quatre systèmes de reproduction différents, c'est-à-dire la sexualité exclusive, la clonalité exclusive, la clonalité partielle acyclique et la parthénogenèse cyclique. Pour ce dernier cas, nous avons également pris en compte les différents nombres de générations clonales et les différentes stratégies d'échantillonnage (avant la phase sexuée, après la phase sexuée). Les résultats montrent que chaque cas est différent.

La deuxième partie des résultats traite de la diversité génétique à un seul locus sous sélection ; nous supposons encore l'asexualité partielle acyclique pour obtenir les résultats présentés dans l'article suivant :

Article IV : La clonalité partielle et la vitesse d'adaptation

Nous avons modélisé les fréquences des génotypes auprès des populations partiellement clonales à des loci sous sélection suivant l'un des quatre scénarios de sélection différents. Les résultats montrent que la clonalité partielle est rarement optimale pour l'adaptation basée sur des loci uniques, à la fois en termes de vitesse et de valeur sélective moyenne de la population au final.

Ce manuscrit est suivi d'un sous-chapitre final qui comprend un premier pas vers une extension du modèle à deux loci.

La thèse se termine par une discussion et une conclusion globale.

Part II Methods

5 Mathematical model

5.1 Model choice

Mathematical models come in many different types; which one to choose depends on the nature of the system that is to be modeled (Otto & Day 2007). The basic dichotomies between different types of models (figure 5.1) are whether the range of possible values for the modeled variables (including time) is discrete ($\in \mathbb{N}$ or $\mathbb{Z}$) or continuous ($\in \mathbb{R}$), and whether the value of the variables will be fully predictable (deterministic model) or to some extent random (stochastic model). The decision usually depends on the scale at which the natural system is analyzed: though most processes in nature are stochastic, they may appear deterministic at a higher scale because the emergent behavior is driven by the mean over many instances of the random lower-scale process. An example from population genetics would be the Hardy-Weinberg equilibrium: though each individual is assumed to mate randomly, the genotype and allele frequencies within the whole population remain constant. Deterministic models are often easier to treat mathematically, but may produce inaccurate results if applied to inherently stochastic systems.

	state & time discrete	state & time continuous
deterministic	*recurrence equation*	*ordinary/partial differential equation*
stochastic	*Markov chain, cellular automaton*	*stochastic differential equation*

Figure 5.1 Schematic overview of different types mathematical models, with examples.

Mathematical models in population genetics typically belong to either of two main groups: so-called "classic" models, which directly describe allele/genotype frequencies or derived quantities (e.g. the fraction of heterozygous genotypes or probabilities of identity between alleles) on a chronological time scale, or coalescent models. Classic models based on genotype frequencies are usually intuitive to construct, but may become very complex: for example, if the frequencies of all diploid genotypes that are possible based on a stretch of 100 DNA base pairs should be modeled, the model would have 10^{100} variables. Therefore, classic population genetic models often look at some sort of subsystem, such as single loci with limited allelic diversity (e.g. Hardy 1908, Weinberg 1908, Wright 1921, de Finetti 1926, 1927), or use some simplification, such as modeling randomly mating, exclusively sexual populations based on allele frequencies rather than genotype frequencies (since the latter can be derived from the Hardy-Weinberg equilibrium; Ewens 2004). Classic models for small, finite populations typically use Markov chain models (more details below), which can be approximated by a diffusion equation (stochastic differential equation, e.g. Ohta & Kimura 1969) if the population size is big, or even by a deterministic model (e.g. Hardy 1908) if the population size (again as a simplification) is assumed to be infinite (Kimura 1964).

Coalescent models (Kingman 1982a, b; Fu & Li 1999) avoid the complexity that limits classic population genetic models by not describing allele (or genotype) frequencies directly, but

concentrating only on those alleles – typically whole DNA sequences (haplotypes) – currently observed to reconstruct their hypothetical genealogy. This reconstruction is based on the paradigm that all alleles in a population are derived from a common ancestor (a concept similar to monophyly/paraphyly in biological systematics); "past" alleles who do not have descendants in the current population can thus be ignored (figure 5.2). Though the branching of different allelic lineages, on a reversed time scale equivalent to their merging or "coalescence", is *per se* independent of mutation, infinite alleles / infinite loci mutation models play a large role in coalescence theory: assuming that each mutation creates a new allele (no back mutation) allows to distinguish a maximal number of allelic lineages and simplifies the derivation of analytical results. These results primarily describe populations in terms of time ("coalescence time" since the universal most recent common ancestor), by a reference population size and measures of allelic (lineage) diversity. Mathematically, coalescence theory typically uses stochastic models with discrete states but continuous time (i.e. Markov processes, Poisson process).

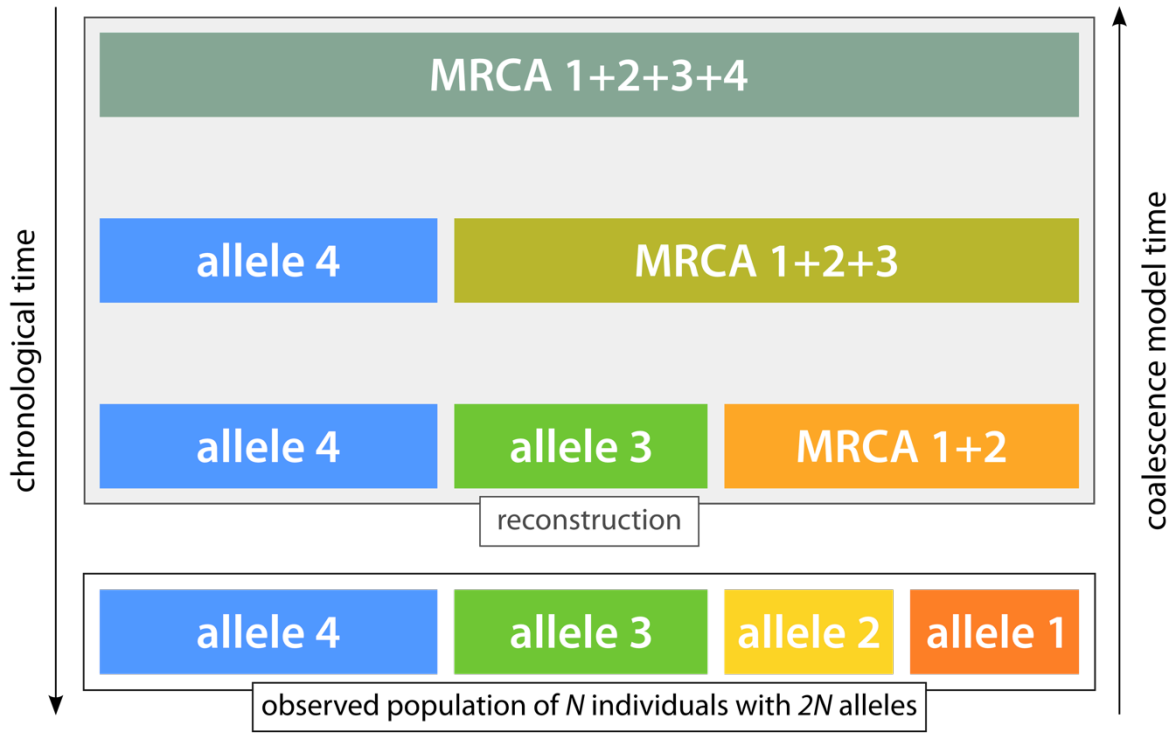

Figure 5.2 Illustration of Kingman's paint box analogy for coalescence (Kingman 1982a, b), with four allelic lineages distinguished by different alleles. MRCA: most recent common ancestor.

This thesis uses a classic model of population genetics. Firstly, this is because the biological system we intend to model is less easily accessible with coalescence theory: individual SNP/microsatellite loci have only a limited allelic diversity (e.g. at most four alleles in the case of an SNP), and back mutations are known to occur (Hile et al. 2000, Estoup et al. 2002). Deriving analytical results with mutation among a finite number of alleles is easier in a time-forward classical model; in coalescence theory, such mutation schemes are as yet only accessible by simulation (Wakeley 2009). Secondly, time-forward models can provide probabilistic predictions for directly observable quantities such as genotype frequencies,

which is especially practical for biological questions relating to a population's future rather than its past, e.g. for studying adaptive processes or conservation genetics. Thirdly, since not much is known about the population genetics of partially asexual species in general, a classic population genetic model describing genotype frequencies can serve as a reference to check whether the results of other, more abstract models are consistent.

5.2 Model assumptions and structure

In terms of biology, our model describes the genotype frequencies within a single, isolated population with a finite, small constant number of individuals (ramets; see chapter 3.1). The individuals are monecious and diploid, and their genotype is based on a single genomic locus with two or more alleles. In the course of one generation, individuals may acquire potentially heritable mutations, usually with a symmetric mutation rate turning the current allele into any other. Then, offspring is produced either exclusively sexually by random mating (including selfing), exclusively asexually, or both in a set proportion (rate of asexuality), from which the requisite number is chosen to replace their parents at the next generation. If natural selection occurs, it mostly affects the reproductive success of the parent generation. This life cycle is schematically represented in figure 5.3. Cyclical parthenogenesis corresponds to several rounds of this cycle, each with either exclusively sexual or exclusively asexual reproduction.

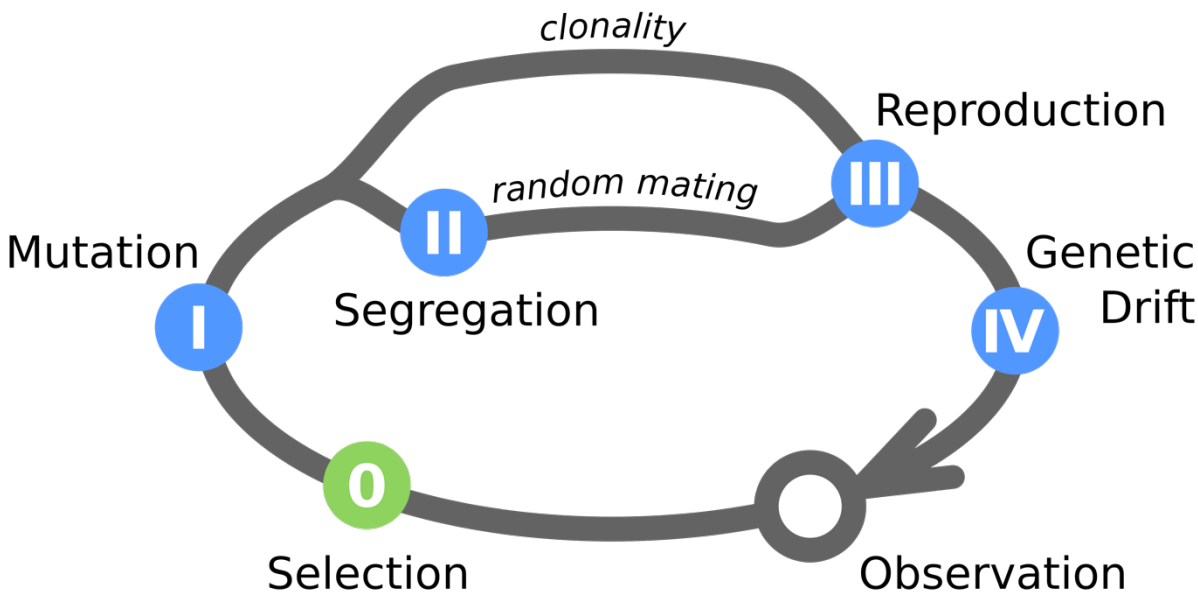

Figure 5.3 Schematic representation of the model life cycle. Step 0 (selection) is not included in the selectively neutral case; cyclic parthenogenesis corresponds to several exclusively asexual/sexual cycles.

In terms of population genetics, our model is a "Wright-Fisher model": there is no survival between generations (in contrast to the Moran model), and the potential number of offspring per individual is unlimited (in contrast to a general Cannings model; Ewens 2004,

compare Der et al. 2011 for a critical appraisal). These two assumptions primarily ease the mathematical description. Survival between generations could be easily introduced by a "survival rate", which keeps a proportion of the offspring at each generation exactly identical (i.e. no mutation) to its "parent", i.e. itself. To take limits in the production of sexual (e.g. number of seeds per plant) and asexual offspring (e.g. number of runners per plant) into account would increase computation time, as the multinomial distribution in our model (chapter 5.3) would have to be replaced by a more complicated hypergeometric distribution. We chose the Wright-Fisher model to start with the most "basic" approach, without introducing additional model parameters (survival rate, maximal number of descendants).

In terms of mathematics, our model is a time and state discrete Markov chain. Markov chains are sequences of stochastic "experiments": As an example (figure 5.4B), imagine a set of six unequally "loaded" dice, labeled one to six. To start the chain, the first die (e.g. number one) is rolled, and the next die is determined by the number which turns up – which might be number one again, which this time turns up a three so that the die is changed, and so on until we stop. The sequence of dice would be the sequence of states of our Markov chain, and the probabilities with which each die turns up different numbers would be its transition probabilities. In the case of our model, the states are different combinations of counts for each genotype – e.g. one individual *aa*, three individuals *aA* and six individuals *AA* (figure 5.4C) – and the transition probabilities between different generations (or observations) are determined by the model parameters N (population size), μ (mutation rate), c (rate of clonality, or fraction of asexual generations per cycle under cyclical parthenogenesis) and, if applicable, s (selection coefficient).

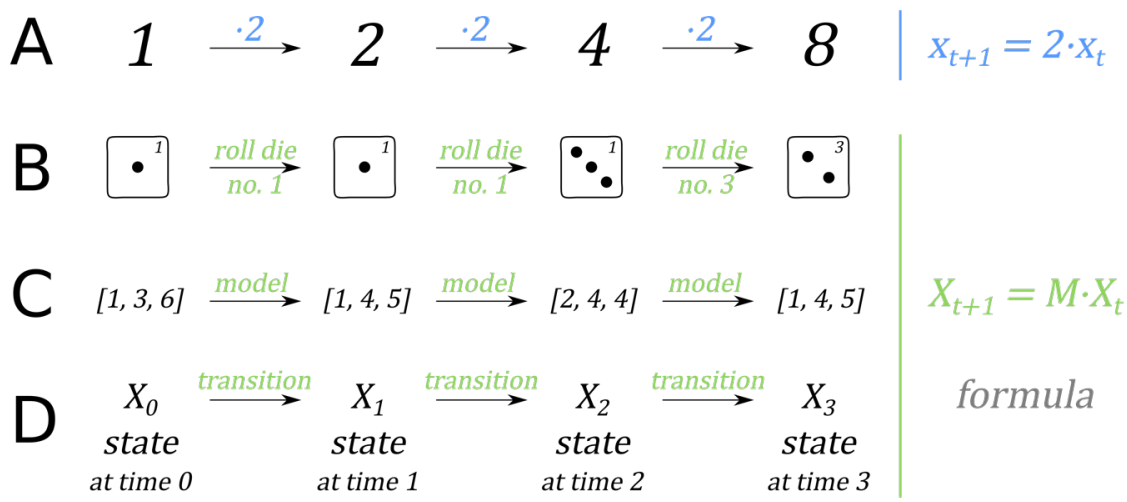

Figure 5.4 Markov chains as sequences of stochastic experiments. A: a geometric progression (a sequence), B: six loaded dice (a Markov chain), C: our model (another Markov chain), D: general form of a Markov chain. M – transition matrix.

5.3 Model equations

We modeled evolution in partially asexual populations by a Markov chain (Markov 1906). Markov chains are defined by their state space S and the pairwise transition probabilities $p(X_{t+1}|X_t)$ between all states X. Four our model, the state space depends on the genetic system that is modeled – number of loci $\mathcal{L}$, number of alleles $\mathcal{A}$, ploidy $\mathcal{P}$, population size N – and the transition probabilities depend on the model parameters – rate of clonality, mutation rate, population size etc.

The state space of our Markov chain consists of all possible combinations of counts $X = (q_{ii}, q_{ij}, \dots)$ with $\sum_i q_{ii} + \sum_{i,j} q_{ij} = N$ for each genotype possible in the modeled genetic system. In general, the number of possible genotypes g is:

$$[1a] \quad g = \prod_{i=1}^{\mathcal{L}} \left\langle \begin{matrix} \mathcal{A}_i \\ \mathcal{P} \end{matrix} \right\rangle = \prod_{i=1}^{\mathcal{L}} \frac{(\mathcal{A}_i + \mathcal{P} - 1)!}{\mathcal{P}! \cdot (\mathcal{A}_i - 1)!}.$$

However, for systems with multiple loci, it is necessary to treat each haplotype as a "composite" allele so that crossing-over can be correctly described, leading to the alternative equation:

$$[1b] \quad g = \left\langle \begin{matrix} \prod_{i=1}^{\mathcal{L}} \mathcal{A}_i \\ \mathcal{P} \end{matrix} \right\rangle = \frac{\left(\prod_{i=1}^{\mathcal{L}} \mathcal{A}_i + \mathcal{P} - 1\right)!}{\mathcal{P}! \cdot \left(\prod_{i=1}^{\mathcal{L}} \mathcal{A}_i - 1\right)!}.$$

In both cases, the number of states |S| corresponds to:

$$[2] \quad |S| = \left\langle \begin{matrix} g \\ N \end{matrix} \right\rangle = \frac{(N + g - 1)!}{N! \cdot (g - 1)!}.$$

As an example, a single locus with two possible alleles {*a, A*} gives rise to three different genotypes {*aa, aA, AA*}, and a "population" of one single individual would thus have three model states to choose from: $S = \{(1,0,0), (0,1,0), (0,0,1)\}$. For two individuals, there are already six possible states in our model: $S = \{(2,0,0), (1,1,0), (0,2,0), (1,0,1), (0,1,1), (0,0,2)\}$. Dividing the combination of counts by the population size gives the corresponding vector of genotype frequencies $\vec{\nu} = [\nu_{aa}, \nu_{aA}, \nu_{AA}]$.

For each state, one can calculate different population genetic parameters such as the frequency of individual alleles ($\nu_a = 1 - \nu_A$), the frequency of heterozygous genotypes ("heterozygosity", H), the mean fitness of the population ($\overline{\Phi}$, based on the genotype fitness values $\varphi_{aa}, \varphi_{aA}, \varphi_{AA}$) or F_{IS}. In the first three cases, this is simply done by multiplying the frequency of each genotype by its contribution to the parameter, e.g.:

$$[3a] \quad \nu_a = \begin{bmatrix} 1 \\ 0.5 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix};$$

$$[3b] \quad H = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix};$$

[3c] $\overline{\Phi} = \begin{bmatrix} \varphi_{aa} \\ \varphi_{aA} \\ \varphi_{AA} \end{bmatrix} \cdot \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}.$

In contrast, the value of F_{IS}, a parameter that describes the relationship between observed and expected (assuming the Hardy-Weinberg equilibrium) heterozygosity (H_o, H_e), is given by the more complex expression:

[3d] $F_{IS} = 1 - \frac{H_o}{H_e} = 1 - \frac{\nu_{aA}}{2\nu_a\nu_A}.$

The transition probabilities $p(X_{t+1}|X_t)$ of our model are based on a multinomial distribution $\mathcal{M}$. In the simplest case, one could imagine that, at each time step, N individuals are randomly chosen (with replacement) to produce one clonal descendant without mutation. In this case, the probability of each genotype in the offspring generation would depend only on its frequency in the parent generation:

[4a] $X_{t+1} \sim \mathcal{M}(N, \vec{\nu}_t)$

and consequently (subscripts ii/ij denoting different genotypes):

[4b] $p(X_{t+1}|X_t) = \frac{N!}{\prod_i (q_{ii,t+1})! \cdot \prod_{i,j} (q_{ij,t+1})!} \cdot \prod_i \nu_{ii,t}^{q_{ii,t+1}} \cdot \prod_{i,j} \nu_{ij,t}^{q_{ij,t+1}}.$

If we integrate other evolutionary processes into the model, the probability of each genotype in the offspring generation changes. The vector of genotype probabilities, for which we had entered $\vec{\nu}_t$, will be transformed according to the action of each evolutionary process: one could imagine that the parent population first gives rise to a "virtual" infinite-sized offspring population, where the frequency of each genotype corresponds to its probability, from which N individuals are then chosen randomly to become the "real" offspring. To determine the effect of each evolutionary process on the genotype probabilities, we will therefore treat these probabilities like genotype frequencies in an infinite population. The order of the transformations corresponds to the order of evolutionary processes in our model as presented in figure 5.3.

0 Selection

To model selection, all genotype frequencies are multiplied by their fitness value, and the resulting vector is rescaled to sum to one:

[5a] $\begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_0 = \frac{1}{\varphi_{aa}\nu_{aa} + \varphi_{aA}\nu_{aA} + \varphi_{AA}\nu_{AA}} \begin{bmatrix} \varphi_{aa}\nu_{aa} \\ \varphi_{aA}\nu_{aA} \\ \varphi_{AA}\nu_{AA} \end{bmatrix}_t$

This generalizes to:

[5b] $\overrightarrow{\nu_0} = (\overrightarrow{\varphi} \cdot \overrightarrow{\nu_t})^{-1} (\overrightarrow{\varphi} \circ \overrightarrow{\nu_t})$

where $\cdot$ denotes the dot product and $\circ$ the Hadamard (element-wise) product of the vectors.

I Mutation

Between two observations, each allele mutates with rate μ, and does not mutate with rate $1-\mu$. For the basic case of one locus with two alleles, this gives:

$$[6a] \quad \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_I = \begin{bmatrix} (1-\mu)^2 & \mu(1-\mu) & \mu^2 \\ 2\mu(1-\mu) & \mu^2+(1-\mu)^2 & 2\mu(1-\mu) \\ \mu^2 & \mu(1-\mu) & (1-\mu)^2 \end{bmatrix} \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_0$$

For more alleles (k-alleles/Jukes-Cantor mutation model), let $\alpha = 1-\mu$ and $\beta = \frac{\mu}{n-1}$ (probability to mutate into one of n alleles), so that this generalizes to:

$$[6b] \quad \begin{cases} \nu_{ii,I} = & \alpha^2 \cdot \nu_{ii,0} + \sum_j \nu_{jj,0} \cdot \beta^2 + \sum_j \nu_{ij,0} \cdot \alpha\beta + \sum_{j,k} \nu_{jk,0} \cdot \beta^2 \\ \nu_{ij,I} = & 2\alpha\beta \cdot (\nu_{ii,0} + \nu_{jj,0}) + 2\beta^2 \sum_k \nu_{kk,0} + (\alpha^2+\beta^2) \cdot \nu_{ij,0} \\ & + (\alpha\beta+\beta^2) \cdot \sum_{k,l} (\nu_{ik,0} + \nu_{jl,0}) + 2\beta^2 \cdot \sum_{k,l} \nu_{kl,0} \end{cases}$$

Asymmetric mutation rates are also possible – here, $n-1$ of n alleles were "lumped" into one to reduce computational effort, giving two alleles *A* and *a* = "*not A*":

$$[6c] \quad \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_I = \begin{bmatrix} \left(1-\frac{\mu}{n-1}\right)^2 & \mu\left(1-\frac{\mu}{n-1}\right) & \mu^2 \\ \left(\frac{2\mu}{n-1}\right)\left(1-\frac{\mu}{n-1}\right) & \mu\left(\frac{\mu}{n-1}\right)+(1-\mu)\left(1-\frac{\mu}{n-1}\right) & 2\mu(1-\mu) \\ \left(\frac{\mu}{n-1}\right)^2 & (1-\mu)\left(\frac{\mu}{n-1}\right) & (1-\mu)^2 \end{bmatrix} \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_0$$

For a two-locus model, the entries in the mutation matrix are the product of the corresponding entries of the mutation matrices for each individual locus. The vector resulting from multiplication with the mutation matrix, $\vec{\nu}_{I^{r=0}}$, is then multiplied with another matrix to model crossing-over at rate r between the two loci in the two double heterozygote genotypes:

$$[6d] \quad \begin{bmatrix} \nu_{aB/Ab} \\ \nu_{ab/AB} \\ \nu_{ab/ab} \\ \vdots \end{bmatrix}_{I^{r=r}} = \begin{bmatrix} 1-r & r & 0 & \cdots \\ r & 1-r & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \nu_{aB/Ab} \\ \nu_{ab/AB} \\ \nu_{ab/ab} \\ \vdots \end{bmatrix}_{I^{r=0}}$$

II Allele segregation / Gamete formation

The frequencies of haploid gamete genotypes simply correspond to the allele frequencies:

$$[7a] \quad \begin{bmatrix} \nu_a \\ \nu_A \end{bmatrix}_{II} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0.5 & 1 \end{bmatrix} \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_I$$

This generalizes to:

$$[7b] \quad \nu_{i,II} = \nu_{ii,I} + 0.5 \sum_j \nu_{ij,I}$$

for all alleles.

III Reproduction

Under acyclic partial asexuality, a proportion c of the offspring is produced clonally and the rest by random mating; for cyclical parthenogenesis, c is either zero (exclusively sexual reproduction) or one (exclusively asexual reproduction). For random mating, the genotype frequencies conform to the Hardy-Weinberg equilibrium based on the allele/gamete frequencies after mutation; for clonal reproduction, the genotype frequencies are only affected by mutation:

$$[8a] \quad \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_{III} = c \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_{I} + (1-c) \begin{bmatrix} \nu_a^2 \\ 2\nu_a\nu_A \\ \nu_A^2 \end{bmatrix}_{II}$$

This generalizes to:

$$[8b] \quad \begin{cases} \nu_{ii,t+1} = & c\nu_{ii,I} + (1-c)\nu_{i,II}^2 \\ \nu_{ij,t+1} = & c\nu_{ij,I} + (1-c)2\nu_{i,II}\nu_{j,II} \end{cases}$$

IV Genetic drift

In analogy to the initial equations [4a], we can now write the final step of the model:

$$[9a] \quad X_{t+1} \sim \mathcal{M}(N, [\nu_{aa}, \nu_{aA}, \nu_{AA}]_{III}),$$

or in general

$$[9b] \quad X_{t+1} \sim \mathcal{M}(N, \vec{\nu}_{III}),$$

which corresponds to the transition probabilities:

$$[10a] \quad p(X_{t+1}|X_t) = \frac{N!}{q_{aa,t+1}! \cdot q_{aA,t+1}! \cdot q_{AA,t+1}!} \cdot \nu_{aa,III}^{q_{aa,t+1}} \cdot \nu_{aA,III}^{q_{aA,t+1}} \cdot \nu_{AA,III}^{q_{AA,t+1}}$$

and

$$[10b] \quad p(X_{t+1}|X_t) = \frac{N!}{\prod_i (q_{ii,t+1})! \cdot \prod_{i,j} (q_{ij,t+1})!} \cdot \prod_i \nu_{ii,III}^{q_{ii,t+1}} \cdot \prod_{i,j} \nu_{ij,III}^{q_{ij,t+1}}.$$

The model description of Stoeckel & Masson (2014) for a single biallelic locus gives exactly the same results as equations [6]-[10], as the equations are only rearranged:

$$(1) \quad \begin{cases} p_{aa}^{n+1} = & (1-\mu)^2 p_{aa}^n + \mu(1-\mu)p_{aA}^n + \mu^2 p_{AA}^n \\ p_{aA}^{n+1} = & 2\mu(1-\mu)p_{aa}^n + [\mu^2 + (1-\mu)^2]p_{aA}^n + 2\mu(1-\mu)p_{AA}^n \\ p_{AA}^{n+1} = & (1-\mu)^2 p_{AA}^n + \mu(1-\mu)p_{aA}^n + \mu^2 p_{aa}^n \end{cases}$$

This corresponds to equation [6a].

$$(2) \quad \begin{cases} q_{aa}^{n+1} = & [(1-\mu)p_{aa}^n + \mu p_{AA}^n + 0.5p_{aA}^n]^2 \\ q_{aA}^{n+1} = & 2[(1-\mu)p_{aa}^n + \mu p_{AA}^n + 0.5p_{aA}^n][(1-\mu)p_{AA}^n + \mu p_{aa}^n + 0.5p_{aA}^n] \\ q_{AA}^{n+1} = & [(1-\mu)p_{AA}^n + \mu p_{aa}^n + 0.5p_{aA}^n]^2 \end{cases}$$

This corresponds to equations [6a] and [7a] inserted into the last vector in equation [8a], so that

$$(3) \quad \pi_{ij}^{n+1} = cp_{ij}^n + (1-c)q_{ij}^n$$

corresponds to the full equation [8a], and their equation:

$$(4) \quad P(s_{aa}, s_{aA}, s_{AA} | r_{aa}, r_{aA}, r_{AA}) = \frac{N!}{s_{aa}! \cdot s_{aA}! \cdot s_{AA}!} \cdot (\pi_{aa}^{n+1})^{s_{aa}} \cdot (\pi_{aA}^{n+1})^{s_{aA}} \cdot (\pi_{AA}^{n+1})^{s_{AA}}$$

is exactly the same as equation [10a].

The basic equation of a Markov chain describes the probabilities $\vec{x}_{t+1}$ of all states at the next time step, given a transition matrix M and a vector of current state probabilities $\vec{x}_t$:

$$[11] \quad \vec{x}_{t+1} = M\vec{x}_t$$

The transition matrix M, which is square and of dimension $|S|$ (number of model states), aggregates all transition probabilities of the chain so that columns correspond to X_t (and consequently sum to one) and rows correspond to X_{t+1} (and need not sum to one). The stochastic column vector $\vec{x}_t$ of dimension $|S|$ is typically a vector of zeros except for the entry corresponding to the current state X_t, which equals one.

The transition matrix M is the same for each generation under acyclic partial clonality, so that the transition probabilities after an arbitrary number of generations k after the current state can be calculated by matrix potentiation:

$$[12] \quad \vec{v}_{t+k} = M^k\vec{x}_t$$

For cyclical parthenogenesis the transition matrix changes during each cycle, as there are two different rates of clonality. However, equation [12] can be used to construct a transition matrix that spans not just from one generation to the next, but across a whole cycle of k asexual and one sexual generation:

$$[13a] \quad M_{CPa} = M_{c=1}^k M_{c=0}$$

$$[13b] \quad M_{CPb} = M_{c=0} M_{c=1}^k$$

Equation [13a] corresponds to observation after, equation [13b] to observation before the sexual generation. The differences between these two matrices and a transition matrix for acyclic partial clonality are discussed in detail in article III.

According to matrix algebra, each transition matrix has $|S|$ different vectors (i.e. they cannot be transformed into each other by multiplication with a common factor or "scalar") for which multiplication with the matrix does not change the relative value of the vector entries, but corresponds to a multiplication with a scalar λ:

$$[14] \quad \lambda\vec{x}_\infty = M\vec{x}_\infty$$

These vectors $\vec{x}_\infty$ are called "eigenvectors", and the associated scalars λ "eigenvalues" of the matrix. The eigenvector with the biggest absolute eigenvalue is the "dominant" eigenvector of the matrix; for our transition matrices, this is always an eigenvector with the eigenvalue one (according to the Perron-Frobenius theorem, Perron 1907). This dominant eigenvector is of special interest for the analysis of the model: it can be demonstrated (von Mises & Pollaczek-Geiringer 1929) that the repeated multiplication of any (non-zero) vector of dimension $|S|$ with the transition matrix M leads to a convergent result, which is equal to

the dominant eigenvector of the matrix. The eigenvector thus describes the probabilities of all model states after an infinite number of generations.

5.4 Model analysis

Improved hardware has made it possible to compute Markov chain models with many more states than in the early days of population genetic research. Yet it also opens up new possibilities for model analysis: calculation-intensive matrix algebra operations (e.g. matrix inversion, finding eigenvalues and eigenvectors) can now be executed in few seconds/minutes by a computer. We used the "classic", well-studied example of an exclusively sexual population of fixed size and including mutation as a reference to explore these new possibilities, but also to probe its limits as the number of states increases (e.g. by increasing the population size).

By translating concepts e.g. from network analysis into the context of our population genetic model, we found several new ways to display and analyze the results of our model. New plotting methods based on *de Finetti* diagrams allow a concise presentation of results across the complete state space of our model, by linking together genotype frequencies, allele frequencies and other population genetic parameters such as F_{IS}. Passing from a "dense" (i.e. nonzero value in every cell) to a "sparse" approximate (i.e. only values which are noticeably different from zero are stored) matrix may make it possible to extend the state space of Markov chain models even further, or at least speed up calculations on computers with less memory. We provide an approximation algorithm that will keep the mathematically important properties of the original "dense" matrix.

The methods presented in the following article will be used throughout the results part of this thesis. Depending on the research question, some were more helpful than others: the "most probable neighbor" and derived "landscape" method seem especially well-adapted to display the dynamic behavior of the Markov chain model, while the "eigenvector" and "time to" methods are more revealing for its long-term and limiting behavior. All results of the thesis were derived without the sparse approximation (either because of the chronology of the work, or because the number of states would still have been too large, e.g. for models with two loci), yet it may still be used in the future, e.g. as part of a data analysis program based on our results.

Analyse du modèle

Du matériel amélioré a permis de calculer les modèles de chaîne de Markov avec beaucoup plus d'états que dans les premiers jours de la recherche en génétique des populations. Pourtant, il ouvre aussi de nouvelles possibilités pour l'analyse des modèles : les opérations de calcul matriciel très exigeants en calcul (par exemple inversion d'une matrice, trouver des valeurs propres et vecteurs propres) peuvent désormais être exécutées en quelques secondes / minutes par un ordinateur. Nous avons utilisé l'exemple « classique » et bien étudié d'une population exclusivement sexuée à taille finie avec mutation comme référence

pour explorer ces nouvelles possibilités, mais aussi de sonder ses limites si le nombre d'états augmente (par exemple en augmentant la taille de la population).

Par exemple, en traduisant des concepts de l'analyse de réseau dans le cadre de notre modèle de génétique des populations, nous avons trouvé plusieurs nouvelles façons d'afficher et d'analyser les résultats de notre modèle. De nouvelles méthodes de traçage basées sur les diagrammes de *de Finetti* permettent une présentation concise des résultats à travers l'espace des états complets de notre modèle, en reliant les fréquences génotypiques, les fréquences des allèles et d'autres paramètres de génétique des populations tels que le F_{IS}. Passant d'une matrice « dense » (c'est-à-dire avec des valeurs non nulles dans chaque cellule) à une matrice « creuse » approximative (c'est-à-dire uniquement des valeurs qui sont sensiblement différentes de zéro sont stockées) peut permettre d'étendre l'espace d'état de modèles de chaîne de Markov encore plus, ou au moins d'accélérer les calculs sur les ordinateurs avec moins de mémoire. Nous fournissons un algorithme d'approximation qui va garder les propriétés mathématiquement importantes de la matrice d'origine « dense ».

Les méthodes présentées dans l'article suivant seront utilisées tout au long de la partie « résultats » de cette thèse. En fonction de la question de recherche, certains étaient plus utiles que d'autres : le « voisin le plus probable » et la méthode « paysage » dérivée semblent particulièrement bien adaptés pour afficher le comportement dynamique du modèle de chaîne de Markov, tandis que les méthodes « vecteur propre » et « temps jusqu'à » sont plus révélatrices pour sa durée et son comportement en limite. Tous les résultats de la thèse ont été obtenus sans l'approximation « creuse » (soit à cause de la chronologie de l'œuvre, ou parce que le nombre d'états aurait toujours été trop grand, par exemple pour les modèles avec deux loci), mais il peut encore être utilisé dans l'avenir, par exemple dans le cadre d'un programme d'analyse de données basé sur nos résultats.

Article I *Outils pour l'interprétation et l'approximation des matrices de transition grandes et denses des chaines de Markov en génétique des populations*

Sommaire de l'article

Contexte – Les chaînes de Markov sont un cadre commun pour des modèles d'état et temps discrets basés sur l'individu en évolution. Bien qu'ils aient joué un rôle important dans le développement de la théorie de base en génétique des populations, l'analyse des scénarios évolutifs plus complexes implique généralement une approximation avec d'autres types de modèles. Comme le nombre d'états augmente, les matrices de transitions grandes et denses impliquées deviennent de plus en plus difficiles à manier. Cependant, le progrès de la technologie en informatique continue de réduire les défis des mégadonnées, donnant ainsi de nouvelles possibilités pour les chaînes de Markov riches en états dans la théorie de la génétique des populations.

Résultats – En prenant un modèle de génétique des populations à la base de fréquences génotypiques comme exemple, nous proposons un ensemble de méthodes pour faciliter le calcul et l'interprétation des matrices de transitions grandes et denses des chaînes de Markov. Avec l'aide de l'analyse de réseau, nous démontrons comment ces matrices peuvent être transformés en graphiques clairs et faciles à interpréter, offrant une nouvelle perspective même sur le cas classique de l'accouplement au hasard en population finie avec mutation. En outre, nous décrivons un algorithme pour économiser de la mémoire dans l'ordinateur en remplaçant la matrice d'origine avec une approximation « creuse » tout en préservant ses propriétés mathématiquement importantes, y compris un vecteur propre dominant (normalisé) correspondant étroitement. Une analyse de sensibilité globale des résultats d'approximation dans notre exemple montre qu'une réduction de la taille de plus de 90% est possible sans affecter de manière significative les résultats du modèle de base. Des implémentations de nos méthodes à titre d'exemple sont collectées dans le module *mamoth* écrit en Python.

Conclusion – Nos méthodes aident à rendre le calcul des modèles stochastiques en génétique des populations impliquant matrices des transition grandes et denses possibles. Nos techniques de visualisation fournissent de nouvelles façons d'explorer ces modèles et de présenter leurs résultats de manière concise. Ainsi, nos méthodes contribueront à établir les chaînes de Markov riches en états comme un complément précieux à la diversité des modèles en génétique des populations actuellement employés, fournissant de nouveaux détails intéressants sur l'évolution dans, par exemple, des systèmes de reproduction non-standard comme la clonalité partielle.

Article I *Interpretation and approximation tools for big, dense Markov chain transition matrices in population genetics*

Katja Reichel*, Valentin Bahier, Cédric Midoux, Nicolas Parisey, Jean-Pierre Masson, Solenn Stoeckel

INRA, UMR1349 Institute for Genetics, Environment and Plant Protection, F-35650, Le Rheu, France

27/10/2015 Algorithms for Molecular Biology, accepted with minor revisions
09/07/2014 older draft version in arXiv.org q-bio: http://arxiv.org/abs/1407.2548

Abstract

Background – Markov chains are a common framework for individual-based state and time discrete models in evolution. Though they played an important role in the development of basic population genetic theory, the analysis of more complex evolutionary scenarios typically involves approximation with other types of models. As the number of states increases, the big, dense transition matrices involved become increasingly unwieldy. However, advances in computational technology continue to reduce the challenges of "big data", thus giving new potential to state-rich Markov chains in theoretical population genetics.

Results – Using a population genetic model based on genotype frequencies as an example, we propose a set of methods to assist in the computation and interpretation of big, dense Markov chain transition matrices. With the help of network analysis, we demonstrate how they can be transformed into clear and easily interpretable graphs, providing a new perspective even on the classic case of a randomly mating, finite population with mutation. Moreover, we describe an algorithm to save computer memory by substituting the original matrix with a sparse approximate while preserving its mathematically important properties, including a closely corresponding dominant (normalized) eigenvector. A global sensitivity analysis of the approximation results in our example shows that size reduction of more than 90% is possible without significantly affecting the basic model results. Sample implementations of our methods are collected in the Python module *mamoth*.

Conclusion – Our methods help to make stochastic population genetic models involving big, dense transition matrices computationally feasible. Our visualization techniques provide new ways to explore such models and concisely present the results. Thus, our methods will contribute to establish state-rich Markov chains a valuable supplement to the diversity of population genetic models currently employed, providing interesting new details about evolution e.g. under non-standard reproductive systems such as partial clonality.

Keywords discrete stochastic model, sparse approximation, eigenvector, network analysis, population genetics, compositional data, *de Finetti* diagram

Background

Natural systems often possess inherently discrete states in space, time or both. Atoms, molecules and cells, organs, individuals, populations and taxa usually appear as distinct entities; along the time axis, the radiation cycles we use as the basis for atomic clocks, neuronal action potentials, developmental stages in an organism's life cycle, generations and the revolutions of the earth around the sun are examples for similar patterns.

Modeling these discrete systems as such can have advantages over continuous approximations. One of the earliest examples comes from thermodynamics (Planck 1900), where heat emission spectra could only be predicted correctly if energy "comes in packets", known as "quanta". This discovery led to the new field of quantum mechanics, which provided the necessary theory for understanding the photovoltaic effect (Einstein 1905), thus proving essential for the invention of solar cells. In biology, the re-discovery of Mendel's rules and thus of the "quantal" nature of genetic heritability, at about the same time as Planck's famous speech, has had a similar impact on the study of evolution as the latter's research has had on thermodynamics (Ewens 2004). While most of the objects of biological research have long been recognised as discrete (e.g., the word *individual* literally means *not dividable*, a notion very similar to that of a *quantum*), we still struggle with understanding the processes, such as evolution, linking them to potential emergent properties (analogous to the physicists' heat spectra) at higher levels. Mathematical models preserving the discrete nature of the biological system are thus an interesting field of study.

Markov chains are a classical framework for modeling state and time discrete stochastic systems. Based on the assumption that the modeled system is *memoryless* (Markov property, Markov 1906), the basic model equation consists in multiplying a "start" vector, providing the state of the system at a given time, with a typically square "step" matrix. This matrix holds the transition probabilities, which depend on the model parameters and typically remain constant through time, between all possible states of the system within one time step. By analyzing the transition matrix, both the "short term" transient behavior and the "long term" limiting behavior of the model can be studied, thus putting the matrix at the center of attention for the biological interpretation of the results. Markov chains and other related forms of matrix-based models, such as Leslie models in population dynamics, are already widely in use (e.g. Feller 1971, Otto & Day 2007, Allen 2011), yet in many cases the number of modeled states is comparatively small and/or a major part of the transitions are considered impossible. The latter property leads to many zeros in the transition matrix, which then becomes *sparse*, as opposed to a *dense* matrix where zeros are rare. Computationally, sparse matrices are advantageous since memory may be saved by storing only those values which are different from zero. Special algorithms exist to carry out standard operations (e.g. matrix multiplication) directly on matrices stored in a sparse format (e.g. Davis 2006, 2011).

In population genetics, state and time discrete Markov chains are known primarily by the example of the classic biallelic Wright-Fisher model (Ewens 2004), which uses a one-dimensional random walk to describe the evolution of allele frequencies under genetic drift. For a population of N diploid organisms, the states of the Markov chain correspond to each

of the $2N + 1$ possible combinations of counts of the two alleles that sum to the constant total $2N$. Accordingly, a square transition matrix (assuming constant population size) would have $(2N + 1)^2$ entries. As the number of states further increases both with the population size and the complexity of the underlying genetic system (number of alleles and loci, table 1), the dynamics of allele frequencies in bigger populations are typically approximated by a continuous diffusion process based on the Fokker-Planck / Kolmogorov equations (Feller 1971), or even by deterministic equations assuming an "infinite" population size (e.g. as for the derivation of the Hardy-Weinberg equilibrium, Hardy 1908, Weinberg 1908). An alternative approach is coalescence theory, which uses re-defined discrete states and a reversed continuous time scale to specifically approximate certain aspects of the original state and time discrete Markov chain (e.g. Orive 1993a, Ceplitis 2003). While each of these approximations has its strengths and weaknesses (e.g. as discussed in Gale 1990, Greenbaum 2015), population genetic models that stay with the classic state and time discrete, chronological framework appear to be rare. One example is the model presented in (Stoeckel & Masson 2014): an extension of a classic biallelic Wright-Fisher model, it is based on genotype rather than allele frequencies. This design appears better adapted for the study of partially clonal populations, but also results in a bigger state space (e.g. for two alleles, combinations of the counts of each of three genotypes rather than those of the two alleles). The technical effort of storing and manipulating the big, dense transition matrices essential to such a model hardly seems to merit the results, which in turn have to be extracted from a great amount of data; adapted methods for interpretation and storage size reduction appear to be missing.

In this article, we provide a set of methods for visualizing and interpreting both the transient and limiting behavior of population genetic models involving state-rich, irreducible, aperiodic and time-homogeneous Markov chains, based on the transition matrix and its dominant eigenvector, as well as a method for approximating a dense transition matrix by a sparse substitute. For the first part, we combine *de Finetti* diagrams (de Finetti 1927) with network analysis, extending both concepts to provide clear and informative diagrams for the analysis of population genetic processes. For the second part, we use a predefined threshold (minimal percentage of information contained in the transition matrix) to keep only the more probable transient behavior of the model, while at the same time ensuring that mathematically important matrix properties are kept. The model presented in (Stoeckel & Masson 2014) serves as an example to illustrate our methods.

Model example

The population genetic model of Stoeckel & Masson (2014) describes the evolution of genotype frequencies based on a single locus with two alleles *a* and *A* in a fixed-size population of diploid, partially asexual organisms. States are defined as assignations of the N individuals in the population on the three possible genotypes (*aa*, *aA*, *AA*). The transition probabilities between the states depend on a symmetric mutation rate μ and a constant rate of asexual reproduction c, defined as the probability that an individual in the next generation was derived clonally from a single parent.

Transition matrices M resulting from this model are generally square, due to the fixed population size (a common feature of many population genetic models, compare (Ewens 2004). They also have a density of one – transitions between all states are possible in one step, although some of them (e.g. all individuals *aa* to all individuals *AA*) are very unlikely. The corresponding Markov chain is thus irreducible (single communicating class, no absorbing states) and aperiodic (period of all states equals one, same state possible in consecutive time steps). Since the mutation rate μ is symmetric, i.e. changes from *a* to *A* are just as likely as the inverse, M is also partially symmetric: if the transition probabilities from one particular state to all others have been calculated, swapping the names of all alleles also gives a correct result (compare figure 1 and 2). The notation in this article assumes left-stochastic matrices (columns represent the transition probabilities from one state to all others and thus sum to one), which implies that the limiting behavior of the Markov chain is described by its transition matrices' (normalized) right eigenvector v to the eigenvalue with the largest absolute value (and multiplicity one, Perron 1907): one.

The number of states in this model, and thus the size of the transition matrix M, depends on the one hand on the population size and on the other hand on the complexity of the genomic system being modeled, in particular the number of different genotypes possible. For a given number of genotypes g, the cardinality of the state space S (respective number of rows and columns in the transition matrix) in a genotype-based discrete stochastic model is:

$$|\mathrm{S}| = \left\langle {\mathrm{g} \atop \mathrm{N}} \right\rangle = \frac{(\mathrm{N}+\mathrm{g}-1)!}{\mathrm{N}! \cdot (\mathrm{g}-1)!}$$

From this equation it follows that the number of states increases exponentially with $1 + (\mathrm{g}-1)/(\mathrm{N}+1)$ for increasing N and with $1 + \mathrm{N}/\mathrm{g}$ for increasing g. For the number of possible genotypes, the ploidy level of the organism $\mathcal{P}$, the number of (partially linked) loci $\mathcal{L}$ and their respective numbers of alleles $\mathcal{A}_\mathrm{i}$, with $\mathrm{i} \in 1 \ldots \mathcal{L}$, need to be taken into account:

$$\mathrm{g} = \prod_{\mathrm{i}=1}^{\mathcal{L}} \left\langle {\mathcal{A}_\mathrm{i} \atop \mathcal{P}} \right\rangle = \prod_{\mathrm{i}=1}^{\mathcal{L}} \frac{(\mathcal{A}_\mathrm{i}+\mathcal{P}-1)!}{\mathcal{P}! \cdot (\mathcal{A}_\mathrm{i}-1)!}$$

Examples for the size of the resulting transition matrices are given in table 1. From these numbers, it is clear that a realistic "base-by-base" model of a full genome is still far beyond the capacity of current computer technology; however, many cases (biallelic SNPs, unlinked loci or blocks of completely linked loci) can already be interpreted based on the very simple *one-locus/two-allele*s model (Brookes 1999). It remains the dependence of $|\mathrm{S}|$ on the population size N, which is fortunately not so strong (for $\mathrm{N} > \mathrm{g}-1$).

To illustrate our methods, we will mostly use transition matrices derived for completely sexual populations ($\mathrm{c} = 0.0$), a case for which both transient and limiting behavior are generally known and interpretations can be easily verified (de Finetti 1927, Ewens 2004). For the mutation rate, $\mu = 10^{-6}$ was chosen as a plausible value based on experimental

Table 1. *Examples of matrix size based on the Stoeckel-Masson model.* Memory sizes are approximate and assume 64-bit accuracy.

N	$\mathcal{P}$	$\mathcal{L}$	$\mathcal{A}$	g	$\lvert S \rvert$	memory use
20	2	1	2	3	231	420 KB
100	2	1	2	3	5 151	205 MB
500	2	1	2	3	125 751	120 GB
1000	2	1	2	3	501 501	2 TB
20	4	1	2	5	10 626	865 MB
20	2	2	2	9	3 108 105	75 TB
20	2	1	4	10	10 015 005	730 TB
20	2	2	4	100	9.8×10^{20}	6.5×10^{21} YB

estimates (Drake 1991, Ellegren et al. 2003, Kronholm et al. 2010). N is either 5 ($|S| = 21$), 20 ($|S| = 231$) or 100 ($|S| = 5\ 151$), for good visibility and easy reproducibility of the results. Our test of the sparse approximation method is based on the limiting distribution of F_{IS}, a population genetic parameter of wide interest (e.g. as discussed in Wright 1922 under the name f, or in Halkett et al. 2005a) that was also analyzed in the original article describing our model example (Stoeckel & Masson 2014). For our example, the definition of F_{IS} based on the allele (ν_a, ν_A) and genotype frequencies ($\nu_{aa}, \nu_{aA}, \nu_{AA}$) is:

$$F_{IS} = 1 - \frac{\nu_{aA}}{2\nu_a\nu_A} = 1 - \frac{\nu_{aA}}{2(\nu_{aa} + 0.5\nu_{aA})(\nu_{AA} + 0.5\nu_{aA})}$$

Results

Working with big, dense transition matrices poses two connected problems: on the one hand, the storage size of the matrix may considerably slow down calculations or be altogether too big for the computer, on the other hand, the relevant information about the model may be difficult to extract from the great amount of data contained in the matrix. Visualization techniques for the interpretation of matrix data can, however, also help to find matrix properties which allow reducing the storage size, such as partial symmetry or the occurrence of many near-zero transition probabilities. We therefore start by describing the visualization techniques in the first part, and then move on to storage size reduction by sparse approximation in the second part of the results.

Visualization

An intuitive first step in analyzing the transient behavior of a Markov chain model is a diagnostic visualization of the transition matrix. By summarizing results in an accessible way, the resulting diagram may ideally also provide a basis for direct biological interpretation. With one exception (landscape plot), all the following visualization methods are available using the functions *histogrid*, *histo3d* and *networkplot* (with its support function *percolation*) in the *mamoth* module.

Heat map

A heat map or histogram of the transition matrix, where the transition probabilities p are symbolized by color / shade or height, is perhaps the easiest way to visualize it (figure 1). The resolution may be enhanced by an appropriate transformation of the range of values for p, for example by using a negative logarithm ($[0; 1] \rightarrow [0; \infty]$) or a *logit* transformation ($[0; 1] \rightarrow [-\infty; \infty]$).

For big matrices, heat maps can be costly to produce (memory size) and are often still not very clear, due to the large number of cases. However, they may help to recognize basic patterns (symmetries, groups of similar / more strongly connected states etc.) of potential value for finding more adapted visualizations / numerical methods.

> In our example, the heat map shows that many of the transition probabilities in the matrix are, though not equal, very close to zero. After re-ordering the states, the partial symmetry of the matrix also becomes visible.

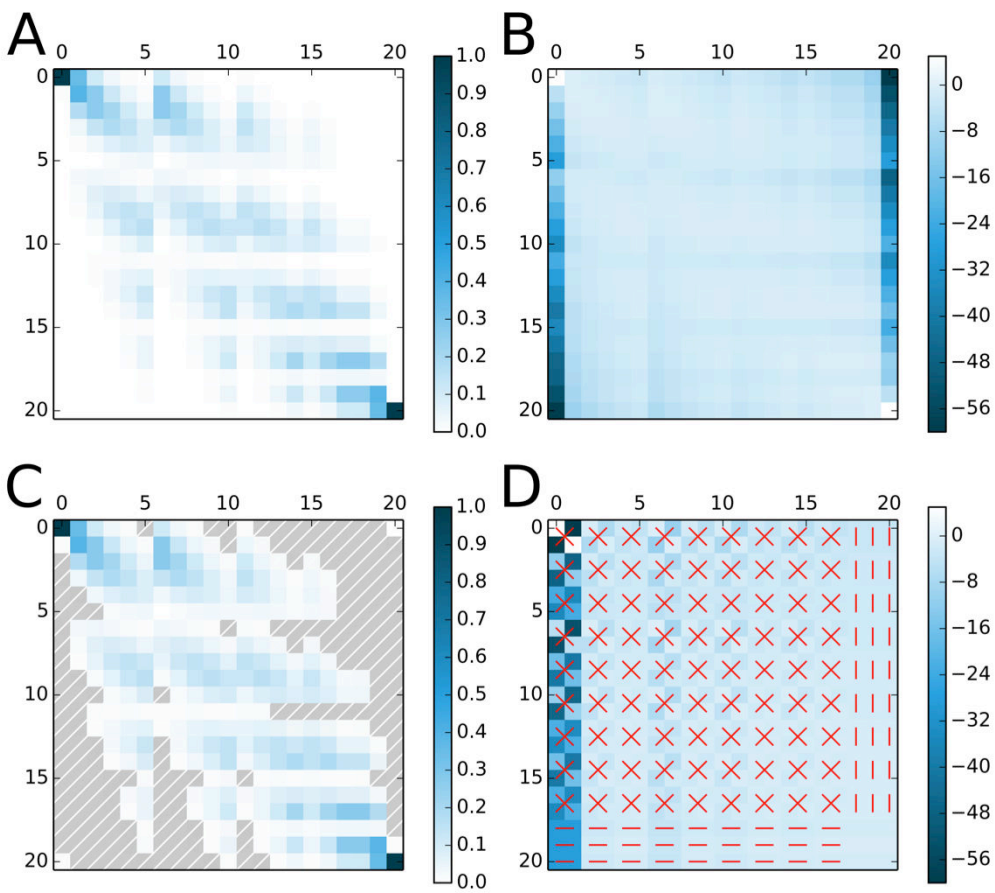

Figure 1. *Heat maps of transition matrices for* $\mathrm{N} = 5, \mu = 10^{-6}, \mathrm{c} = 0.0$. A. original probabilities, dense matrix B. $\mathrm{logit}(10)$ transformed probabilities, dense matrix C. sparse approximate matrix of A, implicitly stored zero values in hatched grey D. as in B, with alternative state order, red lines connect identical values.

Network display

The duality between matrices and graphs (e.g. Allen 2011, Aghagolzadeh et al. 2012) provides an alternative for the visualization and mathematical analysis of either structure. In a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the states of a Markov chain are thus represented as nodes/vertices $\mathcal{V}$ and the transitions as (weighted and directed) edges $\mathcal{E}$ connecting them, which is especially useful for sparse transition matrices.

For big, dense matrices, the number of edges in the resulting complete multidigraph (of edge multiplicity two) equals the number of entries in the transition matrix and is thus too big for easy interpretation. Concepts from network theory can be used to selectively display edges and summarize information about each state of the model system on the nodes. This leads to a variety of very clear synthetic representations constructed with different parameters and taking into account different time scales: from one generation (based on M) across t generations (based on M^{t}) up to the long-time equilibrium (dominant eigenvector of M, v).

To facilitate biological interpretation, we arranged the nodes of the network according to biological "meta data". For our model example where states represent distributions of individuals on three genotypes (*aa, aA, AA*) under a constant population size (compositional data), we placed the nodes in a *de Finetti* diagram (de Finetti 1927, see figure 2), a specialized ternary plot for population genetics.

The following visualization techniques are based on selectively displaying the network's edges:

Most probable neighbor – This is the analog to a *nearest neighbor* if distances (edge weights) represent probabilities. For each state i, there are one or several states j which have the *highest* probability to be the destination of a transition in the next time step; tracing these connections gives the expectation for the one-step transient behavior of the model.
> In our example, the most likely state for the next generation (figure 2) is always on or very near to the Hardy-Weinberg Equilibrium, which is represented by the continuous black curve going through $(1/4;\ 1/2;\ 1/4)$ in the diagram in figure 2A.

Most probable path – This is the counterpart of a *shortest path* if distances (edge weights) represent probabilities. For each non-commutative pair of states i and j, there exists at least one series of consecutive edges connecting i to j along which the *product* of the edge weights is *maximal*. It can be determined by using an "ordinary" shortest path algorithm (e.g. Dijkstra 1959, Biswas et al. 2013) on a negative *log* transform of the transition matrix. The most probable path is the most likely trajectory of the model system to get from one state to another.
> In our example (figure 2), a change from a population with only the *aa* genotype to one with only the *AA* genotype would closely follow the Hardy-Weinberg curve.

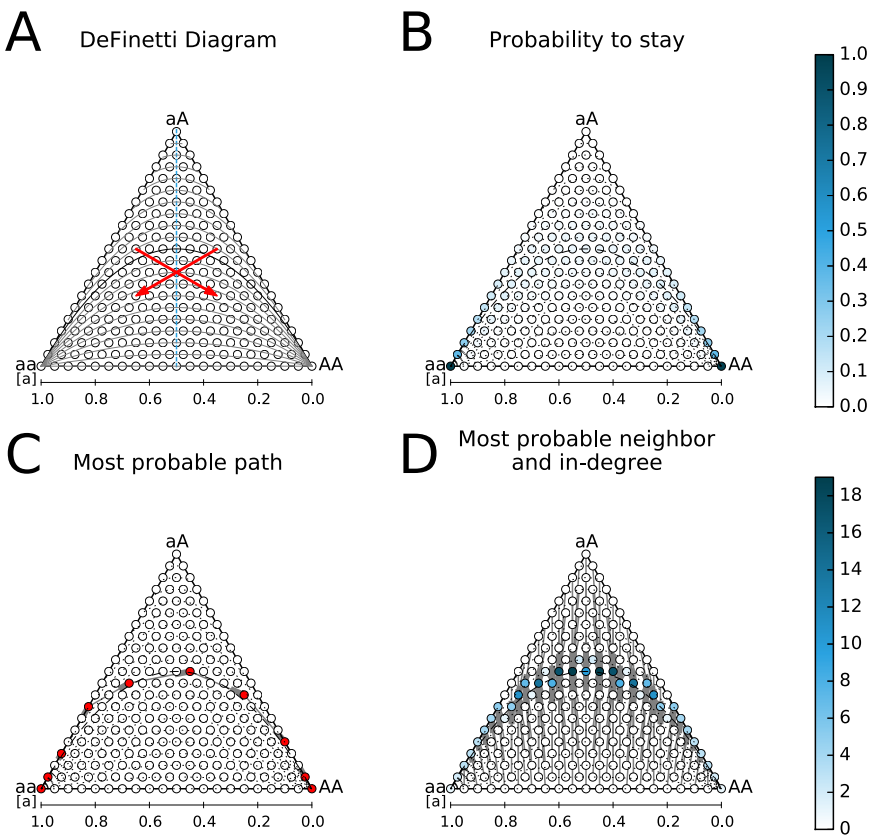

Figure 2. *Network display of transition matrices for* $N = 20, \mu = 10^{-6}, c = 0.0$. A. *De Finetti* diagram showing symmetry (dashed blue axis, red arrows corresponding to identical probabilities) and F_{IS} isocurves (gray and black) B. p_{stay} (node color), probability to stay at each node for one time step C. most probable path connecting $(N, 0, 0)$ to $(0, 0, N)$ D. most probable neighbors (directed edges) and in-degree (node color), i.e. for each node the most likely outbound transition at the next time step and the number of inbound most likely transitions from other states. Enlarged version in additional file 1.

Flow threshold – Using the smallest probability along the most likely path between two nodes i and j as a threshold, very rare transitions can be excluded.
> In our example (additional file 2), horizontal transitions along the base of the triangle, where no heterozygotes are produced despite of two homozygous genotypes being present in the population, would be excluded.

The following visualization techniques are based on changing the appearance of the network's nodes:

Degree – For each node in a graph representing a dense matrix, the number of incoming (*in-degree*) and outgoing (*out-degree*) edges is normally (approximately) equal to the number of nodes (matrix rows/columns). This method should therefore be used in connection with selective edge plotting and interpreted according to context.
> In our example (figure 2), the nodes with the highest in-degree are nearest neighbors to the largest number of nodes; if all states were equally likely at the current generation, those

next to $(0.25;\ 0.5;\ 0.25)$ on the Hardy-Weinberg curve would be the most likely in the next generation.

Betweenness-centrality – Based on the same concept as the *most probable path*, this can be redefined as the number of *most probable paths* passing through each node when connections between each pair of nodes are considered. It can be derived in a similar way as the *most probable path*, by applying a standard algorithm developed for additive distances to a negative *log* transform of the multiplicative probabilities in M. Nodes with a high betweenness-centrality represent frequent transient states.
> In our example, these are all the states along the Hardy-Weinberg curve except for the fixation states (additional file 3).

Probabilities – For each state i in the Markov chain model, several probabilities can be calculated – and displayed on the nodes – to describe both the transient and limiting behavior:

$\mathrm{p_{stay}}$ – *probability to stay for one time step*

$\mathrm{p_{stay}(i) = p_{i,i}}$, the probabilities on the matrix diagonal; for each state i this is the probability that the system remains at state i for the next time step ("stickiness"). This probability allows the easy detection of (near-)absorptive states.
> In population genetics, the fixation states $\{(\mathrm{N}; 0; 0), (0; 0; \mathrm{N})\}$ are typical examples (figure 2).

$\mathrm{p_{out}}$ – *probability to leave in one time step*

$\mathrm{p_{out}(i) = 1 - p_{i,i}}$, the column sums of the matrix without the diagonal; for each state i this is the probability that the system changes state at the next time step ("conductivity"). Being the opposite of $\mathrm{p_{stay}}$, this probability allows the detection of states which are rarely occupied for consecutive time steps.
> In our example, these are the states where the population consists of an approximately even mixture of both homozygotes (central basis of the triangle) or only of heterozygotes (top of the triangle; additional file 2).

In contrast, the row sums of a left-stochastic matrix may exceed one and are thus not probabilities. As a result of the Markov property, a *probability to arrive* always depends on the state at the previous time step, which results in a number of possible definitions.

$\mathrm{p(i|j)}$ – *probability to arrive from state* j *in one time step*

$\mathrm{p(i|j) = p_{j,i}}$, $\mathrm{j} \in \mathrm{S}$, all probabilities in one column of the transition matrix; the probability distribution (mean, variance, skew according to arrangement of nodes) for transitions starting from one particular state. This allows the prediction of the most likely states for the next time step.
> In our example, the variance around the fixation states is much more limited than at the interior states of the triangle (additional file 3).

p_{in} – *probability to arrive in one time step*

$p_{in}(i) = 1/(|S| - 1) \cdot \sum_j p_{j,i}$ for $i \neq j$, the row sums of the matrix divided by the number of other states; probability to arrive at state i if all previous states are equally likely. This shows states which are generally very likely destinations for one-step transitions.
> In our example, these are the states around the Hardy-Weinberg curve (additional file 2).

p_{in}^{∞} – *probability to arrive in an infinite run*

$p_{in}^{\infty}(i) = \sum_j p_{j,i} \cdot v_j$ for $i \neq j$, the sum over the element-wise product of eigenvector and matrix row, without the diagonal; probabilities to arrive at state i if the likelihood of the previous states is distributed according to the limiting distribution. This shows the states which are the most frequent destination of transitions in an infinite run of the model.
> In our example, these are the two states next to the fixation states where there is exactly one "foreign" allele (additional file 3).

p^{∞} – *limiting distribution / eigenvector-centrality*

$p^{\infty}(i) = v_i$, the eigenvector; probability to find the system at state i after infinitely many time steps, or proportion of time spent in each state averaged over infinitely many time steps (limiting distribution). This is the prediction for the most likely states independently of the start state.
> As is well known for our example, these are the fixation states (additional file 2).

Expected time to first passage – To calculate the expected time to arrive at a certain (group of) states from any other, the "target" states are considered absorptive (first passage time, Allen 2011). Based on the sub-matrix M' including only the transition probabilities between non-target states, the times t_{target} are $t_{target} = \mathbf{1}(I - M')^{-1}$ where $\mathbf{1}$ is a row vector of ones matching the dimension of M' and I is the corresponding unit matrix. The first passage times of the target states are zero.
> For our example, plotting the expected time to the fixation states shows that it depends predominantly on the current state's allele frequencies (additional file 3).

Landscape plot

Combining length and direction of the transitions in the most probable neighbor plot (figure 2) gives a three dimensional "landscape" illustrating the most probable dynamics of the Markov chain, similar to the "gravity well" plots known from physics. The expected changes in the genotype frequencies are thus represented in a more intuitive fashion, by imagining the population as a small ball rolling on a "landscape" from "hills" to "valleys". Elevations h are derived from the equality of potential and kinetic energy, which resolves to $h = d^2 \cdot 0.05$ for a single time step, approximating gravitational acceleration by 10. For each model state/node, the distances d are given by the changes in genotype frequencies when moving to the most probable neighbor $d = \sqrt{(\Delta aa)^2 + (\Delta aA)^2 + (\Delta AA)^2}$. The "landscape" is subsequently drawn as a triangular grid, using the elevation at each state/node as support.

To improve readability, h can be rescaled by a constant factor and the landscape colored according to the relative elevation (taking the center of each triangle as reference). The resulting "landscape" shows only the (deterministic) expected dynamics of the Markov chain – one could imagine the accompanying stochastic effects as an "earthquake".

> In our example, the expected dynamics of the genotype frequencies show convergence to the Hardy-Weinberg equilibrium (additional file 4).

Note: because of its dependence on a function or matrix specifying the distances between states, and on the triangular grid-like structure of the state space, this method is not included in the *mamoth* source code.

Approximation

One major drawback of state-rich Markov chain models is that the transition matrix in its full form takes up a lot of memory space (table 1). Beside switching to one of the alternative model types mentioned in the introduction (diffusion approximation, coalescence process), there are multiple computational approaches to addressing this issue while keeping the original state and time discrete framework, including:

- *external memory:* the whole matrix is stored on a (sufficiently large) hard drive, only parts are loaded into active storage when needed (analogous to Dongarra & Sorensen 1986)
- *iterative/selective matrix creation:* the whole matrix is never stored, only parts are created when needed (e.g. in combination with algorithms such as Lehoucq et al. 1997)
- *lumping states based on model properties:* if a group of states has the same (sum of) transition probabilities leading into it and out of it to any other (group of) states and the same analytical meaning (e.g. same value of F_{IS}) they can be combined into one (Kemeny & Snell 1976, Schapaugh & Tyre 2012); other algorithms of state aggregation, such as (Deng 2012), lead to an approximation of the original matrix
- *sparse approximation:* turning a dense matrix into a sparse matrix by approximating very small matrix elements to zero (e.g. as in Kumar et al. 2009, Talwalkar 2010)

Which of the first two options is more appropriate depends both on the available hardware and the nature of the task: if the whole matrix is needed repeatedly, storing it will save the time to recalculate despite increased memory access times, but if calculating the matrix elements is fast, the matrix is needed only once or only some parts of the matrix (e.g. the *most probable neighbor* of each state) are needed, storing the matrix as a whole would be an unnecessary effort.

Because of the symmetry between the two allele frequencies in our model example, almost half of all states could be pairwise lumped, thus reducing matrix size to a little over a quarter of the original. The exception are the states on the symmetry axis of the *de Finetti* diagram (compare figures 1, 2), which do not have a "lumping partner". Symmetry with respect to the allele frequencies is often found in population genetics models (Ewens 2004). However, because of this dependency on model structure a size reduction algorithm based on

lumping would not be applicable to non-symmetric extensions of the original model, e.g. with an asymmetric mutation rate or directional selection. Allele frequencies would have to be analyzed jointly, as the new states retain only the ratio of both; once lumped, "unpacking" the states becomes difficult.

The high number of very small values in the Markov chain transition matrix (figure 1) of our model example suggests that sparse approximation would be very effective. Moreover, as each column of the Matrix corresponds to a probability distribution (constant sum of one) which becomes less uniform as the number of states / population size increases (the expected convergence to a multinormal distribution with variance proportional to $1/\mathrm{N}$ is the underlying principle of the well-known diffusion approximation), the proportion of very small transition probabilities is likely to augment as the matrix size increases. While sparse approximation is independent of model-specific properties such as symmetry and does not change the states as such, it has the disadvantage of changing the actual content of the transition matrix, potentially leading to the loss of relevant properties such left-stochasticity or irreducibility.

The sparse approximation algorithm we propose ensures that the resulting sparse matrix still has all the properties relevant to its function in the Markov chain model. Additionally, it can be executed iteratively so that the complete dense matrix need not be stored. The algorithm iterates over all columns of the transition matrix M and excludes (almost) all values which, in total, contribute less than a threshold value $\mathrm{s} \in [0,1]$ to the column sum:

For all columns $\mathrm{C}^{\mathrm{i}} = \mathrm{M}_{1\ldots|\mathrm{S}|,\mathrm{i}}$ with $\mathrm{i} \in [1, |\mathrm{S}|]$:

1. create a permutation R of the row indices so that the corresponding entries are ranked according to size:
   $\mathrm{R} \leftarrow \mathrm{ordinalrank}(\mathrm{j} | 1 \geq \mathrm{C}^{\mathrm{i}}_{\mathrm{j}} \geq 0)$
2. find the minimal rank (index of R) so the corresponding entries sum at least to the threshold value s:
   $\mathrm{r} \leftarrow \min(\mathrm{k})$ for $\sum_{\mathrm{R}_1}^{\mathrm{R}_\mathrm{k}} \mathrm{C}^{\mathrm{i}}_{\mathrm{R}_\mathrm{k}} \geq \mathrm{s}$
3. keep at least the two biggest values per column:
   $\mathrm{r} \leftarrow \max(2, \mathrm{r})$
4. keep all values of equal rank:
   while $\mathrm{C}^{\mathrm{i}}_{\mathrm{R}_{\mathrm{r}+1}} = \mathrm{C}^{\mathrm{i}}_{\mathrm{R}_\mathrm{r}} : \mathrm{r} \leftarrow \mathrm{r} + 1$
5. round all values with ranks greater then r to zero, but keep those on the main diagonal and the first lower and first upper diagonals:
   $\mathrm{C}^{\mathrm{i}}_{\mathrm{R}_\mathrm{k}} \leftarrow 0$ for all k with $\mathrm{k} > \mathrm{r} \wedge \mathrm{R}_\mathrm{k} \notin \{(\mathrm{i}-1, \mathrm{i}, \mathrm{i}+1) \bmod |\mathrm{S}|\}$
6. rescale the column to sum to one:
   $\mathrm{C}^{\mathrm{i}} \leftarrow \mathrm{C}^{\mathrm{i}}/\mathrm{sum}(\mathrm{C}^{\mathrm{i}})$.

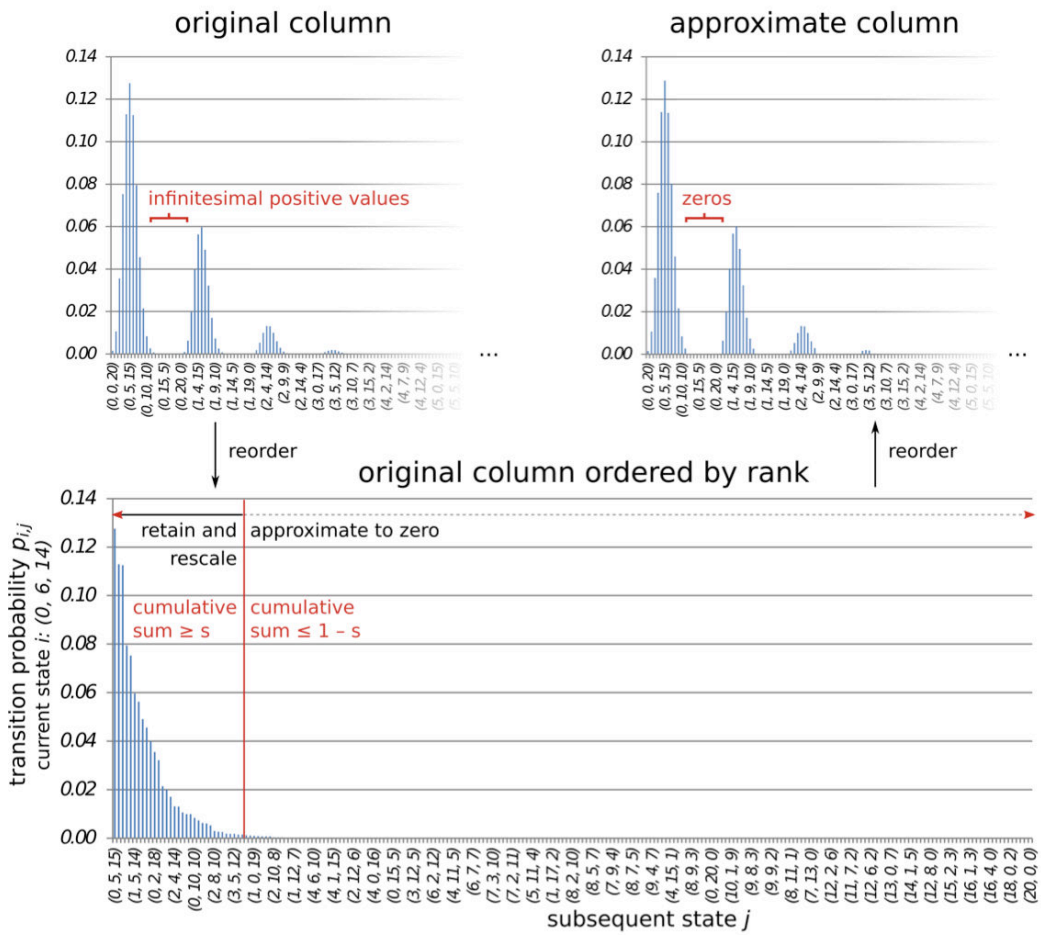

Figure 3. *Illustration of the approximation algorithm* ($s = 0.99$) *for* $N = 20, \mu = 10^{-6}, c = 0.0$ *and the state* (0,6,14). Reordering is based on the relative size of the column entries and their index in the original column, respectively.

The first two steps, together with the rounding in step five, form the core of the algorithm (compare figure 3), steps three and four prevent distortions and steps five and six ensure the continued validity of essential Markov chain transition matrix properties: Irreducibility is assured by keeping at least one outgoing and one incoming transition probability per state in such a way that all states remain connected (step five, first lower and first upper diagonal), aperiodicity by keeping all probabilities to stay at the same state (step five, main diagonal), and the rescaling of each column ensures left-stochasticity of the matrix (step six). In contrast, the property that one-step transitions are possible between all states is deliberately given up. The sparse approximation algorithm is available as the *appromatrix* function in the *mamoth* module.

Both the efficiency, i.e. the density or memory size of the resulting matrix, and the bias vary according to the value of s and the distribution of values in the original matrix. If s is low or the probability distribution in the column is far from uniform, more values will be discarded (compare figure 3). An appropriate value for s has to be determined heuristically by testing successively increasing values, up to the point where the bias due to the approximation no longer interferes with the interpretability of the model results. The sum of the differences between the entries of the approximate and original matrices has a theoretical upper limit

of $(1-s)\cdot|S|$, but the effect of this perturbation on the model output may be more complex.

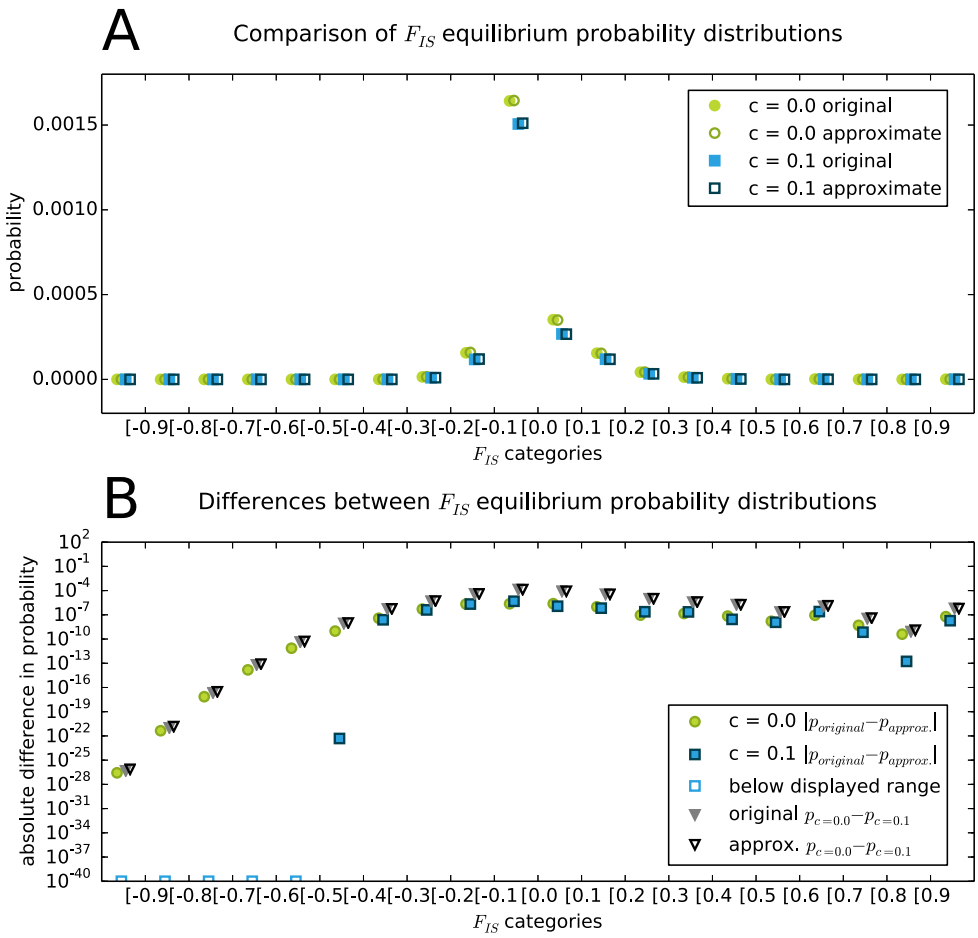

Figure 4. *Comparison of the limiting distribution of* F_{IS} *for* $N = 20, \mu = 10^{-6}, c = \{0.0, 0.1\}$. A. probability distributions based on the original (filled symbols) and the approximate (unfilled symbols) matrix B. pairwise differences between probability distributions, biologically interesting distances marked by triangles.

In our model example, we analysed the effect of sparse approximation on the equilibrium F_{IS} distribution derived from the dominant eigenvector of the transition matrix. The dominant eigenvector of either a sparse or dense matrix can be calculated with the *eigenone* function in *mamoth*, while a comparison between two vectors by a G-Test (correctly omitting infinity values from the test statistic) is implemented in the *testvector* function. A direct comparison between the "original" and "sparse approximate" equilibrium F_{IS} distributions (figure 4) shows a very close fit which does not obscure the biologically relevant changes due to different rates of asexual reproduction. To test if the method gives similarly good results over a wider range of parameters (population size, mutation rate, rate of asexuality and approximation threshold), we performed a Global Sensitivity Analysis (GSA) (Morris 1991, Saltelli 2004, Wainwright et al. 2014) using different divergence statistics to compare the limiting distribution of F_{IS} derived from original and sparse approximate matrix (R Core Team 2013, Pujol et al. 2015) and the density of the sparse matrix.

The results of the GSA show that all four model parameters may generally have non-linear/interacting effects on the quality of the approximation, but in the mean these effects are not very strong (figure 5; the minimal upper bound of the parameters is one). Memory size reduction is highly efficient as the mean density of the sparse matrices was only ≈ 0.11. Individual densities ranged from ≈ 0.42 (small matrix, high threshold) to ≈ 0.03 (big matrix, low threshold), varying most strongly with the population size, though all four parameters have a significant influence. On our reference system (Intel Core i7-3930K 3.2 GHz processor with 64 GB RAM), calculating the sparse approximation based on the original matrix took on average 1.7 s for $\mathrm{N} = 50$ (14.6 s to construct the original), and 31.3 s for $\mathrm{N} = 100$ (221.7 s to construct the original). Finding the dominant eigenvector of sparse approximate and original matrix took on average 0.1 s (sparse) versus 51.7 s (original) for $\mathrm{N} = 50$ and 2.4 s (sparse) versus 7869.1 s (2 h, 11 min, 9.1 s, original) for $\mathrm{N} = 100$, so that in both cases less than one percent of the original runtime was needed with the sparse approximate matrix.

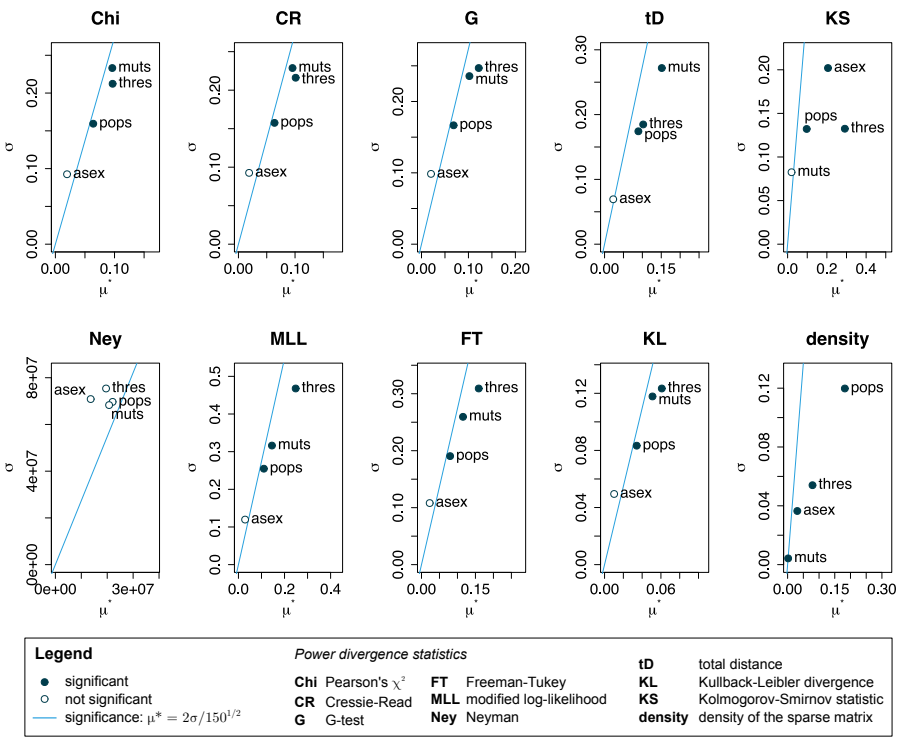

Figure 5. *Global sensitivity analysis of original vs. approximate equilibrium* $\mathrm{F_{IS}}$ *distribution*. Absolute mean μ^* and standard deviation σ of the elementary effects of population size N (pops), mutation rate μ (muts), rate of asexual reproduction c (asex) and sparse approximation threshold s (thres) on the density of the sparse approximate matrix, and on different statistics comparing the limiting $\mathrm{F_{IS}}$ distributions derived from original and sparse approximate matrix. Based on 150 *Morris* samples from the parameter space: population size ($\mathrm{N} = \{10, 20, \ldots, 100\}$), mutation rate ($\mu = \{10^{-12}, 10^{-11}, \ldots, 10^{-3}\}$), rate of asexual reproduction ($c = \{0.1, 0.2, \ldots, 1.0\}$) and approximation threshold ($\mathrm{s} = \{0.8, 0.82, \ldots, 0.98\}$). Infinity values were omitted from the test statistic.

The overall similarity of the original and approximate equilibrium $\mathrm{F_{IS}}$ distributions, measured with different divergence statistics (total distance, Kullback-Leibler divergence,

power divergence statistics, Cressie & Read 1984), is very high: e.g. the mean for the total distance $\sum|f_{orig} - f_{approx}|$ is ≈ 0.06. It is largely independent of the rate of asexual reproduction and depends most strongly on the approximation threshold and the mutation rate. In contrast, the maximal difference (Kolmogorov-Smirnov two-sample test statistic) between classes of the original and approximate equilibrium F_{IS} distribution is hardly affected by the mutation rate, but rather by approximation threshold (high mean effect) and rate of asexual reproduction (strong non-linearity/interaction). Though on average not significant, the Kolmogorov-Smirnov test gave p-values below 0.05 in 20% of the parameter sets sampled. Consequently, the same approximation threshold can be used to compare the overall shape of the distributions across the whole range of rates of asexual reproduction, but it may have to be adapted if mutation rate and population size differ strongly between the modeled scenarios. Care must be taken when individual classes within the distribution (e.g. long-term fixation probability) shall be compared as the probabilities derived from a sparse approximate matrix may then be significantly different from the original.

In conclusion, sparse approximation using our algorithm has the advantage of being easily applicable to all transition matrices independently of the properties of the underlying model, and is well suited to provide an overview of the equilibrium F_{IS} distribution under different rates of asexual reproduction in our model example. However, it needs an initial effort to verify the model results derived from the approximate matrix and to estimate their final bias. For fine-scale analyses, lumping states may provide an approximation-free alternative, but is not always possible as it depends on the model structure.

Discussion

As the technological obstacles of working with "big data" become smaller, new opportunities arise especially for stochastic models, e.g. in population genetics. Yet these opportunities also lead to new challenges: results need to be brought into an interpretable form, and the technological boundaries further pushed back to allow even more complexity. We developed methods to help with the computational analysis and interpretation of state-rich time- and space-discrete Markov chain models in population genetics, focusing on the particularly challenging case of very dense matrices.

Markov chain models are a versatile framework also for population genetic questions, and may often provide a first step in the development of analytic formulae (Ewens 2004). Further relevant parameters such as selection, migration or "unusual" reproductive systems can be easily included in such a model. Yet even for randomly mating population with genetic drift and mutation, a standard case of population genetics, a Markov chain model such as (Stoeckel & Masson 2014) may still yield additional information with the help of our visualization methods: In particular, the short-term dynamics, e.g. probabilistic trajectories connecting a current and a previous or predicted state, and the resulting variation around the expectation of convergence to the Hardy-Weinberg equilibrium are made visible. Especially for small populations, which are highly relevant e.g. for conservation genetics (Ellstrand & Elam 1993), and questions relating to development through time rather than just the long-term equilibrium, such Markov chain models may thus become valuable tools.

Though the size limitation for computational matrix analysis may never be completely removed, we showed that there are ways to circumvent it: even without access to specialized hardware, big, dense transition matrices may be manageable either by lumping states, or by approximating rare transitions to zero with our sparse approximation algorithm. In our model example, the approximation provided sufficiently accurate results for the limiting distribution of $\mathrm{F_{IS}}$. Though there is an initial effort of verification, the advantage of sparse approximate matrices is considerable as they can subsequently be used also on less powerful hardware e.g. to speed up or allow the calculation of eigenvectors on systems incapable of storing the full model. In our model example, the size reduction of sometimes more than 90% would e.g. make it possible to use the equilibrium F_{IS} distributions for the inference of model parameters in an analysis software without having to store a – necessarily incomplete – reference collection of pre-calculated distributions for very big matrices. Moreover, some of our visualization methods (e.g. most probable neighbor, $\mathrm{p_{in}}$, $\mathrm{p_{stay}}$, $\mathrm{p_{out}}$, $\mathrm{p_{i,j}}$) can be used without ever storing the whole matrix, while providing even very powerful conclusions about model behavior. Our sparse approximation method is not intended to substitute other approaches, and we did not test if it outperforms the accuracy of other approximations (e.g. diffusion approximation) for any specific question. Rather, it is a supplement, allowing to keep the structure of the original Markov chain model with the corresponding interpretation techniques beyond the technical limit, and a potential reference for the existing methods.

Individual-based models are becoming more and more popular in biology (Zipkin et al. 2010, Black & McKane 2012), which will further increase the frequency of encountering computationally challenging cases such as the one we presented. In population genetics, modeling more complex evolutionary parameters such as life cycles and reproductive mechanisms, multi-dimensional fitness landscapes or dispersal may often lead to the necessity of extending the traditional models from allele frequencies (Ewens 2004) to genotypes. Due to the diploid/polyploid nature of most higher organisms, this will necessarily increase the size of transition matrices and equation systems to be analyzed. By presenting our approach, we hope to encourage and inspire others to extend and adapt our methods, thus further paving the way for the use of Markov Chain models with big, dense transition matrices.

Conclusion

We described and evaluated a set of tools, implemented in the Python module *mamoth*, for working with state-rich Markov chain models in population genetics. These tools ease the interpretation of model behavior by providing diagnostic visualizations of transition matrices, and allow substituting dense transition matrices with a sparse counterpart by applying an iterative approximation algorithm that is independent of model symmetry. Thus, our methods permit an advanced analysis of increasingly complex Markov chain models in population genetics, without giving up their space and time discrete structure. They may therefore contribute e.g. to the study of the population genetic consequences of partially clonal reproduction.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JPM and SS wrote the population genetic model and proposed first ideas. VB, CM and KR contributed equally to the development of the final methods, which was guided by SS. NP added the sensitivity analysis of the approximation method. SS was in charge of acquiring funds. KR wrote most of the manuscript, to which all authors contributed. All authors read and approved the final manuscript.

Acknowledgements

We thank Jurgen Angst, Sophie Arnaud-Haond, Sina Brunsch, Florent Malrieu, Romuald Rouger and François Timon for constructive discussions, and all reviewers for their helpful criticism. This study is part of the CLONIX project (ANR-11-BSV7-007) financed by the French National Research Agency. Katja Reichel receives a PhD grant by the Région Bretagne and the division "Plant Health and Environment" of the French National Institute of Agricultural Research (INRA), and would like to thank her supervisors Solenn Stoeckel and Jean-Christophe Simon.

Availability and requirements

The methods we described can be easily implemented in any scientific programming environment; we provide sample code for Python for all methods which do not rely on the specific state definitions of our model example.

Project name: mamoth

Project home page: http://www6.rennes.inra.fr/igepp_eng/Productions/Software

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 2.7 or 3.4 and higher, extension modules NumPy/SciPy, matplotlib and NetworkX (Hunter 2007, Oliphant 2007, Hagberg et al. 2008)

License: GNU public license, version 2 (GPL2)

Any restrictions to use by non-academics: see GPL2 license

References

Aghagolzadeh M, Barjasteh I, Radha H. 2012. Transitivity matrix of social network graphs. *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pp. 145–48. IEEE

Allen LJS. 2011. *An introduction to stochastic processes with applications to biology*. Boca Raton, Florida: Chapman & Hall. 2nd ed.

Biswas SS, Alam B, Doja MN. 2013. Generalisation of Dijkstra's algorithm for extraction of shortest paths in directed multigraphs. *Journal of Computer Science*. 9(3):377–382

Black AJ, McKane AJ. 2012. Stochastic formulation of ecological models and their applications. *Trends in Ecology & Evolution*. 27(6):337–345

Brookes AJ. 1999. The essence of SNPs. *Gene*. 234(2):177–186

Ceplitis A. 2003. Coalescence times and the Meselson effect in asexual eukaryotes. *Genetical Research*. 82(3):183–190

Cressie N, Read TRC. 1984. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*. 46(3):440–464

Davis TA. 2006. *Direct methods for sparse linear systems*. Society for Industrial and Applied Mathematics

Davis TA. 2011. Algorithm 915, SuiteSparseQR: Multifrontal multithreaded rank-revealing sparse QR factorization. *ACM Transactions on Mathematical Software*. 38(1):8:1–8:22

De Finetti B. 1927. Conservazione e diffusione dei caratteri Mendeliani. Nota I. Caso panmittico. In *Rendiconti della R. Accademia Nazionale dei Lincei*, Vol. V (11-12), pp. 913–921

Deng K. 2012. *Model reduction in Markov chains with application to building systems*. Doctoral thesis. University of Illinois at Urbana-Champaign

Dijkstra EW. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*. 1:269–271

Dongarra JJ, Sorensen DC. 1986. Linear algebra on high performance computers. *Applied Mathematics and Computation*. 20(1):57–88

Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences*. 88(16):7160–7164

Einstein A. 1905. Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*. 322(6):132–148

Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Current Opinion in Genetics & Development*. 13(6):562–568

Ellstrand NC, Elam DR. 1993. Population genetic consequences of small population size: implications for plant conservation. *Annual Review of Ecology and Systematics*. 217–242

Ewens WJ. 2004. *Mathematical population genetics: I. Theoretical introduction*. New York: Springer. 2nd ed.

Feller W. 1971. *An introduction to probability theory and its applications*. Wiley

Gale JS. 1990. *Theoretical population genetics*. Springer

Greenbaum G. 2015. Revisiting the time until fixation of a neutral mutant in a finite population – A coalescent theory approach. *Journal of Theoretical Biology*. 380:98–102

Hagberg AA, Schult DA, Swart PJ. 2008. Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp. 11–15

Halkett F, Simon J, Balloux F. 2005. Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology & Evolution*. 20(4):194–201

Hardy GH. 1908. Mendelian proportions in a mixed population. *Science*. 49–50

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 9(3):90–95

Kemeny JG, Snell LJ. 1976. *Finite Markov chains*. New York, Berlin, Heidelberg, Tokyo: Springer-Verlag

Kronholm I, Loudet O, de Meaux J. 2010. Influence of mutation rate on estimators of genetic differentiation – lessons from *Arabidopsis thaliana*. *BMC Genetics*. 11(1):33

Kumar S, Mohri M, Talwalkar A. 2009. On sampling-based approximate spectral decomposition. *Proceedings of the 26th International Conference on Machine Learning*. http://www.sanjivk.com/nys_col_ICML.pdf

Lehoucq RB, Sorensen DC, Yang C. 1997. *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* Software, Environment, and Tools Series 6. Philadephia: SIAM

Markov AA. 1906. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. *Proceedings of the Society of Physics and Mathematics at the University of Kazan*. 15(2):135–156

Morris MD. 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics*. 33(2):161–174

Oliphant TE. 2007. Python for scientific computing. *Computing in Science & Engineering*. 9(3):10–20

Orive ME. 1993. Effective population size in organisms with complex life-histories. *Theoretical Population Biology*. 44(3):316–340

Otto SP, Day T. 2007. *A biologist's guide to mathematical modeling in ecology and evolution*. Princetown University Press

Perron O. 1907. Zur Theorie der Matrices. *Mathematische Annalen*. 64(2):248–263

Planck M. 1900. Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*. 2:237–245

Pujol G, Iooss B, Janon A. 2015. *Sensitivity: Sensitivity analysis*. R Package Version 1.11.

R Core Team. 2013. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing

Saltelli A, ed. 2004. *Sensitivity analysis in practice: A guide to assessing scientific models*. Hoboken, New Jersey: Wiley

Schapaugh AW, Tyre AJ. 2012. A simple method for dealing with large state spaces. *Methods in Ecology and Evolution*. 3(6):949–957

Stoeckel S, Masson J-P. 2014. The exact distributions of F_{IS} under partial asexuality in small finite populations with mutation. *PLoS ONE*. 9(1):e85228

Talwalkar A. 2010. *Matrix approximation for large-scale learning*. Courant Institute of Mathematical Sciences New York

Wainwright HM, Finsterle S, Jung Y, Zhou Q, Birkholzer JT. 2014. Making sense of global sensitivity analyses. *Computers & Geosciences*. 65:84–94

Weinberg W. 1908. Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*. 64:368–382

Wright S. 1922. Coefficients of Inbreeding and Relationship. *The American Naturalist*. 56(645):330–338

Zipkin EF, Jennelle CS, Cooch EG. 2010. A primer on the application of Markov chains to the study of wildlife disease dynamics: Modelling disease dynamics with Markov chains. *Methods in Ecology and Evolution*. 1(2):192–198

Additional Files

Additional file 1 – Network display methods 1

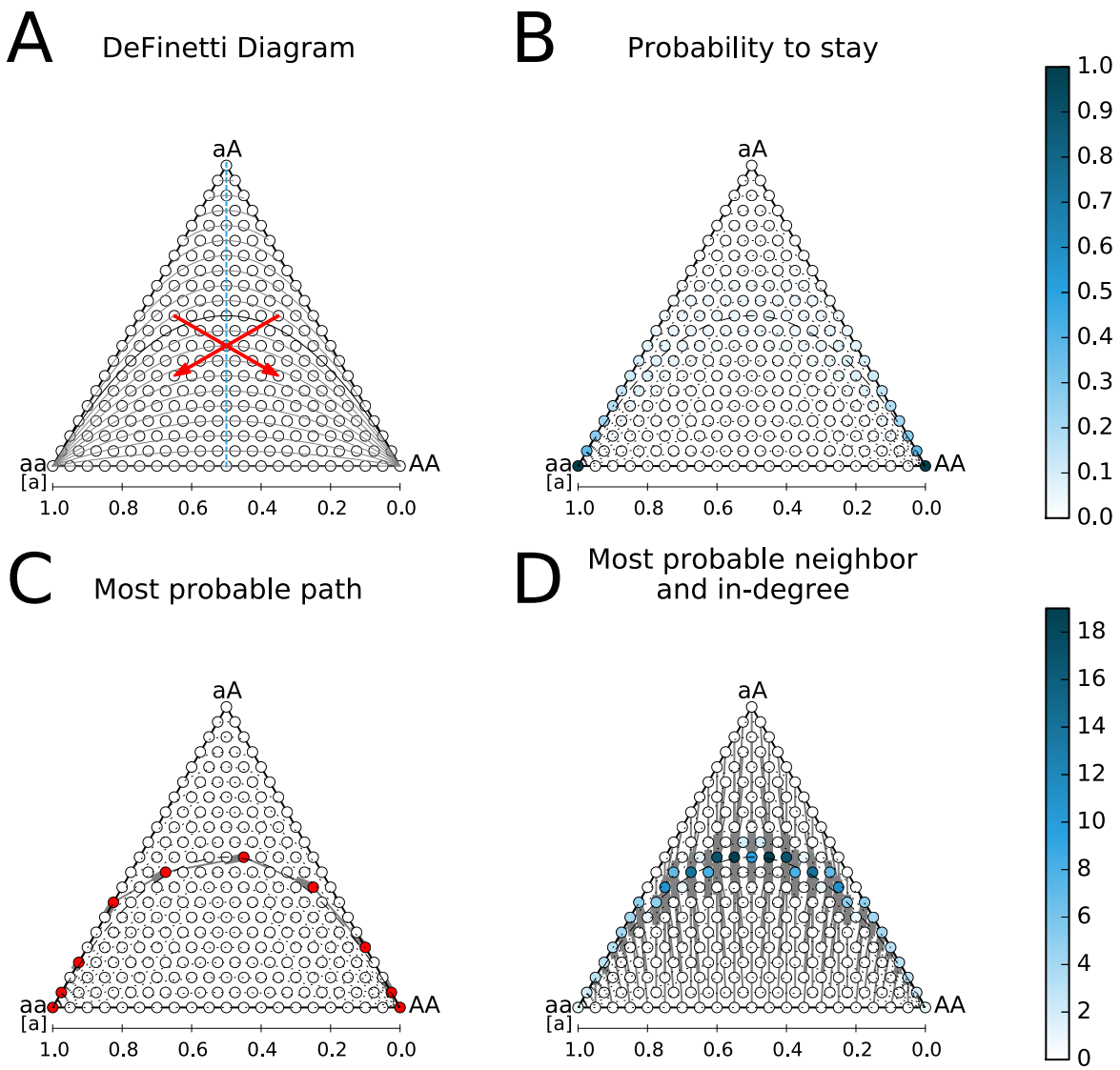

Figure A1. Enlarged version of figure 1. Network display of transition matrices for $N = 20, \mu = 10^{-6}, c = 0.0$. A. *De Finetti* diagram showing symmetry (dashed blue axis, red arrows corresponding to identical probabilities) and F_{IS} isocurves (gray and black) B. p_{stay} (node color), probability to stay at each node for one time step C. most probable path connecting $(N, 0, 0)$ to $(0, 0, N)$ D. most probable neighbors (directed edges) and in-degree (node color), i.e. for each node the most likely outbound transition at the next time step and the number of inbound most likely transitions from other states.

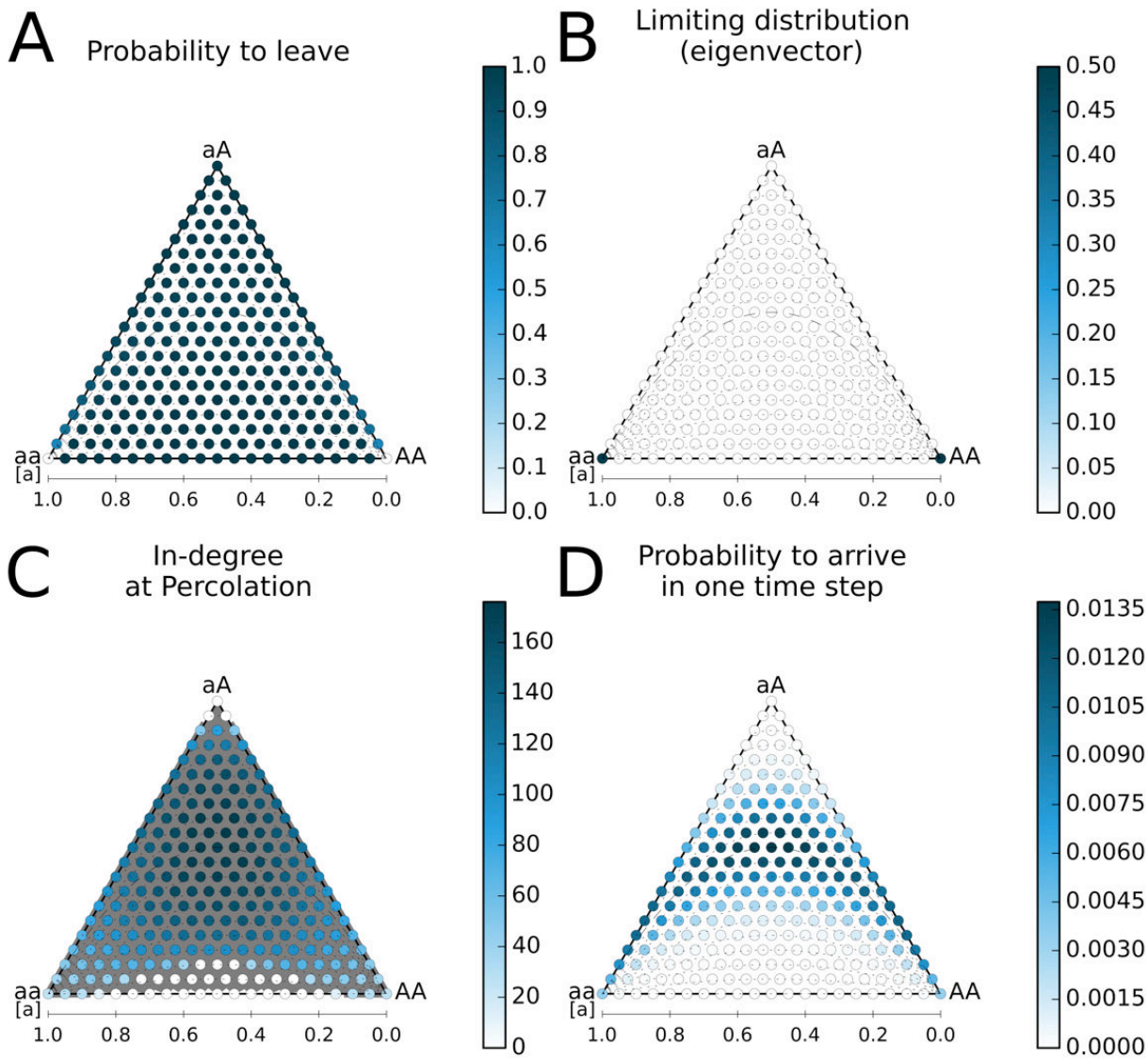

Figure A2. Network display of transition matrices for $N = 20, \mu = 10^{-6}, c = 0.0$. A. p_{out} (node color), probability to leave this node in the next time step B. p^{∞} (node color), limiting probability of each state C. in-degree (node color) at flow between the fixation states (directed edges) D. p_{in} (node color), probability to arrive each state if all previous states are equally probable.

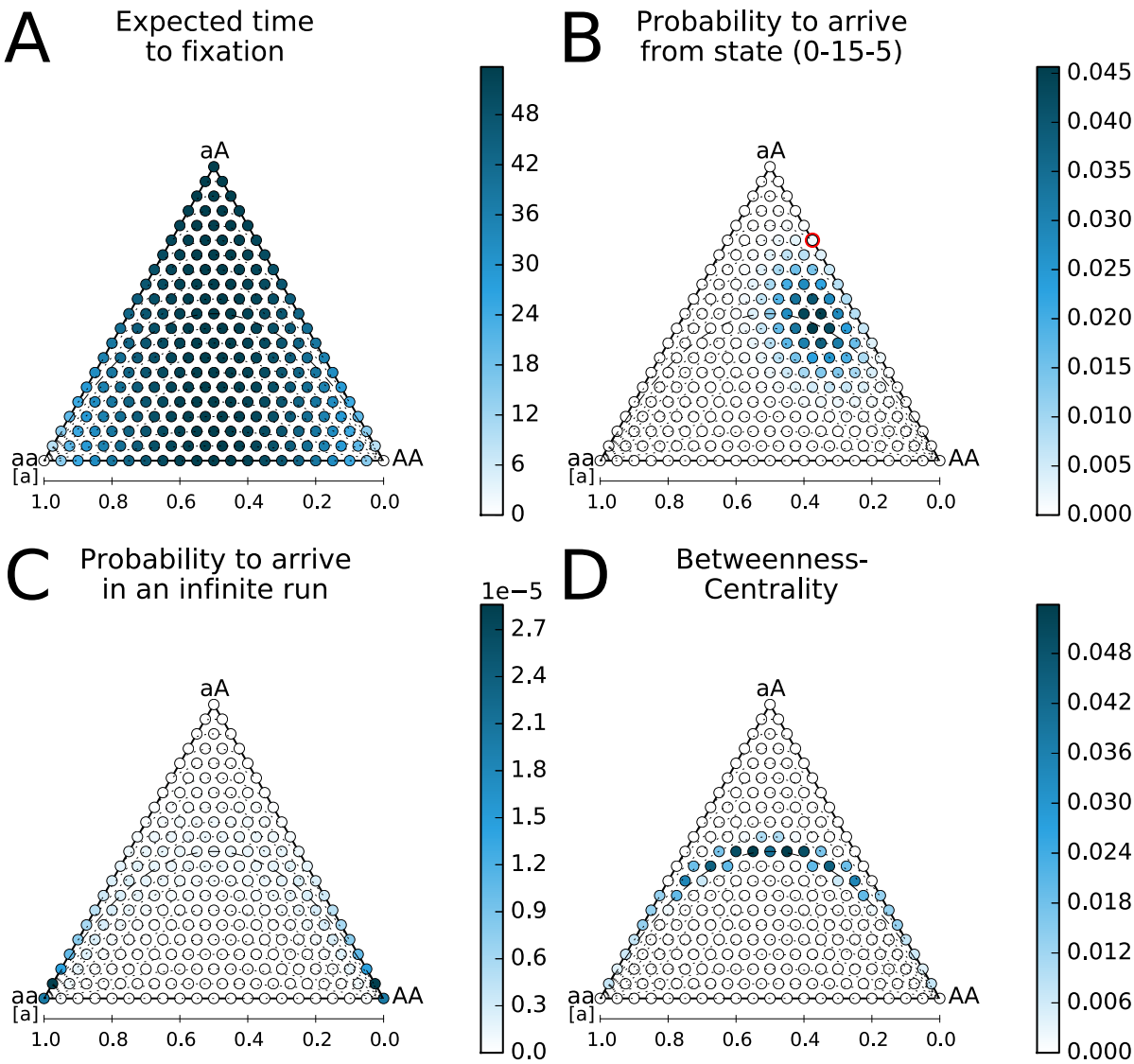

Figure A3. Network display of transition matrices for $N = 20, \mu = 10^{-6}, c = 0.0$. A. expected time to fixation (node color) according to start state B. $p(i|(0,15,5))$ (node color), probabilities of each state if the previous state was (0,15,5) C. p_{in}^{∞} (node color), probability to arrive at each state if the start state probabilities correspond to the limiting distribution D. betweenness-centrality (node color).

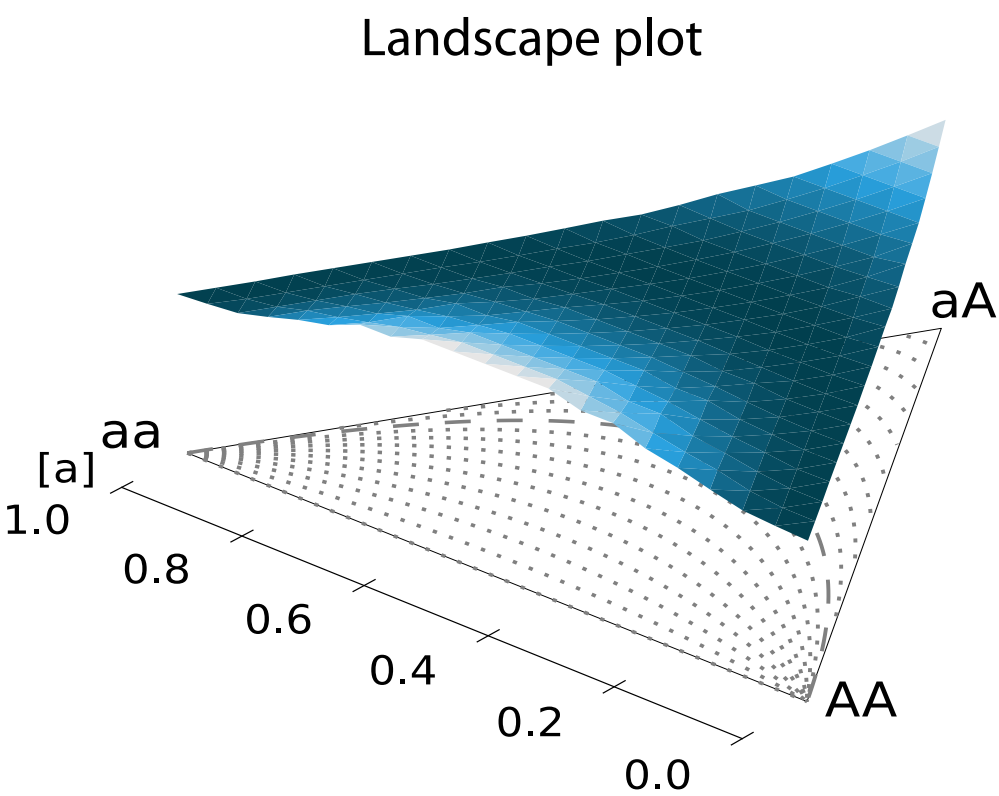

Figure A4. Landscape plot of transition matrix for $N = 20, \mu = 10^{-6}, c = 0.0$. Elevation rescaled by factor 5, color according to relative elevation ("valleys": dark blue, "hills": light grey). The lowest elevation equals zero, the reference *de Finetti* triangle is offset to -0.3.

Additional file 5 – Visualization algorithm runtimes

Table A5. Runtimes in seconds for creating each subfigure of figures 1, 2 and additional files 1-3. Means over three repetitions. Reference system: Intel Core i7-4850HQ 2.3 GHz processor, 16 Gb 1600 MHz DDR3 RAM.

Visualization method	subfigure	runtime [s]
Histogram	1-A	0.001
Histogram, logit(10)	1-B	0.001
Histogram, sparse approximation	1-C	0.055
Histogram, logit(10) reordered	1-D	0.211
De Finetti diagram	*A1-A / 2-A*	*0.405*
Most probable neighbor and in-degree	A1-D / 2-D	0.160
Probability to stay	A1-B / 2-B	0.231
Probability to arrive in an infinite run	A3-C	0.281
Probability to arrive from specified state	A3-B	0.288
Most probable path	A1-C / 2-C	0.339
Expected time to fixation	A3-A	0.380
Probability to arrive in one time step	A2-D	0.409
Limiting distribution (eigenvector)	A2-B	0.455
Probability to leave	A2-A	0.482
In-degree at percolation	A2-C	2.124
Betweenness-Centrality	A3-D	19.063

5.5 Digression: Extending the triangle

As a visual representation of the connections between individual counts, genotype and allele frequencies, heterozygosity and F_{IS}, *de Finetti* triangles could be a useful teaching tool in introductory courses on population genetics. The graphical display may provide an alternative access and could help to "popularize" otherwise somewhat abstract mathematical concepts for biology students. With this motivation, we wanted to see how the diagrams presented in article I could be extended, on the one hand to include more alleles/genotypes, on the other hand to include another F-statistics parameter, F_{ST}.

Adding a third allele to the *de Finetti* triangle turns it into an octahedron (figure 5.5), and the Hardy-Weinberg "curve" becomes a Hardy Weinberg "curved sheet". The apex of this sheet would be higher "up" (more heterozygosity) than the apex of the two allele curve ($\max(H_e) = 1 - 3 \cdot (1/3)^2 = 2/3 > 1/2$), which illustrates that $F_{IS} = 1 - H_o/H_e$ can only ever be minimal (equal minus one) for two equally frequent alleles. However, adding even mode alleles/genotypes is not practical: already for four alleles, the resulting diagram is difficult to imagine and cannot be projected on a plane anymore. A different approach to multiallelic *de Finetti* diagrams may be to "subsume" different alleles into two categories (e.g. selectively advantageous vs. disadvantageous); this corresponds to a "lumping" of states with respect to the mutation part of the model.

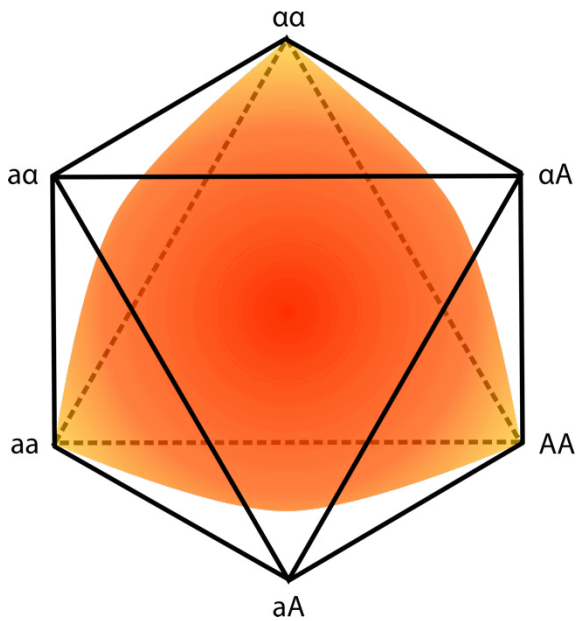

Figure 5.5 Plan view of a *de Finetti* octahedron for three alleles *a*, α, *A*, including the three-dimensional equivalent of the Hardy-Weinberg curve (orange). Heterozygous genotypes are arranged on the frontal ("upper") triangular side, homozygous genotypes on the distal ("lower", dashed lines) triangular side of the octahedron, three of the six enclosing triangular sides correspond to *de Finetti* triangles for one pair of alleles each. The hue (intensity of red) of the "Hardy-Weinberg sheet" corresponds to its distance from the base of the octahedron (expected heterozygosity; low/yellow to high/red).

Plotting F_{ST} on the *de Finetti* triangle is somewhat more complicated, since the "zero" line (mean allele frequencies over all subpopulations) shifts with the data. F_{ST} is calculated from:

$$F_{ST} = \frac{Var_{sub}}{Var_{total}} = 1 - \frac{\overline{\nu_a(1-\nu_a)}}{\overline{\nu_a}(1-\overline{\nu_a})}$$

where Var_{sub} is the variance of allele frequencies within each subpopulation, Var_{total} the variance of allele frequencies over all subpopulations (i.e. treating them as one big population), and ν_a in the second equality stands for the frequency of the *a* allele in our two-allele example. While $\nu_a(1-\nu_a)$ for any ν_a has a simple geometric equivalent (red curve in figure 5.6; vertical arrows pointing towards it represent individual values), the

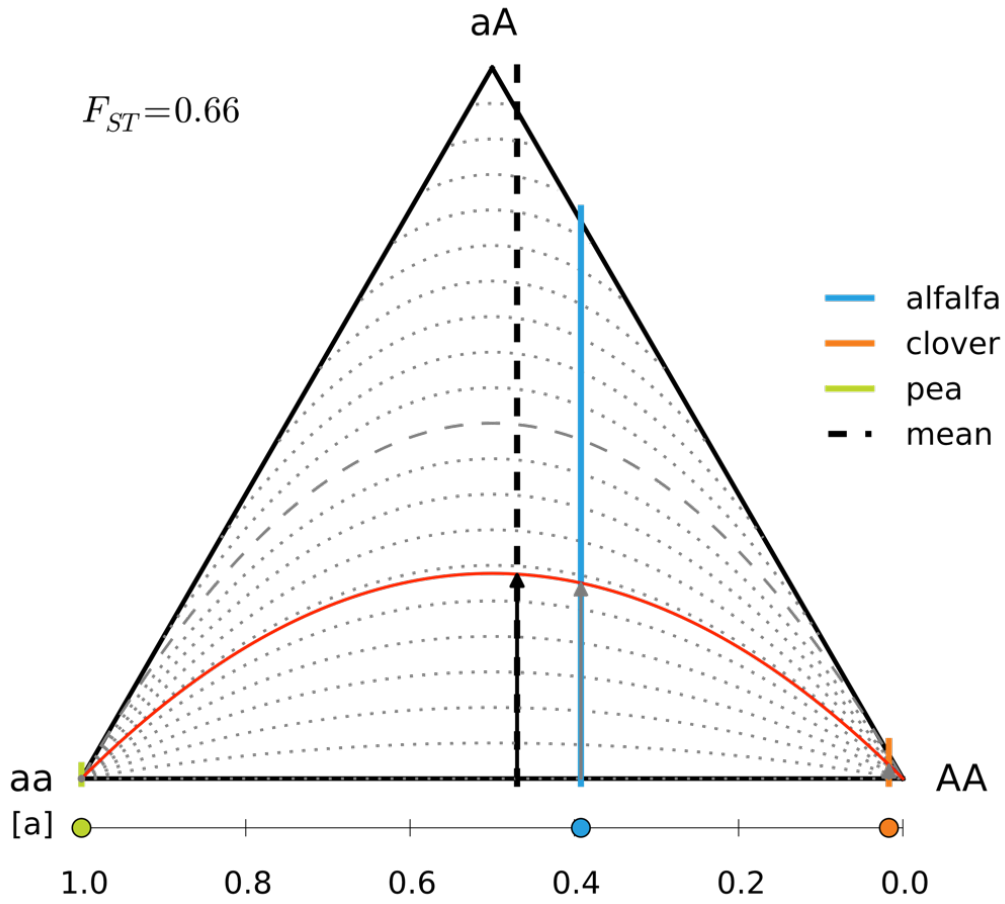

Figure 5.6 Displaying F_{ST} on a *de Finetti* triangle. Green, blue and orange points and bars: allele frequency examples for one SNP in three pea aphid subpopulations on different host plants (with friendly permission of Pierre Nouhaud). Black dashed/dotted lines: Hardy-Weinberg curve / F_{IS} isolines; red dashed line: mean of the allele frequencies; red curve: product of the allele frequencies $\nu_a(1-\nu_a) = \nu_a\nu_A$. Black vertical arrow: $\overline{\nu_a}(1-\overline{\nu_a})$, denominator of the F_{ST} fraction (see text); Grey vertical arrows: $\nu_a(1-\nu_a)$ for each subpopulation, the mean over their lengths gives the numerator of the F_{ST} fraction (see text). Here, the mean of the grey arrow lengths is about one third of the black arrow length; consequently, F_{ST} is about two thirds, which is in accordance with the exact calculation (F_{ST} value in plot).

mean $\overline{\nu_a(1-\nu_a)}$ over the values for each subpopulation is not as easily derived for more than two subpopulations. Still, for two populations one can show e.g. that, if the overall mean allele frequency is close to 0.5, F_{ST} increases the more the subpopulation means diverge from this mean value (reaching $F_{ST} = 1$ if a different allele is fixed in both subpopulations).

In conclusion, higher-dimensional extensions of *de Finetti* diagrams to accommodate more alleles are not very practical, though they may help to understand how the range of F_{IS} values changes compared to the simple two-alleles case. Though F_{ST} cannot be displayed on the *de Finetti* diagram in an equally unambiguous and concise way as F_{IS}, the diagrams could serve to illustrate the complementarity (difference in heterozygosity vs. difference in allele frequencies) of both parameters.

As an example for a ternary diagram, the *de Finetti* diagram also links the analysis of population genetic data to other domains of science, e.g. geology: genotype frequencies are a special kind of "compositional data" (i.e. they sum to a constant: one), and the development of statistical analysis tools (e.g. principal component analysis; Aitchison & Egozcue 2005) for such data are an emerging topic in statistics. Exchange between different subject areas may bring impulses for either field.

Part III Results

6 Selectively neutral diversity

6.1 Neutral diversity under acyclic partial asexuality

The typical aim of population genetic studies is to make inferences about a population's ecology and evolution, such as spatial substructure and sources/directions of migration, population demography or selection for particular traits/loci. Currently, such studies are mostly based on data for single loci: besides AFLPs, which have a complex mutation pattern, do not allow the detection of heterozygosity and shall therefore not be treated here, the currently most popular techniques for population genetic studies are microsatellites (SSRs – short single sequence repeats) and single nucleotide polymorphisms (SNPs).

The most important reference for the patterns of genetic diversity expected at single loci is the Hardy-Weinberg equilibrium (Hardy 1908, Weinberg 1908): due to random mating, at the population level the alleles at each locus should be randomly associated in the absence of all other evolutionary "forces" (e.g. mutation, genetic drift, but also migration, selection). Thus, if a locus is not in Hardy-Weinberg equilibrium, the action of at least one other evolutionary force can be deduced – but which one is often not clear.

The reproductive system is another evolutionary "force" that may lead to deviations from the Hardy-Weinberg equilibrium: mating is not always random, and reproduction may even occur without it. This is the case for partially asexual species. An "update" of the Hardy-Weinberg expectation, ideally already accounting for other "confounding factors" such as mutation and genetic drift due to a finite population size, could therefore greatly help in the analysis of single-locus population genetic data from populations with this reproductive system. However, such a reference might be dependent on the rate of clonality (compare Balloux et al. 2003, Bengtsson 2003, Ceplitis 2003, Stoeckel & Masson 2014), and, moreover, on time (compare Marshall & Weir 1979). In contrast to the previous studies (except Marshall & Weir 1979), we therefore analyzed the dynamics of $\mathrm{F_{IS}}$ in partially clonal organisms.

According to the relative strength of the three evolutionary processes included in our model, based on the parameters population size, mutation rate and rate of clonality, we found three domains for the dynamics of $\mathrm{F_{IS}}$: if either sexual reproduction (low rate of clonality) or mutation (high mutation rate) dominate the dynamics, $\mathrm{F_{IS}}$ converges to zero (Hardy-Weinberg equilibrium), but if genetic drift dominates (small population size, high rate of clonality, low mutation rate), the population loses its genotypic diversity until only one genotype (heterozygous: $\mathrm{F_{IS}} = -1$, homozygous: fixation) is left. Populations with a higher rate of clonality generally take (non-linearly) longer until they have reached their expected mean $\mathrm{F_{IS}}$ value after a deviation and show a greater variation about this mean. Therefore, in contrast to exclusively sexual populations which reach their expectation ($\mathrm{F_{IS}} = 0$) in only one time step, either time series data or information about the history of the population is necessary to correctly interpret instantaneous $\mathrm{F_{IS}}$ values in partially asexual populations.

The dynamics of F_{IS} in partially asexual populations are thus different from those in exclusively sexual or exclusively asexual populations, and change with the rate of clonality. For a case of special interest in the "evolution of sex debate", i.e. big, old exclusively asexual populations, we found that it would be indistinguishable from an exclusively sexual population, which is contrary to some of the previous models ("Meselson effect", compare Bengtsson 2003, Ceplitis 2003; discussed in chapter 8.1.3). Besides a reference for the interpretation of F_{IS} values in partially asexual species, this study also provides suggestions for data collection: firstly, because of the increased variation of F_{IS}, more loci need to be sampled to have an equally accurate estimate of its mean value. Secondly, all samples (not only one per distinct "multilocus" genotype) should be included in the calculation of F_{IS}. Finally, demographic bottlenecks should leave more persistent traces (negative values) in the genetic diversity / F_{IS} under partial asexuality than expected from exclusively sexual reproduction. This effect is further explored in chapter 6.2.

La diversité neutre sous asexualité partielle acyclique

L'objectif typique des études de génétique des populations est d'amener à des conclusions sur l'écologie et l'évolution d'une population, tels que sa sous-structure spatiale et des sources / directions de la migration, la démographie de la population ou de la sélection pour des traits / loci particuliers. Actuellement, ces études sont principalement basées sur les données concernant des loci uniques : outre les AFLP, qui ont un motif de mutation complexe ne permettant pas la détection d'hétérozygotie et par conséquent ne sont pas traitées ici, les techniques actuellement les plus populaires pour les études en génétique des populations sont les microsatellites (SSR – répétitions de séquences simples courtes) et les polymorphismes nucléotidiques simples (SNP).

La référence la plus importante sur les motifs de diversité génétique attendus à un seul locus est l'équilibre de Hardy-Weinberg (Hardy 1908, Weinberg 1908) : en raison de l'accouplement aléatoire, au niveau de la population les allèles à chaque locus doivent être associés de façon aléatoire en l'absence de toutes les autres « forces » de l'évolution (comme par exemple mutation, dérive génétique, mais aussi migration et sélection). Ainsi, si un locus n'est pas en équilibre de Hardy-Weinberg, l'action d'au moins une autre force évolutive peut être déduite – mais il n'est pas souvent clair de déterminer celle dont il s'agit.

Le système de reproduction est une autre « force » évolutive qui peut conduire à des écarts par rapport à l'équilibre de Hardy-Weinberg : l'accouplement n'est pas toujours aléatoire, et la reproduction peut se produire même sans elle. C'est le cas des espèces partiellement asexuées. Une « mise à jour » de l'espérance de Hardy-Weinberg, ayant idéalement déjà tenu compte d'autres « facteurs de confusion », tels que la mutation et la dérive génétique dues à une taille de population finie, pourrait donc grandement contribuer à l'analyse des données de génétique des populations sur un locus unique chez les populations avec ce système de reproduction. Cependant, une telle référence pourrait être dépendante du taux de clonalité (comparer à Balloux et al. 2003, Bengtsson 2003, Ceplitis 2003, Stoeckel & Masson 2014), et également du temps (comparer à Marshall & Weir 1979). Par contraste avec les études précédentes (à l'exception de Marshall & Weir 1979), nous avons dès lors analysé la dynamique de l'F_{IS} chez les organismes partiellement clonaux.

Selon la force relative des trois processus évolutifs inclus dans notre modèle, basée sur les paramètres taille de la population, taux de mutation et taux de clonalité, nous avons trouvé trois domaines de la dynamique du F_{IS} : si la reproduction sexuée (faible taux de clonalité) ou la mutation (taux de mutation élevé) dominent la dynamique, F_{IS} converge vers zéro (équilibre de Hardy-Weinberg), mais si la dérive génétique est dominante (petite taille de la population, taux élevé de clonalité, faible taux de mutation), la population perd de sa diversité génotypique jusqu'à ce que seulement un génotype (hétérozygote: $F_{IS} = -1$, homozygote: fixation) y reste. Les populations avec un taux de clonalité supérieur prennent généralement (et de façon non-linéaire) plus de temps jusqu'à ce qu'ils aient atteint leur valeur F_{IS} attendue en moyenne après une déviation et montrent une plus grande variation autour de cette moyenne. Par conséquent, contrairement aux populations exclusivement sexuées qui atteignent leur attente ($F_{IS} = 0$) en une seule étape de temps, des données en séries chronologiques ou des informations sur l'histoire de la population sont nécessaires pour interpréter correctement les valeurs d' F_{IS} instantanées dans les populations partiellement asexuées.

La dynamique de l'F_{IS} dans les populations partiellement asexuées est donc différente de celles des populations exclusivement sexuées ou exclusivement asexuées, et change avec le taux de clonalité. Dans le cas d'un intérêt particulier pour le débat sur « l'évolution du sexe », à savoir les grandes et anciennes populations exclusivement asexuées, nous avons trouvé qu'il serait impossible de le distinguer d'une population exclusivement sexuée, ce qui est contraire à certains des modèles précédents (« effet Meselson », comparer Bengtsson 2003, Ceplitis 2003; voir chapitre 8.1.3). En plus d'une référence pour l'interprétation des valeurs d' F_{IS} en espèces partiellement asexuées, cette étude fournit également des recommandations pour la collecte des données : d'abord, en raison de la variation accrue du F_{IS}, plus de loci doivent être échantillonnés pour avoir une estimation aussi précise de sa valeur moyenne. Puis, tous les échantillons (non seulement un par génotype à multiples loci distinct) devraient être inclus dans le calcul de F_{IS}. Enfin, les goulots d'étranglement démographiques devraient laisser des traces plus persistantes (valeurs négatives) dans la diversité génétique / F_{IS} sous asexualité partielle que prévues à partir de la reproduction exclusivement sexuée. Cet effet est exploré dans le chapitre 6.2.

Article II *Sexualité rare ou équilibre hors portée ? La dynamique d'F_{IS} chez les organismes partiellement clonaux*

Sommaire de l'article

Contexte – Les organismes capables à la fois de la reproduction sexuée et clonale sont très répandus dans la nature, mais la façon dont leur système reproducteur influe sur la dynamique de leur diversité génétique reste mal comprise. Le coefficient de consanguinité $\mathrm{F_{IS}}$ est un indicateur classique pour les systèmes de reproduction non-standard, qui conduisent à des écarts par rapport à l'équilibre de Hardy-Weinberg ($\mathrm{F_{IS}} = 0$) attendu sous accouplement aléatoire dans des populations sexuées. Les modèles mathématiques incluant la clonalité prédisent des écarts seulement pour une reproduction sexuée extrêmement rare et seulement vers la moyenne d'$\overline{\mathrm{F_{IS}}} < 0$. Pourtant, dans des espèces partiellement clonales, $\mathrm{F_{IS}} \neq 0$ (positif ou négatif) cela est fréquemment observé, également dans les populations où la reproduction sexuée semble par ailleurs significative. La dynamique temporelle encore inconnue de l'$\mathrm{F_{IS}}$ sous clonalité partielle pourrait fournir des explications supplémentaires pour ces départs. Nous avons étudié les effets conjoints de la clonalité partielle, de la mutation et de la dérive génétique avec un modèle de chaîne de Markov discret en temps et états pour comprendre la dynamique de l'$\mathrm{F_{IS}}$ au fil du temps.

Résultats – La clonalité partielle, même à des taux modestes, affecte la dynamique de l'$\mathrm{F_{IS}}$. La clonalité augmente non seulement la variation temporelle de l'$\mathrm{F_{IS}}$, mais réduit également son taux de variation au cours du temps. D'abord, pour des petites populations le temps pour atteindre la valeur moyenne finale $\overline{\mathrm{F_{IS,\infty}}}$ après une perturbation augmente approximativement comme une fonction hyperbolique avec le taux de clonalité. Puis, les valeurs négatives et positives peuvent survenir de façon transitoire, même à des taux intermédiaires de clonalité. La reproduction partiellement clonale par elle-même ralentit la convergence à $\mathrm{F_{IS}} = 0$, mais ne provoque pas de départs de cette valeur. La mutation aléatoire dans des grandes populations conduit finalement à $\mathrm{F_{IS}} = 0$, même en l'absence de l'accouplement aléatoire. La décélération et l'inclinaison vers des valeurs légèrement négatives plutôt que positives proviennent donc principalement de l'interaction entre la clonalité partielle et la dérive génétique.

Conclusion – Nos résultats plaident en faveur d'une interprétation dynamique de l'$\mathrm{F_{IS}}$ en populations partiellement ou purement clonales. Les valeurs négatives ne peuvent pas être interprétées comme une preuve sans équivoque pour la rareté du sexe, mais aussi comme des taux intermédiaires de clonalité dans les populations de taille déterminées, générant des départs transitoires à partir d'$\mathrm{F_{IS}} = 0$. Des observations complémentaires (par exemple distribution des fréquences des génotypes à multiple loci, histoire de la population) ou des données en séries chronologiques peuvent aider à faire la distinction entre différentes conclusions possibles sur l'étendue de la clonalité, lorsque les valeurs moyennes déviant de zéro et / ou une grande variation de l'$\mathrm{F_{IS}}$ à travers des loci sont observées.

Article II *Rare sex or out of reach equilibrium? The dynamics of F_{IS} in partially clonal organisms*

Katja Reichel[1]*, Jean-Pierre Masson[1], Florent Malrieu[2], Sophie Arnaud-Haond[3], Solenn Stoeckel[1]

[1] INRA, UMR1349 IGEPP, F-35650 Le Rheu, France, [2] Université de Tours, CNRS-UMR7350 LMPT, F-37200 Tours, France, [3] IFREMER, UMR5240 MARBEC, F-34203 Sète, France

12/2015 BMC Genetics, resubmission pending

Abstract

Background – Organisms capable of both sexual and clonal reproduction are very common in nature, yet how their reproductive system influences the dynamics of their genetic diversity remains poorly understood. The coefficient of inbreeding F_{IS} is a classic indicator for non-standard reproductive systems, leading to deviations from Hardy-Weinberg equilibrium ($F_{IS} = 0$) expected under random mating in sexual populations. Mathematical models accounting for clonality predict deviations only for extremely rare sex and only towards mean $\overline{F_{IS}} < 0$. Yet in partially clonal species, $F_{IS} \neq 0$ (positive or negative) is frequently observed, also in populations where sexual reproduction seems otherwise significant. The still unknown temporal dynamics of F_{IS} under partial clonality may provide additional explanations for those departures. We studied the joint effects of partial clonality, mutation and genetic drift with a state-and-time discrete Markov chain model to understand the dynamics of F_{IS} over time.

Results – Partial clonality, even at modest rates, affects the dynamics of v. Clonality not only increases the temporal variation of F_{IS}, but also reduces its rate of change over time. First, the time to reach the final mean $\overline{F_{IS,\infty}}$ value after disturbance augments approximately hyperbolically with the rate of clonality in small populations. Secondly, both negative and positive F_{IS} values may arise transiently even at intermediate rates of clonality. Partially clonal reproduction by itself slows down convergence to $F_{IS} = 0$, but does not cause departures from it. Random mutation in large populations eventually leads to $F_{IS} = 0$, even in the absence of random mating. The deceleration and skew toward slightly negative, rather than positive, F_{IS} values thus mainly derive from the interplay between partial clonality and genetic drift.

Conclusion – Our results argue for a dynamical interpretation of F_{IS} in partially and purely clonal populations. Negative values cannot be interpreted as unequivocal evidence for rare sex, but also as intermediate rates of clonality in finite populations, generating transient departures from $F_{IS} = 0$. Complementary observations (e.g. frequency distribution of multilocus genotypes, population history) or time series data may help to discriminate between different possible conclusions on the extent of clonality, when mean $\overline{F_{IS}}$ values deviating from zero and/or a large variation of F_{IS} over loci are observed.

Keywords Partial asexuality, parthenogenesis, mating system, inbreeding coefficient, heterozygote excess, genetic diversity

Background

Reproductive systems impact the evolution of genetic diversity at the population level (Duminil et al. 2007, 2009), making them an important factor for considerations on the evolvability of species. Partially clonal species, i.e. species that are able to reproduce both sexually and clonally, are common across many phyla and ecosystems (de Meeûs et al. 2007) and represent an important part of the global biodiversity. They include many species whose evolution is closely linked to humans, such as cultivated species (McKey et al. 2010), pathogens (Tibayrenc & Ayala 2012), invasive species (Liu et al. 2006), and species threatened by extinction (e.g. Luijten et al. 1996, Sydes & Peakall 1998, Brzosko et al. 2002, Setsuko et al. 2004, Brzyski & Culley 2011). Partially clonal species are therefore frequently the subject of molecular analyses describing their genetic diversity (Schön et al. 2009), and the conclusions drawn depend on a correct understanding of the effects of their reproductive mode on the genetic composition of their populations.

The interpretation of standard population genetic indices from partially clonal populations can be challenging, as expectations may depend on the rate of clonality, which is usually unknown in natural population. Conversely, the estimate of this rate on the basis of indirect approaches such as population genetics analysis remains elusive. One example of an index that has been suggested to change with the rate of clonality is F_{IS} (Balloux et al. 2003, Halkett et al. 2005). In diploid populations, it represents a correlation coefficient among homologous alleles within the same diploid individual at a particular locus, and depends on their tendency to be randomly associated ($F_{IS} = 0$) or more likely identical ($F_{IS} > 0$) or not identical ($F_{IS} < 0$). F_{IS} is defined either based on population heterozygosity (H_e – expected heterozygosity, H_o – observed heterozygosity) or allelic identities/homozygosity (F – allelic identity within individuals, Θ – allelic identity within the population; Balloux et al. 2003):

$$F_{IS} = \frac{H_e - H_o}{H_e} \cong \frac{F - \Theta}{1 - \Theta}, \qquad F_{IS} \in [-1,1]$$

Results from both definitions differ only for loci with just a single allele remaining (fixation), where F_{IS} is usually not defined.

To date, only few mathematical models studying F_{IS} at selectively neutral loci in partially clonal populations have been published. For partially clonal populations otherwise complying with the Hardy-Weinberg conditions, F_{IS} and the underlying genotype frequencies are thought to be identical to those expected for random mating, yet the approach to the Hardy-Weinberg equilibrium (HWE) is slowed down as the rate of clonality increases (Marshall & Weir 1979). If mutation and genetic drift are taken into account (Balloux et al. 2003), very high rates of sexual reproduction are supposed to eventually lead to strongly negative mean F_{IS} values up to $\overline{F_{IS,\infty}} = -1$ for completely clonal populations. In addition to this effect on the mean, a stochastic model (Stoeckel & Masson 2014) showed that also the shape of the expected final (i.e. equilibrium) distribution of F_{IS}, measured by its variance, skewness and kurtosis, changes with the rate of clonality. Based on the results of (Balloux et al. 2003), $\overline{F_{IS}}$ was suggested as an informative parameter to estimate the rate of clonality (Halkett et al. 2005, de Meeûs et al. 2006) in connection with other indices such as linkage disequilibrium or the frequency of repeated multilocus genotypes (Arnaud-

Haond et al. 2007). However, using the mean of the final distribution provided by (Balloux et al. 2003) as a reference for the mean $\overline{F_{IS}}$ values from field studies often pointed to rates of clonality that were at odds with other indices or even direct observation (Stoeckel et al. 2006, e.g. Motoie et al. 2013).

While some previous theoretical studies appear to highlight negative $\overline{F_{IS}}$ as a signature of nearly exclusive clonality (13,14,56), others underline the influence of clonality not only on the final distribution of $\widetilde{F_{IS,\infty}}$ but also on its temporal dynamics in natural population (15,16). We aimed to complement the results of these previous studies by describing the temporal changes of genotype frequencies over time under the influence of partial clonality, mutation and genetic drift. In particular, we looked at how quickly the steady state distribution of F_{IS} is reached after a disturbance (e.g. change of reproductive system, change in demography) depending on rate of clonality, mutation rate and population size. This information could help to explain the departures from $F_{IS} = 0$ observed also in populations thought to have frequent sexual reproduction, which are otherwise unexpected.

We used a stochastic model to follow the neutral dynamics of genotype frequncies in the basic case of a single locus in a diploid, isolated and panmictic population that combines random mating and clonality. To ease the discussion of the full model, we present our results using a "bottom-up" approach starting with the effects of each parameter in isolation, and subsequently connect these partial results to analyze the "complete" system with the joint effects of reproductive system, mutation and genetic drift. Finally, we discuss how our results may assist in the interpretation of field data, based on examples from a literature review, and provide methodological recommendations for data collection and analysis in partially clonal populations.

Methods

Mathematical Model

The biological template for our model is a single population with a finite number of individuals. These individuals correspond to ramets, i.e. factually or potentially physiologically distinct units that may or may not be genetically identical or descended from the same parent. All individuals follow the same life cycle, which consists of a dominant diploid phase during which they can acquire heritable mutations (figure 1). All individuals subsequently produce offspring both clonally and by random mating (including selfing), hereafter referred to as sexuality, from which a fixed number (corresponding to the constant population size) survive randomly to replace their parents in the next generation.

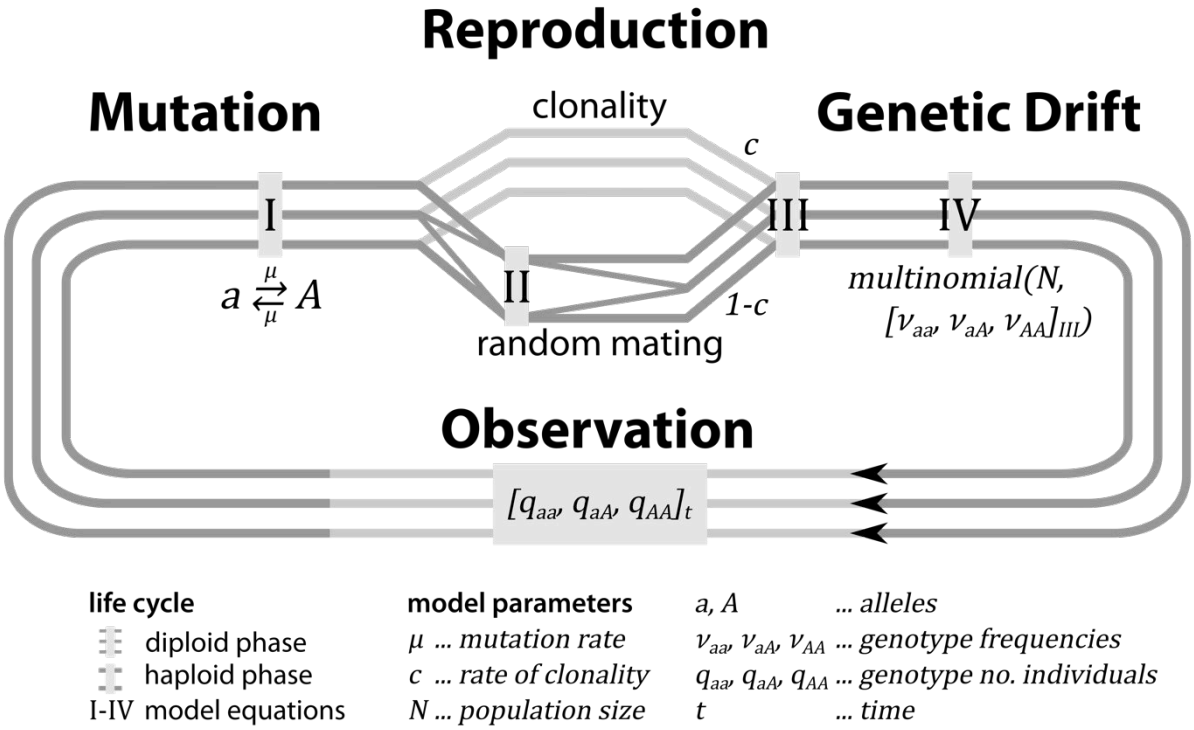

Figure 1. Schematic overview of the mathematical model (example for two alleles). In a dominantly diploid population of fixed size N, the number of individuals/ramets q with a certain genotype (here *aa, aA,* or *AA*) at a particular locus, observed at generation t, and the corresponding genotype frequencies $\nu = q/N$ may change due to mutation (here symmetrical from *a* to *A* and from *A* to *a* with rate μ; see equation I), reproduction (random mating at rate $1 - c$; see equations II and III) or genetic drift (modeled by multinomial drawing of N individuals from the genotype frequency distribution; see equation IV), until observation at the next generation.

We translated this system into a time and state discrete Markov chain model, conceptually similar to (Stoeckel & Masson 2014). Each time step of the model corresponds to one generation, i.e. the time between two consecutive observations of the population (figure 1). The model states represent all possible distributions of the N individuals on g genotypes: For a single locus with two alleles {*a, A*}, there are three different genotypes {*aa, aA, AA*}, and thus $(N + 1)(N + 2)/2$ states in the chain; for a greater number of alleles n, the number of genotypes corresponds to $g = n(n + 1)/2$ and the number of states to $(N + g - 1)!/(N! \cdot (g - 1)!)$. At each time step, the population makes a transition from its current state to a next state (where current and next state can be the same), based on a vector of transition probabilities. These probabilities depend on the genotype frequencies ν_{ii}, ν_{ij} (with the indices $i \neq j \neq k \neq l$ denoting different alleles) derived from the current state, and on the three constant model parameters population size N, mutation rate μ and rate of clonality c, according to the following equations (compare figure 1; all equations for the special case of two alleles are given in additional file 1, part 1.1):

I Mutation. The theoretical frequencies $\nu_{ii,I}, \nu_{ij,I}$ of each genotype after mutation are derived as:

$$\mathrm{I}\begin{cases}\nu_{ii,I} = & \alpha^2\nu_{ii,t} + \alpha\beta\sum_{j\neq i}\nu_{ij,t} + \beta^2\left(\sum_{j\neq i}\nu_{jj,t} + \sum_{k,j\neq i}\nu_{jk,t}\right)\\ \nu_{ij,I} = & (\alpha^2+\beta^2)\nu_{ij,t} + 2\alpha\beta(\nu_{ii,t}+\nu_{jj,t}) + (\alpha\beta+\beta^2)\left(\sum_{k\neq i,j}\nu_{ik,t} + \sum_{l\neq i,j}\nu_{jl,t}\right) + 2\beta^2\left(\sum_{k\neq i,j}\nu_{kk,t} + \sum_{k,l\neq i,j}\nu_{kl,t}\right)\end{cases}$$

where $\alpha = 1 - \mu$, the probability that an allele does not mutate, and $\beta = \mu / (n-1)$, the probability that an allele mutates into one of the $n-1$ others during one generation. This corresponds to a classic k-alleles or Jukes-Cantor substitution model (Jukes & Cantor 1969).

II Gamete formation (allele segregation). The gamete frequencies in the gamete pool after sexual reproduction $\nu_{i,I}$ are calculated as:

$$\mathrm{II}\ \nu_{i,I} = \nu_{ii,I} + \frac{1}{2}\sum_{j\neq i}\nu_{ij,I}$$

There is no difference in the allele frequencies between sexes, mating types etc., and all individuals contribute equally to the gamete pool (pangamy).

III Reproduction (clonality and syngamy). The genotype frequencies $\nu_{ii,III}, \nu_{ij,III}$ after reproduction are calculated as:

$$\mathrm{III}\begin{cases}\nu_{ii,III} = & c\nu_{ii,I} + (1-c){\nu_{i,I}}^2\\ \nu_{ij,III} = & c\nu_{ij,I} + 2(1-c)\nu_{i,I}\nu_{j,I}\end{cases}$$

based on the results from equations I and II. The rate of clonality c thus corresponds to the proportion of offspring per generation that is the result of clonal reproduction. The remainder of the offspring ("rate of sexuality" $(1-c)$) is derived from random mating including selfing (autogamy), assuming that all individuals have the same chance to mate (panmixis).

IV Genetic Drift. The vector of genotype frequencies $\vec{\nu}_{t+1}$ at the next generation, depending on the population size, is derived from:

$$\mathrm{IV}\ \vec{\nu}_{t+1} = \frac{X_{t+1}}{N} \text{ where } X_{t+1} \sim \mathcal{M}(N, \vec{\nu}_{III})$$

where X_{t+1} is the state of the model at the next generation, drawn from a multinomial distribution $\mathcal{M}$ that is based on N, the population size counting all potentially reproducing individuals (mathematically the number of samples), and $\vec{\nu}_{III}$, the vector of genotype frequencies derived from equation III (mathematically the probabilities of the genotype "categories"). Transition probabilities P between any two model states X_t, X_{t+1} can then be calculated based on:

$$P(X_{t+1}|X_t) = \frac{N!}{\prod_i q_{ii,t+1}!\prod_{i,j} q_{ij,t+1}!}\prod_i \nu_{ii,III}^{q_{ii,t+1}}\prod_{i,j}\nu_{ij,III}^{q_{ij,t+1}}$$

where $q_{ii,t+1}, q_{ij,t+1} \in \mathbb{N}_0$ are the natural numbers of individuals per genotype in the presumed next state X_{t+1} and therefore sum to N. Note that our description of genetic drift is based on genotype frequencies rather than allele frequencies. As explained in Ewens (2004), describing population genetic processes based on allele frequencies is a mathema-tical

convenience justified by HWE (i.e. assuming exclusively sexual reproduction), which assures that allele frequencies can always be directly translated into genotype frequencies. For partially clonal populations, we cannot automatically assume HWE and thus modeled all population genetic processes, including genetic drift, at the genotype level.

Model analysis and identification of biological consequences

We analyzed our model with several approaches. First, we studied the effect of each of the three model parameters (c, μ, N) on the genotype frequencies by itself. Setting the other two parameters to have no influence on the model result, i.e. $c = 1, \mu = 0$ and/or $N = \infty$ (or no random drawing in equation IV), and substituting equation II into equation III, the model reduces to one equation per process, i.e. equation I for μ, equation III (with II) for c, and equation IV for N. For each equation/process, we then determined the steady states, i.e. those combinations of genotype frequencies for which $\vec{q}_{t+1} = \vec{q}_t$, and derived the maximal expected convergence times t_c, t_μ and t_N. While t_N could only be approximated from numerical results (Markov chain first passage time approach), for c and μ convergence to the steady states is asymptotic as it can be described by geometric progressions (details of derivation in additional file 1, part 1.2). We therefore defined a universal "acceptable error" $\varepsilon = 1/(2N)$, corresponding to one half the minimal change in genotype frequency that would be measurable by exhaustive sampling in a population of finite size N, below which the distance from the steady states has to pass (convergence criterion). Using the reference times t_c, t_μ and t_N as a measure for the "strength" with which each process acts upon the genotype frequencies, we could then use this analytical basis to partition the parameter space of the full model into regions where either process dominates the genotype frequency dynamics.

Secondly, we approached the full model for the case of two alleles, by following the dynamics of F_{IS} over time from three different start states for combinations of c, μ and N representative of the different regions of the parameter space. Aggregating the transition probabilities between all model states in a transition matrix M (same current state per column, i.e. columns summing to one), the probability distribution of the model states (and consequently the probability distribution of $\widetilde{F_{IS,t}}$) at time t, given by the vector $\vec{x}_t$, is derived by matrix multiplication:

$$\vec{x}_t = M^t \vec{x}_0$$

where $\vec{x}_0$ describes the start state (vector of zeros except for a single one at X_0). We illustrated the numerical result of our model using three start states: $F_{IS,0} \in \{-1; 0; 1\}$ under isoplethic (i.e. equally frequent, $\nu_a = \nu_A = 1/n = 0.5$) allele frequencies, standing for HWE ($F_{IS,0} = 0$) and the most extreme deviations from it (complete homozygosity, $F_{IS,0} = 1$; complete heterozygosity, $F_{IS,0} = -1$). These states were chosen to represent the range of F_{IS} values and not because of their biological significance or frequency in nature. Deviations from the final mean $\overline{F_{IS,\infty}}$ may derive from a recent change in the rate of clonality (e.g. from full sexuality with $F_{IS,0} = 0$), full adaptation to past selection for ($F_{IS,0} = -1$) or against ($F_{IS,0} = 1$) heterozygotes, changes in population size (demographic bottleneck, founder event), secondary contact between two populations in which different alleles got fixed

($F_{IS,0} = 1$) or hybridization with subsequent reproductive isolation from the parents ($F_{IS,0} = -1$). Based on the transition matrix M, we also calculated the time to the final distribution of states (i.e. their "equilibrium" frequencies), which is also the time until the final distribution of $\widetilde{F_{IS,\infty}}$ (Markov chain mixing time, see additional file 1, part 1.5).

To link our results with those obtained by previous authors, we calculated the final mean $\overline{F_{IS,\infty}}$ from equation 10 in Balloux et al. (2003), setting $q_s = 1,\ q_d = 0$ (a finite population, no migration) and $s = 1/N$ (random mating):

$$\overline{F_{IS,\infty}} = \frac{1}{(2N-1) - 2N/c(1-\mu)^2}$$

and the expected time to convergence of F_{IS} iteratively from equation 5 in (Balloux et al. 2003):

$$\begin{bmatrix} 1 - H_{o,t+1} \\ 1 - H_{e,t+1} \end{bmatrix} = \begin{bmatrix} F_{t+1} \\ \Theta_{t+1} \end{bmatrix} = (1-\mu)^2 \left(\begin{bmatrix} c + \frac{1-c}{2N} & (1-c)\left(1 - \frac{1}{N}\right) \\ \frac{1}{2N} & 1 - \frac{1}{N} \end{bmatrix} \begin{bmatrix} F_t \\ \Theta_t \end{bmatrix} + \begin{bmatrix} \frac{1-c}{N} \\ \frac{1}{2N} \end{bmatrix} \right)$$

In contrast to our model, both these equations do not contain the number of alleles, since they are based on an infinite alleles model, and treat the expected and observed hetero-/homozygosity as continuous variables.

Finally, to get a better idea how our theoretical results are comparable to those published for field data, we looked at the sampling effect of using different numbers of polymorphic loci L to estimate the mean $\overline{F_{IS}}$ of the population at time t, $\overline{F_{IS,t,L}}$. Assuming that each locus represents an independent estimate of this mean (no confounding effect of linkage), and that the genotype frequencies are known exactly (exhaustive sampling of all individuals/ramets), it is derived as:

$$\overline{F_{IS,t,L}} = \frac{1}{L} \sum_{z=0}^{L} F_{IS,t,z}$$

Both assumptions are usually violated (Halkett et al. 2005, Arnaud-Haond et al. 2007), so that our results represent a conservative estimate of the true error of this method. We randomly sampled both the steady state distribution $\widetilde{F_{IS,\infty}}$ and the instantaneous distribution $\widetilde{F_{IS,50}}$ of a population that started 50 generations ago with all loci at HWE and equal allele frequencies (isoplethy for two alleles per locus), for the same parameter combinations that we previously used to illustrate the dynamics of the full model. Based on 10^5 random samples of size L, we then calculated the mean signed deviation of the sample means $\dot{F}_{IS,t}$ from the true mean $\overline{F_{IS,t}}$:

$$\Delta\overline{F_{IS,t}} = \begin{cases} \frac{1}{z_1} \sum_{\dot{F}_{IS,t} \geq \overline{F_{IS,t}}} \dot{F}_{IS,t} - \overline{F_{IS,t}} \\ \frac{1}{z_2} \sum_{\dot{F}_{IS,t} < \overline{F_{IS,t}}} \dot{F}_{IS,t} - \overline{F_{IS,t}} \end{cases}, \qquad z_1 + z_2 = 10^5$$

where z_1 and z_2 represent the number of positive and negative deviations, respectively. Loci at or near fixation are typically not used in population genetic studies, since they are

especially affected by genotyping errors. We therefore excluded all loci where the frequency of one allele exceeds $1-\sqrt{1/(2N)}$ (near fixation; see additional file 2, part 2.1 for the derivation of this value, and compare similar considerations in Graffelman & Camarena 2008) from the calculation of values for this analysis.

All computations were performed in Python 2.7 with 64 bit precision, using the modules *numpy, scipy* (Oliphant 2007), *networkx* (Hagberg et al. 2008) and *matplotlib* (Hunter 2007). We illustrate some of our results with *de Finetti* diagrams (de Finetti 1926, Reichel et al. 2014, figures 2A-4A), which are ternary plots of the genotype frequencies $[\nu_{aa}, \nu_{aA}, \nu_{AA}]$ at one locus with two alleles within a population (see additional file 2, figures 2-1 to 2-4 for more information). Details for the literature review in the discussion are given in additional file 3.

Results

The dynamics of F_{IS} and the underlying genotype frequencies through time were affected by the rate of clonality. However, we found that this effect strongly depends on interactions with the mutation rate and population size. We therefore first present the dynamics due to each parameter by itself before analyzing the combination of the three evolutionary forces.

Dynamics of genetic diversity due to each parameter

Partial clonality, c

In this section, only reproduction may change the genotype frequencies (equation II and III, $\mu = 0, N = \infty$), which corresponds to an infinite-sized and non-mutating population. Under exclusively sexual reproduction by random mating, F_{IS} converges to zero in just one time step (figure 2), while exclusively clonal reproduction *per se* produces not change from the parental genotype frequencies and does not change F_{IS}. Between these two extreme cases, there is still convergence towards $F_{IS} = 0$, though it takes longer as the rate of clonal reproduction increases (figure 2B). This result is independent of the number of alleles, as in fact the allele frequencies are not affected by either clonal reproduction or random mating in the absence of genetic drift and mutation.

The maximal time t_c until convergence to $F_{IS} = 0$, due to reproduction only, is directly dependent on the rate of clonal reproduction (derived in additional file 1, part 1.2; compare also equation 5 in Marshall & Weir 1979). It can be approximated as

$$t_c = 1 + \log_c \varepsilon = 1 + \frac{\log \varepsilon}{\log c}$$

with ε corresponding to a small error term, e.g. $1/2N$, as its convergence is asymptotic. Between $t_c = 1$ for exclusively sexual and $t_c = \infty$ for exclusively clonal populations, the dependence of t_c on c is not linear, but increases hyperbolically (figure 2B). Consequently, if the rate of clonal reproduction is low, the same increase in c leads to a much smaller increase in the time to convergence than if clonal reproduction were already comparatively frequent.

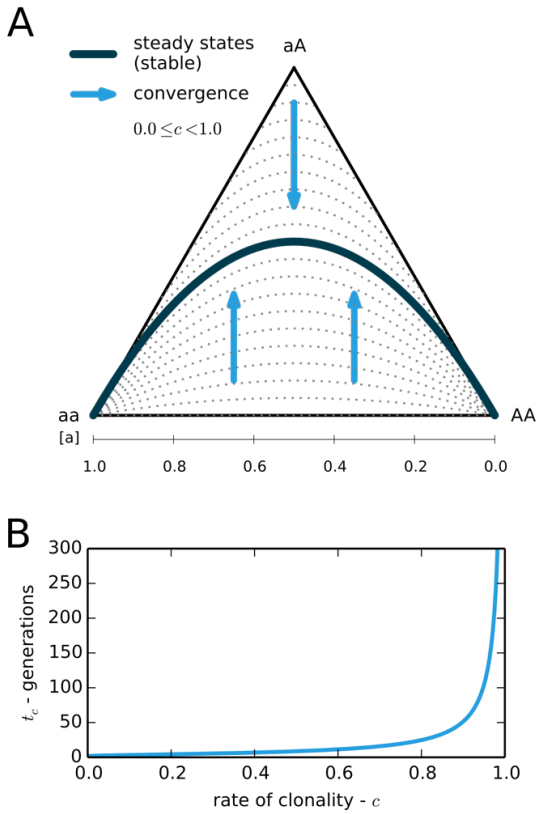

Figure 2. Genotype dynamics due to reproduction (random mating and clonality) only. A: Convergence pattern for $0.0 \leq c < 1.0$, based on figure A2-2 in additional file 2. Arrows indicate the direction of genotype frequency change over time, dark blue line indicates (stable) steady states where genotype frequencies do not change anymore. No genotype frequency changes due to reproduction for $c = 1.0$. Discontinuous grey lines connect states of equal F_{IS} (dashed: $F_{IS} = 0$, dotted: $F_{IS} = \pm 0.1, 0.2 \ldots 1$) B: Maximal expected convergence time t_c in generations for each rate of clonality c.

Mutation μ

In this section, only mutation may change the genotype frequencies (equation I, $c = 1, N = \infty$), which corresponds to an infinite-sized population of non-reproducing and immortal single-celled individuals. Mutation not only introduces diversity in genotypically uniform populations by shifting genotype frequencies away from fixation (figure 3A). It also leads to a convergence to $F_{IS} = 0$ (random association of alleles) if the alleles at both copies of a locus within an individual/cell mutate independently. Assuming equal mutation rates between all alleles, the genotype frequencies thus converge to HWE for isoplethic alleles (for two alleles, $[0.25, 0.5, 0.25]$ or the vertex of the Hardy-Weinberg parabola in figure 3A; for n alleles see additional file 1, part 1.3). However, we also show that other mutation models (e.g. step-wise mutation model for SSRs or transition/transversion models for SNPs, see additional file 1, part 1.3) only affect the final allele frequencies and not F_{IS} as long as mutations are independent of the homologous allele (i.e. excluding for instance gene conversion). Increasing the mutation rate decreases the convergence time (figure 3B).

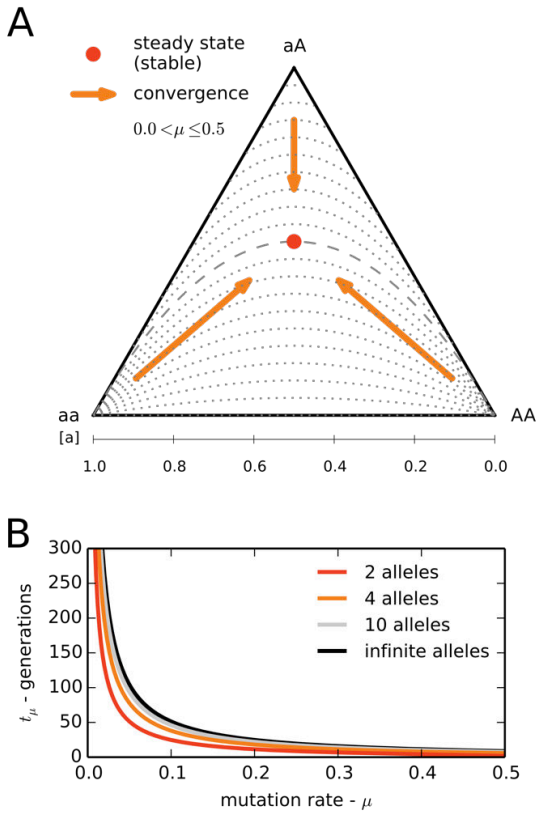

Figure 3. Genotype dynamics due to mutation (k-allele/Jukes-Cantor model) only. A: Convergence pattern for $0.0 < \mu \leq 0.5$, based on figure A2-3 in additional file 2. Arrows indicate the direction of genotype frequency change over time, red dot the (stable) steady state where genotype frequencies do not change anymore. No genotype frequency changes due to mutation for $\mu = 0.0$. Discontinuous grey lines connect states of equal F_{IS} (dashed: $F_{IS} = 0$, dotted: $F_{IS} = \pm 0.1, 0.2 \ldots 1$) B: Maximal expected convergence time t_μ in generations for each rate of mutation μ and different numbers of alleles (red: 2, orange: 4, grey: 10, black: infinite).

The maximal time t_μ until convergence to $F_{IS} = 0$ and equal allele frequencies, due to mutation only, is directly dependent on the mutation rate. It can be approximated as

$$t_\mu = 1 + \log_{\left(1-\mu\frac{n}{n-1}\right)} \varepsilon = 1 + \frac{\log \varepsilon}{\log(1 - \mu\frac{n}{n-1})},$$

which simplifies for two alleles into

$$t_\mu = 1 + \log_{(1-2\mu)} \varepsilon = 1 + \frac{\log \varepsilon}{\log(1 - 2\mu)}.$$

For the highest mutation rate we analyzed, $\mu = 0.5$ (i.e. each allele has the same chance to mutate or not to mutate), the time to convergence t_μ is thus only one generation. Natural mutation rates are typically much lower (Drake et al. 1998, Hile et al. 2000), ranging from 10^{-3} to 10^{-18}. Convergence due to mutation can therefore take very long: by setting $\varepsilon = 0.005$ and assuming a mutation rate of $\mu = 10^{-6}$, it would take up to around $2.6 \cdot \mu^{-1}$ or 2.6 million generations.

Genetic drift N

In this section, only genetic drift may change the genotype frequencies (equation IV, $c = 1, \mu = 0$), which corresponds to an exclusively clonal, non-mutating population of fixed size. Contrary to mutation and sexual reproduction, genetic drift does not lead to a universal convergence of F_{IS} values, but instead to genotypic uniformity (figure 4A). Consequently, F_{IS} either becomes -1 if the remaining genotype is heterozygous, or F_{IS} cannot be defined if the remaining genotype is homozygous. Looking only at the allele frequencies, this leads to the somewhat unusual situation that not just one allele can be stochastically fixed in the population, but also two (or more, depending on the organism's ploidy level) at equal frequencies (compare also explanation to equation IV in the description of our model).

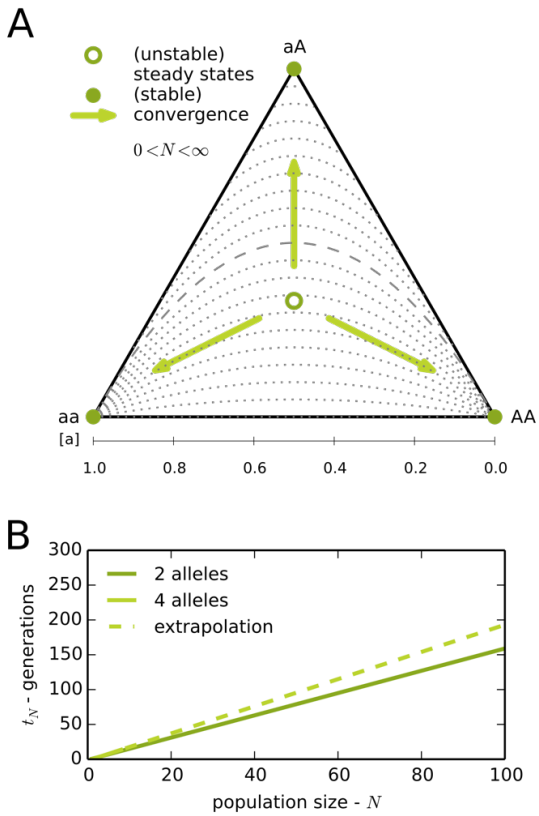

Figure 4. Genotype dynamics due to random genetic drift only. A: Convergence pattern for $0.0 < N < \infty$, based on figure A2-4 in additional file 2. Arrows indicate the direction of genotype frequency change over time, green dots steady states where genotype frequencies do not change in the mean (unstable steady state, unfilled dot), or not at all (stable steady states, filled dots). No genotype frequency changes due to genetic drift for $N = \infty$. Discontinuous grey lines connect states of equal F_{IS} (dashed: $F_{IS} = 0$, dotted: $F_{IS} = \pm 0.1, 0.2 \ldots 1$) B: Maximal expected convergence time t_N in generations for population sizes from 1 to 100, for two different numbers of alleles (darker green: 2, lighter green: 4). Results for four alleles are in part based on an extrapolation (dashed line) from numerical solutions for smaller population sizes.

Which genotype is most likely to finally prevail depends on the initial genotype frequencies: If all genotypes are equally frequent (unstable steady state in figure 4A), all are equally

probable and no prediction can be made. Otherwise, the more frequent one genotype becomes, the less probable it is that it will yet be superseded by a currently rarer genotype. Once one genotype is completely lost from the population, it cannot reappear (still assuming the absence of sexual reproduction or mutation), therefore the convergence to genotypic uniformity is final (stable steady states in figure 4A). The time to convergence decreases with $1/N$, the "quantum step size" of change in genotype frequencies (frequency equivalent of one individual) in a finite population.

To link this result to the dynamics of F_{IS} under genetic drift, we consider the frequencies of different genotypes at $F_{IS} = 0$. For the example of two alleles, the heterozygous genotype is the most frequent if both allele frequencies $\nu_a, \nu_A > 1/3$ (see additional file 1, part 1.4 for n alleles), outside this range one of the homozygous genotypes dominates the population. If both alleles are nearly isoplethic, stochastically increasing heterozygosity and "drifting" towards negative F_{IS} is therefore most likely. For more than two alleles, though an even smaller excess of one allele compared to the others results in a homozygous genotype being the most frequent at $F_{IS} = 0$, this effect can be outweighed by the cumulative frequency of all heterozygous genotypes (expected heterozygosity, $\max(H_e) = (n-1)/n$) still being larger than the cumulative frequency of all homozygous genotypes (expected homozygosity, $1 - H_e$) over an increasing range of allele frequencies.

The maximal expected time t_N required to reach genotypic uniformity, due to genetic drift only, is directly dependent on the population size, but also on the number of genotypes (maximum depending on the number of alleles; figure 4B). For small population sizes, t_N always equals two for $N = 2$, grows approximately linearly with N (additional file 1, part 1.2), and the slope of the linear approximation increases with the number of genotypes/alleles. As an example, in a population of 100 individuals up to about 160 generations are required until only one genotype remains at a locus with two alleles, depending on the start state.

Dynamics under mutation, genetic drift and partial clonality

In this section, genotype frequencies will be affected by mutation, partial clonality and genetic drift together (equations I-IV, $c \in [0,1]$, $\mu \in]0, 0.5]$, $N \in [1, \infty[$), corresponding to the full biological model. Consequently, the dynamics of genotype frequencies and F_{IS} will follow a combination of the patterns we presented for each subsystem. Using the maximal convergence times t_c, t_μ, t_N as a measure of the relative "strength" of each process, the three-dimensional parameter space can be partitioned into different parts where either process dominates (figure 5): If $t_c \ll t_\mu, t_N$ as is usually the case for strictly sexually reproducing populations, genotype frequencies quickly converge to HWE, i.e. $F_{IS,\infty} \cong 0$. For $t_\mu \ll t_c, t_N$, a situation that applies to very big, highly clonal populations, genotype frequencies converge to HWE and the final allele frequencies due to dominant mutation, so that $F_{IS,\infty} \cong 0$ also in this case. Finally, if $t_N \ll t_\mu, t_c$, as in smaller highly clonal populations, dominant genetic drift does not lead to convergence to $F_{IS,\infty} = 0$, but rather to a successive loss of genotypes and eventually to genotypic uniformity (fixation or $F_{IS,\infty} = -1$). Since both random mating and mutation lead to a random association of alleles, the relative "strength" of genetic drift determines the dynamics of F_{IS} in (partially) clonal populations.

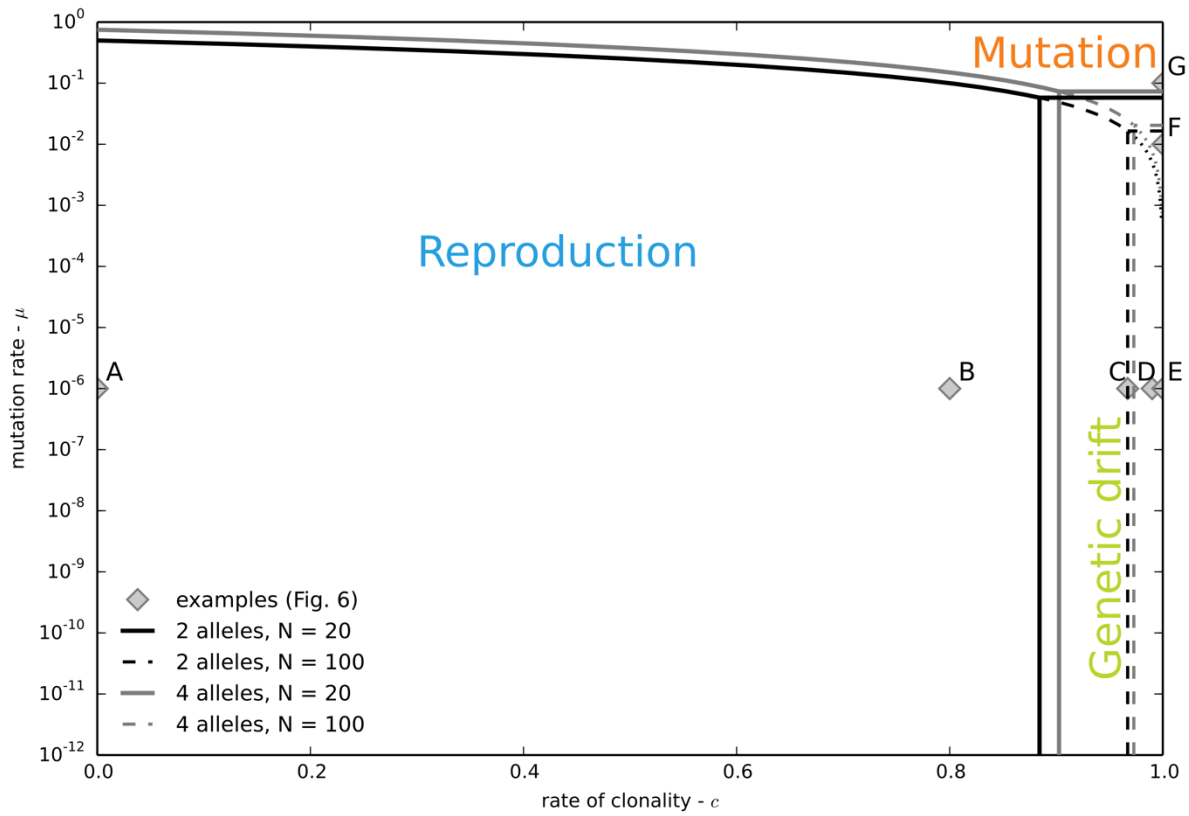

Figure 5. Overview of the model parameter space. With regions where the genotype dynamics are dominated by either reproduction, mutation or genetic drift. Lines correspond to $t_c = t_\mu$, $t_c = t_N$, $t_\mu = t_N$ for two different numbers of alleles (black: 2 alleles, grey: 4 alleles) and two different population sizes (continuous: $N = 20$, dashed: $N = 100$). Labeled dots A-G indicate examples for which the dynamics of F_{IS} are shown in figure 6.

Transitions between the predominance of either process are not abrupt, nor do different processes globally compensate each other, as each convergence pattern is different (figures 2A-4A). Keeping population size and mutation rate constant ($N = 100, \mu = 10^{-6}$) and successively increasing the rate of clonality (figure 6A-E, see also additional file 2 part 2.5) illustrates the changes in the dynamics of F_{IS} as genetic drift takes over: At low to intermediate rates of clonality (figure 6A,B; compare also with figure 5 to see how the range of c to which this applies increases with population size), the dynamics of F_{IS} are almost identical to those expected for a purely sexual population ($c = 0$, figure 6A). While the final mean $\overline{F_{IS,\infty}}$ is yet hardly different from zero for intermediate rates of clonality ($c = 0.8$: $\overline{F_{IS,\infty}} \approx -0.02$, figure 6B), variation around the mean is already increased, and extreme initial $F_{IS,0}$ values can be traced over a certain time. These tendencies – increasingly negative $\overline{F_{IS,\infty}}$, increased variation of F_{IS} values, and increased time/start value dependence – continue until t_c reaches t_N ($c \approx 0.97$: $\overline{F_{IS,\infty}} \approx -0.14$, figure 6C) and then gain even further momentum as sexual reproduction becomes very rare ($c = 0.99$: $\overline{F_{IS,\infty}} \approx -0.33$, figure 6D).

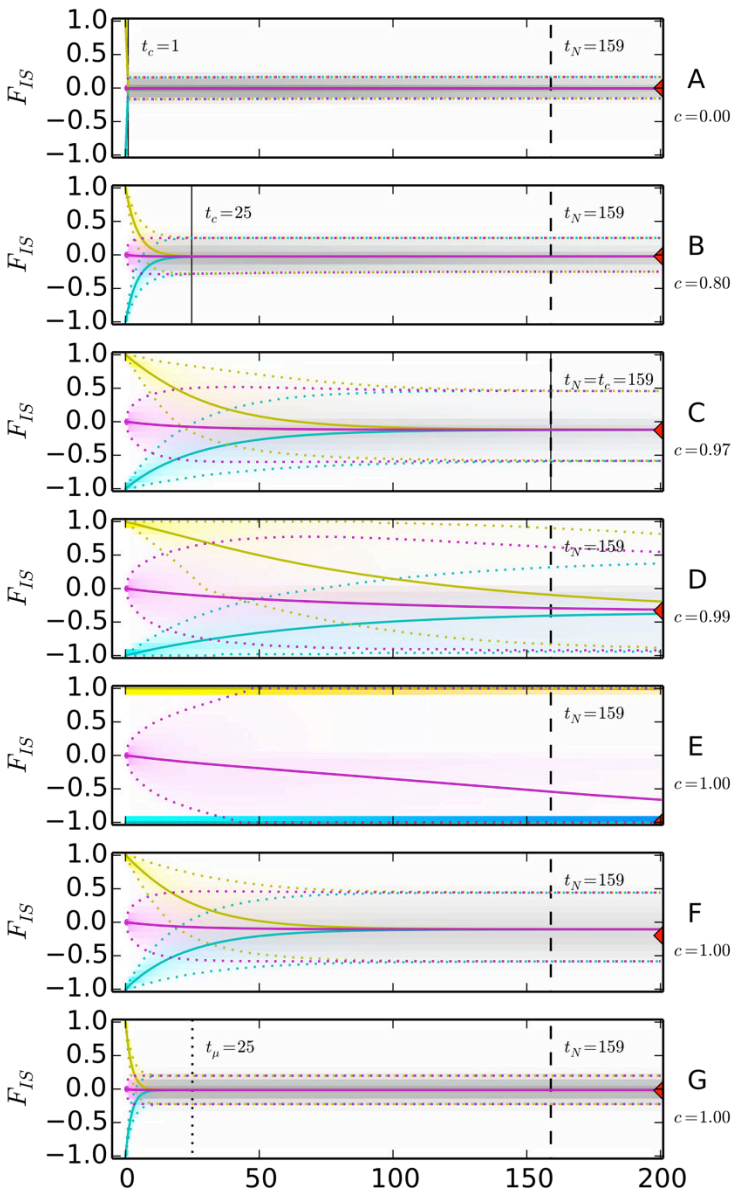

Figure 6. Dynamics of F_{IS} through time for six representative example parameter sets. Single loci with two alleles. Vertical lines represent t_c (continuous), t_N (dashed) and t_μ (dotted), colors represent different start states (yellow: $F_{IS,0} = 1$ for $\nu_a = \nu_A$, magenta: $F_{IS,0} = 0$ for $\nu_a = \nu_A$, cyan: $F_{IS,0} = -1$), with their respective $\widetilde{F_{IS,t}}$ distributions (shading), mean (continuous line) and 95% confidence intervall (dotted lines). Red triangles at $t = 200$ indicate the mean $\overline{F_{IS,\infty}}$ according to Balloux et al. (2003). Model parameters – A: $c = 0$, $\mu = 10^{-6}$, $N = 100$; B: $c = 0.8$, $\mu = 10^{-6}$, $N = 100$; C: $c \approx 0.97$ ($t_c = t_N$), $\mu = 10^{-6}$, $N = 100$; D: $c = 0.99$, $\mu = 10^{-6}$, $N = 100$; E: $c = 1.0$, $\mu = 10^{-6}$, $N = 100$; F: $c = 1.0$, $\mu = 10^{-2}$, $N = 100$; G: $c = 1.0$, $\mu = 10^{-1}$, $N = 100$.

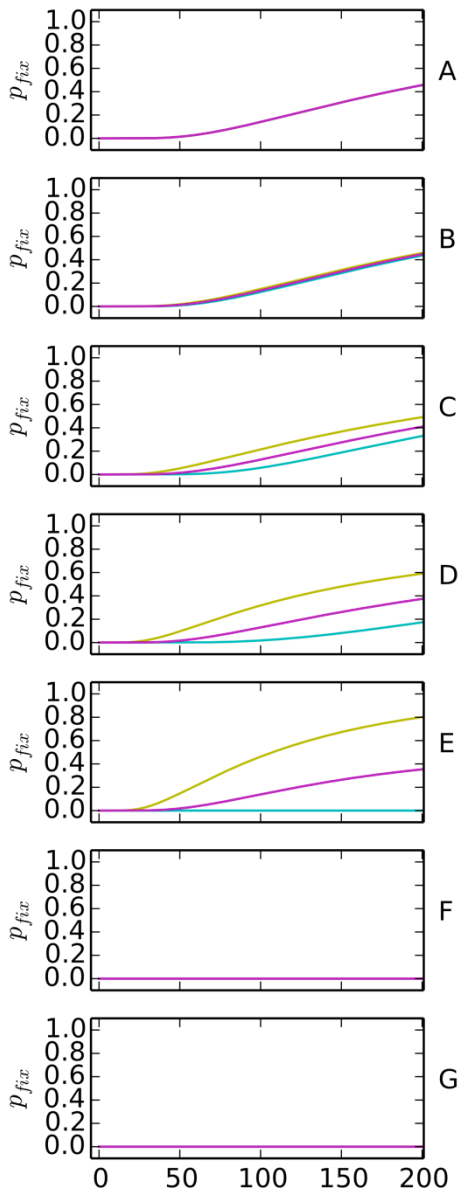

Figure 7. Dynamics of the probability of fixation p_{fix} through time for six representative example parameter sets. Single loci with two alleles, colours represent different start states (yellow: $F_{IS,0} = 1$ for $\nu_a = \nu_A$, magenta: $F_{IS,0} = 0$ for $\nu_a = \nu_A$, cyan: $F_{IS,0} = -1$). Model parameters – A: $c = 0$, $\mu = 10^{-6}$, $N = 100$; B: $c = 0.8$, $\mu = 10^{-6}$, $N = 100$; C: $c \approx 0.97$ ($t_c = t_N$), $\mu = 10^{-6}$, $N = 100$; D: $c = 0.99$, $\mu = 10^{-6}$, $N = 100$; E: $c = 1.0$, $\mu = 10^{-6}$, $N = 100$; F: $c = 1.0$, $\mu = 10^{-2}$, $N = 100$; G: $c = 1.0$, $\mu = 10^{-1}$, $N = 100$.

Finally, for an exclusively clonal population with low mutation rate/small population size ($c \approx 1.0$: $\overline{F_{IS,\infty}} \approx -1.00$, figure 6E), F_{IS} dynamics are governed by genetic drift: Loci with an initial $F_{IS,0} = -1$ (one heterozygous genotype) are effectively "fixed" though there are two different alleles, loci starting from $F_{IS,0} = 1$ at isoplethy rarely acquire any heterozygosity anymore, and loci starting from $F_{IS,0} = 0$ at isoplethy show a wide variety of both positive and negative F_{IS} over the next generations, their mean converging to $\overline{F_{IS,\infty}} = -1.0$. For these latter loci starting out with a high genotypic diversity and a heterozygote genotype being the most frequent, a marked difference in the positive and negative ranges of the F_{IS}

distribution becomes apparent after some generations (t~50, figure 6E and D-B to a lesser extent): the positive values seem to "fade out". The reason for this effect, which also explains the convergence to negative $\overline{F_{IS,\infty}}$ in highly clonal populations, is firstly the trend towards randomly increasing the frequency of the heterozygous genotype(s) at HWE and approximately isoplethic allele frequencies (see previous section and additional file 1, part 1.4). Secondly, it is because the rate of fixation (of one allele) becomes strongly dependent on the initial $F_{IS,0}$ as sexual reproduction becomes rare (figure 7A-E; in plot E, the probability of fixation p_{fix} for $F_{IS,0} = 0$ starts to increase at t~50). Remarkably, the mean $\overline{F_{IS,t}}$ starting from $F_{IS,0} = 0$ (as expected for hitherto not or rarely clonally reproducing populations) has not reached the final $\overline{F_{IS,\infty}}$ even after 200 generations (while $t_N = 159$ – convergence to drift steady states is slowed down by "weak" mutation) in our example.

Highly clonal populations dominated by mutation rather than genetic drift present a very different picture (figure 6F,G, 7F,G: $c = 1.0,\ \mu = \{10^{-2}, 10^{-1}\},\ N = 100$, or simulation for the more realistic conditions $c = 1.0,\ \mu = 10^{-3},\ N = 10^3$ in additional file 2, part 2.5): As in predominantly sexually reproducing populations, F_{IS} values converge to only slightly negative final $\overline{F_{IS,\infty}}$ (figure 6G: $\overline{F_{IS,\infty}} \approx -0.02$) and the variation of F_{IS} values is limited, with the convergence speed depending on t_μ. Yet in contrast, the instantaneous $\overline{F_{IS,t}}$ distributions appear more symmetrical as the fixation of single alleles is very rare (figure 7F,G). If mutations between more than two alleles have to be taken into account, t_μ increases and mutation is accordingly "weakened" in comparison to the other processes (figure 3B; figure 7F, difference between our result and the predicted $\overline{F_{IS,\infty}}$ according to Balloux et al. 2003); however, as the generally close fit between the results of Balloux et al. (2003) between an infinite alleles model and our results suggests, increasing the number of alleles does not profoundly change the dynamics of F_{IS} we illustrated here.

Table 1. Convergence times under partial clonality for six representative example parameter sets. Population size $N = 100$ throughout. *Columns:* c – rate of clonality, μ – mutation rate, t_N – genetic drift maximal expected convergence time, t_c – reproduction maximal convergence time, t_μ – mutation maximal convergence time, t_I – convergence time to the mean $\overline{F_{IS,\infty}}$ based on the model in (Balloux et al. 2003), t_{II} – convergence time to the mean $\overline{F_{IS,\infty}}$ based on our model, t_{III} – convergence time to full final distribution of $\widetilde{F_{IS,\infty}}$. *Rows:* example parameter sets (compare figure 6). Bold: $\min(t_c, t_\mu)$.

$N = 100$	c	μ	t_N	t_c	t_μ	t_I	t_{II}	t_{III}
A	0.0	10^{-6}	159	**1**	2.6×10^6	1	1	2.6×10^6
B	0.8	10^{-6}	159	**25**	2.6×10^6	27	27	2.6×10^6
C	0.97	10^{-6}	159	**159**	2.6×10^6	174	177	2.6×10^6
D	0.99	10^{-6}	159	**529**	2.6×10^6	464	498	2.6×10^6
E	1.0	10^{-6}	159	∞	$\mathbf{2.6\times10^6}$	38 366	≫ 40 000	2.6×10^6
F	1.0	10^{-2}	159	∞	**263**	234	138	264
G	1.0	10^{-1}	159	∞	**25**	25	14	25

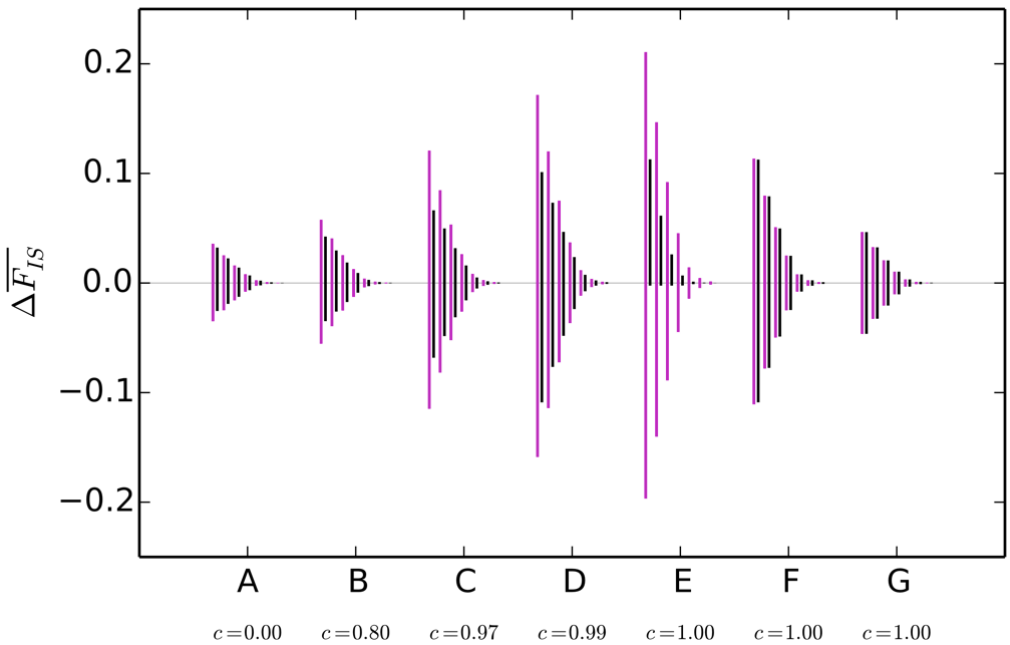

Figure 8. Sampling error of the mean $\overline{F_{IS,t,L}}$ according to number of loci. Mean signed deviation $\Delta\overline{F_{IS}}$ for each parameter set in figure 6, sampling from the $\widetilde{F_{IS,50}}$ distribution at 50 generations after all loci were at $F_{IS,0} = 0$ for $\nu_a = \nu_A$ (left/magenta), or the steady state distribution of F_{IS} ($\widetilde{F_{IS,\infty}}$, right/black). The bars for each parameter set are each based on 10^5 random samples of 5, 10, 25, 100, 1 000, 10 000 and 100 000 loci (left to right). Model parameters – A: $c = 0$, $\mu = 10^{-6}$, $N = 100$; B: $c = 0.8$, $\mu = 10^{-6}$, $N = 100$; C: $c \approx 0.97$ ($t_c = t_N$), $\mu = 10^{-6}$, $N = 100$; D: $c = 0.99$, $\mu = 10^{-6}$, $N = 100$; E: $c = 1.0$, $\mu = 10^{-6}$, $N = 100$; F: $c = 1.0$, $\mu = 10^{-2}$, $N = 100$; G: $c = 1.0$, $\mu = 10^{-1}$, $N = 100$.

The time until the exact final $\widetilde{F_{IS,\infty}}$ distribution is reached typically depends on t_μ (table 1, additional file 1 part 1.5). However, the mean $\overline{F_{IS,\infty}}$ may be reached much earlier (figure 6, table 1). This can be explained by the different patterns of genotype frequency change due to mutation and sexual reproduction: reaching the exact final $\widetilde{F_{IS,\infty}}$ distribution also implies having reached the final allele frequencies (only due to mutation), while reaching the final mean $\overline{F_{IS,\infty}}$ only requires having reached the final heterozygosity (due to mutation and sexual reproduction). When comparing different rates of clonality in populations with the same size and (natural) mutation rate, the increase of the conver-gence time to $\overline{F_{IS,\infty}}$ is therefore directly related to t_c until $t_c > t_\mu$. Compared to the model in (Balloux et al. 2003), convergence times to $\overline{F_{IS,\infty}}$ in our model are consistently higher or equal. In our finite-alleles model, t_μ and consequently the convergence times to $\overline{F_{IS,\infty}}$ would be lower, so that the increased times appear due to the discreteness of hetero-/ homozygosity only included in our model.

The increase in the variation of F_{IS} values accompanying a stronger influence of genetic drift also implies that estimating the exact mean $\overline{F_{IS,t}}$ derived from the full distribution of $\widetilde{F_{IS,t}}$ by the mean over a sample of several loci will become more inaccurate. Based on the same parameter sets as before, we show that this is indeed the case (figure 8). Moreover,

estimates taken from populations that have not yet reached their final distribution of $\widetilde{F_{IS,\infty}}$ may show greater deviations from their current exact mean $\overline{F_{IS,t}}$ value than expected for the final $\overline{F_{IS,\infty}}$ values (figure 8, left/magenta bars; note that the start state for all means taken at $t = 50$ was the same and does not contribute any additional variation). In some highly clonal populations (figure 8D,E at $t = 50$), the mean deviations from the current exact mean $\overline{F_{IS,50}}$ based on ten loci even exceeded ± 0.1.

Discussion

Our results on the dynamics of F_{IS} in partially clonal populations add a new dimension – time – to the description of the final distribution $\widetilde{F_{IS,\infty}}$ and its mean $\overline{F_{IS,\infty}}$ derived from previous models (Balloux et al. 2003, Stoeckel & Masson 2014), thereby providing a missing link with the seminal results of Marshall & Weir (1979). In connection genetic drift and mutation, partial clonality may deeply change the dynamics of F_{IS} and the underlying genotype frequencies over time. We provide a classification of the pattern of genotype frequency change in partially clonal populations (dominated by random mating, genetic drift, or mutation, figure 5), and a way to estimate the time needed for convergence to the final mean $\overline{F_{IS,\infty}}$ after a disturbance: Due to an increased "evolutionary memory" (*sensu* Bengtsson 2003, Desai 2009) for past genotypic diversity in partially clonal populations, population history (e.g. changes in demography, past selection, reproductive system) can produce a transient overrepresentation of F_{IS} values presumed to be very rare based on $\widetilde{F_{IS,\infty}}$. The variation of F_{IS} is increased in partially clonal compared to exclusively sexual populations, so that the mean $\overline{F_{IS,t,L}}$ obtained with the same (low) number of of loci has a greater error. Our findings suggest that the often reported negative mean $\overline{F_{IS}}$ are also compatible with intermediate rates of clonal reproduction and not necessarily a signature of (almost) exclusive clonality. We continue to discuss the mechanism behind the changed dynamics of F_{IS}, its implications and the impact of our results on the interpretation of $\overline{F_{IS}}$ in natural populations of partially clonal organisms.

Why negative F_{IS} in partially clonal populations?

Our results on the dynamics of genotype frequencies due to each evolutionary process (model parameter) separately formally demonstrate that:

- partially clonal reproduction, even with only rare sex ($c \in [0,1[$, figure 2A and additional file 2, part 2.2), leads towards $F_{IS,\infty} = 0$, as clonal reproduction by itself ($c = 1$; additional file 2, part 2.2) does not change genotype frequencies
- mutation, if acting independently at each allelic copy, leads towards $F_{IS,\infty} = 0$ (figure 3A and additional file 2, part 2.3), and
- genetic drift leads towards genotypic uniformity, i.e. either. $F_{IS,\infty} = -1$ (one heterozygous genotype remains) or fixation (one homozygous genotype remains) depending on the initial genotype frequencies (figure 4A).

The first point has important implications for extending our model to include other reproductive modes such as selfing or preferential inbreeding: as reproduction by (partial)

clonality does not result in directed departures from $F_{IS} = 0$, it does not counterbalance the homozygote excess expected for these cases (compare e.g. Wright 1921, Yaglom 1967, Marshall & Weir 1979). Compared to the standard expectation for exclusively sexual populations, clonality only slows down the approach to HWE (figure 2B), granting more influence to other processes such as mutation and genetic drift.

Among the evolutionary processes included in our model, genetic drift alone drives departures from $F_{IS} = 0$ (figure 4A). Therefore the negative $\overline{F_{IS,\infty}}$ suggested as a signature of highly or exclusively clonal populations (Balloux et al. 2003, Halkett et al. 2005, de Meeûs et al. 2006) strongly depend on the relationship between population size and mutation rate (figure 5). Genotype frequency dynamics due to mutation will dominate highly clonal populations if genetic drift is comparatively weak (figure 5, top right part of the diagram), i.e. under very high mutation rates (figure 6F,G) or more commonly for very big populations (additional file 2, part 2.5), and lead to $\overline{F_{IS,\infty}} = 0$ instead. Negative $\overline{F_{IS,\infty}}$ are only expected in smaller clonal populations, where the effects of genetic drift dominate genotype dynamics (figure 6D,E; figure 5, lower right part of the diagram). The "randomizing" effect of mutation on the association of alleles is weakened if the number of possible alleles increases (figure 3B), but exists regardless of this number or the exact mutation scheme (e.g. Estoup et al. 2002, Ellegren 2004 compare additional file 1, part 1.3), except if mutation depends on the alleles at other copies of the same locus within the same individual (e.g. as for gene conversion, McMahill et al. 2007, Flot et al. 2013, which specifically promotes homozygote excess and was not included in our model). According to our results, populations with mutation-dominated genotype dynamics (figure 5, upper part of the diagram) may distinguish themselves from their random mating-dominated counterparts (figure 5, left part of the diagram) by the rarity of loci that are fixed for one allele throughout the whole population rather than by a different mean $\overline{F_{IS}}$.

Our results provide a complementary perspective to the previously proposed explanations of heterozygote excess in highly clonal populations: negative F_{IS} values for clonal lineages have sometimes been associated with the independent accumulation of mutations on homologous chromosomes over long periods of time, commonly known as the "Meselson effect" (Welch & Meselson 2000, mathematical models in Bengtsson 2003, Ceplitis 2003). Here we show that these negative F_{IS} values do not appear at loci with a high rate of mutation and/or in very large ancient populations (figure 6; additional file 2, part 2.5).

F_{IS} dynamics in partially clonal populations: changes to time scale and variation between loci

We demonstrated that the dynamics of genotype frequencies and F_{IS} are slowed down in partially clonal populations, which therefore retain traces of their past for much longer than their exclusively sexual counterparts. This puts the appropriateness of the final $\widetilde{F_{IS,\infty}}$ (Balloux et al. 2003, Stoeckel & Masson 2014) as a reference into question: already for intermediate rates of clonality, the observed genotype frequencies may reflect population history rather than the present reproductive system and dynamics, depending on the time since the last disturbance (figure 6; additional file 2, part 2.5). For example, an almost exclusively clonal population of hybrid origin will maintain excess heterozygosity/lower F_{IS}

for many generations; the re-establishment of expected heterozygosity after secondary contact will take much longer in a highly clonal population, even if the rare sexual reproduction is panmictic and pangamic; exclusively clonal populations that "lost sex" only recently may give the impression that they sometimes reproduce sexually just by inheriting the genotypic diversity of their ancestors.

The deceleration of the dynamics of F_{IS} is connected to the hyperbolical increase of t_c (figure 2B) and thus much stronger under high rates of clonality. We demonstrated that a comparison of the maximal expected convergence times t_c, t_μ, t_N can be an efficient means to predict the overall pattern of F_{IS} dynamics (figure 5-7, additional file 2, part 2.5).The times t_c and t_μ can even be used to estimate convergence times of the complete model (table 1): While the time until the steady state distribution of $\widetilde{F_{IS,\infty}}$ is reached depends on t_μ, the convergence time to the final mean $\overline{F_{IS,\infty}}$ can be estimated by the minimum of t_c and t_μ (i.e. usually t_c in small populations). If $t_N \ll \min(t_c,\ t_\mu)$, loci with different initial genotype frequencies may not converge to the same final $F_{IS,\infty}$ value (convergence to genotypic uniformity), so that the expected final $\overline{F_{IS,\infty}} = -1$ is not reached within biologically realistic time spans (figure 6E,7E, additional file 2). Though not yet included in our model, perenniality (partial survival of the population across generations, as in Orive 1993) is expected to slow down F_{IS} dynamics even further. If disturbances are sufficiently frequent, e.g. in very instable environments or in populations cyclically changing between exclusive sexual and clonal reproduction (Berg & Lascoux 2000, Allen & Lynch 2012), the final $\overline{F_{IS,\infty}}$ and $\widetilde{F_{IS,\infty}}$ based on the currently observed rate of clonality may even never be reached.

Finally, during the approach to the final $\widetilde{F_{IS,\infty}}$, loci may pass through intermediate states (genotype frequencies, F_{IS}; figure 6, additional file 2, part 2.5) that would be unusual according to $\widetilde{F_{IS,\infty}}$. Even at rates of clonality which do not yet affect the expectations for the final $\overline{F_{IS,\infty}}$, the variation of F_{IS} is increased compared to exclusively sexual populations (Balloux et al. 2003, Stoeckel & Masson 2014). Consequently, information about more loci is required to accurately estimate the mean $\overline{F_{IS,t}}$ in partially clonal compared to strictly sexual populations. We found that the variation of F_{IS} observed during the approach to the steady state distribution may be even greater than predicted based on the final $\widetilde{F_{IS,\infty}}$ (figure 8).

Application to field data

We performed a literature analysis, which shows that a very wide variety of F_{IS} values, positive as well as negative, were found in partially clonal populations (figure 9; details in additional file 3). Field data may be influenced by technical biases, including sampling bias due to an unknown spatial structure of clones, missing rare genotypes due to non-exhaustive sampling, genotyping errors (e.g. undetected null alleles for SSRs) or preferential sampling of loci with near-isoplethy (thus increasing the probability to find negative F_{IS}). Moreover, biological processes that are not included in our model may have acted on the data we collected. We applied strict criteria to standardize the dataset we used and retain those that fit best our model, including only studies that did include repeated multilocus genotypes in their calculations, published F_{IS} values (or H_o and H_e) per locus and (clearly isolated) population, and reported on organisms which life cycle fits with our model (i.e.

dominantly diploid, no cyclic clonality as e.g. in aphids). As only few studies matched these strict criteria (Duran et al. 2004, Nagamitsu et al. 2004, Stoeckel et al. 2006, Rougeron et al. 2008, Villate et al. 2010, Corral et al. 2011, Jiang et al. 2011, Liu et al. 2011, Tesson et al. 2011, Vilas et al. 2011, Gao et al. 2012, McInnes et al. 2012, Tew et al. 2012, Barnabe et al. 2013, Motoie et al. 2013), we also kept some for which the mating system departed from random mating, considering that preferential inbreeding may increase the frequency of positive F_{IS} values (compare Yaglom 1967, Marshall & Weir 1979) and preferential outbreeding has been shown to have very little effect on F_{IS} (Navascués et al. 2009).

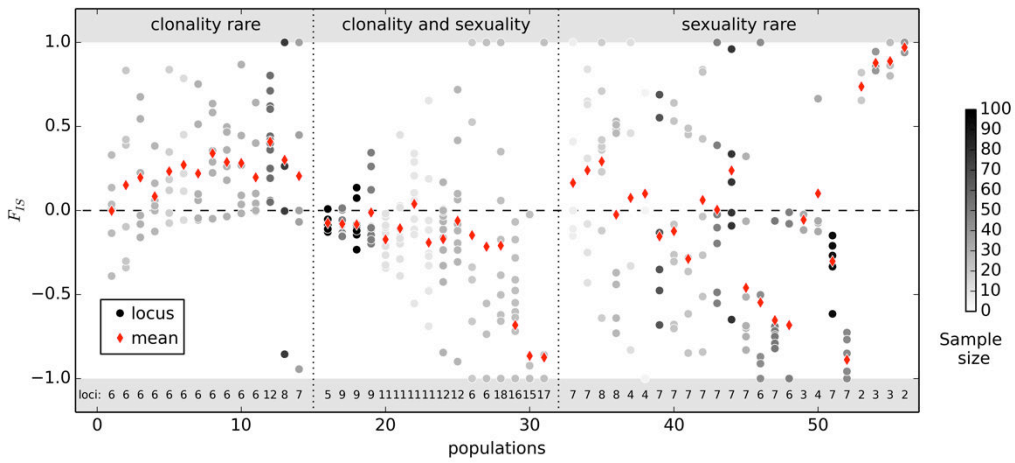

Figure 9: Examples for empirical F_{IS} values of partially clonal populations compiled from field studies. Data for 54 populations (one per column) belonging to 15 species (seven angiosperms, six protists, a sponge and a nematode), based on 15 previous studies (see additional file 3) selected for their near fit with the assumptions of our model from a Web of Science search for [(microsatellite *OR* "SSR" *OR* "simple sequence repeat" *OR* "SNP" *OR* "single nucleotide polymorphism") *AND* (clonal *OR* asexual *OR* vegetative *OR* apomictic *OR* apomixis *OR* agamospermy *OR* parthenogenesis)]. All studies are based on SSR data; for populations 20-31, 39-42 and 49-50, F_{IS} values per locus were calculated from the reported H_o and H_e. Includes populations for which preferential inbreeding (populations 12-14) or outbreeding (self-incompatibility system, populations 16-19) is expected. Dotted lines separate three groups of populations according to the information given by the authors about their putative rate of clonality, i.e. rarely clonal, frequent clonality and sexuality (including unknown), or rarely sexual. Number of sampled loci indicated by numbers at the bottom of the plot, number of samples (individuals/ramets) indicated by the hue of each round dot (light grey: 10, black: > 100). Red lozenges indicate the mean $\overline{F_{IS,t,L}}$ over all sampled loci per population.

The results of our study open up new explanations for the presence, but also the absence, of positive and negative F_{IS} values (both at individual loci or the mean) in partially clonal populations. Values that should be rare based on the steady state distribution $\widetilde{F_{IS,\infty}}$ or its mean may transiently become highly probable already under intermediate rates of clonality: they can be due to the increased variation of F_{IS} or echo a departure from equilibrium due to population history (e.g. demographic bottleneck, change in the rate of clonal reproduction). As an example of application of our results, in some wild cherry populations

(populations 17-19 in figure 9, Stoeckel et al. 2006), slightly negative mean $\overline{F_{IS,t,L}}$ over loci would have suggested almost exclusive clonality when taking the final mean $\overline{F_{IS,\infty}}$ as the reference (e.g. $\overline{F_{IS,t,L}} = -0.083$ for the exhaustively sampled population 18 with $N = 247, \mu \in [10^{-3}, 10^{-12}]$ suggests $c \approx 0.98$). However, the proportion of repeated multilocus genotypes and inferences from parentage analysis suggested only an intermediate rate of clonality ($c \sim 0.5$); based on this value, genotype dynamics would be dominated by random mating ($t_c \approx 10$). Even though *Prunus avium* is a very long-lived species with generation times of ~ 50 years, historical records of the studied populations indicate no major disturbances such as demographic bottlenecks (e.g. due to fires) during the past > 500 years. Thus population history may not be the most likely explanation for deviations from the expected mean $\overline{F_{IS,\infty}} \approx -0.002$. Based on our results and taking into account that only nine loci were analyzed, observed F_{IS} values and their negative mean would be thus best explained by the increased variation of F_{IS} expected already for intermediate rates of clonality.

Our results also suggest ways to further improve the population genetic inferences in natural populations of partial clonal organisms, as proposed in (Halkett et al. 2005). Maximizing the number of loci studied by moving from population genetics to population genomics may help to improve the statistical basis of inferences of population parameters. Rather than focusing exclusively on the mean $\overline{F_{IS,t,L}}$ over loci, the full distribution of $\widetilde{F_{IS,t,L}}$ values per population should be reported and interpreted. Collecting time series of samples may also provide valuable information, as field data normally represent only a "snapshot" of genotype frequencies at a particular point in time, that may or may not be representative of the final distribution of $\widetilde{F_{IS,\infty}}$. Using the Markov chain model implemented here, it is not only possible to statistically analyze example trajectories (e.g. as in figure 2-5 in additional file 2), but also to analytically derive the transition probabilities between two consecutive sets of genotype frequencies for a range of rates of clonality, based on population size, mutation rate and number of generations between the samples. Taking the temporal dynamics of F_{IS} in partially clonal populations into account will help to improve the biological interpretation of F_{IS} values from field data, and could contribute to a refined, unified method for estimating the rate of clonality based on a collection of population genetic indices.

Conclusions

Our results allow reconciling predictions for F_{IS} under partial clonality from theoretical models, which suggest departures from $F_{IS} = 0$ only at nearly pure clonality, with empirical data, also showing such departures under different conditions where sex is known or suspected to be more frequent. These results have three main implications for interpreting F_{IS} in partially clonal populations:

- non-negative F_{IS}, including null values, are not a reliable indicator of the absence of clonal reproduction,
- significant deviations from $F_{IS} = 0$ for multiple loci may either indicate a considerable rate of clonality or a transient departure from the expected equilibrium in populations even when subject to a modest rate of clonality, or result from the

interaction of partial clonality with another evolutionary process that increases/decreases heterozygosity within the population (e.g. non-random mating, Marshall & Weir 1979; gene conversion, McMahill et al. 2007, Flot et al. 2013)

- Increasing the number of loci and studying time series rather than single snapshots of genetic data may improve the accuracy of DNA-based estimates of the rate of clonal reproduction in populations and species.

Availability of supporting data

The data supporting the results of this article is included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JPM and SS laid the foundation for this work. JPM, FM, KR and SS conceived the mathematical model. KR preformed the scientific programming, based on an initial script by SS and did the literature review, the data analyses and illustrations. KR, SAH and SS wrote the manuscript. All authors contributed to editing. SAH and SS were responsible for funding and grant applications. All authors have read and approved the final manuscript.

Acknowledgements

We thank Jurgen Angst, Valentin Bahier, Fabien Halkett, Cédric Midoux and Stéphane De Mita for helpful discussions and comments on the manuscript. We greatly profited from the exchange of ideas within the CLONIX group. We thank the *French National Research Agency* for funding this study as part of the CLONIX project (ANR-11-BSV7-0007), and the *French National Institute for Agricultural Research*, *Plant Health and the Environment Department* (INRA-SPE) as well as the *Région Bretagne* for supporting Katja Reichel during her PhD.

References

Allen DE, Lynch M. 2012. The effect of variable frequency of sexual reproduction on the genetic structure of natural populations of a cyclical parthenogen. *Evolution*. 66(3):919–926

Arnaud-Haond S, Duarte CM, Alberto F, Serrão EA. 2007. Standardizing methods to address clonality in population studies. *Molecular Ecology*. 16(24):5115–5139

Balloux F, Lehmann L, de Meeûs T. 2003. The population genetics of clonal and partially clonal diploids. *Genetics*. 164(4):1635–1644

Barnabe C, Buitrago R, Bremond P, Aliaga C, Salas R, et al. 2013. Putative panmixia in restricted populations of *Trypanosoma cruzi* isolated from wild *Triatoma infestans* in Bolivia. *PLoS ONE*. 8(11):e82269

Bengtsson BO. 2003. Genetic variation in organisms with sexual and asexual reproduction. *Journal of Evolutionary Biology*. 16(2):189–199

Berg LM, Lascoux M. 2000. Neutral genetic differentiation in an island model with cyclical parthenogenesis. *Journal of Evolutionary Biology*. 13(3):488–494

Brzosko E, Wróblewska A, Ratkiewicz M. 2002. Spatial genetic structure and clonal diversity of island populations of lady's slipper (*Cypripedium calceolus*) from the Biebrza National Park (northeast Poland). *Molecular Ecology*. 11(12):2499–2509

Brzyski JR, Culley TM. 2011. Genetic variation and clonal structure of the rare, riparian shrub *Spiraea virginiana* (Rosaceae). *Conservation Genetics*. 12(5):1323–1332

Ceplitis A. 2003. Coalescence times and the Meselson effect in asexual eukaryotes. *Genetical Research*. 82(3):183–190

Corral JM, Molins MP, Aliyu OM, Sharbel TF. 2011. Isolation and characterization of microsatellite loci from apomictic *Hypericum perforatum* (Hypericaceae). *American Journal of Botany*. 98(7):e167–169

De Finetti B. 1926. Considerazioni matematiche sull'ereditarietà Mendeliana. *Metron*. 6(1):3–41

De Meeûs T, Lehmann L, Balloux F. 2006. Molecular epidemiology of clonal diploids: A quick overview and a short DIY (do it yourself) notice. *Infection, Genetics and Evolution*. 6(2):163–170

De Meeûs T, Prugnolle F, Agnew P. 2007. Asexual reproduction: Genetics and evolutionary aspects. *Cellular and Molecular Life Sciences*. 64(11):1355–1372

Desai MM. 2009. Reverse evolution and evolutionary memory. *Nature Genetics*. 41(2):142–144

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics*. 148(4):1667–1686

Duminil J, Fineschi S, Hampe A, Jordano P, Salvini D, et al. 2007. Can population genetic structure be predicted from life-history traits? *The American Naturalist*. 169(5):662–672

Duminil J, Hardy OJ, Petit RJ. 2009. Plant traits correlated with generation time directly affect inbreeding depression and mating system and indirectly genetic structure. *BMC Evolutionary Biology*. 9(1):177

Duran S, Pascual M, Estoup A, Turon X. 2004. Strong population structure in the marine sponge *Crambe crambe* (Poecilosclerida) as revealed by microsatellite markers. *Molecular Ecology*. 13(3):511–522

Ellegren H. 2004. Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*. 5(6):435–445

Estoup A, Jarne P, Cornuet J-M. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*. 11(9):1591–1604

Ewens WJ. 2004. *Mathematical Population Genetics: I. Theoretical Introduction*. New York: Springer. 2nd ed.

Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, et al. 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*. 500(7463):453–457

Gao H, Jiang K, Geng Y, Chen X-Y. 2012. Development of microsatellite primers of the largest seagrass, *Enhalus acoroides* (Hydrocharitaceae). *American Journal of Botany*. 99(3):e99–101

Graffelman J, Camarena JM. 2008. Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity*. 65(2):77–84

Hagberg AA, Schult DA, Swart PJ. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Proceedings of the 7th Python in Science Conference (SciPy2008), pp. 11–15

Halkett F, Simon J, Balloux F. 2005. Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology & Evolution*. 20(4):194–201

Hile SE, Yan G, Eckert KA. 2000. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. *Cancer Research*. 60(6):1698–1703

Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*. 9(3):90–95

Jiang K, Gao H, Xu N-N, Tsang EPK, Chen X-Y. 2011. A set of microsatellite primers for *Zostera japonica* (Zosteraceae). *American Journal of Botany*. 98(9):e236–238

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, Vol. III, ed. HN Munro, pp. 21–132. New York: Academic Press

Liu J, Dong M, Miao SL, Li ZY, Song MH, Wang RQ. 2006. Invasive alien plants in China: role of clonality and geographical origin. *Biological Invasions*. 8(7):1461–1470

Liu W, Zhou Y, Liao H, Zhao Y, Song Z. 2011. Microsatellite primers in *Carex moorcroftii* (Cyperaceae), a dominant species of the steppe on the Qinghai-Tibetan Plateau. *American Journal of Botany*. 98(12):e382–384

Luijten SH, Oostermeijer JGB, van Leeuwen NC, den Nijs HC. 1996. Reproductive success and clonal genetic structure of the rare *Arnica montana* (Compositae) in the Netherlands. *Plant Systematics and Evolution*. 201(1-4):15–30

Marshall DR, Weir BS. 1979. Maintenance of genetic variation in apomictic plant populations. *Heredity*. 42(2):159–172

McInnes LM, Dargantes AP, Ryan UM, Reid SA. 2012. Microsatellite typing and population structuring of *Trypanosoma evansi* in Mindanao, Philippines. *Veterinary Parasitology*. 187(1-2):129–39

McKey D, Elias M, Pujol B, Duputié A. 2010. The evolutionary ecology of clonally propagated domesticated plants: Tansley review. *New Phytologist*. 186(2):318–332

McMahill MS, Sham CW, Bishop DK. 2007. Synthesis-dependent strand annealing in meiosis. *PLoS Biology*. 5(11):e299

Motoie G, Ferreira GEM, Cupolillo E, Canavez F, Pereira-Chioccola VL. 2013. Spatial distribution and population genetics of *Leishmania infantum* genotypes in São Paulo State, Brazil, employing multilocus microsatellite typing directly in dog infected tissues. *Infection, Genetics and Evolution*. 18:48–59

Nagamitsu T, Ogawa M, Ishida K, Tanouchi H. 2004. Clonal diversity, genetic structure, and mode of recruitment in a *Prunus ssiori* population established after volcanic eruptions. *Plant Ecology*. 174(1):1–10

Navascués M, Stoeckel S, Mariette S. 2009. Genetic diversity and fitness in small populations of partially asexual, self-incompatible plants. *Heredity*. 104(5):482–492

Oliphant TE. 2007. Python for scientific computing. *Computing in Science & Engineering*. 9(3):10–20

Orive ME. 1993. Effective population size in organisms with complex life-histories. *Theoretical Population Biology*. 44(3):316–340

Reichel K, Bahier V, Midoux C, Masson J-P, Stoeckel S. 2014. Interpretation and approximation tools for big, dense Markov chain transition matrices in ecology and evolution. *arXiv preprint arXiv:1407.2548*

Rougeron V, Waleckx E, Hide M, de Meeûs T, Arevalo J, et al. 2008. A set of 12 microsatellite loci for genetic studies of *Leishmania braziliensis*. *Molecular Ecology Resources*. 8(2):351–353

Schön I, Martens K, Dijk P, eds. 2009. *Lost sex*. Dordrecht: Springer Netherlands

Setsuko S, Ishida K, Tomaru N. 2004. Size distribution and genetic structure in relation to clonal growth within a population of *Magnolia tomentosa* THUNB. (Magnoliaceae). *Molecular Ecology*. 13(9):2645–2653

Stoeckel S, Grange J, Fernández-Manjarres JF, Bilger I, Frascaria-Lacoste N, Mariette S. 2006. Heterozygote excess in a self-incompatible and partially clonal forest tree species – *Prunus avium* L. *Molecular Ecology*. 15(8):2109–2118

Stoeckel S, Masson J-P. 2014. The exact distributions of F_{IS} under partial asexuality in small finite populations with mutation. *PLoS ONE*. 9(1):e85228

Sydes MA, Peakall R. 1998. Extensive clonality in the endangered shrub *Haloragodendron lucasii* (Haloragaceae) revealed by allozymes and RAPDs. *Molecular Ecology*. 7:87–93

Tesson SVM, Borra M, Kooistra WHCF, Procaccini G. 2011. Microsatellite primers in the planktonic diatom *Pseudo-nitzschia multistriata* (Bacillariophyceae). *American Journal of Botany*. 98(2):e33–35

Tew JM, Lance SL, Jones KL, Fehlberg SD. 2012. Microsatellite development for an endangered riparian inhabitant, *Lilaeopsis schaffneriana* subsp. *recurva* (Apiaceae). *American Journal of Botany*. 99(4):e164–166

Tibayrenc M, Ayala FJ. 2012. Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proceedings of the National Academy of Sciences*. 109(48):E3305–3313

Vilas R, Cao A, Pardo BG, Fernández S, Villalba A, Martínez P. 2011. Very low microsatellite polymorphism and large heterozygote deficits suggest founder effects and cryptic structure in the parasite *Perkinsus olseni*. *Infection, Genetics and Evolution*. 11(5):904–911

Villate L, Esmenjaud D, Van Helden M, Stoeckel S, Plantard O. 2010. Genetic signature of amphimixis allows for the detection and fine scale localization of sexual reproduction events in a mainly parthenogenetic nematode. *Molecular Ecology*. 19(5):856–873

Welch DM, Meselson M. 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science*. 288(5469):1211–1215

Wright S. 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics*. 6(2):124–143

Yaglom IM. 1967. Geometric models of certain genetic processes. *Canadian Journal of Mathematics*. 19(0):1233–1242

Appendix

A1 Mathematical background

A1.1 Model equations

For compound equations describing the concatenation of at least two processes, substitute the variables for time t in the second equation by the result for time $t+1$ from the first equation.

Symbols and abbreviations

N … population size	c … rate of asexual reproduction	μ … mutation rate
n … number of alleles	g … number of genotypes	t…current generation
ν_i … allele frequency	ν_{ij} … genotype frequency	$q_{ii} = N\nu_{ii},\ q_{ij} = N\nu_{ij}$
$i \neq j \neq k \neq l$ … indices referring to alleles	$\alpha = 1-\mu$	$\beta = \mu/(n-1)$
$\mathcal{M}$ … multinomial distribution	X … random variable	P … probability
t_c, t_μ, t_N … max. expected number of generations to convergence		λ … eigenvalue
$\mathcal{S}$ … allele substitution matrix	J … matrix of ones	I … identity matrix
$H = \sum_{i,j} \nu_{ij} = 1 - \sum_i \nu_{ii}$	$H_e = 2\sum_{i,j} \nu_i \nu_j = 1 - \sum_i {\nu_i}^2$	$\forall$ … "for all" sign

ε … approximation bias, set to $\varepsilon = 1/(2N)$ unless specified otherwise

Mutation

- $n = 2$:

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t+1} = \begin{bmatrix}(1-\mu)^2 & (1-\mu)\mu & \mu^2\\ 2(1-\mu)\mu & (1-\mu)^2+\mu^2 & 2(1-\mu)\mu\\ \mu^2 & (1-\mu)\mu & (1-\mu)^2\end{bmatrix}\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_t$$

- $n > 2$:

$$\nu_{ii,t+1} = \nu_{ii,t}\cdot\alpha^2 + \sum_j \nu_{jj,t}\cdot\beta^2 + \sum_j \nu_{ij,t}\cdot\alpha\beta + \sum_{j,k}\nu_{jk,t}\cdot\beta^2$$

$$\begin{aligned}\nu_{ij,t+1} = {} & \left(\nu_{ii,t}+\nu_{jj,t}\right)\cdot 2\alpha\beta + \sum_k \nu_{kk,t}\cdot 2\beta^2 + \nu_{ij,t}\cdot(\alpha^2+\beta^2)\\ & + \sum_{k,l}\left(\nu_{ik,t}+\nu_{jl,t}\right)\cdot(\alpha\beta+\beta^2) + \sum_{k,l}\nu_{kl,t}\cdot 2\beta^2 \quad =\end{aligned}$$

Reproduction

- $n = 2$:

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t+1} = c\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_t + (1-c)\begin{bmatrix}\nu_a^2\\ 2\nu_a\nu_A\\ \nu_A^2\end{bmatrix}_t$$, allele frequencies $\begin{bmatrix}\nu_a\\ \nu_A\end{bmatrix}_t = \begin{bmatrix}1 & 0.5 & 0\\ 0 & 0.5 & 1\end{bmatrix}\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_t$

- $n > 2$:

$$\nu_{ii,t+1} = c\nu_{ii,t} + (1-c)\nu_{i,t}^2 = c\nu_{ii,t} + (1-c)\cdot\left(\nu_{ii,t} + 0.5\sum_j \nu_{ij,t}\right)^2$$

$$\nu_{ij,t+1} = c\nu_{ij,t} + 2(1-c)\nu_{i,t}\nu_{j,t} = c\nu_{ij,t} + 2(1-c)\cdot\left(\nu_{ii,t} + 0.5\sum_k \nu_{ik,t}\right)\left(\nu_{jj,t} + 0.5\sum_l \nu_{jl,t}\right)$$

Genetic drift

Note that all $\nu_{ii,t+1}, \nu_{ij,t+1}$ have to fulfill $N\nu_{ii,t+1}\ (= q_{ii,t+1}), N\nu_{ij,t+1} (= q_{ij,t+1}) \in \mathbb{N}_0$.

- $n = 2$:

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t+1} = X/N \text{ where } X \sim \mathcal{M}\left(N, \left[\nu_{aa,t}, \nu_{aA,t}, \nu_{AA,t}\right]\right),$$

i.e. for $q_{aa,t+1},\ q_{aA,t+1},\ q_{AA,t+1} \in \mathbb{N}_0$ such that $q_{aa,t+1} + q_{aA,t+1} + q_{AA,t+1} = N$:

$$P\left(\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t+1} \middle| \begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t}\right) = \frac{N!}{(q_{aa,t+1})!\cdot(q_{aA,t+1})!\cdot(q_{AA,t+1})!}\cdot \nu_{aa,t}^{q_{aa,t+1}}\cdot \nu_{aA,t}^{q_{aA,t+1}}\cdot \nu_{AA,t}^{q_{AA,t+1}}$$

- $n > 2$:

$$\begin{bmatrix}\nu_{ii}\\ \nu_{ij}\\ \vdots\end{bmatrix}_{t+1} = X/N \text{ where } X \sim \mathcal{M}\left(N, \left[\nu_{ii,t}, \nu_{ij,t}, \dots\right]\right),$$

i.e. for $q_{ii,t+1}, q_{ij,t+1}, \dots \in \mathbb{N}_0$ such that $\sum_i q_{ii,t+1} + \sum_{ij} q_{ij,t+1} = N$:

$$P\left(\begin{bmatrix}\nu_{ii}\\ \nu_{ij}\\ \vdots\end{bmatrix}_{t+1} \middle| \begin{bmatrix}\nu_{ii}\\ \nu_{ij}\\ \vdots\end{bmatrix}_{t}\right) = \frac{N!}{\prod_i (q_{ii,t+1})!\cdot\prod_{i,j}(q_{ij,t+1})!}\cdot\prod_i \nu_{ii,t}^{q_{ii,t+1}}\cdot\prod_{i,j}\nu_{ij,t}^{q_{ij,t+1}}$$

A1.2 Convergence times – individual parameters

Reproduction

As can be easily demonstrated from the reproduction equations in part *A1.1*, neither random mating nor asexual reproduction *per se* change allele frequencies, they only affect the proportion of heterozygous and homozygous genotypes. Let H_t denote the observed heterozygosity at time t, and H_e the expected heterozygosity at $F_{IS} = 0$ (convergence domain) for a given set of allele frequencies.

The sum over all equations $\nu_{ij,t+1} = c\nu_{ij,t} + 2(1-c)\nu_{i,t}\nu_{j,t}$ can be rewritten as $H_{t+1} = cH_t + (1-c)H_e$. Inserting this result into the equation for F_{IS} gives:

$$F_{IS,t+1} = \frac{H_e - H_{t+1}}{H_e} = \frac{H_e - (cH_t + (1-c)H_e)}{H_e} = c\cdot\frac{H_e - H_t}{H_e} = c\cdot F_{IS,t}$$

This recursive relation can be rewritten as a geometric sequence, $F_{IS,t=x} = c^{x-1} \cdot F_{IS,t=0}$. By defining an "acceptable error" ε, we can also calculate the time t_c until it has converged arbitrarily close to $F_{IS} = 0$, starting from $|F_{IS}| = 1$, depending on c: $\varepsilon = c^{t_c-1}$, which transforms to $t_c = 1 + \log_c \varepsilon = 1 + \log \varepsilon / \log c$. For $c = 0$, $t_c = 1$; in contrast, t_c is infinite (no convergence) if $c = 1$. A good value for ε may be $1/2N$, half the frequency corresponding to one individual of the population (detection threshold for deviations from $F_{IS} = 0$).

Mutation

For any n, mutation between genotypes can be described by a matrix similar to the one given in part *A1.1* or table *A1.3-1*. Thus, the genotype frequency vectors issued from the mutation process converge to the matrix' dominant eigenvector (eigenvalue 1), and the convergence can be approximated by a geometric sequence whose common ratio corresponds to the matrix' second largest eigenvalue.

For $n = 2$, the eigenvalues of the mutation matrix can be calculated "by hand": $\lambda = \{1, 1 - 2\mu, (1 - 2\mu)^2\}$, each with a multiplicity of one. The time to convergence, depending on μ, therefore approximately corresponds to $t_\mu = 1 + \log_{(1-2\mu)} \varepsilon = 1 + \log \varepsilon / \log(1 - 2\mu)$. Looking at the extreme values of μ, $t_\mu = 1$ for $\mu = 0.5$, and there is no convergence if $\mu = 0$.

For $n > 2$, the eigenvalues of the mutation matrix can be derived from the eigenvalues of the allele substitution matrix $\mathcal{S}$. If we first consider genotypes as ordered, rather than unordered, pairs of alleles, mutation between them is described by the Kronecker product of $\mathcal{S}$ with itself. The eigenvalues of $\mathcal{S} \otimes \mathcal{S}$ are given by the pairwise products $\lambda_{\mathcal{S},i}\lambda_{\mathcal{S},j}$ (including $i = j$) of the eigenvalues of $\mathcal{S}$. For the *Jukes-Cantor* substitution model, as $\mathcal{S} = \beta J + (\alpha - \beta)I$, the eigenvalues of the allele substitution matrix are $\lambda_{\mathcal{S}} = \{\alpha + (n-1)\beta,\ \alpha - \beta\} = \{1, 1 - \mu\frac{n}{n-1}\}$ with multiplicities $\{1, n-1\}$. The eigenvalues of the mutation matrix are therefore $\lambda = \{1, 1 - \mu\frac{n}{n-1}, (1 - \mu\frac{n}{n-1})^2\}$, and their multiplicities $\{1, n-1,\ n(n-1)/2\}$ after adjusting for the "unorderedness" of the alleles within genotypes.

Thus, for $n > 2$, the second largest eigenvalue reduces to $(1 - n\beta) = (1 - \mu \cdot n/(n-1))$, and its multiplicity increases to $n - 1$. Consequently, the more alleles there are, the longer it takes until mutation has reached its equilibrium.

Genetic drift

If we consider a Markov chain for the states of our model based only on multinomial drawing, all states where only one genotype exists are absorptive. The convergence time t_N for genetic drift thus corresponds to the maximum of the expected time until absorption / the expected time until "fixation" of one genotype starting from any non-absorptive (transient) state.

The vector of expected absorption times for a population can be calculated from its fundamental matrix, i.e. an identity matrix of corresponding size minus the part of the transition matrix describing only transitions between transient states. For the simplest case of $N = 2$ individuals, the fundamental matrix is

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

The expected absorption times of the three transient states $[0.5, 0.5, 0]$, $[0.5, 0, 0.5]$ and $[0, 0.5, 0.5]$ are given by the column sums of the inverse of this matrix, i.e. $(2, 2, 2)$. Consequently, $t_N(N = 2) = 2$, which means that any initially genotypically diverse population would be expected to become uniform after (a maximum of) just two generations.

As N increases, the fundamental matrix is no longer diagonal and the calculation of t_N becomes more complex. Also, the vector of expected absorption times is no longer uniform, but depends on each state's distance to the nearest absorptive state, e.g. $[0.9, 0.1, 0.0]$ being closer to absorption than $[0.4, 0.3, 0.3]$. For a finite N, these distances are divided into quanta of $1/N$ since there can be no fractional individuals; hence the dependence of t_N on N. Based on the numerical results displayed in table *A1.2-1*, the following linear approximations of this dependence can be made (note that t_N not just increases with N, but to a lesser extent also with g or n):

$$t_N(n = 2;\ 2 \leq N \leq 120) = 1.6N - 1$$

$$t_N(n = 3;\ 2 \leq N \leq 15) = 1.8N - 1.5$$

$$t_N(n = 4;\ 2 \leq N \leq 8) = 1.95N - 2.$$

For infinite N, the underlying multinomial distribution turns into a multinormal, and genetic drift becomes a diffusion process.

Table A1.2-1: Numerical results for the drift convergence time t_N, i.e. the maximum of the expected times to absorption / "fixation" of a single genotype through genetic drift, for different population sizes N, numbers of alleles n and resulting numbers of genotypes g.

$n = 2, g = 3$				$n = 3, g = 6$		$n = 4, g = 10$	
N	t_N	N	t_N	N	t_N	N	t_N
2	2.0	20	30.8	2	2.0	2	2.0
3	3.9	30	46.9	3	3.9	3	3.9
4	5.2	40	63.0	4	5.8	4	5.8
5	6.8	50	79.2	5	7.7	5	7.7
6	8.5	60	95.4	6	9.7	6	9.7
7	10.0	70	111.5	7	11.3	7	11.6
8	11.6	80	127.7	8	13.0	8	13.6
9	13.3	90	143.9	9	14.8	–	–
10	14.8	100	160.1	10	16.6	–	–
15	22.8	120	192.4	15	25.7	–	–

A1.3 Mutation and heterozygosity – multiple alleles, asymmetric mutation rate

Multiple alleles

Table A1.3-1: Mutation rates between diploid genotypes at one locus with four possible alleles A, G, C and T (SNP), based on the Jukes-Cantor substitution model. Mutations from "column" to "row" genotype with $\alpha = (1 - \mu)$, $\beta = \mu/3$. Note that all columns sum to one.

	Homozygote genotypes				Heterozygote genotypes					
↲	AA	GG	CC	TT	AG	AC	AT	GC	GT	CT
AA	α^2	β^2	β^2	β^2	$\alpha\beta$	$\alpha\beta$	$\alpha\beta$	β^2	β^2	β^2
GG	β^2	α^2	β^2	β^2	$\alpha\beta$	β^2	β^2	$\alpha\beta$	$\alpha\beta$	β^2
CC	β^2	β^2	α^2	β^2	β^2	$\alpha\beta$	β^2	$\alpha\beta$	β^2	$\alpha\beta$
TT	β^2	β^2	β^2	α^2	β^2	β^2	$\alpha\beta$	β^2	$\alpha\beta$	$\alpha\beta$
AG	$2\alpha\beta$	$2\alpha\beta$	$2\beta^2$	$2\beta^2$	$\alpha^2+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$2\beta^2$
AC	$2\alpha\beta$	$2\beta^2$	$2\alpha\beta$	$2\beta^2$	$\alpha\beta+\beta^2$	$\alpha^2+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$2\beta^2$	$\alpha\beta+\beta^2$
AT	$2\alpha\beta$	$2\beta^2$	$2\beta^2$	$2\alpha\beta$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha^2+\beta^2$	$2\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$
GC	$2\beta^2$	$2\alpha\beta$	$2\alpha\beta$	$2\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$2\beta^2$	$\alpha^2+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$
GT	$2\beta^2$	$2\alpha\beta$	$2\beta^2$	$2\alpha\beta$	$\alpha\beta+\beta^2$	$2\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha^2+\beta^2$	$\alpha\beta+\beta^2$
CT	$2\beta^2$	$2\beta^2$	$2\alpha\beta$	$2\alpha\beta$	$2\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha\beta+\beta^2$	$\alpha^2+\beta^2$

The general equations for mutation (part *A1.1*, compare table *A1.3-1* for $n = 4$) are:

$$\nu_{ii,t+1} = \nu_{ii,t} \cdot \alpha^2 + \sum_j \nu_{jj,t} \cdot \beta^2 + \sum_j \nu_{ij,t} \cdot \alpha\beta + \sum_{j,k} \nu_{jk,t} \cdot \beta^2$$

$$\begin{aligned}\nu_{ij,t+1} &= \left(\nu_{ii,t} + \nu_{jj,t}\right) \cdot 2\alpha\beta + \sum_k \nu_{kk,t} \cdot 2\beta^2 + \nu_{ij,t} \cdot (\alpha^2 + \beta^2) \\ &= + \sum_{k,l} \left(\nu_{ik,t} + \nu_{jl,t}\right) \cdot (\alpha\beta + \beta^2) + \sum_{k,l} \nu_{kl,t} \cdot 2\beta^2\end{aligned}$$

If we sum the frequencies of all homozygous or heterozygous genotypes (e.g. sums over upper and lower part of table *A1.3-1*), we get:

$$\begin{aligned}(1 - H)_{t+1} &= \sum_i \nu_{ii,t+1} \\ &= \alpha^2 \sum_i \nu_{ii,t} + (n-1) \cdot \beta^2 \sum_i \nu_{ii,t} + 2\alpha\beta \sum_{i,j} \nu_{ij,t} + (n-2) \cdot \beta^2 \sum_{i,j} \nu_{ij,t} \\ &= [\alpha^2 + (n-1) \cdot \beta^2] \sum_i \nu_{ii,t} + [2\alpha\beta + (n-2) \cdot \beta^2] \sum_{i,j} \nu_{ij,t}\end{aligned}$$

$$
\begin{aligned}
H_{t+1} = & \sum_{i,j} \nu_{ij,t+1} \\
= & (n-1) \cdot 2\alpha\beta \sum_{i} \nu_{ii,t} + \left(\frac{(n-1)(n-2)}{2}\right) \cdot 2\beta^2 \sum_{i} \nu_{ii,t} + (\alpha^2 + \beta^2) \sum_{i,j} \nu_{ij,t} \\
& + 2(n-2) \cdot (\alpha\beta + \beta^2) \sum_{i,j} \nu_{ij,t} + \left(\frac{(n-2)(n-3)}{2}\right) \cdot 2\beta^2 \sum_{i,j} \nu_{ij,t} \\
= & [(n-1) \cdot 2\alpha\beta + (n-1)(n-2) \cdot \beta^2] \sum_{i} \nu_{ii,t} \\
& + [(\alpha^2 + \beta^2) + 2(n-2) \cdot (\alpha\beta + \beta^2) + (n-2)(n-3) \cdot \beta^2] \sum_{i,j} \nu_{ij,t}
\end{aligned}
$$

As the equilibrium is reached if heterozygosity does not change anymore over time, i.e. transitions from homozygous to heterozygous genotypes are as frequent as the inverse, we can drop the time indices and write:

$$[2\alpha\beta + (n-2) \cdot \beta^2]H = [(n-1) \cdot 2\alpha\beta + (n-1)(n-2) \cdot \beta^2](1-H)$$

After dividing both sides by β and re-substituting $\alpha = (1-\mu)$, $\beta = \mu/(n-1)$, we get:

$$[2(1-\mu) + (n-2)/(n-1) \cdot \mu]H = [2(n-1) \cdot (1-\mu) + (n-2) \cdot \mu](1-H)$$

This simplifies to:

$$[2(n-1) - n\mu](n-1)^{-1}H = [2(n-1) - n\mu](1-H)$$

As $n \geq 2$ and $\mu \geq 0$, the solution for H is:

$$H = (n-1)/n$$

This is exactly identical to the expected heterozygosity under n-allele HWE for isoplethic alleles. For an infinite number of alleles, H converges to one:

$$\lim_{n\to\infty} H = \lim_{n\to\infty} (n-1)/n = 1.$$

Asymmetric mutation rate

To have "manually" verifiable results, we will again use a two-alleles model for illustration: Let *a* and *A* be two different alleles (DNA nucleotides, SSR copy numbers) with mutation rate μ_a for a → A and μ_A for A → a. This corresponds to the following mutation scheme:

↲	a	A
a	$1-\mu_a$	μ_A
A	μ_a	$1-\mu_A$

Mutations between the two alleles can then be described by the allele substitution matrix:

$$\mathcal{S} = \begin{bmatrix} 1-\mu_a & \mu_A \\ \mu_a & 1-\mu_A \end{bmatrix}$$

which has the dominant eigenvector (final allele frequencies):

$$\begin{bmatrix}\nu_{a,\infty}\\ \nu_{A,\infty}\end{bmatrix} = \begin{bmatrix}\frac{\mu_A}{\mu_a+\mu_A}\\ \frac{\mu_a}{\mu_a+\mu_A}\end{bmatrix}, \text{ or } \frac{\nu_{a,\infty}}{\nu_{A,\infty}} = \frac{\mu_A}{\mu_a}.$$

Assuming that each allele mutates independently, i.e. the mutation rates between genotypes are the product of the mutation rates between alleles, this corresponds to the following mutation scheme at the genotype level:

↲	aa	aA	AA
aa	$(1-\mu_a)^2$	$\mu_A(1-\mu_a)$	${\mu_A}^2$
aA	$2\mu_a(1-\mu_a)$	$\mu_a\mu_A+(1-\mu_a)(1-\mu_A)$	$2\mu_A(1-\mu_A)$
AA	${\mu_a}^2$	$\mu_a(1-\mu_A)$	$(1-\mu_A)^2$

Treating the genotypes as "ordered" (i.e. "*aA*" ≠ "*Aa*"), the mutation rates in the genotype mutation scheme can be directly derived from those in the allele mutation scheme – they correspond to the Kronecker product of $\mathcal{S}$ with itself:

$$\mathcal{K} = \mathcal{S} \otimes \mathcal{S} = \begin{bmatrix}(1-\mu_a)^2 & \mu_A(1-\mu_a) & \mu_A(1-\mu_a) & {\mu_A}^2\\ \mu_a(1-\mu_a) & (1-\mu_a)(1-\mu_A) & \mu_a\mu_A & \mu_A(1-\mu_A)\\ \mu_a(1-\mu_a) & \mu_a\mu_A & (1-\mu_a)(1-\mu_A) & \mu_A(1-\mu_A)\\ {\mu_a}^2 & \mu_a(1-\mu_A) & \mu_a(1-\mu_A) & (1-\mu_A)^2\end{bmatrix}$$

To get from $\mathcal{K}$ to the "true" genotype mutation scheme, rows 2 & 3 that describe mutations resulting in either of the two synonymous heterozygous genotypes have to be summed, and one of the two columns 2 or 3 describing mutations from each of the synonymous genotypes towards all others is then discarded.

The eigenvector of the matrix $\mathcal{K}$ equals the Kronecker product of the eigenvectors of $\mathcal{K}$'s factors, the two identical matrices $\mathcal{S}$:

$$\begin{bmatrix}\nu_{a,\infty}\\ \nu_{A,\infty}\end{bmatrix} \otimes \begin{bmatrix}\nu_{a,\infty}\\ \nu_{A,\infty}\end{bmatrix} = \begin{bmatrix}\frac{\mu_A}{\mu_a+\mu_A}\\ \frac{\mu_a}{\mu_a+\mu_A}\end{bmatrix} \otimes \begin{bmatrix}\frac{\mu_A}{\mu_a+\mu_A}\\ \frac{\mu_a}{\mu_a+\mu_A}\end{bmatrix} = \begin{bmatrix}\frac{{\mu_A}^2}{(\mu_a+\mu_A)^2} & \frac{\mu_a\mu_A}{(\mu_a+\mu_A)^2} & \frac{\mu_a\mu_A}{(\mu_a+\mu_A)^2} & \frac{{\mu_a}^2}{(\mu_a+\mu_A)^2}\end{bmatrix}^T$$

Lumping the synonymous heterozygous genotypes together by summing rows 2 & 3 (or columns 2 & 3 of the transposed vector) gives the final genotype frequencies expected under this asymmetric mutation scheme:

$$[\nu_{aa,\infty} \quad \nu_{aA,\infty} \quad \nu_{AA,\infty}]^T = \begin{bmatrix}\frac{{\mu_A}^2}{(\mu_a+\mu_A)^2} & \frac{2\mu_a\mu_A}{(\mu_a+\mu_A)^2} & \frac{{\mu_a}^2}{(\mu_a+\mu_A)^2}\end{bmatrix}^T$$

Consequently, the final genotype frequencies will correspond to:

$$[\nu_{aa,\infty} \quad \nu_{aA,\infty} \quad \nu_{AA,\infty}]^T = [{\nu_{a,\infty}}^2 \quad 2\,\nu_{a,\infty}\nu_{A,\infty} \quad {\nu_{A,\infty}}^2]^T$$

or HWE for the final allele frequencies. The same procedure can be applied to any arbitrary allelic mutation scheme (numerically for higher numbers of alleles). Thus, independently of the actual mutation rates or the number of possible alleles, mutation schemes that act on each allele independently will lead towards a randomization of the combinations of alleles within individuals, i.e. HWE.

A1.4 Genetic drift and heterozygosity – multiple alleles

Whether genetic drift tends to increase or decrease heterozygosity starting from $F_{IS} = 0$ depends on the nature (heterozygous or homozygous) of the most frequent genotype. We shall aim to find the range of allele frequencies for which a homozygous genotype is most frequent.

(I) Without loss of generality, we may assume that $\nu_1 \geq \nu_2 \geq \nu_k \; \forall \; k > 2$, i.e. ν_1 and ν_2 are the frequencies of the two most frequent alleles (equality included).

(II) As all allele frequencies must sum to one, it follows that $\nu_1 + \nu_2 + \sum_k \nu_k = 1$.

(III) As we are only interested in populations at $F_{IS} = 0$ (i.e. in HWE), $\nu_{11} = \nu_1^2$ and $\nu_{12} = 2\nu_1\nu_2$.

Because of (I), ν_{11} will be the frequency of the most frequent homozygote genotype, and ν_{12} will be the frequency of the most frequent heterozygous genotype in the population. With (III), a homozygous genotype will therefore be the most frequent if and only if $\nu_1^2 > 2\nu_1\nu_2$ in a population with $n \geq 2$ alleles at Hardy-Weinberg equilibrium.

Since $\nu_1 > 0$ because of (I), we can divide both sides of the inequality by ν_1 and arrive at the condition $\nu_1 > 2\nu_2$. Following from (I) and (II), $2\nu_2$ is minimal if all allele frequencies except ν_1 are equal; thus we can substitute $\nu_2 = (1 - \nu_1)/(n - 1)$ and resolve the inequality to $\nu_1 > 2/(n + 1)$. Thus, for any given n, the most frequent genotype at $F_{IS} = 0$ will be homozygous if the frequency of the most frequent allele is greater than twice the frequency of the second-most frequent allele, and at least greater than $2/(n + 1)$. As n decreases whenever one allele is lost by genetic drift, this minimal frequency increases and the range where a homozygous genotype is favored decreases.

A1.5 Convergence times – full model

Our model

- Convergence time to the mean $\overline{F_{IS,\infty}}$:

Using the basic equation of our Markov chain model

$$\vec{x}_t = M^t \vec{x}_0$$

where $\vec{x}_0$ is the start state vector, $\vec{x}_t$ the vector of state probabilities and M the transition matrix (based on N, μ, c), we iteratively calculated the difference between the mean $\overline{F_{IS,t}}$ for the two start states $F_{IS,0} = 1, \nu_a = \nu_A$ and $F_{IS,0} = -1$ at each time step. We considered the mean $\overline{F_{IS,\infty}}$ converged when this difference passed below $\varepsilon = 1/(2N)$.

- Convergence time to full final distribution of $\widetilde{F_{IS,\infty}}$:

Similar to our derivation of t_μ, the convergence time of the full model can be approximated using the transition matrices' second largest eigenvalue (Markov chain mixing time approach), λ_2, which we derived numerically. The time to convergence is then $t_{III} = 1 + \log_{\lambda_2} \varepsilon$. Interestingly, we found that $\lambda_2 = (1 - 2\mu)$ in all cases we tested. This appears to be a parallel to the model presented in Balloux et al. 2003 (see below – as explained in part *A1.2*, our value $\lambda_2 = (1 - 2\mu)$ is a special case of $(1 - \mu\frac{n}{n-1})$ which converges to $(1 - \mu)$ for $n \to \infty$ as in the model from Balloux et al. 2003).

Model from Balloux et al. 2003

- Convergence time to the mean $\overline{F_{IS,\infty}}$:

Using the model equation (from Balloux et al. 2003, equation 5 & 6):

$$\begin{bmatrix} F_{t+1} \\ \Theta_{t+1} \end{bmatrix} = (1-\mu)^2 \left(\begin{bmatrix} c + \frac{1-c}{2N} & (1-c)\left(1 - \frac{1}{N}\right) \\ \frac{1}{2N} & 1 - \frac{1}{N} \end{bmatrix} \begin{bmatrix} F_t \\ \Theta_t \end{bmatrix} + \begin{bmatrix} \frac{1-c}{N} \\ \frac{1}{2N} \end{bmatrix} \right)$$

where F_t represents the observed and Θ_t the expected homozygosity at time t, we iteratively calculated the difference between the mean $\overline{F_{IS,t}}$ for the two start states $[F, \Theta]_0 = [1, 0.5](F_{IS,0} = 1, \nu_a = \nu_A)$ and $[F, \Theta]_0 = [0, 0.5](F_{IS,0} = -1)$ at each time step. We considered the mean $\overline{F_{IS,\infty}}$ converged when this difference passed below $\varepsilon = 1/(2N)$.

Note that this equation does not have the structure of a geometric progression, so that the method used above (Markov chain mixing time) cannot be applied.

A2 Additional figures

A2.1 Interpretation of de Finetti *diagrams*

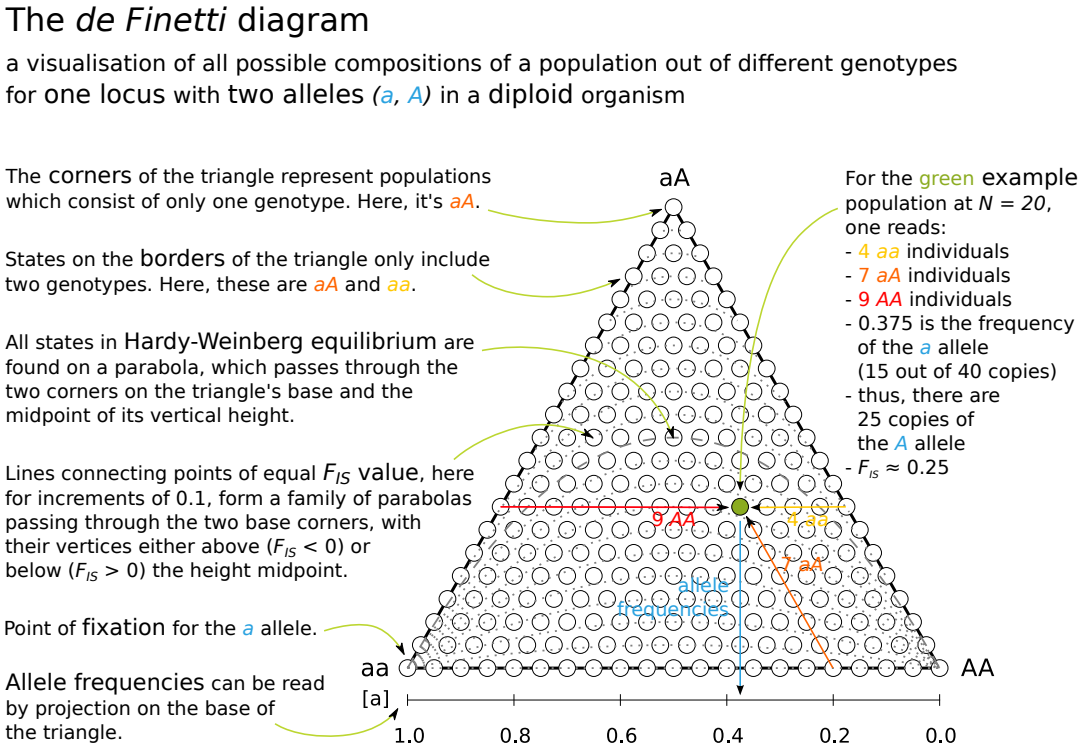

Figure A2-1. Scheme showing how to read *de Finetti* diagrams.

De Finetti diagrams are ternary plots that provide a compact and non-redundant representation of genotype (perpendicular distances from sides), allele (horizontal coordinate) and homo-heterozygote (vertical coordinate) frequencies for a single locus with two alleles. All combinations of genotype counts ((q_{aa}, q_{aA}, q_{AA}), states of our model) correspond to discrete points on the *de Finetti* diagram, with the distance between neighboring states equal to $1/N$. All states for which $F_{IS} = 0$ are on a parabola passing through the fixation states (baseline corners of the triangle) that culminates at the genotype frequencies [0.25, 0.5, 0.25] (vertical height midpoint of the triangle). Points "above" the parabola have negative, "below" the parabola positive F_{IS} values. For each point, the diagram thus allows to simultaneously track e.g. the observed heterozygosity (perpendicular distance from baseline to point), the expected heterozygosity (perpendicular distance from baseline to the Hardy-Weinberg parabola for the given allele frequencies) and the maximum possible heterozygosity (perpendicular distance from baseline to the "upper" side of the triangle for the given allele frequencies), as well as fixation of an allele and the current number of homozygote/heterozygote genotypes (central part / sides / corners of the diagram).

Due to the discreteness of individuals, the exact Hardy-Weinberg genotype frequencies cannot be reached for many combinations of allele frequencies. Instead, the states closest

to HWE possess slight homozygote or heterozygote excess. Near fixation, the expected heterozygosity and the maximum possible heterozygosity converge, so that excess heterozygosity is no longer distinguishable from HWE. This situation first occurs when the difference between maximum possible and expected heterozygosity passes below $1/\mathrm{N}$ (frequency equivalent of one individual). The maximum possible heterozygosity equals $\max(\nu_{\mathrm{aA}}) = 2(1-\nu_{\mathrm{a}})$ if *a* is the most frequent allele, and the expected heterozygosity equals $\exp(\nu_{\mathrm{aA}}) = \mathrm{H_e} = 2\nu_{\mathrm{a}}(1-\nu_{\mathrm{a}})$, so that $1/\mathrm{N} = 2(1-\nu_{\mathrm{a}})^2$ and finally $\nu_{\mathrm{a}} = 1 - \sqrt{1/2\mathrm{N}}$ for the frequency of the most frequent allele. If any one allele exceeds this frequency, it is considered nearly fixed. A similar situation where expected and maximal heterozygosity become indistinguishable occurs when the number of different alleles goes towards $2\mathrm{N}$, its maximum in a finite population: if there are more than N nearly equally frequent alleles, the difference between maximum possible and expected heterozygosity passes below $1/\mathrm{N}$, i.e. graphically the vertex of the (multi-dimensional equivalent of the) Hardy-Weinberg parabola nearly "touches" the states where all individuals of the population are heterozygous.

To visualize the expected changes in the three genotype frequencies (ternary plot coordinates) through time, starting from any possible combination of genotype counts (point/state), we constructed "*de Finetti* landscapes", i.e. three-dimensional plots where the "height" of each point in the landscape is proportional to the sum of squared genotype frequency changes expected per time step when starting from the respective state. As in classical mechanics, the height of each point in the landscape is thus proportional to the square of the speed with which it is left. This is the basis for an analogy with the natural world that makes these plots intuitively interpretable: one can imagine a population as a small ball "rolling" from "hilltops" towards "valleys", changing its genotype frequencies according to this displacement within the ternary plot. The point(s) with zero height thus correspond(s) to the final expected state(s) for the respective parameter combinations (reproductive mode, mutation and genetic drift). The flatter the landscape, the longer it takes to reach these states.

A2.2 De Finetti landscapes for reproduction

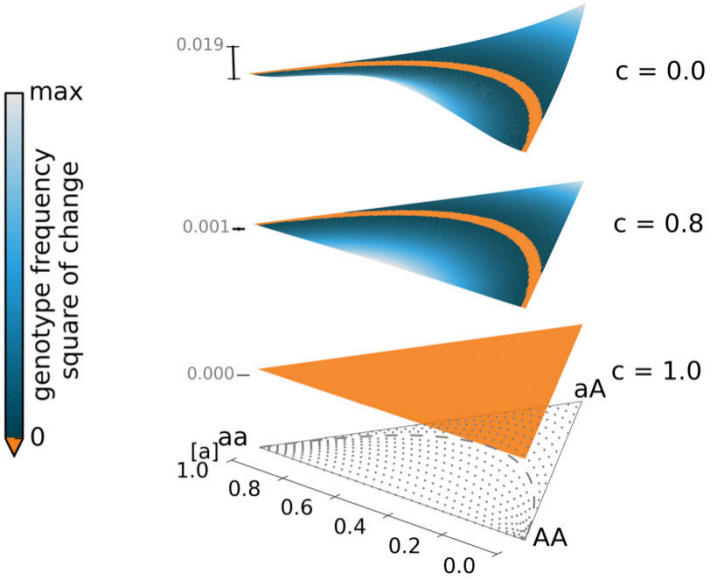

Figure A2-2. *De Finetti* landscapes for reproduction. Increasing the rate of clonality c flattens the landscape (increased time to final expected states), but does not change the final expected states (orange parabola: HWE, $F_{IS} = 0$) except if the population is completely clonal $c = 1.0$.

A2.3 De Finetti landscapes for mutation

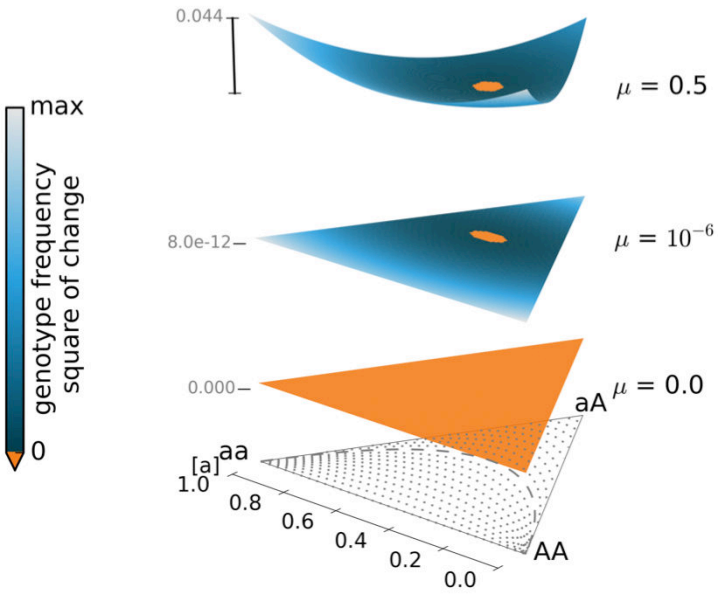

Figure A2-3. *De Finetti* landscapes for mutation. Decreasing the mutation rate μ flattens the landscape (increased time to final expected states), but does not change the final expected states (orange dot: HWE, $F_{IS} = 0$ for equal allele frequencies, $\nu_a = \nu_A = 0.5$) except if there is no mutation, $\mu = 0$.

A2.4 De Finetti *landscapes for genetic drift*

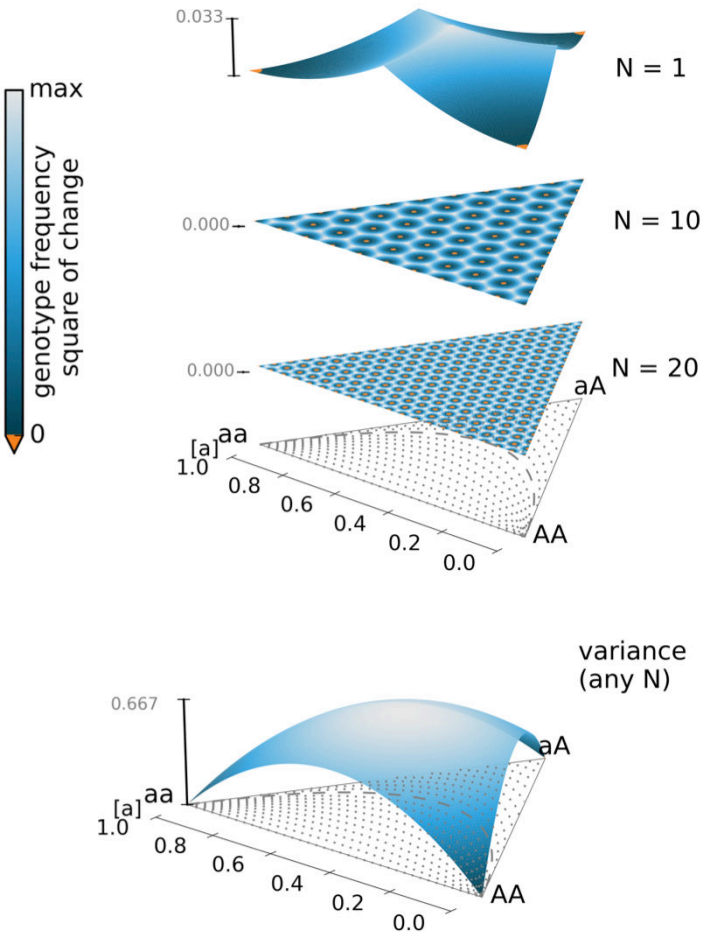

Figure A2-4. *De Finetti* landscapes for genetic drift. Increasing the population size N increases the density of points with zero expected change in the diagram – for any state, the population is most likely to remain where it was (same genotype frequencies) in the next time step. However, genotype dynamics due to genetic drift can be explained by the variance $\mathrm{Var(X)} = \sum_{\mathrm{i}} \nu_{\mathrm{ii}}(1-\nu_{\mathrm{ii}}) + \sum_{\mathrm{i,j}} \nu_{\mathrm{ij}}(1-\nu_{\mathrm{ij}})$ around this expectation, which is highest in the center (genotype frequencies [1/3, 1/3, 1/3]) and zero at the corners of the triangle. This means that the direction of random genotype frequency changes due to genetic drift is least predictable if all genotypes are equally frequent, and all random change will cease if the frequency of one genotype becomes one ("fixation" of a genotype). Note that the corresponding co-variances are usually non-zero, as the genotype frequencies are interdependent.

A2.5 Example trajectories over time

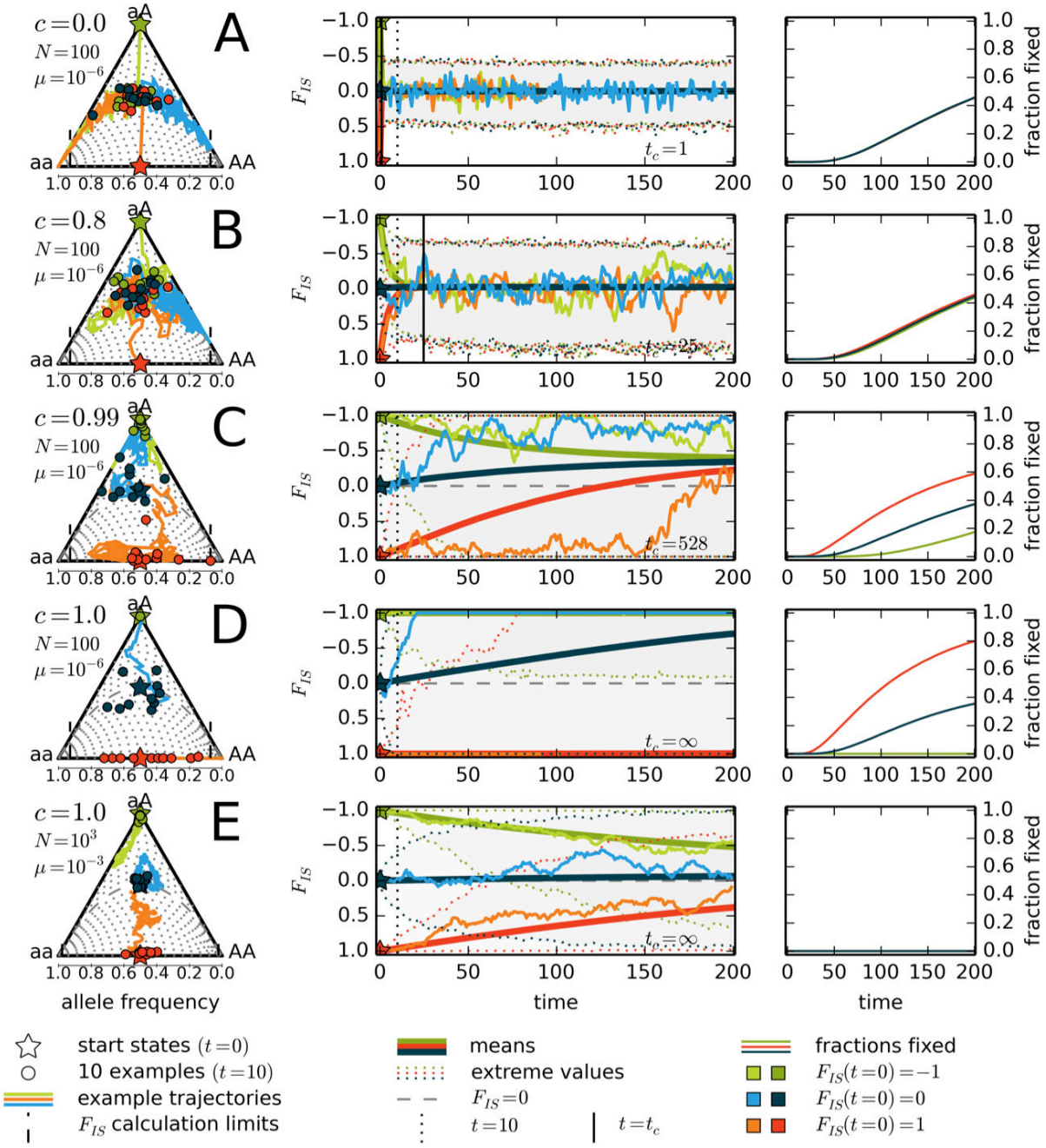

Figure A2-5. Example trajectories over time for different parameter sets (c, μ, N). Legend see next page; overview of interpretation see table A2-1.

Figure A2-5. Example trajectories over time for different parameter sets (c, μ, N). Color codes – start states: In *light/dark green* (lines/dots/stars), trajectories that started at $F_{IS,0} = -1$ (all individuals heterozygotes i.e both allele frequencies 0.5); in *light/dark blue* (lines/dots/stars), trajectories that started at $F_{IS,0} = 0$ and $\nu_{a,0} = \nu_{A,0} = 0.5$ (Hardy-Weinberg proportions, both allele frequencies 0.5); and in *red and orange* (lines/dots/stars), trajectories that started at $F_{IS,0} = 1$ and $\nu_{a,0} = \nu_{A,0} = 0.5$ (all individuals homozygotes, both allele frequencies 0.5). Rows – parameter sets: A, exclusive sexuality, $c = 0.0$; B and C, partial clonality with $c = 0.8$ and $c = 0.99$ respectively; D, exclusive clonality, $c = 1.0$, low mutation rate and small population; E, exclusive clonality, $c = 1.0$, high mutation rate and big population. Columns – diagnostic plots: *Left: De Finetti* diagrams showing one example trajectory (line) traced over 200 generations and ten example states at $t = 10$ (dots) per start state (colors/stars: start states). F_{IS} were not calculated for states outside the vertical dashed black lines (near-fixation, frequency of one allele exceeds $1 - \sqrt{1/(2N)}$). *Central:* corresponding dynamics of F_{IS} over 200 generations illustrated by one example trajectory (thin/light line), the mean over 10^5 trajectories (heavy/dark line) and the range (shaded area delimited by dotted lines) for the three start states (stars); horizontal dashed grey line indicates $F_{IS} = 0$; vertical black lines correspond to $t = 10$ (dotted) and t_c (solid). *Right:* corresponding fraction of trajectories at fixation for one allele, out of 10^5 trajectories over 200 generations for the three start states.

Table A2-1: Effects of different rates of partial clonality on the dynamics of F_{IS}. Text in italics refers to the illustration in figure A2-5.

		Fig. A2-5	**Convergence to a common mean F_{IS}**	**Convergence speed**	**Variation of F_{IS} values**	**Alternation of positive and negative F_{IS}**	**F_{IS} values after ten generations**	**Allele fixation rate**
Rate of clonal reproduction		*Where to look*	*Central plot thick lines*	*Central plot thick lines and t_c*	*Central plot grey area / thin lines, left plot thin lines*	*Central plot thin lines*	*Left plot dots*	*Right plot lines*
Exclusively sexual to intermediate rate		*Row A*	Yes, $F_{IS} = 0$	One/very few generations	Depends on N only	High frequency	independent of start F_{IS}	Identical for all start F_{IS}
Intermediate to high rate		*Row B*	Yes, $F_{IS} \approx 0$	Several generations	Increased	Low frequency	slightly biased according to start F_{IS}	Slight dependence on start F_{IS}
High to very high rate		*Row C*	Yes, $F_{IS} < 0$	Many generations	Full range	Very low frequency	distinct according to start F_{IS}	Pronounced dependence on start F_{IS}
Very high rate to exclusively clonal	Genetic drift dominates (N small)	*Row D*	No, fixation or $F_{IS} = -1$	Depends on N	Depends on start F_{IS}	Rare except for recent loss of sex or near fixation	distinct according to start F_{IS}	$F_{IS} = -1$ hardly fixes, $F_{IS} = 0$ at reduced rate, $F_{IS} = 1$ very fast
	Mutation dominates (N very big)	*Row E*	Yes, $F_{IS} = 0$	Depends on μ	Depends on N	Depends on μ & N	Depends on μ	Depends on μ & N

A3 Literature review – references

We conducted a Web of Science search with the exact search term [(microsatellite OR "SSR" OR "simple sequence repeat" or "SNP" or "single nucleotide polymorphism") AND (clonal OR asexual OR vegetative OR apomictic OR apomixis OR agamospermy OR parthenogenesis)], which yielded 5480 references. Screening the 2000 most recent references yielded 21 studies that were accessible and reported relevant data (F_{IS} values; no known population substructure, no filtering of repeated MLGs, no cyclical parthenogenesis, dominantly diploid life cycle). Six of these studies reported only mean F_{IS} values over all loci, the remaining 15 are listed below.

Table A3-1: References and supplementary data for the literature review. "PN" refers to the number of the dataset(s) in figure 9, *: F_{IS} was calculated from H_e and H_o.

Reference	Details	PN
S. Duran, M. Pascual, A. Estoup, X. Turon (2004): Strong population structure in the marine sponge *Crambe crambe* (Poecilosclerida) as revealed by microsatellite markers	*Crambe crambe,* Porifera Mediterranean + Atlantic	1-11
V. Rougeron, E. Waleckx, M. Hide, T. de Meeûs, J. Arevalo, A. Llanos-Cuentas, A.L. Bañuls (2008): A set of 12 microsatellite loci for genetic studies of *Leishmania braziliensis*	*Leishmania braziliensis*, Euglenozoa Peru	12
G. Motoie, G. E. M. Ferreira, E. Cupolillo, F. Canavez, V. L. Pereira-Chioccola (2013): Spatial distribution and population genetics of *Leishmania infantum* genotypes in São Paulo State, Brazil, employing multilocus microsatellite typing directly in dog infected tissues	*Leishmania infantum*, Euglenozoa Brazil	13-14
T. Nagamitsu, M. Ogawa, K. Ishida, H. Tanouchi (2004): Clonal diversity, genetic structure, and mode of recruitment in a *Prunus ssiori* population established after volcanic eruptions	*Prunus ssiori*, Angiospermae Japan	16
S. Stoeckel, J. Grange, J. Fernandez-Manjarres, I. Bilger, N. Frascaria-Lacoste, S. Mariette (2006): Heterozygote excess in a self-incompatible and partially clonal forest tree species – *Prunus avium* L.	*Prunus avium*, Angiospermae France	17-19
J. M. Corral, M. Puente Molins, O. M. Aliyu, T. F. Sharbel (2011): Isolation and characterization of microsatellite loci from apomictic *Hypericum perforatum* (Hypericaceae)	*Hypericum perforatum*, Angiospermae USA, Czech Republic, Germany	20*-23*
K. Jiang, H. Gao, N.-N. Xu, E. P. Keung Tsang, X. Chen (2011): A set of microsatellite primers for *Zostera japonica* (Zosteraceae)	*Zostera japonica*, Angiospermae China, Taiwan	24*-25*
J. M. Tew, S. L. Lance, K. L. Jones, S. D. Fehlberg (2012): Microsatellite development for an endangered riparian inhabitant, *Lilaeopsis schaffneriana* subsp. *recurva* (Apiaceae)	*Lilaeopsis schaffneriana* ssp. *recurva*, Angiospermae USA, Mexico	26*-27*

W. Liu, Y. Zhou, H. Liao, Y. Zhao, Z. Song (2011): Microsatellite primers in *Carex moorcroftii* (Cyperaceae), a dominant species of the steppe on the Qinghai-Tibetan Plateau	*Carex moorcroftii*, Angiospermae China	28*-31*
C. Barnabe, R. Buitrago, P. Bremond, C. Aliaga, R. Salas, P. Vidaurre, C. Herrera, F. Cerqueira, M.-F. Bosseno, E. Waleckx, S. F. Breniere (2013): Putative panmixia in restricted populations of *Trypanosoma cruzi* isolated from wild *Triatoma infestans* in Bolivia	*Trypanosoma cruzi*, Euglenozoa Bolivia	33-38
S. W. M. Tesson, M. Borra, W. H. C. F. Kooistra, G. Procaccini (2011): Microsatellite primers in the planktonic diatom *Pseudo-nitzschia multistriata* (Bacillariophyceae)	*Pseudo-nitzschia multistriata*, Heterokonta Italy	39*-42*
L. Villate, D. Esmenjaud, M. van Helden, S. Stoeckel, O. Plantard (2010): Genetic signature of amphimixis allows for the detection and fine scale localization of sexual reproduction events in a mainly parthenogenetic nematode	*Xiphinema index*, Nematoda France	43-48
H. Gao, K. Jiang, Y. Geng, X.-Y. Chen (2012): Development of microsatellite primers of the largest seagrass, *Enhalus acoroides* (Hydrocharitaceae)	*Enhalus acoroides*, Angiospermae China	49*-50*
L. M. McInnes, A. P. Dargantes, U .M. Ryan, S. A. Reid (2012): Microsatellite typing and population structuring of *Trypanosoma evansi* in Mindanao, Philippines	*Trypanosoma evansi*, Euglenozoa Philippines	51-52
R. Vilas, A. Cao, B. G. Pardo, S. Fernández, A. Villalba, P. Martínez (2011): Very low microsatellite polymorphism and large heterozygote deficits suggest founder effects and cryptic structure in the parasite *Perkinsus olseni*	*Perkinsus olseni*, Alveolata Spain	53-56

6.2 Multilocus simulation of a demographic bottleneck

Co-authors: Clément Barthélémy, Romuald Rouger, Solenn Stoeckel

A main result of the preceding article was that the dynamics of F_{IS} are slowed down under partial asexuality, which means that extreme deviations from the mean will take longer to disappear. Demographic bottlenecks, a more or less abrupt diminution of the population size and subsequent re-expansion, are one way how such extreme deviations may arise: As an example, imagine a series of epidemics or storms which devastate a landscape, leaving only some random survivors. Over the subsequent years, the population grows again to its previous size, but the catastrophe leaves traces: genotypes/alleles may have disappeared, or randomly changed their frequencies. While the big partially clonal population that lived through the bottleneck may initially have had all loci in Hardy-Weinberg equilibrium with isoplethic alleles, this is most probably not the case during the size reduction. Our question is, how long does the recovery of $F_{IS} = 0$ take, depending on the population's rate of asexuality?

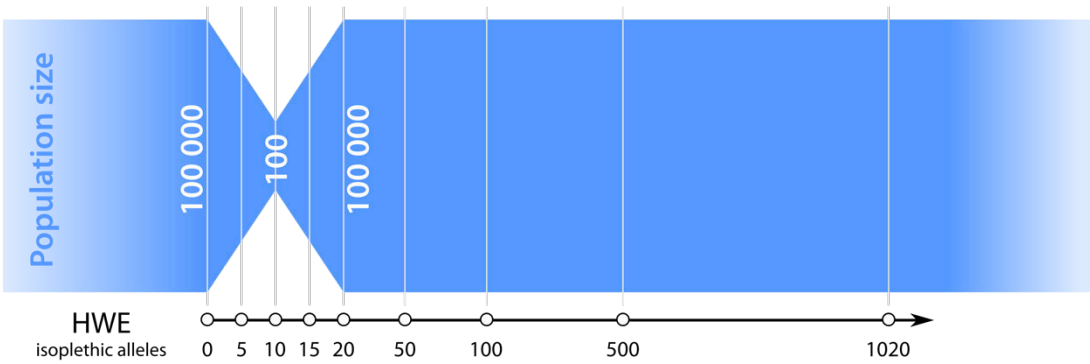

Figure 6.1 Schematic representation of the simulated demographic bottleneck effect. Numbers below arrow: times (numbers of generations after the start) at which the distribution of F_{IS} was sampled.

We simulated a demographic bottleneck on a large scale: A population of initially/finally 10^5 individuals suffered a linear decrease to only 100 individuals over ten generations, and afterwards underwent a symmetrical linear increase of its population size (figure 6.1). Each individual had 100 (physically unlinked, but co-inherited during asexual reproduction) selectively neutral polymorphic loci, each with either two (bi-allelic SNP), four (tetra-allelic SNP or SSR) or ten (typical SSR) possible alleles, among which mutation occurred with a symmetric rate $\mu = 10^{-3}$. Loci in populations with a high rate of asexuality were thus mutation-dominated (convergence towards $F_{IS} = 0$) except at the nadir of population size. The distribution of F_{IS} values, both among loci within the same simulation and between 100 independent repeats, was recorded five (middle of decrease), ten (minimal population size), 15 (middle of increase), 20 (end of demographic bottleneck, original population size), 50 and 100 generations, and for exclusively asexual populations also 500 and 1020 generations after the start.

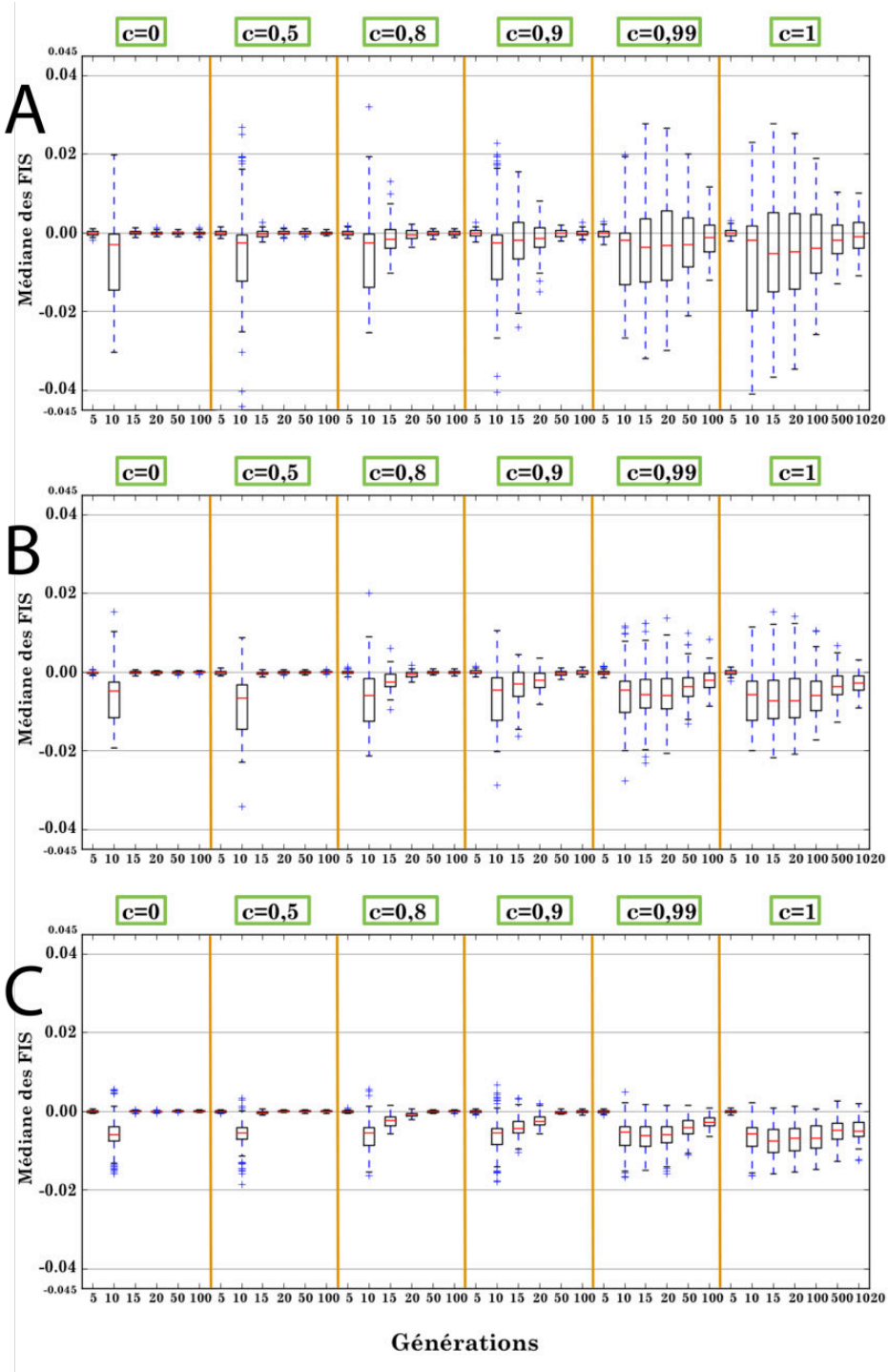

Figure 6.2 Distributions of F_{IS} (means over 100 independent simulations) during and after a demographic bottleneck, in populations with different rates of asexuality, for markers with a different maximal number of alleles. c: rate of asexuality; "Générations" (generations): number of generations since the beginning of the simulation, population sizes are $N_5 = 50\,050$, $N_{10} = 100$, $N_{15} = 50\,050$, $N_{\geq 20} = 10^5$. A: maximal two alleles per locus, B: maximal four alleles per locus, C: maximal ten alleles per locus.

The results confirmed our expectations from small populations and single loci (figure 6.2): the higher the rate of asexual reproduction, the longer it takes until $F_{IS} = 0$ is reached again after a demographic bottleneck. Under the set conditions, the demographic bottleneck

barely leaves a trace in F_{IS} for exclusive sexuality and 50% asexuality except at the minimal population size, yet the predominantly negative F_{IS} values observed there take at least until the end of the bottleneck ($c = 0.8$) or even after ($c = 0.9$) at higher rates of asexuality. At $c \geq 0.99$, where the return to $F_{IS} = 0$ depends (almost) completely on mutation, F_{IS} even at first continues its negative trend into the population growth phase; respectively 100 and 1020 generations are then not enough to eliminate the strong bias towards negative F_{IS} values. Results are very similar across different numbers of alleles, yet the spread of the simulated F_{IS} distributions is smaller with more alleles, i.e. the effect is easier to observe.

This preliminary study is a persuasive illustration of the relevance of our results for single loci at a larger scale. Though this simulation involved multiple (neutral) loci, the results were well explicable based on the previously described single-locus model (article II). Moreover, it may have great importance for the interpretation of field data: while the methods currently used to detect demographic bottlenecks in population genetic data (e.g. BOTTLENECK, Piry et al. 1999) largely rely on allele frequencies rather than observed F_{IS} values, and may thus not be directly compromised by the effect we observed, the time scales for the detection (recent vs. historic bottleneck) may be different between exclusively sexual and partially asexual populations. However, a more detailed analysis of the compatibility of the currently available software with partial asexuality is still pending.

A further integration of population genetic and demographic models would be especially interesting in the context of the next chapter, which deals with cyclical parthenogenesis: beside its peculiar alternation of sexual and asexual reproduction, this reproductive system is also usually characterized by strong population growth during the asexual phase, followed by a sudden diminution during the sexual phase. This dynamic is not yet integrated into the next article, but would certainly affect the results presented there.

6.3 Neutral diversity under cyclical parthenogenesis

Cyclic parthenogenesis is a form of partial asexuality that appears to be especially common among animals. It was first described for aphids (Bonnet 1745, Owen 1849), but is also found in other arthropods such as *Daphnia* (Decaestecker et al. 2009). In contrast to "acyclic" partial asexuality, which was the subject of the previous article, in this reproductive system clonal and sexual offspring are not produced simultaneously at a constant rate, but sequentially: typically, one generation of sexual reproduction is cyclically followed by several generations of asexual reproduction. The change between asexual and sexual reproduction may be linked to the seasonality of habitats (e.g. as in aphids and *Daphnia*), or the whole life cycle may be substructured by host changes, as for parasitic rust fungi (e.g. poplar leaf rust, Barrès et al. 2012) or parasitic flatworms (Prugnolle et al. 2005a). In plants, a similar situation could be found in partially clonal species with synchronized mass flowering, such as bamboo (Makita 1998; see also discussion in Muirhead & Lande 1997).

Depending on the author's views, previous studies have sometimes included cyclical parthenogenesis either with the acyclic case (e.g. Balloux et al. 2003 implicitly by comparing their results with those of Berg & Lascoux 2000) or treated it as equivalent to sexual

reproduction (e.g. Jaquiéry et al. 2014). In view of the results presented in article II, both these hypotheses seem questionable – however, they can be easily tested with our model. As a first step, we compared the equilibrium distributions of F_{IS} across exclusive sexuality, exclusive asexuality, acyclic partial clonality and cyclic parthenogenesis, taking two different sampling schemes (directly before/after the sexual generation) and different cycle lengths (numbers of asexual generations) into account. To our knowledge, this is the first theoretical study establishing a reference for F_{IS} under cyclical parthenogenesis.

We found that all four reproductive systems produce different patterns of genetic diversity. Though the F_{IS} distributions from cyclically parthenogenetic populations sampled directly after the sexual generation closely correspond to those expected for an exclusively sexual population, this similarity is lost during the asexual phase. Before sexual reproduction, the F_{IS} distributions are more similar (but not identical) to their counterpart assuming acyclic partial clonality. The closeness of these similarities depends on the strength of genetic drift (number of clonal cycles) during the asexual phase. Looking directly at the genotype frequencies, it becomes clear why: though sexual reproduction regularly "resets" F_{IS} to zero, it does not reset the change in allele frequencies during the asexual phase.

Genotype frequency dynamics under cyclical parthenogenesis can be somewhat predicted from the results presented in article II for exclusively sexual and exclusively asexual reproduction: Cyclically parthenogenetic populations sampled at some time during the asexual phase are similar to "recently clonal" populations, which may retain their ancestral diversity over several generations. However, as a next step it would be interesting to look more closely at the dynamics of allele frequencies under cyclical parthenogenesis. The results of previous studies on spatially substructured populations (e.g. Berg & Lascoux 2000) suggest that population differentiation and F_{ST} values increase more quickly (for the connection between allele frequencies and F_{ST}, see chapter 5.5). How many generations of clonality would be needed until clonal erosion (Vanoverbeke & De Meester 2010) could already happen by genetic drift alone? Moreover, the first results presented in article III do not yet take the characteristic population dynamics of partially clonal organisms into account (see chapter 6.2). By establishing cyclical parthenogenesis as a reproductive system in its own right, also distinct from acyclic partial asexuality, we hope that our results will increase the awareness of field biologists (sampling strategy) and the interest of theoreticians, leading to more development in this area.

La diversité neutre sous parthénogenèse cyclique

La parthénogenèse cyclique est une forme de l'asexualité partielle qui semble être particulièrement fréquente chez les animaux. Elle a été décrite pour la première fois chez les pucerons (Bonnet 1745, Owen 1849), mais se retrouve également dans d'autres arthropodes tels que les daphnies (Decaestecker et al. 2009). Contrairement à l'asexualité "acyclique" partielle, qui a fait l'objet de l'article précédent, dans ce système reproducteur la progéniture clonale et sexuelle ne se produit pas au même temps et à un taux constant, mais de manière séquentielle : généralement, une génération de la reproduction sexuée est cycliquement suivie par plusieurs générations de reproduction asexuée. Le changement entre la reproduction sexuée et asexuée peut être lié à la saisonnalité des habitats (par exemple

comme chez les pucerons et les daphnies), où le cycle de vie entier peut être sous-structuré par des changements d'hôte, comme chez les champignons parasitaires de la rouille (par exemple la rouille du peuplier, Barrès et al. 2012) ou chez les vers plats parasitiques (Prugnolle et al. 2005a). Chez les plantes, une situation similaire pourrait être trouvée dans les espèces partiellement clonales avec une floraison de masse synchronisée, comme le bambou (Makita 1998 ; voir également la discussion dans Muirhead & Lande 1997).

Selon le point de vue de l'auteur, des études antérieures ont parfois inclus la parthénogenèse cyclique soit avec le cas acyclique (comme fait par exemple par Balloux et al. 2003 implicitement en comparant leurs résultats avec ceux de Berg & Lascoux 2000) soit traité comme équivalent à la reproduction sexuée (par exemple Jaquiery et al. 2014). Compte tenu des résultats présentés dans l'article II, ces deux hypothèses semblent discutables – cependant, elles peuvent être facilement testées avec notre modèle. Dans un premier temps, nous avons comparé les distributions d'équilibre de l' F_{IS} à travers la sexualité exclusive, l'asexualité exclusive, la clonalité partielle acyclique et la parthénogenèse cyclique, prenant en compte deux plans d'échantillonnage différents (directement avant / après la génération sexuée) et des longueurs de cycle différentes (nombre de générations asexuées). À notre connaissance, cette étude est la première à établir une référence théorique pour l'F_{IS} sous parthénogenèse cyclique.

Nous avons pu constater que les quatre systèmes de reproduction produisent des motifs différents de diversité génétique. Bien que les distributions de l' F_{IS} en populations cycliquement parthénogénétiques échantillonnées directement après la génération sexuée soient très proches à celles attendues pour une population exclusivement sexuée, cette similitude est perdue lors de la phase asexuée. Avant la reproduction sexuée, les distributions de l' F_{IS} sont plus similaires (mais pas identiques) à leur homologue en supposant la clonalité partielle acyclique. La proximité de ces similitudes dépend de la force de la dérive génétique (nombre de cycles clonales) pendant la phase asexuée. Si on regarde directement les fréquences génotypiques, l'explication est claire : si la reproduction sexuée « réinitialise » régulièrement l'F_{IS} et le remet à zéro, elle ne réinitialise pas le changement dans les fréquences des allèles pendant la phase asexuée.

La dynamique des fréquences génotypiques sous parthénogenèse cyclique peut être prédite à peu près à partir des résultats présentés dans l'article II pour la reproduction exclusivement sexuée et exclusivement asexuée : les populations parthénogénétiques cycliques échantillonnées à un certain moment au cours de la phase asexuée sont similaires aux populations « récemment clonales », qui peuvent conserver leur diversité ancestrale durant plusieurs générations. Cependant, lors d'une prochaine étape, il serait intéressant de regarder de plus près la dynamique des fréquences alléliques sous parthénogenèse cyclique. Les résultats des études antérieures sur les populations avec sous-structure spatiale (par exemple Berg & Lascoux 2000) suggèrent que la différenciation des populations et les valeurs d'F_{ST} augmentent plus rapidement (pour la connexion entre les fréquences alléliques et l'F_{ST}, voir le chapitre 5.5). Combien de générations de clonalité seraient nécessaires jusqu'à ce que l'érosion clonale (Vanoverbeke & De Meester 2010) puisse déjà exister à cause de la dérive génétique seule ? En outre, les premiers résultats présentés dans l'article III ne prennent pas encore la dynamique caractéristique des

populations d'organismes partiellement clonaux en compte (voir chapitre 6.2). En établissant la parthénogenèse cyclique comme un système de reproduction à son propre intérêt, et également distinct de l'asexualité partielle acyclique, nous espérons que nos résultats vont augmenter sa prise en compte par les biologistes de terrain (stratégie d'échantillonnage) et l'intérêt des théoriciens, conduisant à plus de développement dans ce domaine.

Article III Effets des cycles de vie complexes sur la diversité génétique : le cas de la parthénogenèse cyclique

Sommaire de l'article

Les modèles neutres de la diversité génétique des populations en espèces avec des cycles de vie complexes sont souvent difficiles à anticiper. La parthénogenèse cyclique, caractérisant les organismes qui présentent plusieurs cycles de reproduction clonale suivie par un événement sexuel, est un tel cycle de vie. Plusieurs espèces, y compris les ravageurs des cultures (pucerons), les parasites humains (trématodes) ou des modèles en sciences de l'évolution (daphnies), pratiquent la parthénogenèse cyclique. Il est donc essentiel de comprendre l'impact d'un tel cycle de vie sur la diversité génétique des populations. Nous proposons un modèle de chaîne de Markov de la parthénogenèse cyclique permettant d'analyser les distributions exactes du F_{IS} sous différents niveaux de clonalité. Notre analyse montre tout d'abord qu'un écart de la sexualité exclusive est observé après seulement quelques générations de clonalité, même si un grand nombre de générations de clones ne suffit pas pour la distribution de l'F_{IS} exacte à converger vers des résultats à pleine clonalité. Puis, pour les nombres petits et modérés de générations clonales, l'événement sexuel de la parthénogenèse cyclique réinitialise la population à l'égard des prévisions à partir de la sexualité exclusive : mais ce n'est pas le cas lorsque le nombre de générations de clones dans le cycle précédent a été suffisant pour fixer une proportion importante des hétérozygotes dans la population. Enfin, la clonalité partielle acyclique, correspondant au cycle de vie où une proportion fixe des individus est produite par clonage à chaque génération, ne donne pas les mêmes effets sur la diversité génétique que la parthénogenèse cyclique par rapport au même taux de clonalité. Des simulations individus-centrées de populations plus grandes ont confirmé les résultats obtenus par notre modèle de chaîne de Markov. Cette étude fournit la première étape vers un outil d'inférence permettant de quantifier le niveau de clonalité chez les espèces qui se servent de la parthénogenèse cyclique.

Article III *Effects of complex life-cycle on genetic diversity: The case of cyclical parthenogenesis*

Romuald Rouger[1], Katja Reichel[1], Florent Malrieu[2], Jean-Pierre Masson[1], Solenn Stoeckel[1]*

[1] *INRA, UMR1349 Institute for Genetics, Environment and Plant Protection, F-35650, Le Rheu, France*
[2] *Laboratoire de Mathématiques et Physique Théorique (UMR CNRS 7350) & Fédération Denis Poisson (FR CNRS 2964), Université François Rabelais, Parc de Grandmont, F-37200 Tours, France*

09/2015 Heredity, currently under review

Abstract

Neutral patterns of population genetic diversity for species showing complex life-cycles are often difficult to anticipate. Cyclical parthenogenesis, characterizing organisms displaying several rounds of clonal reproduction followed by a sexual event, is one such life-cycle. Several species, including crop pests (aphids), human parasites (trematodes) or models in evolutionary sciences (*Daphnia*), are cyclical parthenogens. It is therefore crucial to understand the impact of such a life-cycle on population genetic diversity. We propose a Markov chain model of cyclical parthenogenesis permitting to analyse exact distributions of F_{IS} under various levels of clonality. Our analysis firstly demonstrates that departures from full sexuality are observed after only few generations of clonality, yet a high number of clonal generations is not enough for the exact F_{IS} distribution to converge towards results under full clonality. Secondly, for small to moderate numbers of clonal generations, the sexual event of cyclical parthenogenesis resets the population towards predictions under full sexuality; but not when the number of clonal generations in the preceding cycle was large enough to fix a substantial proportion of heterozygotes in the population. Finally, acyclic partial clonality, corresponding to the life-cycle where a fixed proportion of individuals reproduce clonally within each generation, does not yield the same effects on genetic diversity than cyclical parthenogenesis when compared at the same ratio of clonality. Individual-based simulations of larger populations confirmed the results obtained by our Markov chain model. This study provides the first step towards an inference tool permitting to quantify the level of clonality in species displaying cyclical parthenogenesis.

Keywords cyclical parthenogenesis, Markov chains, F_{IS} distribution, *de Finetti* diagrams

Introduction

For decades, population genetic models using idealised populations have proven their efficiency to describe how genetic diversity should be distributed in natural populations according to a given range of assumptions (e.g. Hardy-Weinberg principle, Wright-Fisher model). When matching empirical data, such an approach has the great advantage of reducing the complexity of a biological system to a much more convenient approximation obeying to a known and fixed range of parameters (Hamilton 2009). Conversely, if deviations are observed, models may be refined by relaxing one or more assumptions in order to accurately describe the population genetics of the organism under scrutiny (e.g. mutation level, mating pattern).

Strict sexuality is a very common feature of idealised model populations; however, clonality is also a widespread mode of reproduction across all kinds of organisms (de Meeûs et al. 2007). Numerous field observations of intra-population genetic diversity in organisms using clonal reproduction showed strong deviations from predictions given by strictly sexual models (Ellstrand & Roose 1987, Delmotte et al. 2002, Papura et al. 2003, Stoeckel et al. 2006, Kanbe & Akimoto 2009, Allen & Lynch 2012, Aradottir et al. 2012). The main observed effects of clonal reproduction can be summarized as (1) a decrease in genotypic diversity, (2) an excess of heterozygotes resulting in strongly negative F_{IS}, and (3) an increased linkage disequilibrium due to the non-independent segregation of alleles between loci (Halkett et al. 2005).

Mathematical models and simulations relaxing the assumption of strict sexuality were developed in order to understand the effects of clonality on genetic diversity (Marshall & Weir 1979, Balloux et al. 2003, Bengtsson 2003, de Meeûs & Balloux 2004, de Meeûs & Balloux 2005, Prugnolle et al. 2005b). Besides giving a precise description of the case of full clonality, these models agreed that the effect of clonality on parameters of genetic diversity is difficult to distinguish from strict sexuality for moderate to intermediate levels of clonality (de Meeûs et al. 2006). For instance, only a small amount of sexuality in a mainly clonal population maintains a high level of genotypic diversity (Bengtsson 2003). Similarly, mean F_{IS} values obtained in mainly clonal populations are nearly indistinguishable from mean F_{IS} value obtained under panmixia (Balloux et al. 2003). Concerning linkage disequilibrium, the analytical exploration of models investigating the effects of clonality is complex (de Meeûs & Balloux 2004). Nevertheless, individual-based simulations highlighted the incoherent behaviour of several linkage disequilibrium estimators in response to increasing level of clonality (de Meeûs & Balloux 2004). Without finer predictions, the use of genetic diversity estimators in order to infer the level of clonality is therefore limited.

The abovementioned models report the mean value for a parameter of interest (e.g. F_{IS}); however, they fail to predict its full distribution, which would be a necessary prerequisite for any inferences from field data. Up to now, the probabilistic distribution could only be estimated using individual-based simulations, impeding the formulation of exact predictions (Balloux et al. 2003). Recently, Stoeckel & Masson (2014) proposed a stochastic model that permits an exact description of the full probabilistic distribution of F_{IS} in partially clonal organisms. This approach confirmed the small, and nearly indistinguishable,

difference existing between moderate levels of clonality and strict sexuality concerning mean F_{IS} values. Additionally, it permitted to highlight notable effects of moderate levels of clonality on both the probability of positive F_{IS} and the dynamic of F_{IS} from one generation to the next.

To tackle the effects of partial clonality on genetic diversity, most of these models were designed to fit populations for which clonal and sexual reproduction co-occur in time. This type of life-cycle, thereafter referred to as *Acyclic Partial Clonality* (*APC*), is common in many plants for instance (Vallejo-Marín et al. 2010). However, many species, especially in the animal kingdom, use a variation of this life-cycle, conventionally called *Cyclical Parthenogenesis (CP)*. In *CP*, clonal and sexual reproduction alternate in time with one to many generations of clonality followed by an event of sexual reproduction (figure 1). Analysing how *CP* impacts intra-population genetic diversity is particularly relevant given that crop pests (e.g. aphids), human parasites (e.g. trematodes) and classical biological models in evolution (e.g. *Daphnia*) are found among the organisms performing this life-cycle. To date, models studying *CP* primarily focused on detecting the intra-population genetic effect of both migration and variance in reproductive success (Prugnolle et al. 2005a, b) or on the amount of genetic differentiation between populations of cyclical parthenogens (Berg & Lascoux 2000). Complementarily, simulations on a limited number of neutral markers showed that the parthenogenetic phase reduces clonal diversity within populations and has an impact on mean F_{IS} when the number of parthenogenetic generations is high enough (Vanoverbeke & De Meester 2010). However, no model was designed to describe the exact probabilistic distribution of intra-population genetic diversity depending on the level of clonality in *CP*.

In this paper, we aim to investigate the population genetic effects of various levels of clonality in *CP* using an adaptation of the stochastic model of (Stoeckel & Masson 2014). Firstly, we describe how genetic diversity is affected by the number of clonal generations in a cycle; our predictions are that increasing the number of clonal generations in a cycle will amplify departures from the full sexuality scenario while increasing convergence towards results under full clonality. Secondly, we quantify how genetic diversity seasonally varies in *CP*, the sexual event being supposed to reset genotype frequencies in the entire population to Hardy-Weinberg proportions. Thirdly, as *APC* model outputs are often used to discuss results obtained for cyclical parthenogens, probability distributions of genetic diversity under *APC* and *CP* for similar levels of clonality are compared. Finally, we test the level of clonality needed to empirically tell apart datasets produced under *CP* from full sexuality, full clonality or *APC*.

Methods

The model used in this study is a strict adaptation of the mathematical model developed by Stoeckel & Masson (2014). This model is based on a biallelic system (*A* and *a*) in a population of N diploid individuals for which genotypic frequencies rather than allele frequencies are computed. The number of individuals of each possible genotype are $r_{aa} \in \mathbb{N}$, $r_{Aa} \in \mathbb{N}$ and $r_{AA} \in \mathbb{N}$ respectively. Their frequency at a time t is therefore $p_{ij}^{t} = r_{ij}/N$ where i and j are alleles *A* or *a*. At each time step, genotypic frequencies are only modified

by the action of reciprocal mutation between the two alleles (occurring at a rate μ), genetic drift and reproductive mode. In *APC*, clonal reproduction occurs for each time step at a rate c. In *CP*, each time step either represents an event of strict clonal reproduction ($c = 1$) or an event of strict sexual reproduction ($c = 0$). The number of successive clonal generations in a cycle is n_{clonal} and the number of sexual generations, n_{sex}, is set to 1 as observed in most biological systems using *CP* (figure 1). Panmixia is assumed during sexuality.

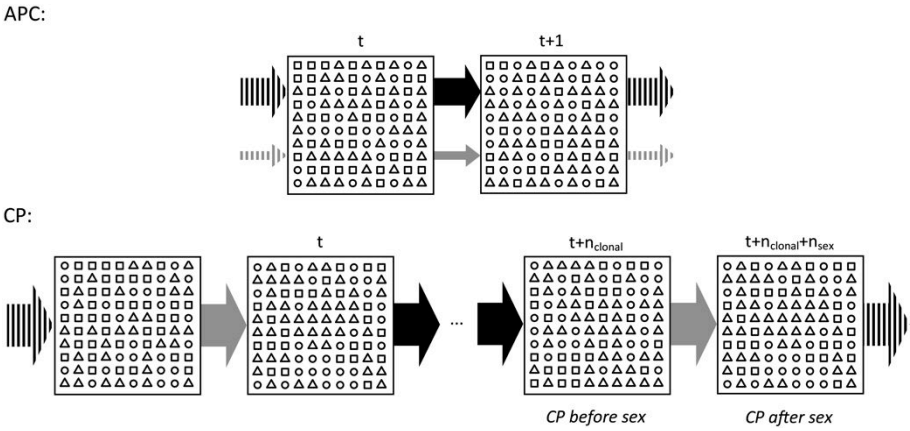

Figure 1. *Description of* Acyclic Partial Clonality (*APC*) *and* Cyclical Parthenogenesis (*CP*). At t the population consists of 100 individuals of genotype *AA* (circle), *Aa* (square) and *aa* (triangle). In *APC*, a fraction c of the population reproduce clonally (black arrow) and a fraction $1 - c$ reproduce sexually (grey arrow). In *CP* at t, the entire population undergo n_{clonal} events of clonal reproduction before a single event of sexual reproduction. *CP before sex* and *CP after sex* indicate the points at which we calculate the distribution of genetic diversity in the system.

The genotypic frequencies at $t + 1$ as functions of genotypic frequencies at t under clonal (p_{ij}^{t+1}) and sexual (q_{ij}^{t+1}) reproduction were given, respectively, by equation 1 and 2 of Stoeckel & Masson (2014; see also supplementary information 1).

Transition matrices

In *APC*, the overall genotypic frequencies at $t + 1$ (π_{ij}^{t+1}) are functions of p_{ij}^{t+1} and q_{ij}^{t+1} given the clonality rate c:

$$\pi_{ij}^{t+1} = c\, p_{ij}^{t+1} + (1 - c)\, q_{ij}^{t+1}$$

(1; Stoeckel & Masson 2014)

The transition probability from each current genotypic state (r_{aa}, r_{Aa}, r_{AA}) to each next state (s_{aa}, s_{Aa}, s_{AA}) is therefore calculated using the multinomial expression for every combination of two states:

$$p[(s_{aa}, s_{Aa}, s_{AA})|(r_{aa}, r_{Aa}, r_{AA})] = \frac{N!}{s_{aa}!\, s_{Aa}!\, s_{AA}!} (\pi_{aa}^{t+1})^{s_{aa}} (\pi_{Aa}^{t+1})^{s_{Aa}} (\pi_{AA}^{t+1})^{sAA}$$

(2; Stoeckel & Masson 2014)

Knowing the probability distribution of all previous states in a column vector $P^t(r_{aa}, r_{Aa}, r_{AA})$, the probability distribution of the next states $P^{t+1}(s_{aa}, s_{Aa}, s_{AA})$ can be written as the recurrence relation:

$$P^{t+1}(s_{aa}, s_{Aa}, s_{AA}) = P_{APC} \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

(3; Stoeckel & Masson 2014)

where P_{APC} is the transition matrix from previous to next states. Obviously, this matrix must be orientated so that the sum of each column is

$$\sum_{\substack{i=(s_{aa},s_{Aa},s_{AA}) \\ (s_{aa},s_{Aa},s_{AA})=N}} p[i|(r_{aa}, r_{Aa}, r_{AA})] = 1.$$

In *CP*, the probabilities of transitions from previous to next states were computed separately for sexual (p_{sex}) and clonal (p_{clonal}) modes of reproduction:

$$p_{sex}[(s_{aa}, s_{Aa}, s_{AA})|(r_{aa}, r_{Aa}, r_{AA})] = \frac{N!}{s_{aa}!\, s_{Aa}!\, s_{AA}!}(q_{aa}^{t+1})^{s_{aa}}(q_{Aa}^{t+1})^{s_{Aa}}(q_{AA}^{t+1})^{sAA}$$

$$p_{clonal}[(s_{aa}, s_{Aa}, s_{AA})|(r_{aa}, r_{Aa}, r_{AA})] = \frac{N!}{s_{aa}!\, s_{Aa}!\, s_{AA}!}(p_{aa}^{t+1})^{s_{aa}}(p_{Aa}^{t+1})^{s_{Aa}}(p_{AA}^{t+1})^{sAA}$$

(4)

Recurrence equations were also expressed for each mating system:

$$P^{t+1}(s_{aa}, s_{Aa}, s_{AA}) = P_{sex} \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

$$P^{t+1}(s_{aa}, s_{Aa}, s_{AA}) = P_{clonal} \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

(5)

While only one sexual event happens for each cycle of *CP*, n_{clonal} parthenogenetic generations occur. The probability distribution of the next states after n_{clonal} parthenogenetic generations $P^{t+n_{clonal}}(s_{aa}, s_{Aa}, s_{AA})$ can be inferred from the recurrence equation:

$$P^{t+n_{clonal}}(s_{aa}, s_{Aa}, s_{AA}) = (P_{clonal})^{n_{clonal}} \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

(6)

Finally, the probability distribution of next states after n_{clonal} parthenogenetic generations followed by one sexual generation ($n_{sex} = 1$) is:

$$P^{t+n_{clonal}+n_{sex}}(s_{aa}, s_{Aa}, s_{AA}) = [P_{sex} \cdot (P_{clonal})^{n_{clonal}}] \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

(7)

Similarly, the probability distribution of next states after one sexual generation followed by n_{clonal} parthenogenetic generations is:

$$P^{t+n_{sex}+n_{clonal}}(s_{aa}, s_{Aa}, s_{AA}) = [(P_{clonal})^{n_{clonal}} \cdot P_{sex}] \cdot P^t(r_{aa}, r_{Aa}, r_{AA})$$

(8)

A graphical explanation of the ordering of matrix multiplication in equation 7 and 8 is given in supplementary information 2.

The transition matrix in equation (7), $P_{CP\ after\ sex} = [P_{sex} \cdot (P_{clonal})^{n_{clonal}}]$, therefore permits to study genotypic frequencies of populations considered just after the sexual phase of *CP* (*CP after sex*). Similarly, the transition matrix in equation (8), $P_{CP\ before\ sex} = [(P_{clonal})^{n_{clonal}} \cdot P_{sex}]$, permits to study genotypic frequencies of populations considered at the end of their clonal phase (i.e. before the sexual event, *CP before sex*; figure 1).

Through generations (i.e. when $t \to \infty$), the probability distribution of next states $P^t(s_{aa}, s_{Aa}, s_{AA})$ will converge towards a *s*tationary *p*robability *d*istribution (*spd* hereafter) of genotypic states. As all generated transition matrices belong to irreducible and ergodic Markov chains, the *spd* of genotypic states is given by the eigenvector corresponding to the largest eigenvalue as stated by the Perron-Frobenius theorem (Li & Schneider 2002). The transition matrices and resulting exact *spd* were calculated using Python 2.7 (Van Rossum 2007), NumPy 1.9.1 (Van der Walt et al. 2011) and SciPy 0.15.0 (Oliphant 2007).

Parameter sets

The sets of parameters to be used for the calculation of each *CP* transition matrix were selected based on life-cycle descriptions of common cyclical parthenogens. The number of clonal generations in each cycle was set to $n_{clonal} \in \{1;\ 9;\ 99;\ 999\}$ permitting to study all cases from short cycles comprising only a few clonal events (e.g. Cynipidae, some aphid lineages) to long cycles comprising a large number of clonal events (e.g. assumed in permanent populations of cladocerans or monogonont rotifers). Comparisons of *CP* with *APC* are based on *APC* transition matrices with a rate of clonality $c = n_{clonal}/(n_{sex} + n_{clonal}) \in \{0.5; 0.9; 0.99; 0.999\}$. Additionally, transition matrices for full sexuality $(c = 0)$ and full clonality $(c = 1)$ were computed. The mutation rate was set to $\mu = 10^{-6}$ and the population size to $N = 200$.

Genetic diversity

The probability of fixation (i.e. the probability of fixing genotypes *aa* or *AA*) as well as the probability of heterozygote fixation (i.e. the probability of fixing genotype *Aa*) were obtained from the *spd* of each scenario. Monomorphic markers being non-informative in empirical population genetics, states where an allele is fixed in the population were therefore removed from the *spd* in subsequent analyses. The stationary probability of each genotypic state was rescaled so the vectors of stationary probability sum to 1.

Given the predicted response of the inbreeding coefficient to clonality, F_{IS} was calculated for each genotypic state following Rousset (2002):

$$F_{IS} = \frac{F - \theta}{1 - \theta}$$

where F is the average allelic identity within individuals in the population and θ is the average allelic identity between a pair of individuals in the population. The exact F_{IS} distribution based on the *spd* of genotypic states was inferred for each scenario. The first four moments (i.e. mean, variance, skewness and kurtosis) of each exact F_{IS} distribution were calculated together with the probability of getting positive F_{IS} (table 1). For the purpose of visualisation, the exact F_{IS} distribution was approximated using weighted kernel

density estimation, where each genotypic state represents a F_{IS} observation weighted by its exact stationary probability. This weighted kernel density estimation used a Gaussian kernel function with a bandwidth set to 0.05. The F_{IS} density distribution was then estimated at 512 equally spaced points across the F_{IS} range (i.e. $[-1,1]$) using the R function *density* (R Core Team 2013).

Synthetic parameters such as F_{IS} are almost always used in population genetics studies to describe a genetic dataset. Although convenient to use, summarizing the genetic information using such indices causes a loss of the information encompassed in a full dataset. In our case, the *spd* of genotypic states comprises the total genetic information available. The *spd* of genotypic states was therefore visualised using "*de Finetti* diagrams" (de Finetti 1927). The stationary probability of each genotypic state was represented by a colour scale, using the R package *ggtern* (Hamilton 2015). Pairwise divergence between scenarios were quantified using the Jensen-Shannon measure of divergence between pairs of *spd* (D_{JS}), where each genotypic state represented a discrete class.

Discrimination between scenarios

Full F_{IS} distributions, *de Finetti* diagrams and measures of divergence are all calculated based on the exact *spd* of genotypic states. Observing this *spd* in a biological system implies considering an infinite number of markers. Going one step further than only describing theoretical exact distribution of genetic diversity, we aimed to assess the level of clonality needed to correctly discriminate scenarios.

Firstly, we tested whether discrimination between scenarios is possible given a fair number of loci. For each scenario, we sampled 10 000 genotypic states based on their stationary probability. Empirically, this procedure corresponds to the screen of 10 000 independent biallelic SNP in a "stationary population". 10 000 loci are not enough to get a fine enough approximation of the exact *spd* of genotypic states; calculating divergence between scenario using such an estimation leads to an overestimation of D_{JS}. Kolmogorov Smirnov statistic (D_{KS}) based on F_{IS} was therefore preferred over D_{JS} (based on genotypic states) in order to compare scenarios. To this aim, the F_{IS} value of each sampled genotypic state was used to build the *e*mpirical *c*umulative *d*istribution *f*unction (*ecdf* hereafter) of F_{IS} for each scenario. Significant differences between pairs of scenarios were checked using a two-sample Kolmogorov-Smirnov test, as implemented in the R library *dgof* (Arnold & Emerson 2011).

Secondly, we tested whether discrimination between scenarios was still possible given a greater population size. Unfortunately, the number of possible genotypic states increases exponentially when population gets larger; the calculation of the exact *spd* then becomes rapidly intractable for any computer (Reichel et al. submitted a). Individual-based simulations following our model were therefore launched to approximate the full F_{IS} distribution of large population sizes. They were based on a population of 10 000 organisms using either *CP* or *APC*. The analysed parameter sets were identical to those for the exact distribution model, with $\mu = 10^{-6}$, $n_{clonal} \in \{1;\ 9;\ 99;\ 999\}$ in *CP*, $c \in \{0{,}5;\ 0{,}9;\ 0{,}99;\ 0{,}999\}$ in *APC*, $c = 0$ for full sexuality and $c = 1$ for full clonality. Each

simulation comprised a single biallelic locus and was repeated 10 000 times. Initial alleles and genotypes were randomly assigned for each individual and repetition. F_{IS} values for each locus were calculated after a burn-in period of 10 000 generations. For *APC*, F_{IS} values were acquired at generations 10 001. For *CP*, they were collected before and after the next sexual event (e.g. generation 10 009 and 10 010 if $n_{clonal} = 9$). Significant differences between scenarios were checked by a two-sample Kolmogorov-Smirnov test as described earlier. A Kernel density estimation based on all repetitions (i.e. 10 000 F_{IS} observations) was used to visualise the approximate F_{IS} distribution for each simulation.

Results

Effect of clonality during cyclical parthenogenesis (CP)

As expected for populations of finite size, modelled by discrete genotypic states, the mean F_{IS} value in the full sexuality scenario is slightly negative (see supplementary information 4 for an explanation), and the probability of getting negative F_{IS} is higher than the probability of getting positive F_{IS} (table 1). Deviations from this baseline scenario were

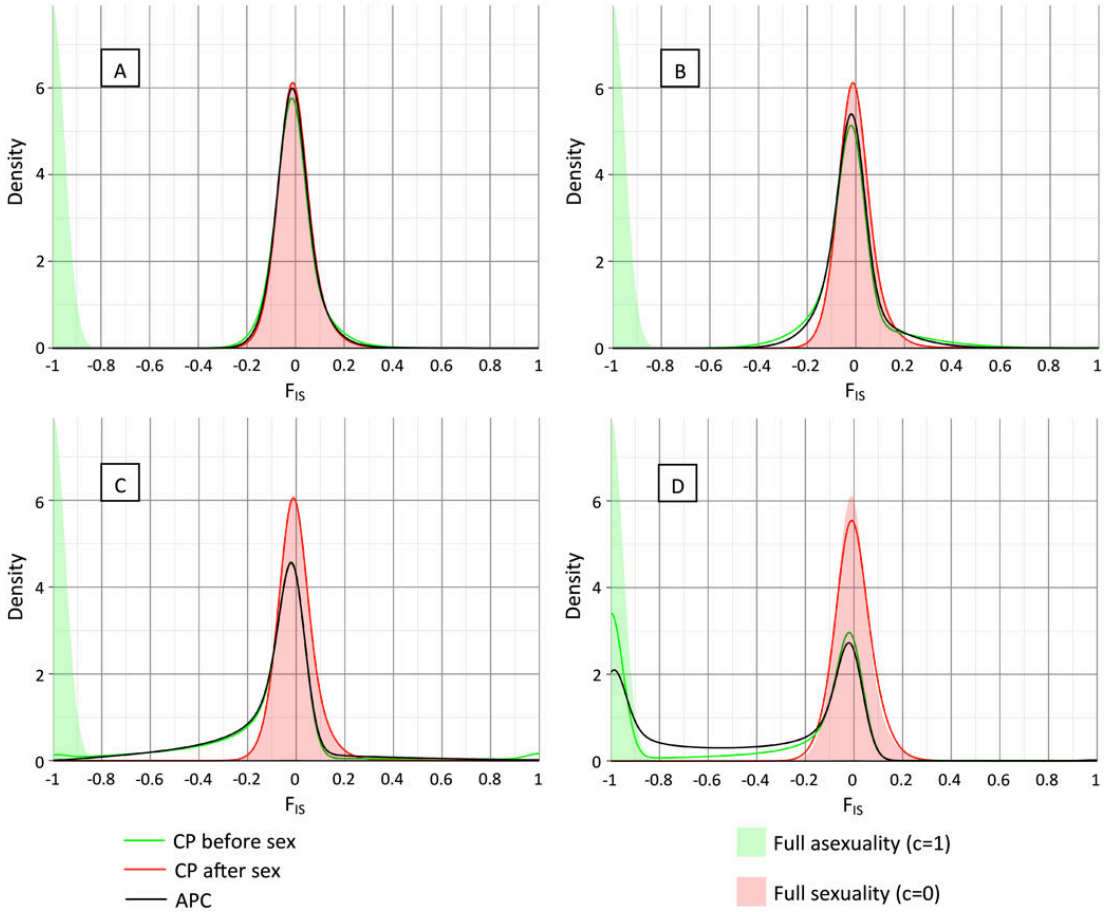

Figure 2. *Weighted kernel density estimation of full F_{IS} probability density function* in function of levels and modes of clonality ($N = 200, \mu = 10^{-6}$). A: $n_{clonal} = 1$ in *CP* and $c = 0.5$ in *APC*; B: $n_{clonal} = 9$ in *CP* and $c = 0.9$ in *APC*; C: $n_{clonal} = 99$ in *CP* and $c = 0.99$ in *APC*; D: $n_{clonal} = 999$ in *CP* and $c = 0.999$ in *APC*.

Table 1. *Descriptive statistics of F_{IS} distribution in function of levels and modes of clonality* ($N = 200$, $\mu = 10^{-6}$). In *APC*, c is the fraction of individuals in the population reproducing clonally, in *CP*, c is the proportion of clonal generation in a cycle ($c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$); $p_{F_{IS}>0}$ is the probability of observing positive F_{IS}; p_{fixation} is the probability of fixing one of the allele in the population and p_{het} is the probability of fixing heterozygotes in the population.

c	Mode	$p_{F_{IS}>0}$	p_{fixation}	p_{het}	Mean	Variance	Skewness	Kurtosis
0	Full sexuality	0.23906	0.99468	0	-0.00251	0.00371	2.78976	28.45313
0.5	*APC*	0.22437	0.99467	0	-0.00434	0.00432	2.35436	21.73083
	CP after sex	0.23910	0.99468	0	-0.00251	0.00371	2.78926	28.44413
	CP before sex	0.21360	0.99468	0	-0.00439	0.00627	2.42837	20.70856
0.9	*APC*	0.16736	0.99461	0	-0.01518	0.01083	1.51667	12.44997
	CP after sex	0.23994	0.99464	0	-0.00251	0.00372	2.77488	28.24919
	CP before sex	0.15258	0.99464	0	-0.01580	0.02008	1.68986	12.15171
0.99	*APC*	0.07512	0.99418	0	-0.08887	0.05072	-0.14673	7.55953
	CP after sex	0.24989	0.99439	0	-0.00251	0.00378	2.60582	26.45150
	CP before sex	0.06317	0.99439	0.00007	-0.09135	0.08699	0.32198	7.75835
0.999	*APC*	0.01109	0.99055	0.00153	-0.44071	0.18498	-0.08717	1.72032
	CP after sex	0.32637	0.99163	0	-0.00251	0.00419	1.44645	15.60082
	CP before sex	0.00136	0.99163	0.00356	-0.49267	0.21512	-0.06478	1.21910
1	Full clonality	0.00000	0.49762	0.49762	-0.99471	0.00447	13.49324	188.08319

visible in the results for *CP before sex* even in short parthenogenetic cycles (figure 2, green distributions). The probability of getting positive $\mathrm{F_{IS}}$ started decreasing from $\mathrm{n_{clonal}} = 1$ and reached its smallest value for $\mathrm{n_{clonal}} = 999$ (table 1). As a result, the mean $\mathrm{F_{IS}}$ also moved towards negative values. Contrastingly, the mode of the $\mathrm{F_{IS}}$ distribution stayed unchanged for small to intermediate numbers of clonal generations ($\mathrm{n_{clonal}} \in \{1; 9; 99\}$), therefore increasing the variance of the distribution (figure 2, table 1). In these scenarios, the skewness also decreased, driven by the increasing weight of the left tail of the distribution (figure 2 A, B and C). This negative tail can in fact be seen as the proportion of trajectories departing from $\mathrm{F_{IS}} \approx 0$ towards heterozygous fixation in the population because of genetic drift affecting whole genotypes rather than single alleles during the parthenogenetic phase (figure 2). At intermediate levels of clonality (i.e. $\mathrm{n_{clonal}} \in \{9; 99\}$), these trajectories rarely have the time to reach heterozygous fixation, thus increasing the probability of transient states (i.e. negative tails, figure 2) without increasing the probability of heterozygous fixation (table 1). On the other hand, the increasing probability of getting negative $\mathrm{F_{IS}}$ value was inversely correlated to the probability density around the mode of the distribution. This flattening of the distribution was quantified by the decrease of the kurtosis as the number of clonal generations in *CP* increased (table 1).

Although the probability of heterozygote excess increased for intermediate numbers of parthenogenetic generations, cycles with long periods of clonality are needed for genetic drift to fix the heterozygote genotype in the population. At $n_{\text{clonal}} = 999$, most trajectories departing from Hardy-Weinberg towards negative F_{IS} values reach heterozygous fixation, thus increasing the probability of getting F_{IS} values equal to -1 while decreasing the

probability of transient states (figure 2, table 1). A characteristic bimodal distribution was then observed (figure 2). The variance of this distribution stayed high, the mean F_{IS} value being comprised between the two modes of the distribution (table 1). Similarly, the skewness of the distribution remained negative, as the highest density of the distribution was still located around $F_{IS} \approx 0$. Finally, bimodality also kept the kurtosis at low values.

Despite the increase of the probability of fixing heterozygotes when the number of parthenogenetic generations reaches large values, this probability is still relatively small in comparison to the probability of homozygous fixation (table 1). Contrarily to the states of homozygous fixation (i.e. $r_{AA} = 200$ or $r_{aa} = 200$), the state of heterozygous fixation (i.e. $r_{Aa} = 200$) is regularly broken by sexual reproduction. The stationary probability of fixing homozygotes by genetic drift is therefore still largely dominant. Once sexual reproduction is totally removed from the system (full clonality scenario), the state where all individuals are heterozygous, i.e. heterozygotes are fixed, becomes as stable as states where homozygotes are fixed; the probability of completely heterozygous loci becomes as high as that of completely homozygous loci (table 1). Consequently, all moments of the F_{IS} distribution change drastically, the mean F_{IS} comes close to the mode of the distribution, both located near – 1 (mean F_{IS} =– 0.99471), therefore causing the variance to be very small. The entire distribution is tightly grouped around the minimum F_{IS} value, and therefore right-skewed, causing high value of skewness and kurtosis.

The representation of the *spd* using *de Finetti* diagrams complemented the description of the F_{IS} distributions (figure 3). Graphically, *de Finetti* diagrams confirmed that deviations from full sexuality scenario were visible in *CP before sex* even in short parthenogenetic cycles. Measures of divergence between *spd* quantitatively validated the deviation (table 2). When $n_{\text{clonal}} = 1$, all highly probable states are grouped along the F_{IS} isocline $F_{IS} = 0$ (figure 3, *CP before sex*, $c = 0.5$); however, the spread of high probabilities around this isocline was larger than under full sexuality. The dispersal of the highest probabilities was amplified when $n_{\text{clonal}} = 9$ (figure 3, *CP before sex*, $c = 0.9$). For both $n_{\text{clonal}} = 1$ and $n_{\text{clonal}} = 9$, the stationary probabilities of all genotypic states along a given isocline were approximately of the same order of magnitude.

Table 2. *Jensen-Shannon measures of divergence* (D_{JS}) between pairs of exact *spd* of selected scenarios ($N = 200, \mu = 10^{-6}$).

Reference scenario	Test scenario	n_{clonal}			
		1	9	99	999
Full sexuality; $c = 0$	*CP before sex*	0.0235	0.1634	0.3523	0.5193
Full clonality; $c = 1$	*CP before sex*	0.9823	0.9786	0.9196	0.3460
Full sexuality; $c = 0$	*CP after sex*	<0.0001	<0.0001	0.0013	0.1880
CP after sex	*CP before sex*	0.0235	0.1637	0.3634	0.6201
APC; $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP before sex*	0.0094	0.0316	0.0855	0.1449
APC; $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP after sex*	0.0038	0.0852	0.2894	0.6127

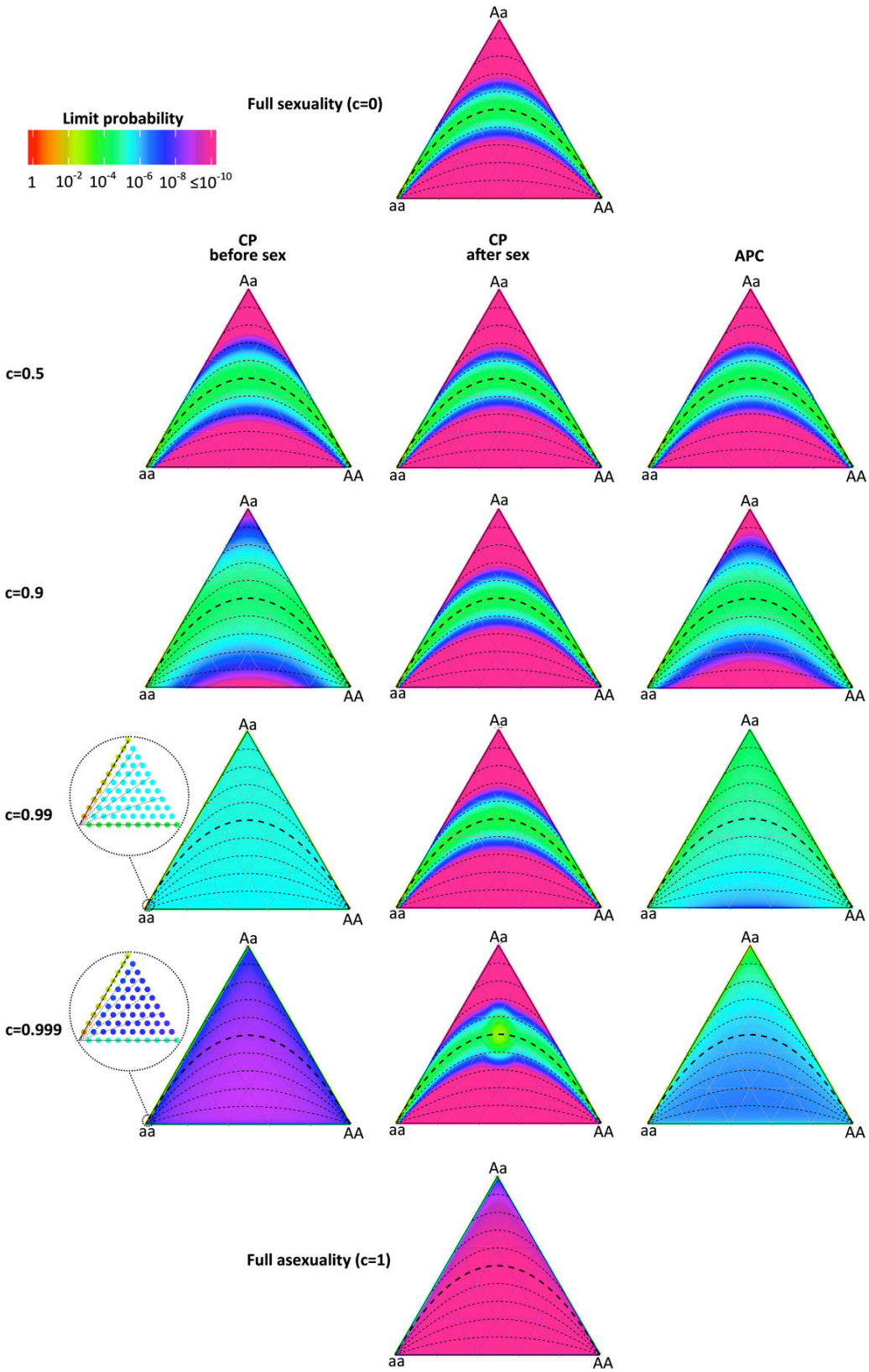

Figure 3. De Finetti *diagrams illustrating* spd *of genotypic states in function of modes and levels of clonality* ($N = 200$, $\mu = 10^{-6}$). In *APC*, *c* is the fraction of individuals in the population produced clonally, while in *CP*, *c* is the proportion of clonal generations in a cycle ($c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$). Clarification on how to read *de Finetti* diagrams can be found in supplementary information 3.

Strong differences to this pattern occurred when increasing the number of clonal generations in the cycle. As typified when zooming in on the corner of the diagram for $n_{\text{clonal}} = 99$, genetic drift "pulled" the highest stationary probabilities towards the loss of one genotype (i.e. the edge of the *de Finetti* diagram, figure 3, *CP before sex*, $c = 0.99$). Here, stationary probabilities of states along the same isoclines cannot be considered equiprobable. In fact, the F_{IS} distribution depicted in figure 2 relies entirely on the aggregated probabilities of the genotypic states for which one of the homozygote genotypes was lost. Interestingly, the stationary probabilities of each genotypic state not located on the edges of the diagram are small but approximately equal, irrespectively of their F_{IS} value. This overall pattern was strengthened when the number of clonal generations reaches extreme values. For $n_{\text{clonal}} = 999$, the stationary probabilities of genotypic states combining together at least one copy of each genotype (i.e. interior of the *de Finetti* diagram) were almost negligible.

Measures of divergence between the *spd* of full sexuality and *CP before sex* also indicated quick departure from the expectation for full sexuality ($D_{JS} = 0.0235$ for $n_{\text{clonal}} = 1$), and increased with the number of clonal generations in the cycle ($D_{JS} = 0.5193$ for $n_{\text{clonal}} = 999$) as expected. In fact, the *spd* of *CP before sex* slowly converges towards the *spd* of the full clonality scenario as the number of clonal generations increases in the cycle. However, the divergence was still high even when considering the longest cycle of *CP* in our analysis ($D_{JS} = 0.3460$ for $n_{\text{clonal}} = 999$).

Effect of seasonality in CP

The shape of the F_{IS} distribution radically changes just after the occurrence of the sexual event of *CP* (figure 2, red distributions). As expected, panmixia in the population resets the genetic diversity in the population close to the distribution observed under full sexuality for short to intermediate cycle length. When $n_{\text{clonal}} \in \{1,9,99\}$, F_{IS} distributions expected for completely sexual organisms are nearly undistinguishable from the F_{IS} distribution for cyclical parthenogens collected after the sexual event (figure 2). Differences only start to appear when the number of clonal generations preceding the sexual event is high ($n_{\text{clonal}} = 999$). In that case, the density around the mode of the distribution is smaller than in the full sexuality scenario and the distribution gets slightly more spread out. Both variance and kurtosis confirmed this visual interpretation but the most pronounced effect concerned the increased probability of getting positive F_{IS} (table 1). Interestingly, at any level of clonality considered, the probability of fixation stays the same after and before the sexual event of *CP* (table 1).

Measures of divergence failed to discriminate between *spd* produced by the full sexuality scenario and *CP after sex* for $n_{\text{clonal}} \in \{1; 9\}$ ($D_{JS} < 0.0001$ in both cases, table 2). For $n_{\text{clonal}} = 99$, the Jensen-Shannon index started to indicate divergence between the two distributions ($D_{JS} = 0.0013$, table 2). However, this divergence was very hard to notice graphically when looking at the corresponding *de Finetti* diagram (figure 3). At $n_{\text{clonal}} = 999$, divergence between the two *spd* is stronger ($D_{JS} = 0.1880$) and is also graphically noticeable on *de Finetti* diagrams (figure 3). The analysis of this diagram further permits to

explain the observed shape of the F_{IS} distribution described earlier. The probability of heterozygote fixation before the sexual event following 999 clonal generations is high; in consequence, the state where both alleles are equifrequent after sexual reproduction (i.e. $(r_{aa}, r_{Aa}, r_{AA}) = (50, 100, 50)$) gets a high stationary probability. Additionally, multinomial sampling (genetic drift) randomly increases the probability of neighbouring states (e.g. $(r_{aa}, r_{Aa}, r_{AA}) = (47, 101, 52)$) explaining the high probability region in the centre of the *de Finetti* diagram (figure 3) and the increased variance of F_{IS} (figure 2, table 1, *CP after sex*, $n_{\text{clonal}} = 999$).

At any level of clonality considered, the deviations from the full sexuality scenario obtained after vs. before sexual reproduction under *CP* are by no means comparable. Jensen-Shannon divergence between *spd* inferred just before and just after the sexual event of *CP* permitted to quantify this "seasonality effect" (table 2). For $n_{\text{clonal}} = \{1,9,99\}$, levels of divergence inferred between these two *spd* are very close to the divergence calculated between *CP before sex* and the full sexuality scenario. For $n_{\text{clonal}} = 999$, the divergence between *CP* before and after sex become higher than the divergence observed between *CP before sex* and the full sexuality scenario ($D_{JS} = 0.6201$ and 0.5193 respectively).

Comparison between CP and APC

Differences between F_{IS} distributions produced under *APC* or *CP* depend on both the time of sampling in *CP* (after or before sex) and the number of clonal generations in the cycle. At $n_{\text{clonal}} = 1$ (or $c = 0.5$), the F_{IS} distribution under *APC* graphically represents an intermediate case between the two distributions obtained under *CP* after and before sex (figure 2, black distribution). For intermediate levels of clonality ($n_{\text{clonal}} \in \{9; 99\}$ or $c \in \{0.9; 0.99\}$) the F_{IS} distribution under *APC* is tightly related to the F_{IS} distribution under *CP before sex* (figure 2). Although variance of F_{IS} under *APC* is lower than under *CP before sex* (table 1), visual discrimination between the two distributions is difficult (figure 2). In contrast, strong differences between *APC* and *CP before sex* are noticeable when the level of clonality is high ($n_{\text{clonal}} = 999$ or $c = 0.999$, figure 2). Main dissimilarities concern the probability of getting positive F_{IS}, which is almost ten times lower in *CP before sex* than in *APC* (0.00136 and 0.01109 respectively), and the probability of fixing heterozygote genotypes in the population, which is higher in *CP before sex* than in *APC* (table 1).

Similarly to the F_{IS} distributions, *de Finetti* diagrams and Jensen-Shannon indices of divergence all indicated that *spd* produced under *APC* at $c = 0.5$ represent an in-between distribution between *CP before* and *after sex* when $n_{\text{clonal}} = 1$ (figure 3, table 2). For $n_{\text{clonal}} \in \{9; 99; 999\}$ (or $c \in \{9; 99; 999\}$), D_{JS} also validated that *spd* under *APC* were closer to *spd* under *CP before sex* than *spd* under *CP after sex* (table 2). However, the *de Finetti* diagrams showed more divergence between *spd* under *APC* and *CP* than anticipated based on the F_{IS} distributions. Visually, differences produced under *APC* and *CP before sex* appeared from $n_{\text{clonal}} = 9$ (or $c = 0.9$, figure 3). The increased variance of F_{IS} calculated in *CP before sex* was illustrated by the larger dispersion of high probabilities around the isocline $F_{IS} = 0$. At $n_{\text{clonal}} = 99$ (or $c = 0.99$), the *de Finetti* diagrams obtained under *APC* did not show all the highest probabilities completely "pulled" towards the edge of the diagram as previously demonstrated under *CP before sex*. In *APC*, genotypic states along the same negative F_{IS}

isocline got stationary probabilities of approximately the same order of magnitude. This last instance is very representative of the possibility of getting two almost identical F_{IS} probability distribution despite getting strong differences in terms of genetic/genotypic diversity in the population. A similar pattern occurred at $n_{\text{clonal}} = 999$ (or $c = 0.999$). Interestingly, the more homogeneous stationary probabilities along the same isocline made the probability of getting states located in the middle of the diagram more likely in *APC* than in *CP before sex*.

Discrimination between scenarios

At $N = 200$, sampling 10 000 independently evolved loci based on their stationary probability was enough to statistically discriminate F_{IS} distributions between scenarios (table 3).The F_{IS} *ecdf* obtained under *CP before sex* was significantly different from the *ecdf* of both the full sexuality and full clonality scenarios at any considered level of clonality. In contrast, a large number of clonal generations ($n_{\text{clonal}} = 999$) were needed to detect significant differences between F_{IS} *ecdf* obtained under *CP after sex* and the full sexuality scenario. By extension, F_{IS} *ecdf* of *CP* before and after sex were also found to be significantly different, as were the F_{IS} *ecdf* under *APC* and under *CP* for any period of sampling (before or after sex) and level of clonality. Interestingly, the Kolmogorov-Smirnov statistic (D_{KS}) between *APC* and *CP before sex* did not steadily increase with the number of clonal generations, as observed with the Jensen-Shannon index of divergence (table 2). However, correlation between the Kolmogorov-Smirnov statistic and Jensen-Shannon divergence was high ($R^2 = 0.96, p < 0.001$), indicating that D_{KS} between F_{IS} *ecdf* may be a good approximation of D_{JS} between exact *spd* of genotypic states.

Table 3. *Two-sample Kolmogorov-Smirnov test between empirical cumulative distribution functions (*ecdf*) of selected scenarios* for $N = 200$ and $\mu = 10^{-6}$. The *ecdf* were obtained by sampling 10 000 genotypic states in the *spd* of each scenario. D_{KS}: Kolmogorov-Smirnov statistic; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Reference scenario	Test scenario	n_{clonal}			
		1	9	99	999
Full sexuality $c = 0$	*CP before sex*	$D_{KS} = 0.0404$ ***	$D_{KS} = 0.1378$ ***	$D_{KS} = 0.2868$ ***	$D_{KS} = 0.5540$ ***
Full clonality $c = 1$	*CP before sex*	$D_{KS} = 0.9950$ ***	$D_{KS} = 0.9938$ ***	$D_{KS} = 0.9782$ ***	$D_{KS} = 0.5672$ ***
Full sexuality $c = 0$	*CP after sex*	$D_{KS} = 0.009$ $(p = 0.8127)$	$D_{KS} = 0.0150$ $(p = 0.2106)$	$D_{KS} = 0.0151$ $(p = 0.2043)$	$D_{KS} = 0.097$ ***
CP after sex	*CP before sex*	$D_{KS} = 0.0436$ ***	$D_{KS} = 0.1369$ ***	$D_{KS} = 0.2853$ ***	$D_{KS} = 0.5447$ ***
APC $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP before sex*	$D_{KS} = 0.0224$ *	$D_{KS} = 0.0313$ ***	$D_{KS} = 0.0236$ **	$D_{KS} = 0.2615$ ***
APC $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP after sex*	$D_{KS} = 0.0262$ **	$D_{KS} = 0.1189$ ***	$D_{KS} = 0.2927$ ***	$D_{KS} = 0.5706$ ***

F_{IS} distributions obtained from individual-based simulations at $N = 10\,000$ naturally showed some divergence from full F_{IS} distributions obtained at $N = 200$ (figure 4). This change is due to the randomizing effect of reciprocal mutations, whose number per generation increases with the number of individuals, thus impeding the fixation of a single genotype by genetic drift. The visual relationships between the distributions of different scenarios remain, however, the same as described for $N = 200$. Accordingly, two-sample Kolmogorov-Smirnov tests still permitted to detect significant differences between F_{IS} *ecdf* of different scenarios (table 4). Significant differences were even highlighted between scenarios of full sexuality and *CP after sex* from $n_{\text{clonal}} = 9$ although no differences between the density functions were graphically noticeable (figure 4).

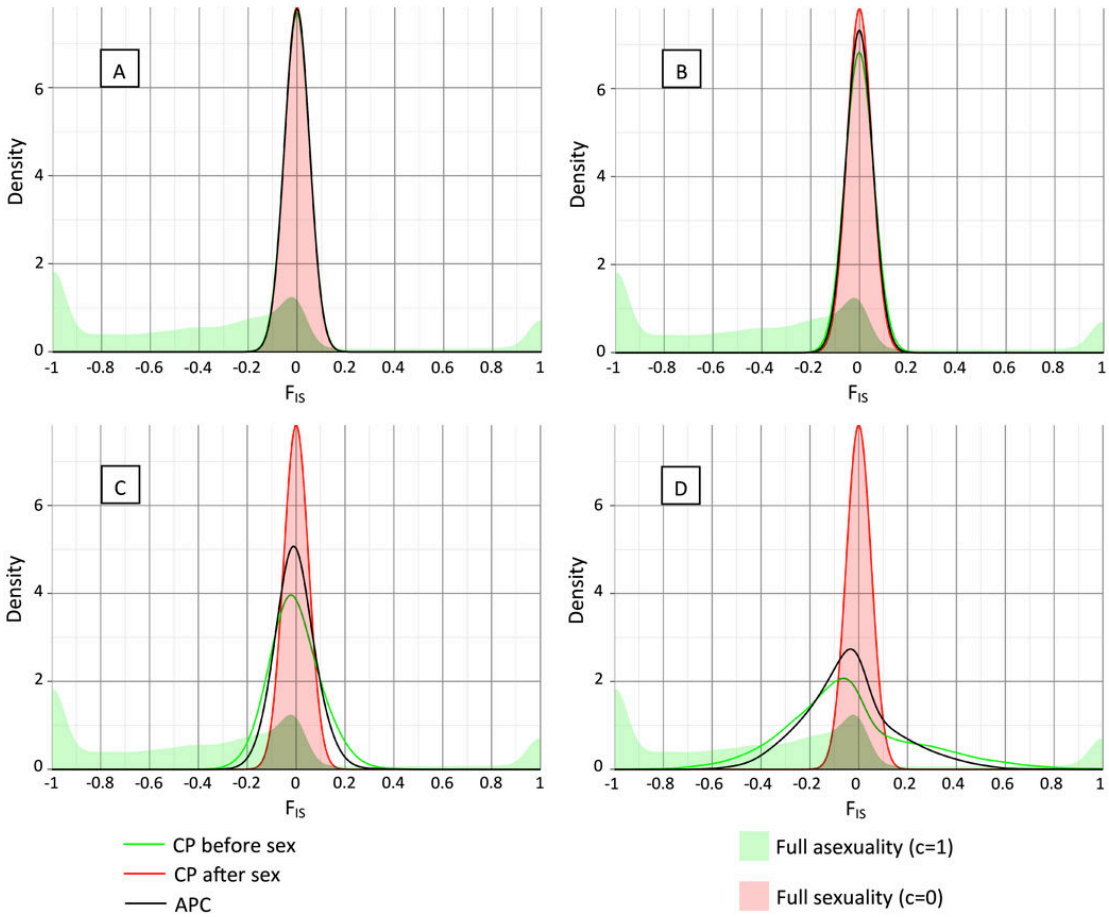

Figure 4. *Weighted kernel density estimation of F_{IS} based on 10 000 simulations*, in function of levels and modes of clonality ($N = 10\,000$, $\mu = 10^{-6}$). A: $n_{\text{clonal}} = 1$ in *CP* and $c = 0.5$ in *APC*; B: $n_{\text{clonal}} = 9$ in *CP* and $c = 0.9$ in *APC*; C: $n_{\text{clonal}} = 99$ in *CP* and $c = 0.99$ in *APC*; D: $n_{\text{clonal}} = 999$ in *CP* and $c = 0.999$ in *APC*.

Table 4. *Two-sample Kolmogorov-Smirnov test between empirical cumulative distribution functions (*ecdf*) of selected scenarios* for $N = 10\,000$ and $\mu = 10^{-6}$. The *ecdf* were obtained by simulating 10 000 F_{IS} trajectories. In *CP*, F_{IS} values were recorded after and before the next sexual event, following a burn-in period of 10 000 generations. For

APC, full sexuality and full clonality, F_{IS} values were recorded at generation 10 001. D_{KS}: Kolmogorov-Smirnov statistic; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Reference scenario	Test scenario	n_{clonal}			
		1	9	99	999
Full sexuality $c = 0$	*CP before sex*	$D_{KS} = 0.0821$ ***	$D_{KS} = 0.2551$ ***	$D_{KS} = 0.4409$ ***	$D_{KS} = 0.5914$ ***
Full clonality $c = 1$	*CP before sex*	$D_{KS} = 0.7141$ ***	$D_{KS} = 0.6787$ ***	$D_{KS} = 0.5800$ ***	$D_{KS} = 0.3981$ ***
Full sexuality $c = 0$	*CP after sex*	$D_{KS} = 0.0136$ ($p = 0.3443$)	$D_{KS} = 0.0321$ ***	$D_{KS} = 0.0654$ ***	$D_{KS} = 0.0678$ ***
CP after sex	*CP before sex*	$D_{KS} = 0.0773$ ***	$D_{KS} = 0.2501$ ***	$D_{KS} = 0.4406$ ***	$D_{KS} = 0.5921$ ***
APC $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP before sex*	$D_{KS} = 0.0558$ ***	$D_{KS} = 0.0857$ ***	$D_{KS} = 0.0932$ ***	$D_{KS} = 0.0882$ ***
APC $c = n_{\text{clonal}}/(n_{\text{clonal}} + n_{\text{sex}})$	*CP after sex*	$D_{KS} = 0.0307$ ***	$D_{KS} = 0.1835$ ***	$D_{KS} = 0.3804$ ***	$D_{KS} = 0.5463$ ***

Discussion

Several studies have quantified and formalised the effect of clonality on genetic diversity in species for which sexual and clonal reproduction co-occur in time (Marshall & Weir 1979, Orive 1993, Balloux et al. 2003, Bengtsson 2003, de Meeûs & Balloux 2004, de Meeûs & Balloux 2005, Stoeckel & Masson 2014). Despite many examples of *CP* in the wild, analytical models or simulations looking at the effect of this reproductive system on parameters of genetic diversity are rarer (Berg & Lascoux 2000, Prugnolle et al. 2005a, b; Vanoverbeke & De Meester 2010). The model presented here provides the first full probability distributions of genetic diversity depending on level of clonality in cyclical parthenogenesis.

Few generations of clonality cause departures from the assumption of full sexuality

Our model first confirms results obtained by previous investigators showing that low levels of clonality has little effect on mean F_{IS} value (Balloux et al. 2003). The novelty of our results lies in the fact that few generation of clonality, even when preceded by regular sexual events, are enough to substantially impact the full probability distribution of F_{IS} (*CP before sex*). The main effects of successive clonal generations can be summarized by a flattening of the distribution around $F_{IS} = 0$ together with a spread of the F_{IS} distribution towards negative values, the extremity of this negative tail eventually reaching $F_{IS} = -1$ (i.e. heterozygote fixation) and causing the distribution to become bimodal. This result stresses once again the importance of taking the time of sampling into account when analysing population genetic datasets of cyclical parthenogens (Berg & Lascoux 2000).

As intuited in previous studies (Pfrender & Lynch 2000, De Meester et al. 2006, Allen & Lynch 2012), our model confirms that a single bout of sexual reproduction following a low number of clonal generations rearranges genetic diversity towards predictions under full sexuality

(*CP after sex*). However, this pattern is not entirely established when the number of clonal generations preceding the sexual event is large enough for genetic drift to shift genotypic frequencies away from distributions under full sexuality. Such neutral deviations predicted under *CP* may result, for example, in erroneous identification of loci as "outliers" and their over-interpretation as being under selective pressure. This slight effect may be difficult to observe with a limited number of classical population genetic markers (e.g. microsatellites), but it is particularly relevant for the analysis of large scale population genomic data from cyclical parthenogens (Orsini et al. 2011, Routtu et al. 2014).

Long period of clonality is not equivalent to full clonality

Although already few generations of clonality cause departure from full sexuality scenario, long periods of clonality following rare sexual events do not yield a similar distribution as full clonality. Under full clonality, when the effects of genetic drift dominate over those of mutation (e.g. $N = 200$ and $\mu = 10^{-6}$), the F_{IS} probability distribution is centred around $F_{IS} = -1$ (heterozygote fixation). Oppositely, in *CP before sex*, the F_{IS} probability distribution at the end of the longest clonal phase investigated here ($n_{\text{clonal}} = 999$) is still distinctively bimodal around $F_{IS} = -1$ and $F_{IS} \approx 0$. Simulations at $N = 10\,000$ and $\mu = 10^{-6}$ also confirmed this difference, although both distributions looked different due to the change in the mutation/drift balance. The contrasting F_{IS} distributions under full clonality versus long cycle of *CP* provide an additional and seducing population genetic tool to test the absence of sex in natural or experimental populations of putatively ancient asexuals (Danchin et al. 2011).

CP and APC yield different distribution of genetic diversity

Probability distributions of F_{IS} under *CP after* or *before sex* are different from those observed under *APC*. Furthermore, the F_{IS} probability distribution obtained under *APC* does not represent an average situation between the two seasonal distributions of *CP*. The F_{IS} probability distribution under *APC* is closely linked to the distribution obtained under *CP before sex*. However, the flattening of the distribution and its spread towards negative F_{IS} values is quicker in *CP before sex,* where genetic drift only interacts with mutation, than in *APC* where low levels of sexual reproduction additionally reduce the effects of genetic drift (see also Reichel et al. submitted b).

Because of the absence of an explicit model describing the effect of clonality on genetic diversity in CP, many studies had to rely on predictions from models of genetic diversity under *APC* to discuss their *CP* data outputs (Halkett et al. 2005b, Vorburger 2006, Kanbe & Akimoto 2009, Allen & Lynch 2012). Especially designed to fit the *CP* lifecycle, our model permits to refine these predictions and is particularly suited to test more precise hypothesis about genetic diversity in cyclical parthenogens.

Discrimination between scenarios

Two-sample Kolmogorov-Smirnov tests permitted to discriminate F_{IS} *ecdf* obtained between most of the scenarios at both $N = 200$ and $N = 10\,000$ (full sexuality, full asexuality, *CP before sex*, *CP after sex* and *APC*). Interestingly, significant divergence between

F_{IS} *ecdf* produced by scenarios of full sexuality and *CP after sex* were only detected at $n_{\text{clonal}} = 999$ for $N = 200$, whereas nine generations of asexuality were enough for $N = 10\ 000$. This result is attributable to the critical value of D_{KS} being lower at $N = 10\ 000$ than $N = 200$ (critical $D_{KS}(p = 0.05)$ at $N = 200$: 0.136; at $N = 10\ 000$: 0.019).

Although two-sample Kolmogorov-Smirnov tests based on F_{IS} probability distributions are highly informative in our case, comparing *de Finetti* diagrams between *CP before sex* and *APC* perfectly demonstrates that probability distributions of F_{IS} do not exhaustively describe the effect of clonality on genetic diversity. On the contrary, the Jensen-Shannon index of divergence (D_{JS}) uses the totality of the information comprised within the full *spd* of genotypic states to calculate the exact divergence between two scenarios. Unfortunately, D_{JS} is, to date, difficult to compute when working on empirical datasets (see Methods).

Nevertheless, our approach represents a new step towards inferring levels and modes of clonality in natural systems from population genetic data. To further achieve this goal, a likelihood ratio approach is currently under construction based on indices of divergence (D_{KS} and/or D_{JS}) calculated between theoretical predictions and empirical data for which the mode and level of clonality are to be assessed.

Conclusion and perspectives

The model presented here permitted exact predictions of genetic diversity in a context of cyclical parthenogenesis. It also demonstrated the differences between cyclical parthenogenesis and other reproductive modes: full sexuality, full clonality, and acyclic partial clonality. We hope that this research will lead to further investigations to refine predictions of genetic diversity in populations of cyclical parthenogens taking into account other classical aspects of their life-cycles (*e.g.* population bottleneck following sexual reproduction, effect of selection).

Acknowledgements

The authors would like to thank Jurgen Angst, Pierre Nouhaud and Jean-Christophe Simon for helpful discussions, the Région Bretagne and the the Plant Health and Environment (SPE) division of the French National Institute for Agricultural Research (INRA) for financing Katja Reichel, as well as the French National Research Agency (ANR) for funding within the CLONIX project (ANR-11-BSV7-0007).

Conflict of interest

All authors declare no conflicts of interest.

Data archiving

All data are included in the article or can be reproduced by the program code provided in supplementary information. Supplementary information is available on Heredity's website.

References

Allen DE, Lynch M. 2012. The effect of variable frequency of sexual reproduction on the genetic structure of natural populations of a cyclical parthenogen. *Evolution*. 66(3):919–926

Aradottir GI, Hanley SJ, Collins CM, Dawson KJ, Karp A, et al. 2012. Population genetics of *Tuberolachnus salignus*, an obligate parthenogenetic aphid. *Agricultural and Forest Entomology*. 14(2):197–205

Arnold TB, Emerson JW. 2011. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*. 3(2):34–39

Balloux F, Lehmann L, de Meeûs T. 2003. The population genetics of clonal and partially clonal diploids. *Genetics*. 164(4):1635–1644

Bengtsson BO. 2003. Genetic variation in organisms with sexual and asexual reproduction. *Journal of Evolutionary Biology*. 16(2):189–199

Berg LM, Lascoux M. 2000. Neutral genetic differentiation in an island model with cyclical parthenogenesis. *Journal of Evolutionary Biology*. 13(3):488–494

Danchin EGJ, Flot J-F, Perfus-Barbeoch L, Doninck KV. 2011. Genomic perspectives on the long-term absence of sexual reproduction in animals. In *Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution*, ed. P Pontarotti, pp. 223–242. Springer Berlin Heidelberg

De Finetti B. 1927. Conservazione e diffusione dei caratteri Mendeliani. Nota I. Caso panmittico. In *Rendiconti della R. Accademia Nazionale dei Lincei*, Vol. V (11-12), pp. 913–921

Delmotte F, Leterme N, Gauthier J-P, Rispe C, Simon J-C. 2002. Genetic architecture of sexual and asexual populations of the aphid *Rhopalosiphum padi* based on allozyme and microsatellite markers. *Molecular Ecology*. 11(4):711–723

De Meester L, Vanoverbeke J, De Gelas K, Ortells R, Spaak P. 2006. Genetic structure of cyclic parthenogenetic zooplankton populations – a conceptual framework. *Archive für Hydrobiologie*. 167(1-4):217–244

De Meeûs T, Balloux F. 2004. Clonal reproduction and linkage disequilibrium in diploids: A simulation study. *Infection, Genetics and Evolution*. 4(4):345–351

De Meeûs T, Balloux F. 2005. F-statistics of clonal diploids structured in numerous demes. *Molecular Ecology*. 14(9):2695–2702

De Meeûs T, Lehmann L, Balloux F. 2006. Molecular epidemiology of clonal diploids: A quick overview and a short DIY (do it yourself) notice. *Infection, Genetics and Evolution*. 6(2):163–170

De Meeûs T, Prugnolle F, Agnew P. 2007. Asexual reproduction: Genetics and evolutionary aspects. *Cellular and Molecular Life Sciences*. 64(11):1355–1372

Ellstrand N, Roose M. 1987. Patterns of genotypic diversity in clonal plant species. *American Journal of Botany*. 74(1):123–131

Halkett F, Simon J-C, Balloux F. 2005. Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology & Evolution*. 20(4):194–201

Hamilton M. 2009. *Population Genetics*. Wiley-Blackwell: New Jersey, USA

Hamilton N. 2015. *Ggtern: An extension to "ggplot2", for the creation of ternary diagrams.* R Package, Version 1.0.6.0.

Kanbe T, Akimoto S. 2009. Allelic and genotypic diversity in long-term asexual populations of the pea aphid, *Acyrthosiphon pisum*, in comparison with sexual populations. *Molecular Ecology*. 18(5):801–816

Li C-K, Schneider H. 2002. Applications of Perron-Frobenius theory to population dynamics. *Journal of mathematical biology*. 44(5):450–462

Marshall DR, Weir BS. 1979. Maintenance of genetic variation in apomictic plant populations. *Heredity*. 42(2):159–172

Oliphant TE. 2007. Python for scientific computing. *Computing in Science & Engineering*. 9(3):10–20

Orive ME. 1993. Effective population size in organisms with complex life-histories. *Theoretical Population Biology*. 44(3):316–340

Orsini L, Jansen M, Souche EL, Geldof S, De Meester L. 2011. Single nucleotide polymorphism discovery from expressed sequence tags in the waterflea *Daphnia magna*. *BMC Genomics*. 12(1):309

Papura D, Simon J-C, Halkett F, Delmotte F, Le Gallic J-F, Dedryver C-A. 2003. Predominance of sexual reproduction in Romanian populations of the aphid *Sitobion avenae* inferred from phenotypic and genetic structure. *Heredity*. 90(5):397–404

Pfrender ME, Lynch M. 2000. Quantitative genetic variation in *Daphnia*: Temporal changes in genetic architecture. *Evolution*. 54(5):1502–1509

Prugnolle F, Liu H, de Meeûs T, Balloux F. 2005a. Population genetics of complex life-cycle parasites: An illustration with trematodes. *International journal for parasitology*. 35(3):255–263

Prugnolle F, Roze D, Théron A, de Meeûs T. 2005b. F-statistics under alternation of sexual and asexual reproduction: A model and data from schistosomes (platyhelminth parasites). *Molecular Ecology*. 14(5):1355–1365

R Core Team. 2013. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reichel K, Bahier V, Midoux C, Parisey N, Masson J-P, Stoeckel S. submitted a. Interpretation and approximation tools for big, dense Markov chain transition matrices in population genetics. *Algorithms for Molecular Biology*.

Reichel K, Masson J-P, Malrieu F, Arnaud-Haond S, Stoeckel S. submitted b. Rare sex or out of reach equilibrium? The dynamics of F_{IS} in partially clonal organisms. *BMC Genetics*.

Rousset F. 2002. Inbreeding and relatedness coefficients: What do they meassure? *Heredity*. 88:371–380

Routtu J, Hall MD, Albere B, Beisel C, Bergeron RD, et al. 2014. An SNP-based second-generation genetic map of *Daphnia magna* and its application to QTL analysis of phenotypic traits. *BMC Genomics*. 15(1):1033

Stoeckel S, Grange J, Fernández-Manjarres JF, Bilger I, Frascaria-Lacoste N, Mariette S. 2006. Heterozygote excess in a self-incompatible and partially clonal forest tree species – *Prunus avium* L. *Molecular Ecology*. 15(8):2109–2118

Stoeckel S, Masson J-P. 2014. The exact distributions of F_{IS} under partial asexuality in small finite populations with mutation. *PLoS ONE*. 9(1):e85228

Vallejo-Marín M, Dorken ME, Barrett SCH. 2010. The ecological and evolutionary consequences of clonality for plant mating. *Annual Review of Ecology, Evolution, and Systematics*. 41(1):193–213

Van der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy array: A structure for efficient numerical computation. *Computing in Science Engineering*. 13(2):22–30

Vanoverbeke J, De Meester L. 2010. Clonal erosion and genetic drift in cyclical parthenogens – the interplay between neutral and selective processes. *Journal of Evolutionary Biology*. 23(5):997–1012

Van Rossum G. 2007. *Python programming language*. USENIX Annual Technical Conference.

Vorburger C. 2006. Temporal dynamics of genotypic diversity reveal strong clonal selection in the aphid *Myzus persicae*. *Journal of Evolutionary Biology*. 19(1):97–107

Appendix

Supplementary information 1 – Transition equations from Stoeckel & Masson (2014):

Genotypic frequency at $t+1$ (p^{t+1}, superscript does not denote exponentiation) in function of genotypic frequencies at t (p^t) after a clonal event (μ is the reciprocal mutation rate):

$$p_{aa}^{t+1} = (1-\mu)^2 p_{aa}^t + \mu(1-\mu)p_{Aa}^t + \mu^2 p_{AA}^t$$

$$p_{Aa}^{t+1} = 2\mu(1-\mu)p_{aa}^t + [\mu^2 + (1-\mu)^2]p_{Aa}^t + 2\mu(1-\mu)p_{AA}^t$$

$$p_{AA}^{t+1} = \mu^2 p_{aa}^t + \mu(1-\mu)p_{Aa}^t + (1-\mu)^2 p_{AA}^t$$

Genotypic frequencies at $t+1$ (q^{t+1}) in function of genotypic frequencies at t (p^t) after a sexual event:

$$q_{aa}^{t+1} = \left[(1-\mu)p_{aa}^t + \frac{1}{2}p_{Aa}^t + \mu\, p_{AA}^t\right]^2$$

$$q_{Aa}^{t+1} = 2\left[(1-\mu)p_{aa}^t + \frac{1}{2}p_{Aa}^t + \mu\, p_{AA}^t\right]\left[\mu\, p_{aa}^t + \frac{1}{2}p_{Aa}^t + (1-\mu)\, p_{AA}^t\right]$$

$$q_{AA}^{t+1} = \left[\mu\, p_{aa}^t + \frac{1}{2}p_{Aa}^t + (1-\mu)\, p_{AA}^t\right]^2$$

Supplementary information 2 – Explanation of matrix multiplication in eq. 7 and 8:

Consider the two transition matrix P_{sex} and P_{clonal} for which each column obey to the expression $\sum_{\substack{i=(s_{aa},s_{Aa},s_{AA}) \\ (s_{aa},s_{Aa},s_{AA})=N}} p[i|(r_{aa},r_{Aa},r_{AA})] = 1$:

P_{sex}:

States at $t+1$ (s_{aa},s_{Aa},s_{AA}) \ States at t (r_{aa},r_{Aa},r_{AA})	A	B	C	D
A	$p_{\text{sex}}[A\|A]$	$p_{\text{sex}}[A\|B]$	$p_{\text{sex}}[A\|C]$	$p_{\text{sex}}[A\|D]$
B	$p_{\text{sex}}[B\|A]$	$p_{\text{sex}}[B\|B]$	$p_{\text{sex}}[B\|C]$	$p_{\text{sex}}[B\|D]$
C	$p_{\text{sex}}[C\|A]$	$p_{\text{sex}}[C\|B]$	$p_{\text{sex}}[C\|C]$	$p_{\text{sex}}[C\|D]$
D	$p_{\text{sex}}[D\|A]$	$p_{\text{sex}}[D\|B]$	$p_{\text{sex}}[D\|C]$	$p_{\text{sex}}[D\|D]$

P_{clonal}:

States at $t+1$ (s_{aa},s_{Aa},s_{AA}) \ States at t (r_{aa},r_{Aa},r_{AA})	A	B	C	D
A	$p_{\text{clonal}}[A\|A]$	$p_{\text{clonal}}[A\|B]$	$p_{clonal}[A\|C]$	$p_{\text{clonal}}[A\|D]$
B	$p_{\text{clonal}}[B\|A]$	$p_{\text{clonal}}[B\|B]$	$p_{clonal}[B\|C]$	$p_{\text{clonal}}[B\|D]$
C	$p_{\text{clonal}}[C\|A]$	$p_{\text{clonal}}[C\|B]$	$p_{clonal}[C\|C]$	$p_{\text{clonal}}[C\|D]$
D	$p_{\text{clonal}}[D\|A]$	$p_{\text{clonal}}[D\|B]$	$p_{clonal}[D\|C]$	$p_{\text{clonal}}[D\|D]$

The logical path needed to calculate the probability distribution of states after *one sexual generation followed by one clonal generation* is better understood through an example. In the case presented below we want to calculate the transition probability to state D from state B, this probability is equal to the expression:

$$p(B \to D) = p_{sex}[A|B]p_{clonal}[D|A] + p_{sex}[B|B]p_{clonal}[D|B] + p_{sex}[C|B]p_{clonal}[D|C] + p_{sex}[D|B]p_{clonal}[D|D]$$

This probability is an entry in the matrix resulting of multiplication $P_{clonal} \cdot P_{sex}$ as shown below (and not $P_{sex} \cdot P_{clonal}$):

P_{sex}

	A	B	C	D
A	$p_{\text{sex}}[A\|A]$	$p_{\text{sex}}[A\|B]$	$p_{\text{sex}}[A\|C]$	$p_{\text{sex}}[A\|D]$
B	$p_{\text{sex}}[B\|A]$	$p_{\text{sex}}[B\|B]$	$p_{\text{sex}}[B\|C]$	$p_{\text{sex}}[B\|D]$
C	$p_{\text{sex}}[C\|A]$	$p_{\text{sex}}[C\|B]$	$p_{\text{sex}}[C\|C]$	$p_{\text{sex}}[C\|D]$
D	$p_{\text{sex}}[D\|A]$	$p_{\text{sex}}[D\|B]$	$p_{\text{sex}}[D\|C]$	$p_{\text{sex}}[D\|D]$

P_{clonal}

	A	B	C	D
A	$p_{\text{clonal}}[A\|A]$	$p_{\text{clonal}}[A\|B]$	$p_{\text{clonal}}[A\|C]$	$p_{\text{clonal}}[A\|D]$
B	$p_{\text{clonal}}[B\|A]$	$p_{\text{clonal}}[B\|B]$	$p_{\text{clonal}}[B\|C]$	$p_{\text{clonal}}[B\|D]$
C	$p_{\text{clonal}}[C\|A]$	$p_{\text{clonal}}[C\|B]$	$p_{\text{clonal}}[C\|C]$	$p_{\text{clonal}}[C\|D]$
D	$p_{\text{clonal}}[D\|A]$	$p_{\text{clonal}}[D\|B]$	$p_{\text{clonal}}[D\|C]$	$p_{\text{clonal}}[D\|D]$

	A	B	C	D
A	$p(A \to A)$	$p(B \to A)$	$p(C \to A)$	$p(D \to A)$
B	$p(A \to B)$	$p(B \to B)$	$p(C \to B)$	$p(D \to B)$
C	$p(A \to C)$	$p(B \to C)$	$p(C \to C)$	$p(D \to C)$
D	$p(A \to D)$	$p(B \to D)$	$p(C \to D)$	$p(D \to D)$

Supplementary information 3 – de Finetti *diagram*

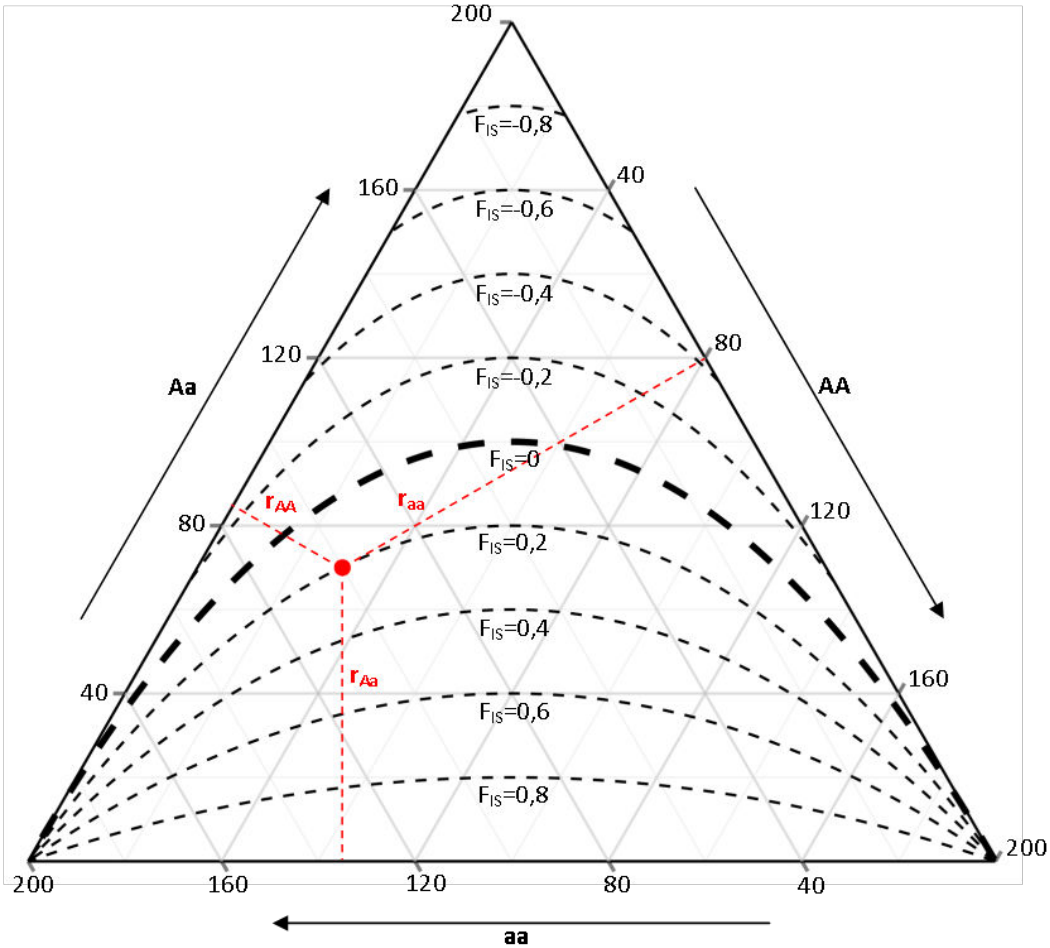

Figure A1. A *de Finetti* diagram is a classical ternary diagram which permits to simultaneously display the frequency of the three genotypes *AA, Aa* and *aa*. Here, the axis labels are adjusted for a population size of $N = 200$. The red dot represents the genotypic state $(r_{AA}, r_{Aa}, r_{aa}) = (30, 70, 100)$. The length of the red dotted lines permits to graphically infer the frequency of each genotype. Black dotted lines symbolises F_{IS} isocline, i.e. all possible states having the same F_{IS} value.

Supplementary information 4 – Explanation of the slightly negative mean F_{IS} observed in finite full sexual populations

In our exact population genetics model, the discrete nature of individuals $(N, r_{aa}, r_{Aa}, r_{AA}) \in \mathbb{N}$ has meaningful consequences on the distributions of individuals among the possible genotypes. Indeed, in the case where ${f_a}^2 \cdot N$ (i.e. proportion of homozygotes *aa* assuming Hardy-Weinberg proportions) is not a natural number, the population will not be in a state with precisely the Hardy-Weinberg proportions. Therefore, if evolutionary forces drive the population to HWE, it will oscillate by sampling the nearest discrete possible genotypic states which are more likely to yield negative F_{IS} values. This effect can be geometrically approached using *de Finetti* diagrams. Indeed, F_{IS} isoclines are convex functions toward lower F_{IS} values. As an example, imagine a population made of $N = 200$ individuals with 66 copies of the *a* allele ($f_a = 0.33$) and 144 copies of the *A* allele ($f_A = 0.67$). In this case, the exact state at Hardy-Weinberg equilibrium would be ($f_{aa} = 21.78$; $f_{Aa} = 88.44$; $f_{AA} = 89.78$) and cannot be reached in our model as in reality (individuals cannot be decimal quantities). The population will oscillate between the nearest genotypic states made of natural numbers which are ($f_{aa} = 21$; $f_{Aa} = 89$; $f_{AA} = 90$) with $F_{IS} = -0.0102$, ($f_{aa} = 22$; $f_{Aa} = 88$; $f_{AA} = 90$) with $F_{IS} = 0.0050$ and ($f_{aa} = 22$; $f_{Aa} = 89$; $f_{AA} = 89$) with $F_{IS} = -0.0025$ therefore causing negative F_{IS} on average.

Thus, all along the isocline $F_{IS} = 0$ on the *de Finetti* diagram, the discrete nature of individuals results in sampling more genotypic states standing for slightly negative F_{IS}than for slightly positive ones. Taking into account within our model the discrete nature of biological systems makes our results differ from results obtained in earlier models based on allelic identities (Balloux et al. 2003) where allelic and genotypic frequencies are assumed to be continuous.

7 Diversity under selection

7.1 Effects of selection under acyclic partial asexuality

In the last chapter, we analyzed the genotype frequency dynamics at selectively neutral loci. But what happens if genotype frequencies within a partially asexual population are also subject to selection? Based on the results from article II, one could suppose that their dynamics are generally slower in (partially) asexual than in exclusively sexual populations, and that adaptation is therefore always fastest with sexual reproduction. Alternatively, the speed of adaptation may depend on the nature of the genotypes that are selected (for/against). As a third hypothesis, the speed of adaptation may always be optimal under partially asexual reproduction, based on a "best of both worlds" argument.

We studied the genotype dynamics in acyclic partially asexual populations at a single locus under four basic selection scenarios: selection for a dominant or recessive "beneficial" allele, and selection for or against a heterozygous genotype. Selection was "non-lethal", i.e. the selectively least advantageous genotype could still survive and reproduce (compare e.g. Lokki 1976), keeping the population size constant. We also included scenarios with multiple alleles where the genotype fitness is determined by the dosage of a single allele. For each selection scenario, we determined the genotype frequency combinations that would increase their probability compared to neutral expectations, i.e. whose observation might serve to distinguish neutral and selected loci. We also compared the expected time until a population first reaches maximal mean fitness (time to adaptation) across different rates of clonality.

We found that the genotype frequency combinations that distinguish neutrality and selection not only depend on the selective scenario, but also on the rate of clonality. The same applies for the expected time to adaptation, which is generally quite long (hundreds to thousands of generations) except for small populations and very strong selection. However, exclusively sexual populations did not always have the shortest time to adaptation: depending on the selection scenario, the mutation rate and if the selectively most advantageous genotype(s) already existed in the population at the beginning of a selective sweep, exclusively asexual populations could also be faster to adapt. But for one exception (selection for heterozygous genotype, which did not yet exist at the beginning of the sweep) where rare sex was the fastest way to adapt, the times to adaptation of partially asexual populations were an intermediate between those of exclusive sexuality and exclusive asexuality.

The methods that are currently used to detect selection from genomic data are very much oriented towards selection for or against dominant/recessive alleles (compare Vitti et al. 2013). This article takes a wider view, considering selection at the genotype level, even though still at a single locus only. Preliminary studies also included intermediates between the different scenarios; it would be interesting to extend the analyses for such cases. Previous results for infinite populations (Marshall & Weir 1979, Overath & Asmussen 1998;

see also for a preliminary comparison between selection under partial asexuality and partial selfing), suggested that the characteristic genotype frequency combinations for partially asexual populations only change if the fitness of the heterozygous genotype is higher or equal compared to the fitnesses of the homozygous genotypes (the exclusion of equality in (Overath & Asmussen 1998) is probably an oversight). However, this distinction may become less clear in finite populations.

This article draft does not yet include a detailed analysis of the dynamics/genotype frequency changes during the adaptive process (compare figures 6 and 7 in article II). These data will be important for understanding selective dynamics under cyclical parthenogenesis: As an example, under the dominant selection scenario (*aa* genotype less fit than the *aA* and *AA* genotypes) the genotype combinations that are expected to be more frequent than at neutrality are very different between exclusively sexual and the exclusively asexual populations. Depending on the number of asexual generations, and depending on the moment during the life cycle when a selectively advantageous/ disadvantageous mutation appears, this may lead to situations where the genotype frequency combinations that are increased under selection are highly different between cyclical parthenogenesis and exclusively sexual reproduction.

Effet de la sélection sous asexualité partielle acyclique

Dans le dernier chapitre, nous avons analysé la dynamique de fréquences de génotype aux loci sélectivement neutres. Mais qu'advient-il si les fréquences génotypiques dans une population partiellement asexuée sont également soumises à la sélection ? Basé sur les résultats de l'article II, on pouvait supposer que leurs dynamiques sont généralement plus lentes dans les populations (partiellement) asexuées que dans les exclusivement sexuées, et que l'adaptation est donc toujours la plus rapide avec la reproduction sexuée. Alternativement, la vitesse d'adaptation peut dépendre de la nature des génotypes sélectionnés (positivement / négativement). En une troisième hypothèse, la vitesse d'adaptation pourrait toujours être optimale avec la reproduction partiellement asexuée, à partir de l'argument que c'est la combinaison du « meilleur des deux mondes ».

Nous avons étudié la dynamique de génotypes dans les populations partiellement asexuées acycliques à un seul locus selon quatre scénarios de base de la sélection : la sélection pour un allèle « bénéfique » dominant ou récessif, et la sélection pour ou contre un génotype hétérozygote. La sélection a été « non létale », c'est-à-dire que le génotype sélectivement moins avantageux pourrait encore survivre et se reproduire (comparer par exemple avec Lokki 1976), maintenant ainsi la taille de la population constante. Nous avons également inclus des scénarios avec plusieurs allèles où la valeur sélective de chaque génotype est déterminée par le dosage d'un seul allèle. Pour chaque scénario de sélection, nous avons déterminé des combinaisons de fréquences génotypiques qui augmenteraient leurs fréquences par rapport aux attentes neutres, c'est-à-dire dont l'observation peut servir à distinguer des loci neutres et sélectionnés. Nous avons également comparé le temps prévu jusqu'à ce qu'une population atteigne pour la première fois son maximum de la valeur sélective moyenne (temps d'adaptation) à travers différents taux de clonalité.

Nous avons constaté que les combinaisons des fréquences génotypiques qui distinguent la neutralité de la sélection ne dépendent pas seulement du scénario sélectif, mais aussi du taux de clonalité. Il en va de même pour la durée prévue pour l'adaptation, qui est généralement assez longue (des centaines de milliers de générations) à l'exception des populations moins grandes et de la sélection très forte. Cependant, les populations exclusivement sexuées ne présentent pas forcément des délais les plus courts dans l'adaptation : selon le scénario de sélection, selon le taux de mutation et selon le fait que si le génotype(s) sélectivement le(s) plus avantageuse(s) existai(en)t déjà dans la population au début d'un balayage sélectif, les populations exclusivement asexuées pourraient aussi s'adapter plus rapidement. Mais à une exception près (la sélection pour le génotype hétérozygote, qui n'existait pas encore au début du balayage) où le sexe rare est le meilleur moyen de s'adapter, la durée de l'adaptation des populations partiellement asexuées est un intermédiaire entre celui de la sexualité exclusive et de l'asexualité exclusive.

Les méthodes qui sont actuellement utilisées pour détecter la sélection à partir de données génomiques sont très orientées vers la sélection pour ou contre des allèles dominants / récessifs (comparer Vitti et al. 2013). Cet article a une vision plus large, en tenant compte de la sélection au niveau du génotype, bien qu'encore à seulement un locus unique. Des études préliminaires incluent également des intermédiaires entre les différents scénarios ; il serait intéressant d'étendre les analyses pour de tels cas. Les résultats précédents des populations infinies (Marshall & Weir 1979, Overath & Asmussen 1998 ; le voir aussi pour une comparaison préliminaire entre la sélection sous asexualité partielle et autofécondation partielle) ont suggéré que les combinaisons des fréquences génotypiques caractéristiques pour les populations partiellement asexuées ne changent que si la valeur sélective du génotype hétérozygote est supérieure ou égale aux valeurs des génotypes homozygotes (l'exclusion de l'égalité dans Overath & Asmussen 1998 est probablement un oubli). Toutefois, cette distinction peut devenir moins prononcée dans les populations de taille finie.

Ce projet d'article ne comprend pas encore une analyse détaillée de la dynamique / des changements des fréquences génotypiques au cours du processus d'adaptation (comparer les figures 6 et 7 de l'article II). Ces données seront importantes pour la compréhension de la dynamique sélective sous la parthénogenèse cyclique : par exemple, dans le scénario de sélection dominante (valeur sélective du génotype *aa* moins que ceux des génotypes *aA* et *AA*) les combinaisons de génotypes qui devraient être plus fréquentes que dans la neutralité sont très différentes entre les populations exclusivement sexuées et exclusivement asexuées. Selon le nombre de générations asexuées, et selon le moment du cycle de vie où une mutation sélective avantageuse / désavantageuse apparaît, cela peut conduire à des situations où les combinaisons de fréquences génotypiques qui sont augmentées sous sélection sont très différents entre la parthénogenèse cyclique et la reproduction exclusivement sexuée.

Article IV Clonalité partielle et vitesse d'adaptation

Sommaire de l'article

Détecter des loci sous sélection devient rapidement une application standard de la théorie de génétique des populations. Pourtant, pour trouver ce que l'on cherche, il faut savoir où chercher. Les systèmes de reproduction sont bien connus pour affecter la génétique des populations, et de fait les études précédentes ont montré que la clonalité partielle affecte la distribution d'équilibre et la dynamique de l'F_{IS}. Nous avons étudié la dynamique des fréquences génotypiques et la valeur sélective moyenne des populations lors de la sélection basée sur un seul locus, en prenant en compte des différents scénarios de sélection, de mutation et de dérive génétique. En particulier, nous avons comparé les équilibres (fréquences génotypiques, valeur sélective moyenne des populations) et la durée d'adaptation entre les différents taux de clonalité partiels.

Nous avons constaté que différents scénarios de sélection peuvent changer leur « signature », à savoir les combinaisons de fréquences des génotypes qui sont censées être plus fréquentes sous sélection, en fonction du taux de clonalité. Cela est particulièrement visible dans les scénarios de sélection où les génotypes hétérozygotes ont des valeurs sélectives les plus élevées, et si la reproduction sexuée est rare. Pour la plupart des scénarios que nous avons analysés, le temps d'adaptation est optimal soit sous reproduction exclusivement sexuée soit exclusivement asexuée. La seule exception est la sélection d'un génotype hétérozygote pas encore présent dans la population, où le sexe rare accélère l'adaptation.

Nous avons montré que ni la reproduction exclusivement sexuée ni la clonalité partielle est toujours la voie la plus rapide à l'adaptation. Nos résultats peuvent être utilisés pour adapter les méthodes de détection des loci sous sélection chez les espèces partiellement clonales, et pour améliorer l'interprétation des données de terrain.

Article IV *Partial asexuality and the speed of adaptation*

Katja Reichel*, Jean-Pierre Masson, Solenn Stoeckel

INRA, UMR1349 Institute for Genetics, Environment and Plant Protection, F-35650, Le Rheu, France

unsubmitted draft

Abstract

Detecting loci under selection is rapidly becoming a standard application for population genetic theory. Yet to find what one is looking for, one has to know where to look. Reproductive systems are well known to affect population genetics, and indeed previous studies showed that partial clonality affects the equilibrium distribution and dynamics of F_{IS}. We studied the dynamics of genotype frequencies and population mean fitness during selection based on a single locus, taking different selection scenarios, mutation and genetic drift into account. In particular, we compared the equilibria (genotype frequencies, population mean fitness) and the time to adaptation across different rates of partial clonality.

We found that different selection scenarios may change their "signature", i.e. the combinations of genotype frequencies that are expected to be more frequent under selection, according to the rate of clonality. This is especially noticeable for selection scenarios where heterozygote genotypes have the highest mean fitness, and if sexual reproduction is rare. For most scenarios we analyzed, the time to adaptation is optimal either under exclusively sexual or exclusively asexual reproduction. The only exception is selection for a heterozygous genotype not yet present in the population, where rare sex speeds up adaptation.

We showed that neither exclusively sexual reproduction nor partial clonality is always the fastest way to adaptation. Our results may be used to adapt methods for detecting loci under selection in partially clonal species, and improve the interpretation of field data.

Keywords selection, asexual reproduction, SNP, SSR, *de Finetti* diagram

Introduction

Adaptive evolution, the process whereby organisms change in response to natural selection, is a major research topic in evolutionary biology at least since the work of Charles Darwin (1872, Grant & Grant 2003). Natural selection acts on phenotypes, increasing (positive selection) or decreasing (negative selection) their relative potential to survive and/or reproduce (fitness). The trait values that determine the fitness of an individual have a heritable and a non-heritable component. The chosen mode of reproduction can have a great impact on the way the heritable fitness component is passed on from one generation to the next; indeed, the discussion about the evolution of different reproductive systems is largely based on their potential to influence adaptive evolution, especially its speed (Otto 2009).

The evolution of reproductive systems is, however, not the only link between reproduction and adaptation. A great deal of research in evolutionary ecology is currently directed at finding the mechanisms underlying the heritable variability of selective traits (e.g. Jaquiéry et al. 2014, Lamichhaney et al. 2015): The aim is to connect environmental conditions to phenotypes and genotypes, so that adaptive evolution can be studied directly "in action". Detecting candidate regions within genomes that may be involved in adaptive processes often relies on population genetic "signatures" of selection (Vitti et al. 2013) – typically, natural selection leads to the enrichment of a "beneficial" allele at a particular locus, which can then be detected e.g. as increased differentiation (high F_{ST}) between different populations. Yet other evolutionary processes, most notably reproduction, can modify the neutral reference (e.g. Stoeckel & Masson 2014) and potentially also the expected selection signatures. A sound theoretical reference, describing the patterns of genetic variation expected under different selection scenarios, could therefore prove essential for studying adaptation in populations with non-standard reproductive systems.

We were interested in adaptive evolution in partially clonal populations, where each individual may reproduce both clonally and sexually by random mating. Although widespread both throughout the earth's biomes and on the tree of life, partially clonal species have hitherto received comparatively little attention from theoretical population geneticists. They have repeatedly, though not consistently (e.g. Ryndin et al. 2001, Hartfield et al. 2012), been included in models about the evolution of sexual reproduction. Outside of this context, only few studies compared adaptive processes under combined sexual/clonal reproduction (e.g. Marshall & Weir 1979, Muirhead & Lande 1997, Overath & Asmussen 1998). Most of these focused on some particular selection scenario and on the outcome of selective sweeps rather than on their dynamic, or did not include any evolutionary processes beside reproduction and selection.

Recent research on the neutral variation of single loci (Reichel et al. submitted, Stoeckel & Masson 2014) showed that genotype frequency dynamics in finite partially clonal populations may differ from those under exclusive sexuality: attraction to the Hardy-Weinberg equilibrium is slowed down, which grants more influence to evolutionary processes leading away from it (e.g. genetic drift). These results also raise new questions about adaptive evolution in partially clonal populations: Firstly, how do the expected

distributions of genotype frequencies ("signature") and the mean fitness of the population change under different rates of partial asexuality? And secondly, if the neutral dynamics of F_{IS}/heterozygosity are slowed down in partially clonal populations, how does this affect the expected time until adaptation (i.e. maximal mean fitness of the population) under different selection scenarios?

To answer these questions, we extended the Markov chain model of Reichel et al. (submitted) to include selection. We studied four different basic scenarios: selection for a recessive allele *A* (or against a dominant allele *a*), selection for a dominant allele *A* (or against a recessive allele *a*), selection for a heterozygous genotype *aA* and divergent selection for different homozygous genotypes *aa* and *AA*. For each scenario, we first compared the expected distribution of genotype frequencies to its neutral counterpart to see which combinations would become more/less frequent under selection. Secondly, we determined the expected final mean fitness of the population to see if different rates of clonal reproduction can lead to "better" adaptation, in terms of this parameter. Finally, we calculated the expected time to adaptation (maximal mean fitness) under different rates of asexuality for "hard" selective sweeps, starting from minimal mean fitness with all alleles present, and "soft" selective sweeps, starting from Hardy-Weinberg equilibrium with all alleles equally frequent; though the latter combination of genotype frequencies becomes less common at selectively neutral loci as the rate of clonality increases, it represents a situation where all genotypes are present prior to the selective sweep. Thus, we could compare the speed of adaptation in partially clonal populations under the different scenarios to the expectations for exclusively sexual or exclusively asexual reproduction.

Methods

The biological basis of our model is a single, isolated population of constant finite size N (Wright-Fisher model; compare Reichel et al. submitted). Individuals correspond to ramets (i.e. physiological rather than genetic units), are diploid and may undergo somatic mutations that can be passed on to their offspring at a mutation rate μ. For each new generation, a fraction c of the offspring is formed by clonal reproduction (i.e. genetically identical to its parent except for somatic mutations), whereas the rest (no survival between generations) derives from sexual reproduction by random mating including selfing at rate $1/N$.

We extended the Markov chain model for genotype frequencies in a small finite population described in Reichel et al. (submitted, based on Stoeckel & Masson 2014) by adding selection before mutation in the life cycle (figure 1). Biologically, this corresponds to selection acting mainly on the reproductive success ("seed census") and less on the establishment of individuals ("adult census", as discussed in Overath & Asmussen 1998). Equations I-IV from the original model are provided for reference in additional file 1. The general form of the new equation 0 for selection is:

$$\overrightarrow{v_0} = (\vec{\varphi} \cdot \overrightarrow{v_t})^{-1}(\vec{\varphi} \circ \overrightarrow{v_t})$$

where $\overrightarrow{v_t}$ is the vector of genotype frequencies observed at time t, $\overrightarrow{v_0}$ the vector of genotype frequencies after selection, $\vec{\varphi}$ the vector of genotype fitness values, $\cdot$ denotes the

scalar product (sum over elementwise multiplication) and ∘ the Hadamard/Schur product (elementwise multiplication).

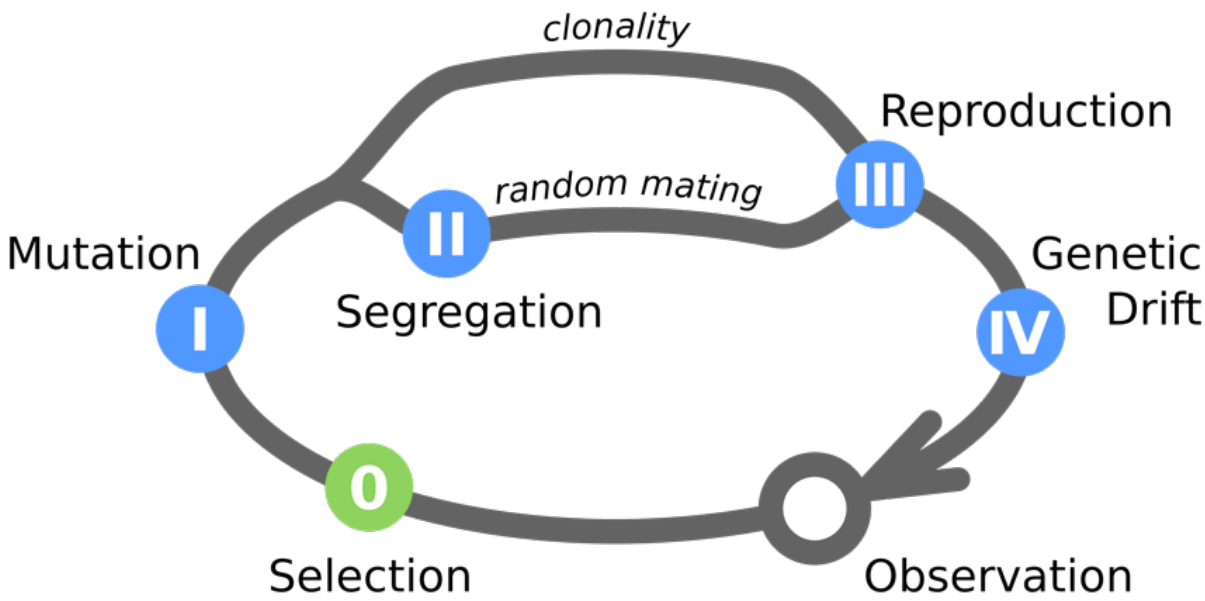

Figure 1. *Schematic representation of the model life cycle.* Steps I to IV (blue) are identical to the model in Reichel et al. (submitted), step 0 (green, selection) has been added. The haploid phase of the life cycle is between steps II and III.

The different selection scenarios correspond to different parameterizations of the vector of genotype fitness values $\varphi \in [0,1]$ in equation 0, as presented in table 1, based on a selection coefficient $s \in [0,1]$. All asymmetries in the scenarios are formulated in favor of the *A* allele. As our definition of genotype fitness is relative (constant population size, rescaling by $(\vec{\varphi} \cdot \vec{v_t})^{-1}$ in equation 0) and the maximal genotype fitness is always one, selection "for" particular genotypes is exactly equivalent to selection "against" the others. Consequently, though the scenarios' names are based on positive selection, the results apply equally for the complementary scenario of negative selection, e.g. selection for a recessive allele *A* is equivalent to selection against a dominant allele *a*.

Table 1. Overview of selection scenarios. Fitness values φ for each genotype depending on the selection parameter s. Initial states (v_{aa}, v_{aA}, v_{AA}) for hard selective sweeps; soft selective sweeps start from $(N/4, N/2, N/4)$ for all scenarios.

Scenario	short	φ_{aa}	φ_{aA}	φ_{AA}	Initial state H
Neutral	N	1	1	1	*any of below*
Recessive	R	$1-s$	$1-s$	1	$(N-1,1,0)$
Dominant	D	$1-s$	1	1	$(N-1,1,0)$
Overdominant	O	$1-s$	1	$1-s$	$(N/2,0,N/2)$
Underdominant	U	1	$1-s$	1	$(0,N,0)$

For loci with more than two alleles, *a* stands for all alleles that are not the favored allele *A*. Consequently, v_{aA} is the sum of the frequencies of all genotypes that have exactly one *A* allele, and v_{aa} is the sum of the frequencies of all (homozygous and heterozygous)

genotypes without the *A* allele. All subsumed genotypes have the same fitness value, i.e. a genotype's fitness depends only on the number of *A* alleles it possesses. This means that mutation between *A* and "*a*" in becomes asymmetric, with $\mu_{A\to a} = \mu$ and $\mu_{a\to A} = \mu/(n-1)$, where μ is the "global" mutation rate (assuming a k-alleles/Jukes-Cantor mutation model) and n the number of alleles. Equation I (mutation) thus becomes:

$$\begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_I = \begin{bmatrix} \left(1-\frac{\mu}{n-1}\right)^2 & \mu\left(1-\frac{\mu}{n-1}\right) & \mu^2 \\ \left(\frac{2\mu}{n-1}\right)\left(1-\frac{\mu}{n-1}\right) & \mu\left(\frac{\mu}{n-1}\right)+(1-\mu)\left(1-\frac{\mu}{n-1}\right) & 2\mu(1-\mu) \\ \left(\frac{\mu}{n-1}\right)^2 & (1-\mu)\left(\frac{\mu}{n-1}\right) & (1-\mu)^2 \end{bmatrix} \begin{bmatrix} \nu_{aa} \\ \nu_{aA} \\ \nu_{AA} \end{bmatrix}_0$$

All other model equations remain unchanged.

Based on equations 0 to IV, we constructed the transition matrix M of the Markov chain model: this is a square matrix containing all transition probabilities $p(q_{t+1}|q_t)$, where q_t denotes the set of numbers of individuals $q \in \mathbb{N}_0$ with $\sum q = N$ possessing each respective genotype ("state" of the Markov chain) at time t.

The expected frequency with which each model state would be observed if all model parameters (selection scenario, s, μ, c, N) had always been the constant ("equilibrium") is derived from the dominant eigenvector of the transition matrix. The mean fitness of the population $\overline{\Phi}$ and the mean frequency of the *A* allele $\overline{\nu_A}$ are derived by summing over the product of expected state frequency with the mean fitness / allele frequency for each state.

The expected time to adaptation is calculated as the first passage time from a specified initial state until the population first reaches a state of maximal mean fitness ($\overline{\Phi} = 1$; all individuals have a genotype with $\varphi = 1$). For each selection scenario, we looked at two particular initial states: a subscript S denotes soft selection acting on a population in Hardy-Weinberg equilibrium with equal allele frequencies $(N/4, N/2, N/2)$, and a subscript H denotes hard selection acting on a population in which both alleles are present but which otherwise has the lowest possible mean fitness (see table 1).

Results

Expected genotype frequencies and population mean fitness

Natural selection evidently changes the genotype frequencies, but which combinations become more or less common may depend on the rate of clonality as well (figure 2). Under the recessive scenario, a population exclusively of *AA* individuals would be both most fit and most likely under all rates of clonality; yet though its gain in probability is largest under exclusive asexuality, stochastic flux out of it ($\overline{\Phi} < 1$) is also greatest then (figure 2R).

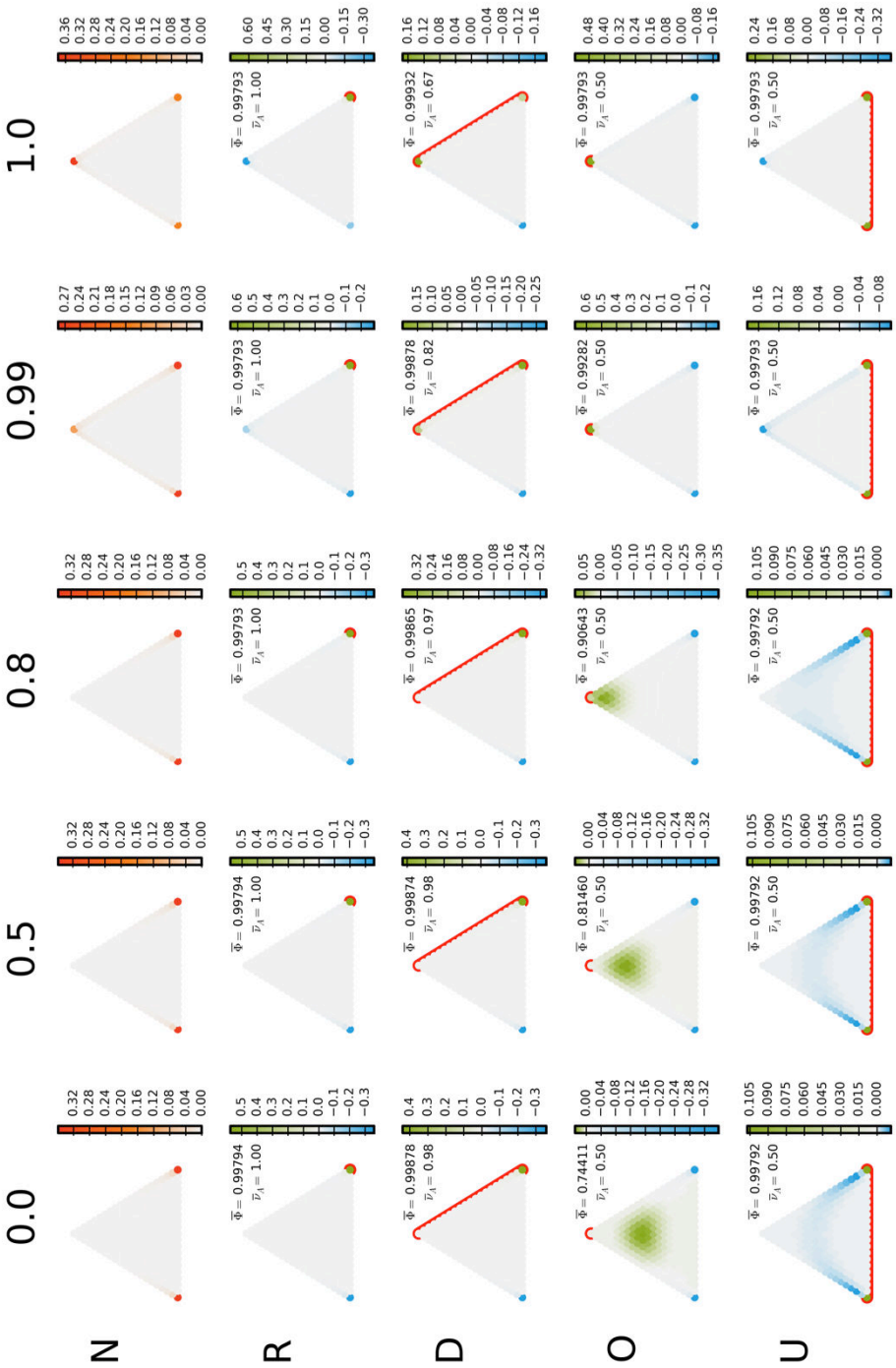

Figure 2. Equilibrium genotype frequencies under selection compared to neutral expectation. Neutrality and four selection scenarios (see table 1) with $s = 0.5$ and five rates of partial clonality c $(0.0, 0.5, 0.8, 0.99, 1.0)$ in a population with $N = 20$ and $n = 2$ alleles. $\overline{\Phi}$: mean population fitness at the equilibrium, $\overline{\nu_A}$: mean frequency of allele *A* at the equilibrium. Colors: state probability at neutrality (orange/red, scale: probability) and combinations of genotype frequencies that are more probable (green), ca. equally probable (white) and less probable (blue) under selection than at neutrality (scale: difference in probability). *De Finetti* diagrams with genotype *aa* in the lower left corner (more information in additional file 2); red brackets enclose states of maximal mean fitness. Figures for $s = 0.5, N = 20, n = 10$ and $s = 0.01, N = 100, n = 2$ in additional file 3.

For the dominant scenario, any state where all individuals have at least one *A* allele has maximal mean fitness, but not all of them are more probable than under neutrality (figure 2D). In an exclusively sexual population, the characteristic "pattern" looks almost identical to the recessive case, yet as sexual reproduction becomes less frequent the equilibrium mean frequency of the *A* allele diminishes. As c comes close to one, the balance finally flips: the greatest increase in probability is then at a population where all individuals are heterozygous with one *A* allele.

For the overdominant scenario, the state where all individuals are heterozygous with one *A* allele has the highest mean fitness (figure 2O). Still, this state is only reached at nearly exclusive clonality, otherwise it is not found any more often than under neutrality. Even though, the mean fitness at equilibrium augments already from $c \geq 0$ onwards, as the states with increased probability shift towards ever higher heterozygosity.

For the underdominant scenario, any state without heterozygotes in the population has maximal mean fitness (figure 2U). However, only the fixation states noticeably increase their probability here. As the rate of asexuality increases, the differences between neutral and selection probabilities become greater.

Augmenting the number of alleles makes the two symmetric selection scenarios (U and O) asymmetric, but otherwise hardly changes the patterns (appendix, figure A3.1). Increasing the population size "sharpens" the contours of the states with increased probability; yet lowering the selection coefficient generally decreases the probability differences between selective and neutral case, so that selection generally becomes harder to detect (appendix, figure A3.2).

Expected time to adaptation

Each selection scenario has a characteristic shape of the "time to adaptation" curve (figure 3): for the recessive and underdominant cases, it curves upward (longer time to adapt) at high rates of clonality, while for the dominant and overdominant cases the lower rates of clonality take longest to reach a state of maximal mean fitness. The times to adaptation during a soft selective sweep are always shorter than or equal to those for the hard selective sweep. For the recessive and dominant case, the difference in time between fastest and slowest adapting rate of asexuality is simply smaller during a soft selective sweep. However, for the overdominant and underdominant cases, the shape of the curve is different under high rates of clonality and high selection coefficients: for hard selective sweeps, the time to adaptation increases as $c \longrightarrow 1.0$, but for a soft selective sweep it decreases. Consequently, the time needed to form a new advantageous genotype (homozygote or heterozygote) from the alleles already present in the population increases the time to adaptation in those cases.

Increasing the population size (figure 4) also increases the time to adaptation. While the principal pattern stays the same in the recessive and dominant cases, the times to adaptation in the overdominant case increase so dramatically, that it is safe to assume that the state of maximal mean fitness is only ever reached under very high rates of clonality.

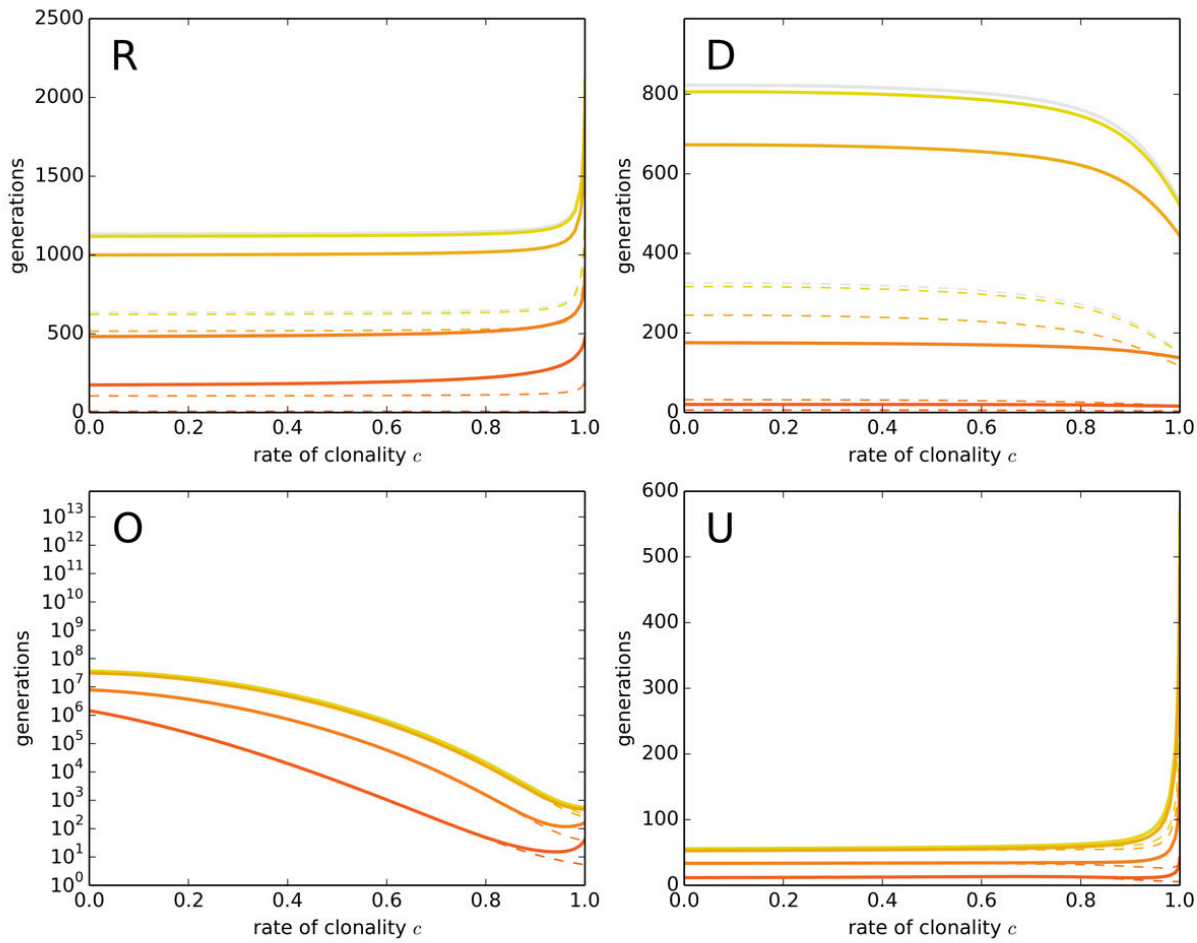

Figure 3. Time to maximal mean fitness under different rates of partial clonality. Four selection scenarios (see table 1). Colors (dark to light): selection coefficients s $(0.5, 0.1, 0.01, 0.001)$, neutral case in light grey. Lines for higher s may hide those for lower s. Continuous lines: hard selective sweep, dashed lines: soft selective sweep. Population with $N = 20, \mu = 10^{-3}$ and $n = 2$ alleles.

For the underdominant case and high selection coefficients, the time to adaptation at first slightly augments with the rate of clonality, but then drops again for very high rates of clonality. As this pattern is similar for both soft and hard selective sweeps, it might be due to a sub-optimal "detour" of the adaptation process. Decreasing the mutation rate or increasing the number of alleles further increases the time to adaptation, but otherwise produces the same pattern.

If the mutation rate is increased (figure 5), the situation changes dramatically. Now, the time to adaptation is always minimal in completely clonal populations, regardless of the selection scenario. Apparently, very frequent mutation randomizing the association of alleles can have the same benefit for the time to adaptation as – at least occasional – sexual reproduction.

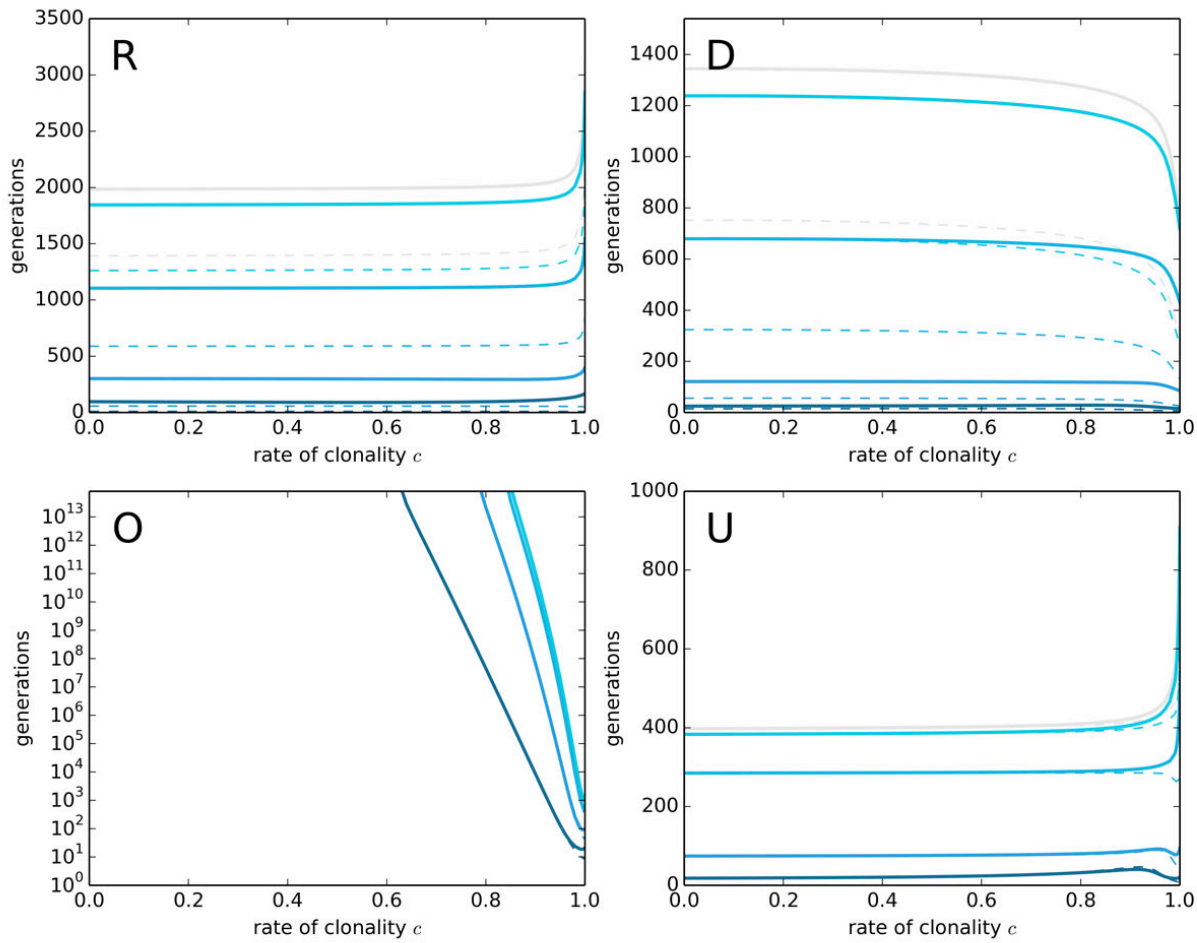

Figure 4. Time to maximal mean fitness under different rates of partial clonality. Four selection scenarios (see table 1). Colors (dark to light): selection coefficients s $(0.5, 0.1, 0.01, 0.001)$, neutral case in light grey. Lines for higher s may hide those for lower s. Continuous lines: hard selective sweep, dashed lines: soft selective sweep. Population with $N = 100, \mu = 10^{-3}$ and $n = 2$ alleles.

Discussion

We analyzed adaptive evolution, based on a single locus under selection, in partially clonal populations. To our knowledge, this is the first such study that also accounts for mutation and genetic drift, and includes all four principal selection scenarios. Our results may serve as a reference for improving the detection of loci under selection in field studies of partially clonal organisms and widen the discussion about the speed of adaptation due to different reproductive systems.

Our results suggest that partial clonality may indeed change the combinations of genotype frequencies that indicate selection at single loci: As an example, under exclusively sexual reproduction selection for a "beneficial" allele leads to increased fixation of this allele, regardless if its effect is dominant or recessive. In contrast, highly clonal populations show high heterozygosity rather than fixation if the "beneficial" allele is dominant, a "signature" similar to that of selection for the heterozygote genotype. Methods for the

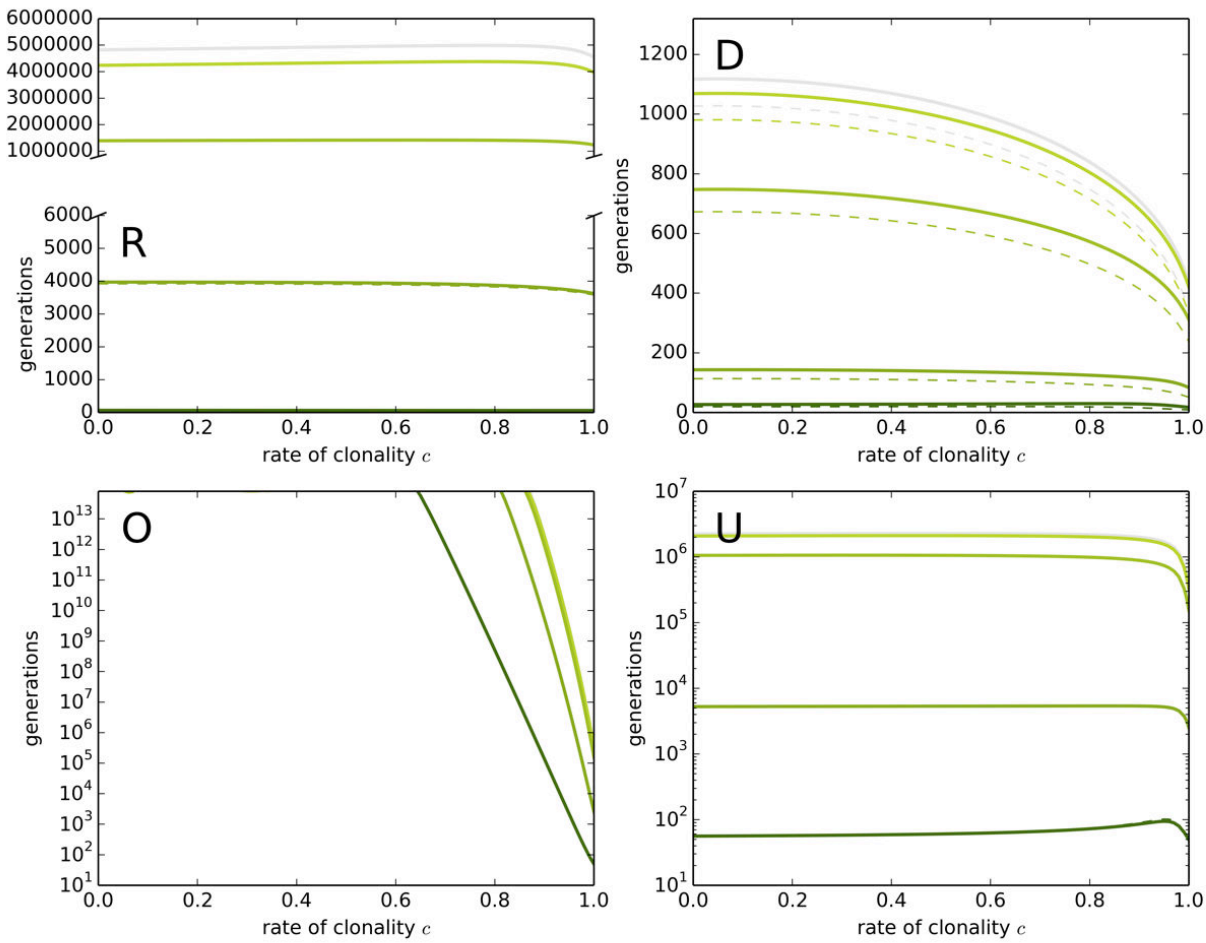

Figure 5. Time to maximal mean fitness under different rates of partial clonality. Four selection scenarios (see table 1). Colors (dark to light): selection coefficients s $(0.5, 0.1, 0.01, 0.001)$, neutral case in light grey. Lines for higher s may hide those for lower s. Continuous lines: hard selective sweep, dashed lines: soft selective sweep. Population with $N = 100$, $\mu = 10^{-2}$ and $n = 2$ alleles.

detection of selection that are based on a comparison of allele frequencies (F_{ST}) would miss such selection scenarios. Based on our expected distributions of genotype frequencies, it might be possible to develop new detection methods for selection candidate regions, that are not only better adapted for partially clonal species but could also pick up e.g. selection for heterozygosity or weak selection in exclusively sexual populations.

During selective sweeps based on single loci under selection, partially clonal species are hardly ever fastest to reach maximal mean fitness – in most cases either exclusively sexual or exclusively asexual reproduction would take less long, with partial clonality somewhere in between. We did not find a direct relationship with the slowed-down dynamics of F_{IS} observed in the absence of selection. Adaptation under exclusive asexuality is sometimes slowed down because the most advantageous "recombinant" genotype does not arise, though all necessary alleles are available. This leads to a remarkable exception under selection for a heterozygous genotype, where adaptation is speeded up by rare sexual reproduction first creating the most fit genotype in an otherwise highly clonal population.

In general, the expected number of generations until the whole population has a maximal-fitness genotype is very long, especially for hard selective sweeps. Populations that reached the maximal mean fitness are therefore probably rare in nature, in particular among big populations.

The results of this study should not be over-interpreted in terms of evolutionary advantages/disadvantages of particular reproductive systems: at the present moment, it is not clear whether a higher speed of adaptation or a higher mean population fitness under some specific selection scenario are relevant criteria for the evolution of the whole reproductive system. Moreover, selection on only a single locus, without taking the whole genomic context into account, is certainly an over-simplification of the complexity of adaptive evolution. Still, we demonstrated that studying a variety of selection scenarios may lead to a more complete picture: perhaps the "mixture" of different selection scenarios throughout the evolutionary history of different taxa could also have contributed to the evolution of its reproductive system? Improving the population genetic tools to detect selection also in partially clonal populations will hopefully lead to a better understanding of field data, which may in turn provide more answers about the evolution of reproductive systems.

Conclusion

We provide expectations for the "signatures" of different selection scenarios at single loci under partial clonality, and for the times until all individuals within the population have a maximal-fitness genotype. Different rates of clonal reproduction change the genotype frequency combinations most likely to indicate selection. Adaptation in exclusively clonal populations may sometimes be speeded up by rare sex, but partial clonality is generally not the fastest way to adapt.

Acknowledgements

We are indebted to F. Bonhomme, R. Rouger, D. Roze, J.-C. Simon and our collaborators within the CLONIX project (ANR-11-BSV7-0007) for helpful discussions. KR would like to thank the Région Bretagne and the department Plant Health & Environment for financing her work.

References

Darwin CR. 1872. *The origin of species by means of natural selection or the preservation of favoured races in the struggle for life.* New York: Mentor (reprint 1958). 6th ed.

Grant RB, Grant PR. 2003. What Darwin's finches can teach us about the evolutionary origin and regulation of biodiversity. *BioScience*. 53(10):965–75

Hartfield M, Otto SP, Keightley PD. 2012. The maintenance of obligate sex in finite, structured populations subject to recurrent beneficial and deleterious mutation. *Evolution*. 66(12):3658–69

Jaquiéry J, Stoeckel S, Larose C, Nouhaud P, Rispe C, et al. 2014. Genetic control of contagious asexuality in the pea aphid. *PLoS Genetics*. 10(12):e1004838

Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, et al. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*. 518(7539):371–375

Marshall DR, Weir BS. 1979. Maintenance of genetic variation in apomictic plant populations. *Heredity*. 42(2):159–172

Muirhead CA, Lande R. 1997. Inbreeding depression under joint selfing, outcrossing, and asexuality. *Evolution*. 51(5):1409–1415

Otto SP. 2009. The evolutionary enigma of sex. *The American Naturalist*. 174(s1):S1–14

Overath RD, Asmussen MA. 1998. Genetic diversity at a single locus under viability selection and facultative apomixis: Equilibrium structure and deviations from Hardy-Weinberg frequencies. *Genetics*. 148(4):2029–2039

Reichel K, Masson J-P, Malrieu F, Arnaud-Haond S, Stoeckel S. submitted. Rare sex or out of reach equilibrium? The dynamics of F_{IS} in partially clonal organisms. *BMC Genetics*

Ryndin A, Kirzhner V, Nevo E, Korol A. 2001. Polymorphism maintenance in populations with mixed random mating and apomixis subjected to stabilizing and cyclical selection. *Journal of Theoretical Biology*. 212(2):169–181

Stoeckel S, Masson J-P. 2014. The exact distributions of F_{IS} under partial asexuality in small finite populations with mutation. *PLoS ONE*. 9(1):e85228

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annual Review of Genetics*. 47(1):97–120

Appendix

Additional file 1 – Model equations

The following equations are based on two alleles; for more alleles, *a* stands for all alleles that are not *A*.

0 Selection

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_0 = \left(\begin{bmatrix}\varphi_{aa}\\ \varphi_{aA}\\ \varphi_{AA}\end{bmatrix}\cdot\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_t\right)^{-1}\left(\begin{bmatrix}\varphi_{aa}\\ \varphi_{aA}\\ \varphi_{AA}\end{bmatrix}\circ\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_t\right)$$

I Mutation

two alleles:

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_I = \begin{bmatrix}(1-\mu)^2 & \mu(1-\mu) & \mu^2\\ 2\mu(1-\mu) & \mu^2+(1-\mu)^2 & 2\mu(1-\mu)\\ \mu^2 & \mu(1-\mu) & (1-\mu)^2\end{bmatrix}\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_0$$

n alleles:

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_I = \begin{bmatrix}\left(1-\frac{\mu}{n-1}\right)^2 & \mu\left(1-\frac{\mu}{n-1}\right) & \mu^2\\ \left(\frac{2\mu}{n-1}\right)\left(1-\frac{\mu}{n-1}\right) & \mu\left(\frac{\mu}{n-1}\right)+(1-\mu)\left(1-\frac{\mu}{n-1}\right) & 2\mu(1-\mu)\\ \left(\frac{\mu}{n-1}\right)^2 & (1-\mu)\left(\frac{\mu}{n-1}\right) & (1-\mu)^2\end{bmatrix}\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_0$$

II Allele segregation / Gamete formation

$$\begin{bmatrix}\nu_a\\ \nu_A\end{bmatrix}_{II} = \begin{bmatrix}1 & 0.5 & 0\\ 0 & 0.5 & 1\end{bmatrix}\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_I$$

III Reproduction

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{III} = c\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_I + (1-c)\begin{bmatrix}\nu_a^2\\ 2\nu_a\nu_A\\ \nu_A^2\end{bmatrix}_{II}$$

IV Genetic drift

$$\begin{bmatrix}\nu_{aa}\\ \nu_{aA}\\ \nu_{AA}\end{bmatrix}_{t+1} = X_{t+1}/N \text{ where } X_{t+1} \sim \mathcal{M}(N, [\nu_{aa}, \nu_{aA}, \nu_{AA}]_{III})$$

The *de Finetti* diagram

a visualisation of all possible compositions of a population out of different genotypes for one locus with two alleles *(a, A)* in a diploid organism

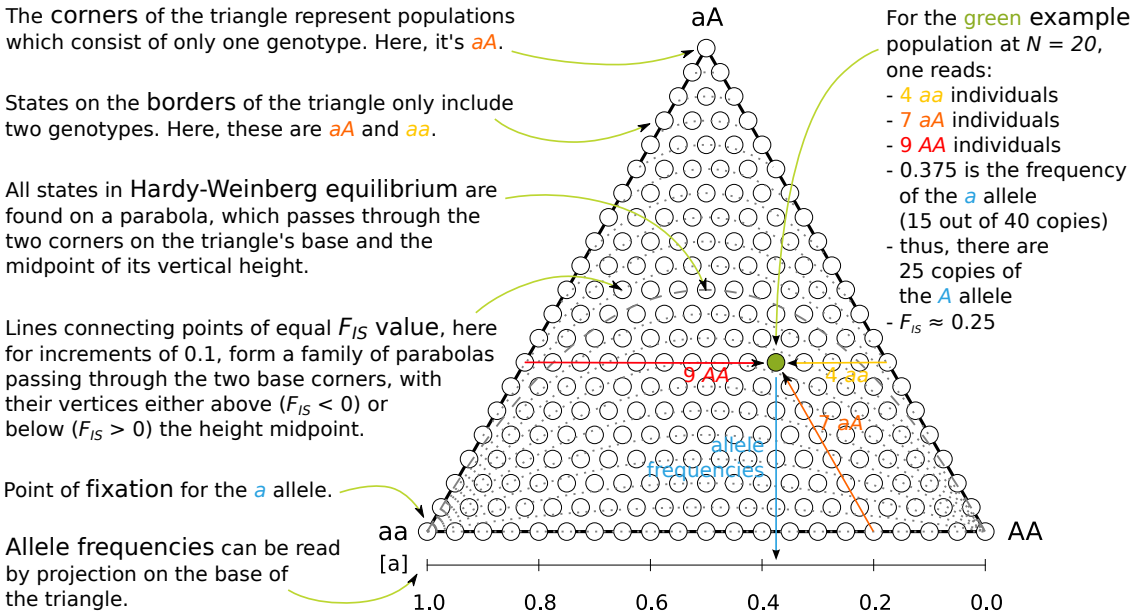

Figure A2.1: Schematic explanation of a *de Finetti* diagram. For more than two alleles, *a* represents all alleles that are not *A*, i.e. the frequency of *"aa"* is a sum over the frequencies of several heterozygous and homozygous genotypes; the F_{IS} / Hardy-Weinberg parabolas are only valid for the two-allele case.

Additional file 3 – Supplementary figures

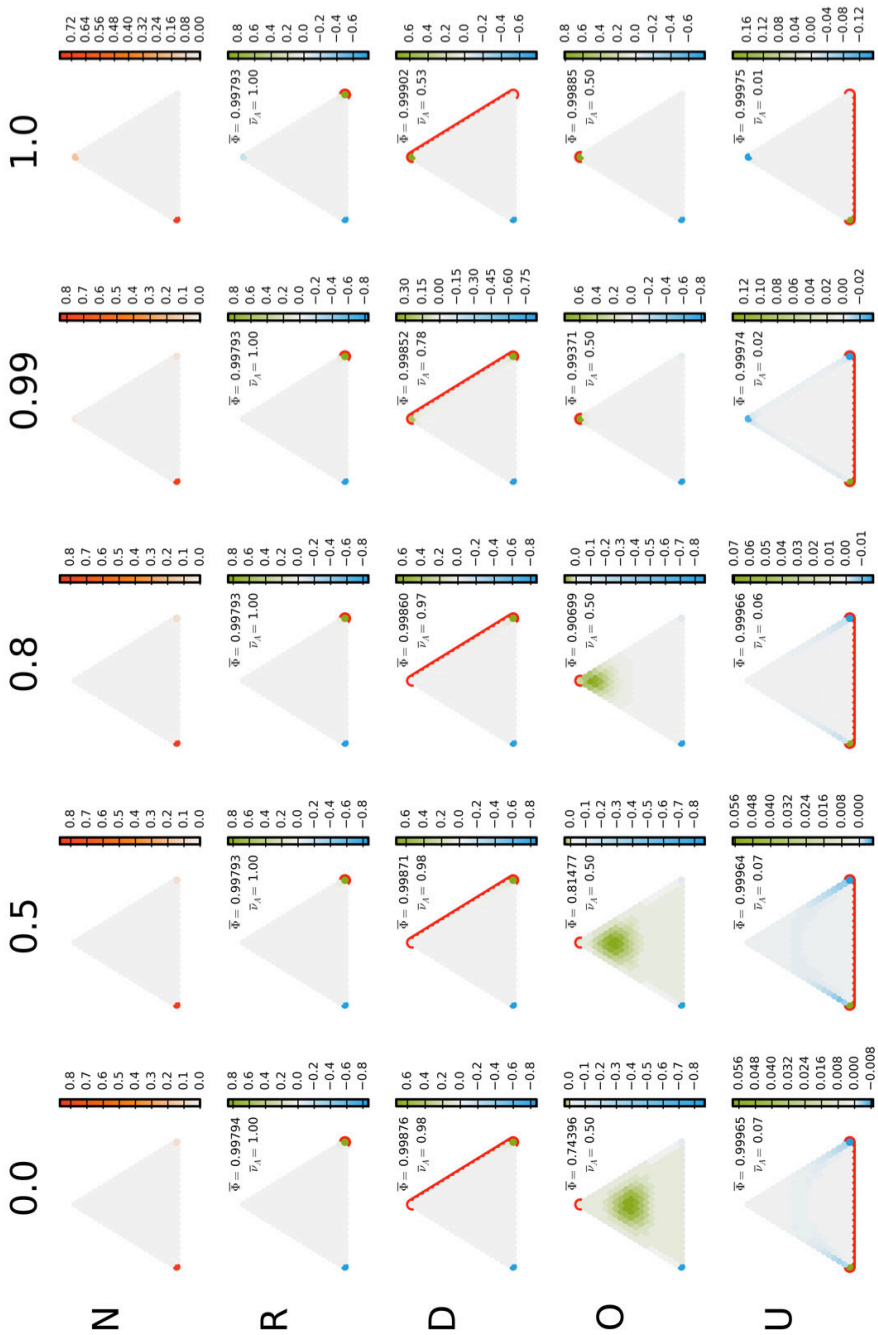

Figure A3.1 Equilibrium genotype frequencies under selection compared to neutral expectation. Neutrality and four selection scenarios (see table 1) with $s = 0.5$ and five rates of partial clonality c $(0.0, 0.5, 0.8, 0.99, 1.0)$ in a population with $N = 20$ and $n = 10$ alleles. $\overline{\Phi}$: mean population fitness at the equilibrium, $\overline{\nu_A}$: mean frequency of allele *A* at the equilibrium. Colors: state probability at neutrality (orange/red, scale: probability) and combinations of genotype frequencies that are more probable (green), ca. equally probable (grey) and less probable (blue) under selection than at neutrality (scale: difference in probability). *De Finetti* diagrams with genotype *aa* in the lower left corner (additional file 2); red brackets enclose states of maximal mean fitness.

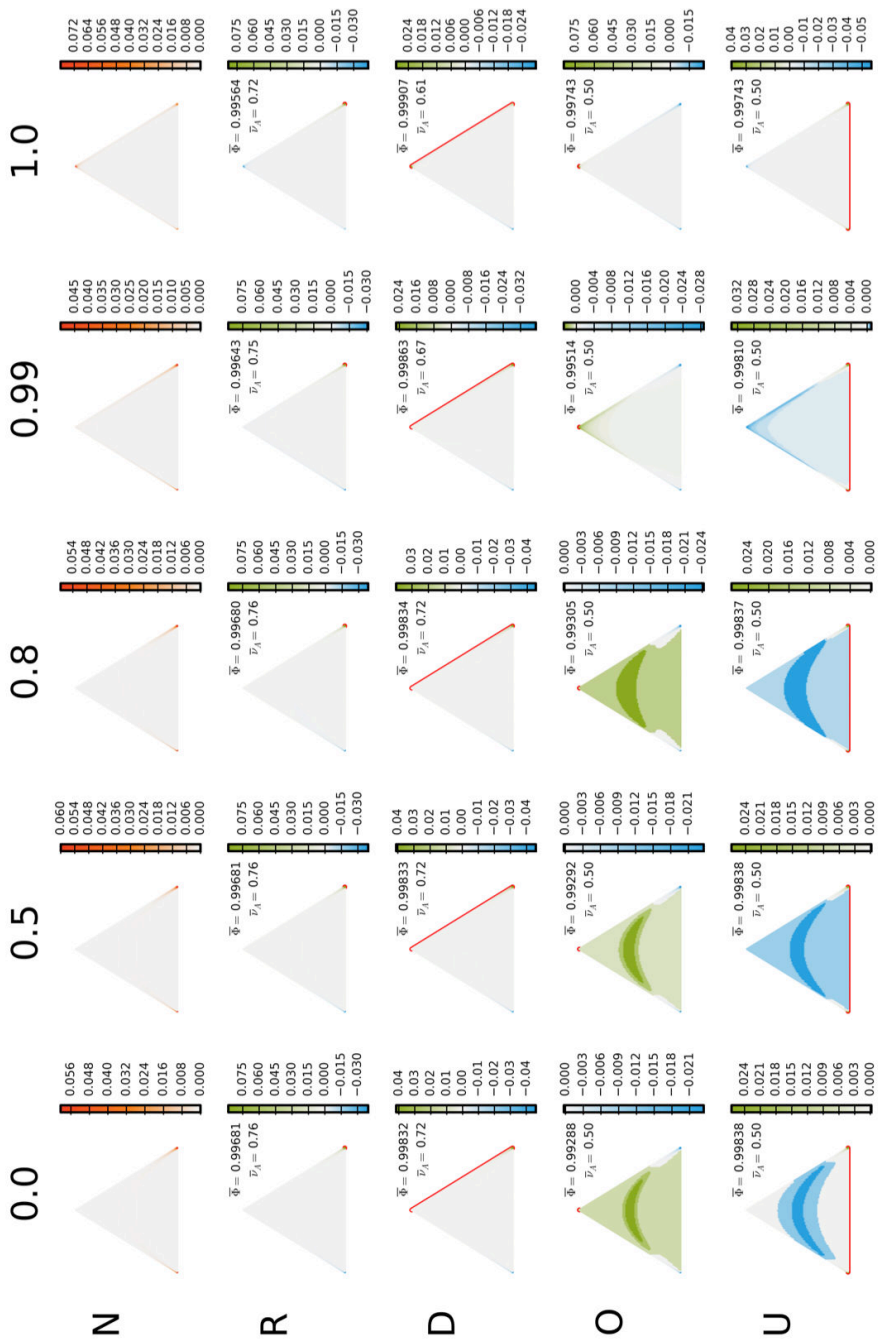

Figure A3.2 Equilibrium genotype frequencies under selection compared to neutral expectation. Neutrality and four selection scenarios (see table 1) with $s = 0.01$ and five rates of partial clonality c $(0.0, 0.5, 0.8, 0.99, 1.0)$ in a population with $N = 100$ and $n = 2$ alleles. $\overline{\Phi}$: mean population fitness at the equilibrium, $\overline{\nu_A}$: mean frequency of allele *A* at the equilibrium. Colors: state probability at neutrality (orange/red, scale: probability) and combinations of genotype frequencies that are more probable (green), ca. equally probable (grey) and less probable (blue) under selection than at neutrality (scale: difference in probability). *De Finetti* diagrams with genotype *aa* in the lower left corner (additional file 2); red brackets enclose states of maximal mean fitness.

7.2 Linkage

Having so far looked only at single loci, we neglected one other effect of clonality on genetic diversity: the increase co-inheritance of genotypes at different loci, as measured by indices of linkage disequilibrium (Halkett et al. 2005). *Linkage disequilibrium* means that the association of alleles at two different loci is not random: as an example, consider two loci with two possible alleles *a/A*, *b/B* each. If the alleles are transmitted randomly and independently to the offspring, the frequencies of the four possible haplotypes (combinations of alleles at locus *a* and *b*) should equal the product of the allele frequencies at each locus (e.g. $\nu_{a+b} = \nu_a \nu_b$). However, if the two loci are close to each other on the same chromosome (see figure 7.1), resulting in no segregation during meiosis and only a low probability of crossing-over changing the association of alleles, a situation called "physical linkage", the haplotype frequencies will differ, i.e. be in linkage disequilibrium. Physical linkage is, however, not the only mechanism that leads to linkage disequilibrium: others include selection ("functional linkage") and clonal reproduction. Because of this connection between linkage, selection and clonality, many population genetic models (e.g. Roze 2014) and theories (e.g. Felsenstein 1974) about adaptation under different rates of clonality involve at least two loci.

Figure 7.1 A pair of homologous chromosomes, showing a double heterozygote genotype at two physically linked loci *a* and *b*. *a/A, b/B:* alleles; *r:* genetic distance between the two loci (effective rate of crossing-over).

Under exclusively clonal reproduction, the offspring genotype is identical to that of its parent except for mutations, which means that only mutation may change the association of alleles at the two loci and lead to the disappearance of an initial linkage disequilibrium. Yet as we have seen for the association of alleles at single loci (F_{IS}, article II), the randomization of initial non-random associations by mutation can take very long. Under partial asexuality, the different mechanisms of recombination (reassortment of chromosomes, crossing over) can contribute to this randomization, but may not be as effective as in an exclusively sexual population (compare de Meeûs & Balloux 2004). Here, we quantified the expected effect of partially clonal reproduction on the presence and strength of linkage disequilibrium.

We compared the probability p_{co} that a pair of alleles at two different loci within the same chromosome of a diploid organism is co-transmitted to its offspring, under different rates of

partially clonal reproduction c, to the expectation for p_{co} under exclusively sexual reproduction by random mating. In doing so, we took into account different "genetic distances" between the two loci, measured by the effective rate *r* of meiotic crossing-over (i.e. excluding double, quadruple etc. crossing-over events) between them, and symmetric mutation at rate μ at both loci.

The equation for p_{co} is one minus the probability of a mutation at either locus, minus the probability of a meiotic crossing-over between both loci, i.e.

$$p_{co} = 1 - 2\mu - (1 - c)r$$

This relationship is plotted in figure 7.2; note that mutation has only a very small influence on p_{co}, so that the decay of linkage disequilibrium due to mutation only in exclusively clonal populations would take very long (similar to the results for F_{IS}, article II).

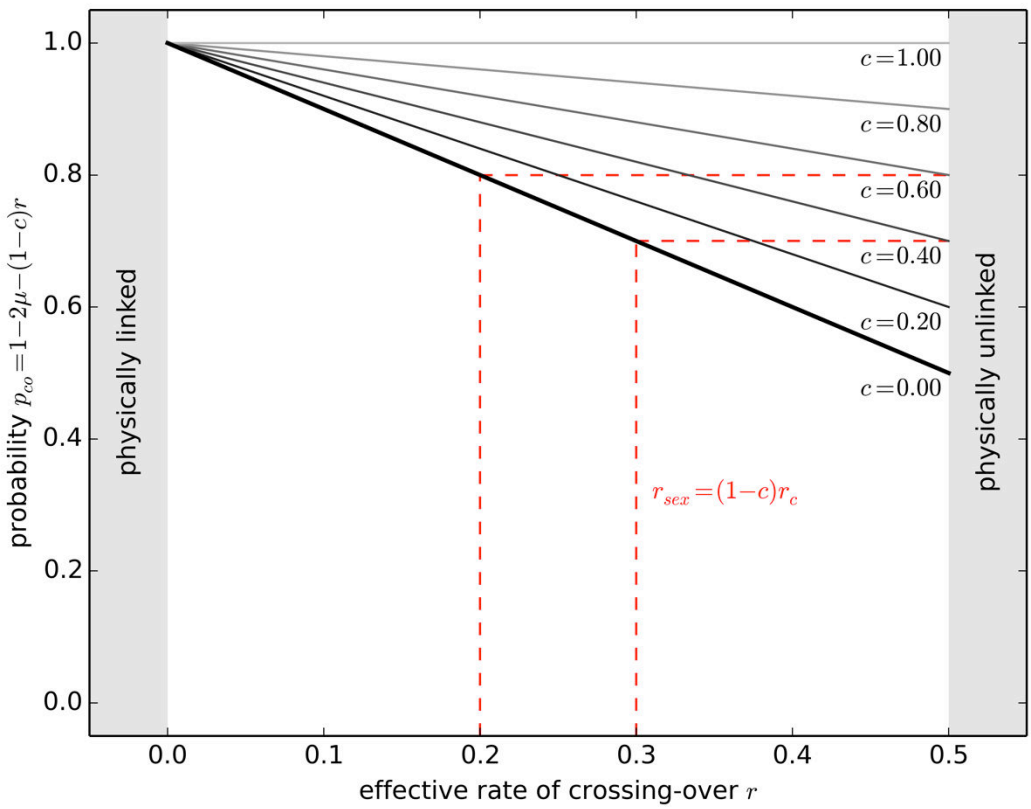

Figure 7.2 Probability of the co-inheritance of the alleles at two loci with different genetic distance (effective rate of crossing-over r) under different rates of clonality c. See text for details.

Comparing p_{co} for arbitrary c to $p_{co}(c = 0)$ for exclusively sexual reproduction gives the simple linear relationship

$$r_{sex} = (1 - c)r_c \text{ or } c = 1 - r_{sex}/r_c$$

As an example, two physically unlinked loci in a population with 60% clonal reproduction will show exactly the same pattern of co-inheritance as two loci in an exclusively sexual

population which have an effective rate of crossing over of 0.2, assuming the same mutation rate in each case (figure 7.2). Thus, the expectation for the behavior of different linkage disequilibrium estimators can be directly transferred from the usually better-known exclusively sexual case to the less well-known partially asexual case (figure 7.3; compare de Meeûs & Balloux 2004).

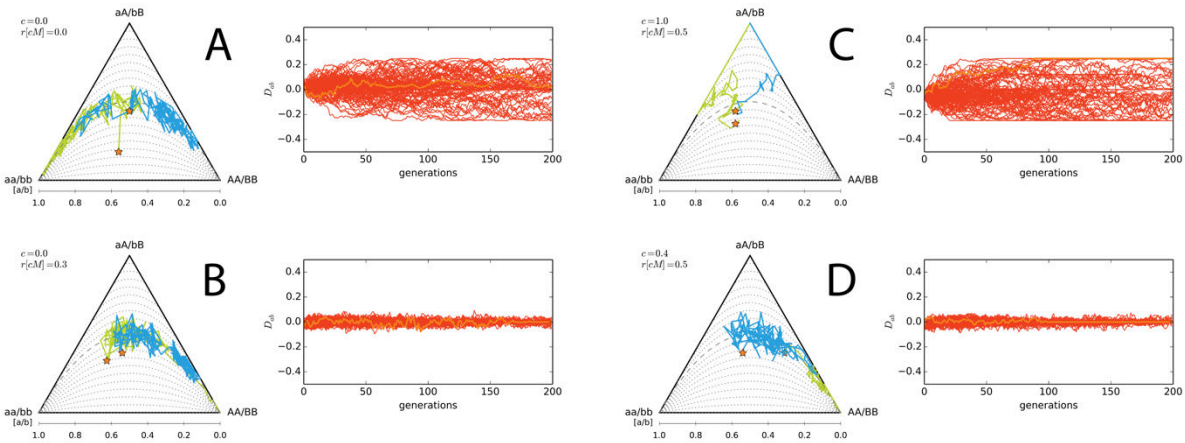

Figure 7.3 Comparison of the two-locus linkage disequilibrium $D_{ab} = \nu_{ab} - (\nu_a \nu_b)$ (see Weir 1996) for physically unlinked loci under (partial) asexuality (C,D) and corresponding physically (partially) linked loci under exclusive sexuality (A,B). A: $c = 0, r = 0$ B: $c = 0, r = 0.3$ C: $c = 1.0, r = 0.5$ D: $c = 0.4, r = 0.5$. Random start (orange stars) and 100 simulations each, orange line corresponds to example population (green and blue line: genotype frequencies at locus *a* and *b*, respectively) shown in *de Finetti* diagram.

Based on this result, it appears as though *c* in a partially clonal population could easily be measured by comparing the effective rates of crossing-over between pairs of loci under exclusively sexual reproduction (e.g. from crossing experiments as for the genetic mapping of genomes) to those observed "in the wild", i.e. for the whole partially clonal population. However, firstly estimating r_{sex} will only be possible if sexual and clonal offspring can be easily distinguished (e.g. for plants with vegetative reproduction, but exclusively sexual seed formation). Secondly, estimating r_c from field data is usually not straightforward, as all parentage relationships would have to be known – this is effectively the same as already knowing the rate of clonality, at least for one generation. Substituting the effective rates of crossing-over by estimators of linkage disequilibrium (similar to the approach proposed in de Meeûs & Balloux 2004, Halkett et al. 2005) may improve the situation only partially: all estimators suffer from the problem that the linkage phase (i.e. which pair of alleles at the two loci *a* and *b* is actually on the same chromosome; alleles *a*+*b* vs. alleles *a*+*B* in figure 7.1) in double heterozygote genotypes is unknown. As the exact haplotype frequencies can therefore not be calculated (compare discussion in Weir 1996), estimators of linkage disequilibrium potentially introduce a significant error (compare results of de Meeûs & Balloux 2004).

As we have also seen in previous multilocus simulation results (chapter 6.2), the changed pattern of co-inheritance due to partial asexuality appears not to affect the predicted outcome for the range of F_{IS} values under neutral conditions: starting with a random association of alleles between loci (no linkage disequilibrium), the direction and strength of

subsequently acquired disequilibria is also random. For physically unlinked loci, the effect of partial asexuality on the genotype frequency / F_{IS} dynamics at single loci and on the observed linkage disequilibrium should still be independent; a potential influence of co-inheritance on the distribution of F_{IS} values across loci, in particular on its mean, has not yet been quantified (compare discussion about the reliability of the mean F_{IS} across loci in article II).

Our result also has implications for linkage disequilibrium-based methods to detect genomic regions under selection (compare Vitti et al. 2013): the higher the rate of clonality, the more difficult it will become to delimit selected regions as the "background" linkage disequilibrium in partially clonal populations is already high. In an extreme example, under exclusive clonality, a selective sweep caused by a beneficial mutation at one locus may lead to the predominance of the whole genome in which the mutation first occurred within the population. As the randomization of alleles at the non-selected loci by mutation is slow, the particular locus that caused the sweep would retrospectively be virtually undetectable. Time series data, as previously discussed for F_{IS} (article II), might also be helpful in this situation: by comparing allelic diversity before and after the sweep, at least some loci could be excluded. However, this assumes either luck or clairvoyance, as data collection would have to start before the actual event (selective sweep) takes place.

To conclude, the pattern of linkage disequilibrium in partially asexual species, brought about by the *en bloc* inheritance of the parental genome under asexual reproduction, can be easily predicted based on a linear relationship to what is known about linkage in exclusively sexual populations. Applying this knowledge to estimate rates of clonality may be technically difficult, and partially clonal reproduction may interfere with linkage disequilibrium-based methods for detecting selection. Time series data could resolve some of these difficulties.

Part IV Discussion and conclusion

8 Synthesis

8.1 Main findings

8.1.1. General remarks

Based on a mathematical model, this thesis demonstrated that the patterns and dynamics of genetic diversity in partially asexual populations are indeed generally different from those expected under exclusively sexual or exclusively asexual reproduction. Moreover, we described what these differences are and how they originate in a basic genetic system. The results and conclusions reached in this thesis are a step towards a general extension of population genetic theory for partially asexual species.

This thesis showed that life cycles matter: Genotype frequency dynamics in populations that may produce sexual and asexual offspring in parallel are different from those of populations cyclically alternating between sexual and asexual reproduction. This adds another level of complexity to the development of population genetic theory for partially asexual species, as not only different rates of clonality, but also different partially asexual life cycles need to be accounted for (article III).

Time is an important factor for the population genetics of partially asexual species. We advocate a dynamic view of population genetic processes, as it allows treating complex life cycles (such as cyclical parthenogenesis) as a composite of less complex parts (single steps of sexual/asexual reproduction). To interpret population genetic data, a description of short-term change could be more helpful than long-term equilibria, since natural populations may often be confronted with an inconstant environment.

The results of this thesis suggest that allele frequencies alone are not sufficient to describe some specific problems in population genetics. We based our model of the population genetics of partially asexual species on genotype frequencies, which allowed a description of genetic drift in populations without a universal gamete stage. Much of population genetics is currently dominated by an “allele frequency” view (compare e.g. Gale 1990, Ewens 2004, Wikipedia contributors 2015). This view may have to be revised to accommodate partial asexuality (see also Ceplitis 2003, Hartfield et al. 2015).

Remarques générales

Basée sur un modèle mathématique, cette thèse a démontré que les caractéristiques et la dynamique de la diversité génétique dans les populations partiellement asexuées sont en effet généralement différentes de celles attendues avec la reproduction exclusivement sexuée ou exclusivement asexuée. En outre, nous avons décrit ce que ces différences sont et comment elles se produisent dans un système génétique de base. Les résultats et les conclusions de cette thèse sont une étape vers une généralisation de la théorie de génétique des populations pour les espèces partiellement asexuées.

Cette thèse a montré que les cycles de vie sont importants : la dynamique de la fréquence du génotype dans les populations qui peuvent produire une descendance sexuée et asexuée en parallèle est différente de celles des populations qui alternent cycliquement entre la reproduction sexuée et asexuée. Cela ajoute un autre niveau de complexité à l'élaboration de la théorie de la génétique des populations pour les espèces partiellement asexuées, parce que les taux de clonalité sont non seulement différents, mais aussi les différents cycles de vie partiellement asexués doivent être pris en compte (article III).

Le temps est un facteur important pour la génétique des populations des espèces partiellement asexuées. Nous préconisons une vision dynamique des processus de génétique des populations, car elle permet de traiter les cycles de vie complexes (comme la parthénogenèse cyclique) comme un composite de pièces moins complexes (les étapes simples de la reproduction sexuée / asexuée). Pour interpréter les données de génétique des populations, une description des changements à court terme pourrait être plus utile que les équilibres à long terme, étant donné que les populations naturelles peuvent souvent être confrontées à un environnement changeant.

Les résultats de cette thèse suggèrent que les fréquences alléliques ne suffisent pas pour décrire certains problèmes spécifiques en génétique des populations. Nous avons basé notre modèle de la génétique des populations des espèces partiellement asexuées sur les fréquences génotypiques, ce qui a permis une description de la dérive génétique dans les populations sans phase gamétique universelle. Une grande partie de la génétique des populations est actuellement dominée par une vue centrée sur les fréquences alléliques (comparer par exemple Gale 1990, Ewens 2004, Wikipedia contributors 2015). Ce point de vue doit être révisé pour tenir compte de l'asexualité partielle (voir également Ceplitis 2003, Hartfield et al. 2015).

8.1.2. Genotype frequency dynamics under partial asexuality

In acyclic partially asexual populations, the attraction of the Hardy-Weinberg equilibrium is weakened. As the associations of alleles are only partially randomized at each generation, reaching complete randomness takes longer. This effect, which was first described by Marshall & Weir (1979), is the common cause of all changes in the genotype frequency dynamics of acyclic partially asexual populations, compared to their exclusively sexual (random mating) counterparts. It contrasts with the consequences of other deviations from random mating, such as partial selfing: there, the equilibrium for the association of alleles (heterozygosity) is shifted, but the speed of approach remains the same (Marshall & Weir 1979).

The observed changes in genotype frequency dynamics under partial clonality depend on evolutionary processes other than reproduction. The results of Marshall & Weir (1979) probably gave the impression that "nothing changes" if asexual reproduction is added to a randomly mating system, since the equilibrium stays the same. However, this is no longer true if other evolutionary processes leading away from the Hardy-Weinberg equilibrium (in our case, genetic drift and/or selection) are taken into account.

The relationship between the rate of asexuality and its effect on a population genetic parameter may differ. Consequently, not only populations with an extremely high rate of clonality are different from the exclusively sexual case. For the mean F_{IS} at selectively neutral loci under acyclic partial clonality, we found a hyperbolical relationship (article II). Yet the times to adaptation under different selection scenarios echoed this pattern only in part (article IV). The probability of co-inheritance (chapter 7.2) appears to be linearly correlated to the rate of clonality, yet parameters measuring linkage disequilibrium may follow a different scheme (compare de Meeûs & Balloux 2004). As the concomitant effect of different evolutionary processes can be very different for populations that do not have to conform to the Hardy-Weinberg equilibrium, the effect of partial asexuality on each population genetic parameter should be modeled individually.

Under selectively neutral conditions, partial asexuality increases the "evolutionary memory" of populations for past events that affected their genotypic diversity. We showed that bottleneck effects take much longer to wear off under partial asexuality (chapter 6.2), and that populations originating from hybridization or a highly homozygous population may take very long to reach their expected equilibrium distribution of F_{IS} values (article II). In cyclically parthenogenetic populations, the number of asexual generations in each cycle and the sampling time determine how similar the observed distribution of F_{IS} values is to the expectation for exclusive sexuality (article III).

Adaptation to selection at a single locus is neither always fastest in exclusively sexual populations, nor under partial asexuality. We showed that the rate of clonality that "optimizes" the time to adaptation depends on the selective scenario, the mutation rate and the genotypic diversity previous to a selective sweep (article IV). Moreover, selection may leave different signatures in exclusively sexual and partially asexual populations, which make it more difficult to be detected in genomic data: distances over which linkage phenomena can be observed are extended (chapter 7.2), and different genotype frequencies may be selectively enhanced (article IV).

La dynamique des fréquences génotypiques sous asexualité partielle

Dans les populations partiellement asexuées acycliques, l'attraction à l'équilibre de Hardy-Weinberg est affaiblie. Comme les associations d'allèles ne sont que partiellement randomisées à chaque génération, atteindre la mixité aléatoire complète prend plus de temps. Cet effet, qui a d'abord été décrit par Marshall & Weir (1979), est la cause commune de tous les changements en dynamique des fréquences génotypiques des populations partiellement asexuées acycliques, par rapport à leurs homologues exclusivement sexués (avec accouplement aléatoire). Elle contraste avec les conséquences d'autres écarts par rapport à l'accouplement au hasard, comme l'autofécondation partielle : là, l'équilibre de l'association des allèles (hétérozygotie) est décalé, mais la vitesse de l'approche reste la même (Marshall & Weir 1979).

Les changements observés dans la dynamique des fréquences génotypiques sous clonalité partielle dépendent de processus évolutifs autres que la reproduction. Les résultats de Marshall & Weir (1979) ont probablement donné l'impression que « rien ne change » si la

reproduction asexuée est ajoutée à un système d'accouplement au hasard, car l'équilibre reste le même. Toutefois, cela n'est plus vrai si d'autres processus évolutifs menant loin de l'équilibre de Hardy-Weinberg (dans notre cas, la dérive et / ou la sélection génétique) sont pris en compte.

La relation entre le taux de l'asexualité et son effet sur un paramètre de génétique des populations peut différer. Par conséquent de ce résultat, les populations avec un taux de clonalité intermédiaire (et non seulement ceux avec un taux de clonalité extrêmement élevé, voir les exclusivement clonales / asexuées) peuvent également différer du cas exclusivement sexué. Pour la moyenne d'F_{IS} aux locus sélectivement neutres sous clonalité partielle acyclique, nous avons trouvé une relation hyperbolique (article II). Pourtant, les durées d'adaptation dans différents scénarios de sélection font en partie seulement l'écho de ce principe (article IV). La probabilité de co-héritage (chapitre 7.2) semble être linéairement corrélée aux taux de clonalité, mais les paramètres de mesure du déséquilibre de liaison peuvent encore suivre un régime différent (comparer de Meeûs & Balloux 2004). Comme l'effet concomitant des différents processus de l'évolution peut être très différent pour les populations qui ne se conforment pas nécessairement à l'équilibre de Hardy-Weinberg, l'effet de l'asexualité partielle sur chaque paramètre de génétique des populations doit être modélisé individuellement.

Dans des conditions sélectivement neutres, l'asexualité partielle augmente la « mémoire évolutive » des populations pour les événements passés qui ont affecté leur diversité génotypique. Nous avons montré que les effets de goulot d'étranglement prennent beaucoup plus de temps à se dissiper sous l'asexualité partielle (chapitre 6.2), et que les populations originaires de l'hybridation ou une population très homozygote peuvent prendre beaucoup de temps pour atteindre leur distribution des valeurs d'F_{IS} attendues à l'équilibre (article II). Dans les populations cycliquement parthénogénétiques, le nombre de générations asexuées dans chaque cycle et le moment d'échantillonnage déterminent le degré de similitude de la distribution observée des valeurs d'F_{IS} et des attentes à partir de la sexualité exclusive (article III).

L'adaptation à la sélection à un seul locus n'est pas toujours la plus rapide, ni dans les populations exclusivement sexuées, ni sous asexualité partielle. Nous avons montré que le taux de clonalité qui « optimise » la durée de l'adaptation dépend du scénario sélectif, du taux de mutation et de la diversité génotypique avant un balayage sélectif (article IV). En outre, la sélection peut laisser des signatures différentes dans les populations exclusivement sexuées et partiellement asexuées, ce qui la rend plus difficile à détecter dans les données génomiques : les distances sur lesquelles les phénomènes de liaison peuvent être observés sont étendues (chapitre 7.2), et différentes fréquences génotypiques peuvent être augmentées de manière sélective (article IV).

8.1.3. Meselson effect

The results for exclusively clonal populations presented in this thesis turned out to be different from those presented elsewhere. We therefore discussed partially asexual populations mainly with reference to exclusive sexuality. However, the genotype frequency dynamics in exclusively clonal populations are both interesting in themselves and as a reference for partially asexual populations, which is why they shall be discussed here.

According to our model, genotype frequency dynamics in small exclusively clonal populations are dominated by genetic drift, and in large clonal populations by mutation. The distinction between sizes is based on the relation between population size and mutation rate (see article II). In the first case, exclusively asexual populations successively loose any initial genotypic diversity they may have had. It is this scenario which leads to highly negative mean F_{IS} values, modeled both by Balloux et al. (2003) (compare also their figure 6, which shows that the effective numbers of alleles and genotypes obtained by simulation converge to two and one, respectively; i.e. a single heterozygous genotype) and us. In the second case, all loci will eventually be at Hardy-Weinberg equilibrium ($F_{IS} = 0$) with the allele frequencies that result from the underlying mutation scheme: as an example, the observed homozygosity at each SNP in an exclusively clonal population should converge to $4 \cdot (1/4)^2 = 1/4$ (or one half for a biallelic SNP), assuming the Jukes-Cantor (or k-alleles) mutation model. This means that very old, large and exclusively asexual populations should be virtually indistinguishable from exclusively sexual populations, based only on genetic data.

This result contrasts with the popular "Meselson effect" hypothesis about genetic diversity in exclusively clonal populations. The hypothesis predicts that, as each allelic copy in an exclusively clonal population "accumulates" mutations independently and there is no genetic exchange between individuals, the two homologous allelic copies within each individual may become much more different than in an exclusively sexual population (Birky 1996, Mark Welch & Meselson 2000). In consequence, very old and exclusively clonal populations should be completely heterozygous and have $2N$ highly different alleles. A "Meselson effect" at almost exclusive clonality is part of the results of all coalescence models for the population genetics of partially clonal populations (Bengtsson 2003, Ceplitis 2003, Hartfield et al. 2015).

What is the basis of this discrepancy? A comparison of our model with the published coalescence models of partial clonality reveals a number of differences in the assumptions. In the coalescence models, the population size is assumed to be comparatively large, while our model does not impose such limitations. The time scales considered for coalescence to the most recent common ancestor of all alleles/individuals may be much longer than those in our model, which is based only the common ecological definition of a population (see chapter 3.1) without making any specific asumptions about its origin. Finally, the mutation schemes are very different: our model assumes memory-less mutation between a finite number of alleles at a single SSP or SNP locus, i.e. including back mutations, whereas the coalescence models are based on infinite alleles / infinite loci assumption, i.e. no back mutations and/or homoplasy between alleles – which correspond to DNA sequences – are

possible. This difference between the mutation schemes appears to be the most likely explanation why our model does not produce a "Meselson effect".

We hold that none of the models proposed so far adequately describes sequence divergence in natural populations, as would be required to test the "Meselson effect" hypothesis. With our model of single loci, we cannot directly say how many "multilocus alleles" (haplotypes) there would be in an exclusively sexual or exclusively clonal population. Assuming that each locus has the same mutation rate, a sequence length/number of neutrally polymorphic loci of more than the inverse of the per-locus mutation rate would ensure that, at the equilibrium of the mutation process, each haplotype within a population is different from all others at least at one individual locus, with only a small chance to create the same haplotype by two independent mutations. However, this consideration applies regardless of the reproductive system. In contrast, the coalescence models with their infinite alleles/loci lead to the unrealistic result that, after a long enough time, two randomly picked sequences should be 100% divergent. However, as discussed by Birky (1996), sequence divergence can never exceed 75%, as there is a probability of 0.25 that two DNA bases are identical by chance. Though the probability of back mutations can be very low for some mutation processes, such as sequence inversions or deletions, these lead to homology problems (Rivas & Eddy 2008) and are usually not considered in population genetic studies.

An "accumulation of mutations" scenario in exclusively clonal populations can be observed for one special case, a single clonal lineage of recent origin. Here, the results from our model and the coalescence models coincide, as back mutations will be rare within a short time span. However, the increasing divergence among the offspring haplotypes corresponds to the dynamics after a demographic bottleneck, i.e. it merely restores the equilibrium allelic diversity. Considering a finite alleles model, mutations cease to "accumulate" as soon as the associations between alleles (within & between loci) are "randomized".

8.2 Practical implications

This thesis established a reference for the interpretation of genotype frequencies in partially asexual populations. However, it also showed that, based on the standards that have been developed for exclusively sexual populations, the sampling methods for partially asexual organisms should be adapted.

In contrast to exclusively sexual populations, the changed genotype frequency dynamics and increased variation under acyclic partial asexuality make the expectations for F_{IS} less clear-cut and harder to test. Although it seems counter-intuitive (after all, "clones" should all be similar and thus provide no new information; compare "saturation" graph figure B1.1 in Arnaud-Haond et al. 2007), collecting more data is the primary solution for this problem. Especially when the rate of clonality is expected to be high, exhaustive sampling of individuals and analyzing as many polymorphic loci as possible would be ideal to make sure that the population's genetic diversity is not underestimated. This may not be possible in all cases; still, the higher the precision of the measured data (e.g. mean F_{IS}), the more statistical power any test will have.

To compare field data to our results, F_{IS} should be calculated with each sample/ramet included. Our model is based on single loci (equivalent to the "unlinked" loci in Balloux et al. 2003), so that the overall frequencies of genotypes based on several loci (multilocus genotypes) cannot be inferred with it. Removing apparent "duplicates" from the data thus introduces a random bias, which may moreover increase with the rate of clonality (due to the increased probability of co-inheritance, chapter 7.2).

Sampling time is important for population genetic studies of partially asexual organisms: for cyclical parthenogenesis, we demonstrated that data collected just before and just after the sexual generation may be widely different (article III). Consequently, the sampling time relative to the organism's life cycle should be kept as similar as possible across different populations, and should be published with the results to allow repeatability and meta-analyses. In acyclic partial asexuals, the relation between past demographic events (e.g. change of rate of asexuality, change of population size) and the sampling time will determine if population history has to be considered when interpreting the currently observed genotype frequencies.

The time dependence of genotype frequencies under partial asexuality suggests another sampling technique: time series of data from the same population. Though population genetics is explicitly concerned with the change of genotype frequencies, time series and temporal comparisons are not very commonly used. Our results suggest that they may be particularly useful in partially asexual populations, to discriminate between historic and ongoing processes. With our model, it is possible to estimate an average rate of clonality based on two or more successive samples if the number of generations, population size and mutation rate stay constant.

Finding a way to estimate the rate of clonality from population genetic data is a strong motive for the development of population genetic theory. This rate is often not directly observable, either because of technical constraints (e.g. parasites reproducing inside their host) or because the population size is too big. Our results at once give hope and point out potential pitfalls: though time series data may help to provide more reliable estimates, the increased sampling effort it requires both for reliably estimating genotype frequencies and for collecting the time series data itself may make this method either impractical or inaccurate. Combining estimates from multiple population genetic parameters, e.g. F_{IS}, linkage disequilibrium, clonal heterogeneity or frequency distribution of multilocus genotypes (compare Halkett et al. 2005, Arnaud-Haond et al. 2007), may be a way to improve accuracy. However, it could also be that the ranges for which the discriminative power of each parameter is lowest coincide (e.g. low rates of clonality – rare clonal offspring has to be found), so that some cases remain indistinguishable.

8.3 Contribution to evolution of sex debate

Though explaining the evolution of sex was not the primary aim of this thesis, we nonetheless produced some results that could be of interest in this context. The debate has traditionally revolved around models where each individual reproduces either only sexually

or only asexually (e.g. Agrawal & Chasnov 2001, Otto 2009), with one reproductive mode eventually taking over. However, this scenario may be rare in nature (Lovell et al. 2014); transitions from exclusively sexual to exclusively + partially asexual (e.g. Koltunow et al. 2011) or from partially clonal to partially clonal + exclusively clonal (e.g. Jaquiéry et al. 2014) may be much more common. As our results showed, evolution in partially clonal organisms is generally different from evolution in exclusively sexual populations, and this difference should therefore be taken into account. Also considering the long history of mitosis and meiosis in the evolution of eukaryotes (Cavalier-Smith 2002), a different question could be asked: Why is it that some species, humans among them, appear to have entirely given up their inherited potential for asexual reproduction in the first place? Or to put it more polemically: Why should humans not be cloned?

Partially asexual populations have already been used in experimental studies on the evolution of sex (e.g. Goddard et al. 2005, Becks & Agrawal 2012, Gray & Goddard 2012) though in some cases may have been confused with partial automicts (D'Souza & Michiels 2007, 2010). However, the focus appears to have been more on "cylical parthenogenesis"-like systems where sexual reproduction could be induced by an external signal. Other systems may be more difficult to handle, yet there is clearly a need for more field data also for them. Moreover, it would be interesting to know more about the exact mechanisms of the adaptation processes studied in such experimental examples – according to our results for the time to adaptation in partial asexuals based on single loci (article IV), the conclusion that "sex speeds up adaptation" may not be the whole truth. As suggested by (partially) asexual lineages that originated from hybridization (e.g. Beck et al. 2012), where the hybrid genotype had a selective advantage ("heterosis") over its parents (Grant 1976, Hörandl 2006), the advantageousness of a reproductive system could depend on the predominant selection scenario.

Our results for a mutation model with a finite number of alleles could also be of interest in the discussion about the evolution of sex. As an example, the "Muller's ratchet" hypothesis, which predicts that exclusively asexual populations may accumulate slightly deleterious dominant mutations by genetic drift, is based on the assumption that there is no back mutation. With back mutation, any slightly deleterious (almost neutral) mutation would be at least as likely (since also selected against) to disappear again, so that at any time the amount of slightly deleterious mutations within the genome should be limited. To our knowledge, such a scenario has not yet been modeled.

In our finite-alleles mutation model, mutation and random mating are similar in that they both randomize the combinations of alleles within populations, except that random mating is faster and cannot create new alleles. However, mating need not be random – to use the "card game" analogy proposed by Otto (2009), "players" (organisms) may have some way to announce some of their "cards" to each other (e.g. morphology, behavior), so that only cards of the same color (species, ecotype) are swapped and the risk of dramatically decreasing the value of the hand (offspring fitness) is limited. Moreover, diploid or polyploid organisms have the chance to retain some (typically one half) of their parental "cards" (genes) or hide some that do not fit with the current hand (gene regulation). In contrast to incessantly randomizing mutation, sexual reproduction may have evolved initially as a way to keep

basic "functioning" genotypes together (speciation) and, on a larger scale, reduce the variation of offspring. A similar hypothesis was proposed by Gorelick & Heng (2011), yet again a mathematical description, including parameter ranges for which it might apply, is still missing.

Changing scale from species to populations and individuals to cells, our results for cyclical parthenogenesis could appear in a new light: As suggested by Hastings (1991), even humans are cyclically parthenogenetic at the cell level, since our germ line cells undergo several rounds of mitosis before finally engaging in meiosis to form haploid germ cells. At the cellular vantage point, sex is only employed when useful – in multicellular life forms, only when a new organism is initiated (which raises the question why there are no natural chimaeras with somatic cells from different events of syngamy). Clearly, sex is also costly at the cellular level, and the multiple rounds of mitosis are an economical way to increase germ cell production. Cyclical parthenogenesis appears to be the repetition of this principle at a higher level, which may make "cost of sex" related models especially pertinent compared to other systems. For example, aphids or daphnia replicate organisms without sex while conditions are approximately constant and produce a sexual generation at the end of the season, thereby keeping the "cost" of reproduction low (Maynard Smith 1978). However, other cyclically parthenogenetic organisms such as trematodes use clonal reproduction to migrate between different hosts, thus potentially passing through highly different environments. To understand why only some species "transferred" the principle of cyclical parthenogenesis from the cellular to the individual level, it could be interesting to compare these different cases, especially with respect to demographic bottlenecks and selection (inconstancy of the environment).

Eventually, the evolution or maintenance of sexual and/or asexual reproduction may have more than one reason, it may have different reasons in different circumstances/species, and it may not be connected to population genetic diversity at all – for the vegetatively and sexually reproducing European beachgrass (*Ammophila arenaria*; see chapter 2.4), vegetative reproduction could be primarily an ecological asset, since it allows it to inhabit an otherwise inaccessible habitat, sand dunes. Connecting ecology and evolution by uncovering the underlying functional genetics and physiological processes may help to extend our view and make the patterns clearer. We hope that the reference provided by our model may ease this process.

8.4 Perspectives

As we deliberately wanted to keep our model simple, we did not include a number of details which may be relevant to particular partially asexual species, including e.g. sexual reproduction other than random mating (in particular selfing – though a comparison with Marshall & Weir (1979) may give an idea of the result), survival between generations and limited numbers of offspring per parent. Similar to selfing, gene conversion would increase the number of homozygotes. Including mutation schemes with unequal mutation rates between (a finite number of) alleles may change the allele frequencies at the equilibrium of mutation, but otherwise would not significantly alter the results of our model, as discussed in article II. Additional mutation during meiosis may lead to a faster convergence of the allele frequencies to their equilibrium, but should have comparatively little impact on heterozygosity. We already provided the first steps for extending our model to multiple loci, including (partial) physical linkage/co-inheritance between loci (chapter 7.2). Extensions to multiple populations connected by migration already exist (e.g. Berg & Lascoux 2000, Balloux et al. 2003), though the range of different schemes of migration (island model, stepping-stone model; unequal population sizes) is not yet fully explored. These previous results suggest that the effect of acyclic partial clonality on the final mean F_{ST} is similar to that on the final mean F_{IS} (i.e. only affected by very high rates of clonality, towards less differentiation), but that cyclical parthenogenesis can lead to increased inter-population differentiation (high F_{ST}).

An important step towards providing a population genetic reference for all partially asexual species would be an extension to other ploidy levels and life cycles. Dominantly haploid (haplontic) organisms, which typically reproduce clonally during their multicellular haploid phase, include Bryophytes and Ascomycetes. The latter include important plant pests and pathogens (e.g. *Fusarium, Ascochyta, Ophiostoma, Cryphonectria*), but also some species that are used by humans (e.g. several species of *Penicillium*). In dominantly haploid organisms, the most important signature of asexuality will be a changed pattern of linkage/co-inheritance; a comparison with already existing models for viruses (e.g. Neher 2013) and models for the evolution of sex assuming haploidy (e.g. Roze 2014) may provide valuable leads. Combining the results for haplontic and diplontic partially asexual life cycles should lead to models for complex haplodiplontic life cycles with asexual reproduction during both phases, as observed for some algae (e.g. Couceiro et al. 2015). In contrast, population genetic models for polyploid partially asexual species have only little previous work to build on (Asher & Nace 1971, Overath & Asmussen 2000b), as the theory is still not even well developed for exclusively sexual polyploids (Dufresne et al. 2014). The first step toward such a model might be an inventory of the different mechanisms of inheritance (compare figure 3.1) involved in polyploid sexual reproduction.

The most important perspective of this thesis is, however, the application of its results for the collection and interpretation of field data in the many partially asexual species. We provided an example for the impact of our results on the interpretation of mean F_{IS} values (article II). As our model is deductive, its results cannot be “proven” or “disproven” by field data; rather, such comparisons will help us to understand if the evolution of a particular

population in nature is influenced only by its reproductive system, mutation, genetic drift and selection according to the assumptions of our model, or if other, as yet unexplored evolutionary processes (e.g. non-random mating, migration) play a role as well. In this way, we hope that our results may contribute to a better understanding of the role of different reproductive systems, and the importance of genetic diversity for evolution as such.

Perspectives

Comme nous voulions délibérément garder notre modèle simple, nous n'avons pas inclus un certain nombre de détails qui peuvent être pertinents pour certaines espèces partiellement asexuées, y compris par exemple la reproduction sexuée autre que l'accouplement aléatoire (en particulier l'autofécondation – même si une comparaison avec Marshall & Weir (1979) peut donner une idée du résultat), la survie entre les générations et un nombre limité d'enfants par parent. Semblable à l'autofécondation, la conversion génique augmenterait le nombre d'homozygotes. L'inclusion des régimes de mutation avec des taux de mutation inégaux entre (un nombre fini de) allèles pourrait modifier les fréquences des allèles dans l'équilibre de la mutation, mais autrement les résultats de notre modèle ne seraient pas modifiés de manière significative, comme indiqué dans l'article II. De la mutation supplémentaire lors de la méiose peut conduire à une convergence plus rapide des fréquences des allèles à leur équilibre, mais devrait avoir relativement peu d'impact sur l'hétérozygotie. Nous avons déjà fourni les premières étapes pour étendre notre modèle à des loci multiples, y compris la liaison physique / co-héritage (partielle) entre loci (chapitre 7.2). Des extensions à des populations multiples connectées par la migration existent déjà (par exemple Berg & Lascoux 2000, Balloux et al. 2003), bien que la diversité des schémas de migration (modèle en îles, modèle « stepping stone » ; tailles des populations inégales) n'ait pas encore été complètement explorée. Les résultats précédents suggèrent que l'effet de la clonalité partielle acyclique sur la moyenne finale de l'F_{ST} est similaire à celui de la moyenne finale de l'F_{IS} (c'est-à-dire seulement affectée par des taux de clonalité très élevés, vers moins de différenciation), mais que la parthénogenèse cyclique pourrait conduire à une augmentation de la différenciation entre populations (F_{ST} haute).

Une étape importante vers une référence sur la génétique des populations de toutes les espèces partiellement asexuées serait une extension à d'autres niveaux de ploïdie et cycles de vie. Les organismes majoritairement haploïdes (haplophasiques / haplodiplophasiques avec dominance de la phase haploïde), qui se reproduisent généralement par clonage pendant leur phase haploïde multicellulaire, comprennent les bryophytes et les ascomycètes. Chez ces derniers sont inclus les ravageurs de plantes importants et des agents pathogènes (par exemple *Fusarium, Ascochyta, Ophiostoma, Cryphonectria*), mais aussi quelques espèces qui sont utilisées par les humains (par exemple, plusieurs espèces de *Penicillium*). Chez les organismes majoritairement haploïdes, la signature la plus importante de l'asexualité sera un changement des motifs de liaison / co-héritage ; une comparaison avec les modèles déjà existants pour les virus (par exemple Neher 2013) et les modèles pour l'évolution du sexe qui supposent l'haploïdie (par exemple Roze 2014) peut fournir des pistes précieuses. En combinant les résultats des cycles de vie partiellement asexués les haplophasiques et les diplophasiques devraient conduire à des modèles de

cycles de vie haplodiplophasiques complexes avec une reproduction asexuée pendant les deux phases, comme observé pour certaines algues (par exemple Couceiro et al. 2015). Par contre, les modèles en génétique des populations pour les espèces partiellement asexuées et polyploïdes n'ont pas encore fait l'objet de beaucoup de recherches sur lesquelles on pourrait se baser (Asher & Nace 1971, Overath & Asmussen 2000b), parce que la théorie concernant des polyploïdes exclusivement sexués n'est même pas encore tout à fait développée (Dufresne et al. 2014). La première étape vers un tel modèle pourrait être un inventaire des différents mécanismes de l'hérédité (comparer la figure 3.1) impliqués dans la reproduction sexuée des polyploïdes.

La perspective la plus importante de cette thèse est cependant l'application de ses résultats à la collecte et à l'interprétation des données de terrain auprès de nombreuses espèces partiellement asexuées. Nous avons fourni un exemple de l'impact de nos résultats sur l'interprétation des valeurs moyennes d'F_{IS} (article II). Comme notre modèle est déductif, ses résultats ne peuvent pas être « prouvés » ou « réfutés » par des données de terrain ; ces comparaisons vont plutôt nous aider à comprendre si l'évolution d'une population particulière dans la nature est influencée seulement par son système de reproduction, sa mutation, sa dérive génétique et sa sélection en fonction des hypothèses de notre modèle, ou si d'autres processus évolutifs (par exemple accouplement non aléatoire, migration) encore inexplorés jouent également un rôle. De cette façon, nous espérons que nos résultats pourront contribuer à une meilleure compréhension du rôle des systèmes de reproduction différents, et de l'importance de la diversité génétique dans l'évolution en tant que telle.

Part V Bibliography

Agrawal AF, Chasnov JR. 2001. Recessive mutations and the maintenance of sex in structured populations. *Genetics*. 158(2):913–917

Aitchison J, Egozcue JJ. 2005. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*. 37(7):829–850

Aristotle 384-322 BCE. 2002. *On the generation of animals*. Adelaide, Australia: eBooks@Adelaide, The University of Adelaide Library

Arnaud-Haond S, Duarte CM, Alberto F, Serrão EA. 2007. Standardizing methods to address clonality in population studies. *Molecular Ecology*. 16(24):5115–5139

Asher JH. 1970. Parthenogenesis and genetic variability. II. One-locus models for various diploid populations. *Genetics*. 66(2):369–391

Asher JH Jr, Nace GW. 1971. The genetic structure and evolutionary fate of parthenogenetic amphibian populations as determined by Markovian analysis. *American Zoologist*. 11(2):381–398

Avise JC. 2015. Evolutionary perspectives on clonal reproduction in vertebrate animals. *Proceedings of the National Academy of Sciences*. 112(29):8867–8873

Baird AH, Guest JR, Willis BL. 2009. Systematic and biogeographical patterns in the reproductive biology of scleractinian corals. *Annual Review of Ecology, Evolution, and Systematics*. 40(1):551–571

Balloux F, Lehmann L, de Meeûs T. 2003. The population genetics of clonal and partially clonal diploids. *Genetics*. 164(4):1635–1644

Barrès B, Dutech C, Andrieux A, Halkett F, Frey P. 2012. Exploring the role of asexual multiplication in poplar rust epidemics: Impact on diversity and genetic structure. *Molecular Ecology*. 21(20):4996–5008

Beck JB, Alexander PJ, Allphin L, Al-Shehbaz IA, Rushworth C, et al. 2012. Does hybridization drive the transition to asexuality in diploid *Boechera*? *Evolution*. 66(4):985–995

Becks L, Agrawal AF. 2012. The evolution of sex is favoured during adaptation to new environments. *PLoS Biology*. 10(5):e1001317

Bengtsson BO. 2003. Genetic variation in organisms with sexual and asexual reproduction. *Journal of Evolutionary Biology*. 16(2):189–199

Berg LM, Lascoux M. 2000. Neutral genetic differentiation in an island model with cyclical parthenogenesis. *Journal of Evolutionary Biology*. 13(3):488–494

Birky CW. 1996. Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics*. 144(1):427–437

Bogdanov YF. 2003. Variation and evolution of meiosis. *Russian Journal of Genetics*. 39(4):363–381

Bonen L, Bogart JP, Bi K, Fu J, Noble DWA, Niedzwiecki J. 2007. Unisexual salamanders (genus *Ambystoma*) present a new reproductive mode for eukaryotes. *Genome*. 50(2):119–136

Bonnet C. 1745. *Traité d'insectologie; ou Observations sur les pucerons*. Vol. 1. Paris: Durand

Brookes AJ. 1999. The essence of SNPs. *Gene*. 234(2):177–86

Brzyski JR, Culley TM. 2011. Genetic variation and clonal structure of the rare, riparian shrub *Spiraea virginiana* (Rosaceae). *Conservation Genetics*. 12(5):1323–1332

Cavalier-Smith T. 2002. Origins of the machinery of recombination and sex. *Heredity*. 88(2):125–141

Ceplitis A. 2003. Coalescence times and the Meselson effect in asexual eukaryotes. *Genetical Research*. 82(3):183–190

Channing A, Edwards D. 2013. Wetland megabias: Ecological and ecophysiological filtering dominates the fossil record of hot spring floras. *Palaeontology*. 56(3):523–556

Chapman H, Houliston GJ, Robson B, Iline I. 2003. A case of reversal: The evolution and maintenance of sexuals from parthenogenetic clones in *Hieracium pilosella*. *International Journal of Plant Sciences*. 164(5):719–728

Collado-Vides L. 2001. Clonal architecture in marine macroalgae: ecological and evolutionary perspectives. *Evolutionary Ecology*. 15(4-6):531–545

Couceiro L, Le Gac M, Hunsperger HM, Mauger S, Destombe C, et al. 2015. Evolution and maintenance of haploid-diploid life cycles in natural populations: The case of the marine brown alga *Ectocarpus*. *Evolution*. 69(7):1808–1822

Dajdok Z, Nowak A, Danielewicz W, Kujawa-Pawlaczyk J, Bena W. 2011. *NOBANIS – Invasive alien species fact sheet Spiraea tomentosa*. Online Database of the North European and Baltic Network on Invasive Alien Species – NOBANIS. www.nobanis.org. [access 05/02/2014]

Darwin CR. 1860. *Charles Darwin to Charles Giles Birdle Daubeny, 16 July*. Darwin Correspondence Project: letter 2869A. www.darwinproject.ac.uk. [access 23/10/2015]

Decaestecker E, Meester LD, Mergeay J. 2009. Cyclical parthenogenesis in *Daphnia*: Sexual versus asexual reproduction. In *Lost Sex*, eds. I Schön, K Martens, P Dijk, pp. 295–316. Springer Netherlands

De Finetti B. 1926. Considerazioni matematiche sull'ereditarietà Mendeliana. *Metron*. 6(1):3–41

De Finetti B. 1927. Conservazione e diffusione dei caratteri Mendeliani. Nota I. Caso panmittico. In *Rendiconti della R. Accademia Nazionale dei Lincei*, Vol. V (11-12), pp. 913–921

De Meeûs T, Balloux F. 2004. Clonal reproduction and linkage disequilibrium in diploids: A simulation study. *Infection, Genetics and Evolution*. 4(4):345–351

De Meeûs T, Balloux F. 2005. F-statistics of clonal diploids structured in numerous demes. *Molecular Ecology*. 14(9):2695–2702

De Meeûs T, Lehmann L, Balloux F. 2006. Molecular epidemiology of clonal diploids: A quick overview and a short DIY (do it yourself) notice. *Infection, Genetics and Evolution*. 6(2):163–170

De Meeûs T, Prugnolle F, Agnew P. 2007. Asexual reproduction: Genetics and evolutionary aspects. *Cellular and Molecular Life Sciences*. 64(11):1355–1372

Der R, Epstein CL, Plotkin JB. 2011. Generalized population models and the nature of genetic drift. *Theoretical Population Biology*. 80(2):80–99

Dostál O. 2015. *Mendel and his life*. Conference presentation. Research in Plant Genetics – From Mendel's Peas to the Present. Brno, Czech Republic.

D'Souza TG, Michiels NK. 2007. Correlations between sex rate estimates and fitness across predominantly parthenogenetic flatworm populations. *Journal of Evolutionary Biology*. 21:276-286.

D'Souza TG, Michiels NK. 2010. The costs and benefits of occasional sex: Theoretical predictions and a case study. *Journal of Heredity*. 101(Supplement 1):S34–41

Dufresne F, Stift M, Vergilino R, Mable BK. 2014. Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*. 23(1):40–69

Duminil J, Fineschi S, Hampe A, Jordano P, Salvini D, et al. 2007. Can population genetic structure be predicted from life-history traits? *The American Naturalist*. 169(5):662–672

Duminil J, Hardy OJ, Petit RJ. 2009. Plant traits correlated with generation time directly affect inbreeding depression and mating system and indirectly genetic structure. *BMC Evolutionary Biology*. 9(1):177

Durka W. 2002. Blüten- und Reproduktionsbiologie. In *BIOLFLOR - Eine Datenbank Zu Biologisch-Ökologischen Merkmalen Der Gefäßpflanzen in Deutschland.*, Vol. 38, eds. S Klotz, I Kühn, W Durka, pp. 133–75. Bonn: Bundesamt für Naturschutz

Estoup A, Jarne P, Cornuet J-M. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*. 11(9):1591–1604

Ewens WJ. 2004. *Mathematical Population Genetics: I. Theoretical Introduction*. 2nd ed. New York: Springer

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics*. 78(2):737–756

Fryxell PA. 1957. Mode of reproduction in higher plants. *The Botanical Review*. 23(3):135–233

Fu Y-X, Li W-H. 1999. Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theoretical Population Biology*. 56:1–10

Gale JS. 1990. *Theoretical Population Genetics*. Springer

Goddard MR, Godfray CJ, Burt A. 2005. Sex increases the efficacy of natural selection in experimental yeast populations. *Nature*. 434:636–640

Gorelick R, Heng HHQ. 2011. Sex reduces genetic variation: A multidisciplinary review. *Evolution*. 65(4):1088–1098

Grant V. 1976. *Artbildung bei Pflanzen*. Berlin / Hamburg, Germany: Verlag Paul Parey

Gray JC, Goddard MR. 2012. Sex enhances adaptation by unlinking beneficial from detrimental mutations in experimental yeast populations. *BMC evolutionary biology*. 12(1):43

Halkett F, Simon J, Balloux F. 2005. Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology & Evolution*. 20(4):194–201

Hand ML, Koltunow AMG. 2014. The genetic control of apomixis: Asexual seed formation. *Genetics*. 197(2):441–450

Hardy GH. 1908. Mendelian proportions in a mixed population. *Science*. 49–50

Hartfield M, Otto SP, Keightley PD. 2012. The maintenance of obligate sex in finite, structured populations subject to recurrent beneficial and deleterious mutation. *Evolution*. 66(12):3658–3669

Hartfield M, Wright SI, Agrawal AF. 2015. Coalescent times and patterns of genetic diversity in species with facultative sex: effects of gene conversion, population structure and heterogeneity. *bioRxiv*. 019158

Hastings IM. 1991. Germline selection: Population genetic aspects of the sexual/asexual life cycle. *Genetics*. 129:1167–1176

Helsen P, Browne RA, Anderson DJ, Verdyck P, Van Dongen S. 2009. Galapagos' *Opuntia* (prickly pear) cacti: extensive morphological diversity, low genetic variability. *Biological Journal of the Linnean Society*. 96(2):451–461

Hile SE, Yan G, Eckert KA. 2000. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. *Cancer Research*. 60(6):1698–1703

Hilton MJ. 2006. The loss of New Zealand's active dunes and the spread of marram grass (*Ammophila arenaria*). *New Zealand Geographer*. 62(2):105–120

Hojsgaard D, Klatt S, Baier R, Carman JG, Hörandl E. 2014. Taxonomy and biogeography of apomixis in angiosperms and associated biodiversity characteristics. *Critical Reviews in Plant Sciences*. 33(5):414–427

Honnay O, Bossuyt B. 2005. Prolonged clonal growth: Escape route or route to extinction? *Oikos*. 108(2):427–432

Hörandl E. 2006. The complex causality of geographical parthenogenesis. *New Phytologist*. 171:525–538

Hörandl E. 2009. Geographical parthenogenesis: Opportunities for asexuality. In *Lost Sex*, eds. I Schön, K Martens, P Dijk, pp. 161–186. Springer Netherlands

Jahn I. 2004. *Geschichte der Biologie*. 3rd ed. Hamburg, Germany: Nikol

Jaquiéry J, Stoeckel S, Larose C, Nouhaud P, Rispe C, et al. 2014. Genetic control of contagious asexuality in the pea aphid. *PLoS Genetics*. 10(12):e1004838

Judson OP, Normark BB. 1996. Ancient asexual scandals. *Trends in Ecology & Evolution*. 11(2):41–46

Kimura M. 1964. Diffusion models in population genetics. *Journal of Applied Probability*. 1(2):177–232

Kingman JFC. 1982a. The coalescent. *Stochastic processes and their applications*. 13(3):235–248

Kingman JFC. 1982b. On the genealogy of large populations. *Journal of Applied Probability*. 19:27–43

Klimeš L, Klimešová J. 1999. CLO-PLA2 – a database of clonal plants in central Europe. *Plant Ecology*. 141(1-2):9–19

Klimeš L, Klimešová J, Hendriks R, Van Groenendael J. 1997. Clonal plant architecture: A comparative analysis of form and function. In *The Ecology and Evolution of Clonal Plants*, eds. H De Kroon, J Van Groenendael, pp. 1–29. Leiden, The Netherlands: Blackhuys Publishers

Koltunow AMG, Johnson SD, Okada T. 2011. Apomixis in hawkweed: Mendel's experimental nemesis. *Journal of Experimental Botany*. 62(5):1699–1707

Le Trionnaire G, Hardie J, Jaubert-Possamai S, Simon J-C, Tagu D. 2008. Shifting from clonal to sexual reproduction in aphids: Physiological and developmental aspects. *Biology of the Cell*. 100(8):441–451

Lokki J. 1976. Genetic polymorphism and evolution in parthenogenetic animals. VII. The amount of heterozygosity in diploid populations. *Hereditas*. 83:57–64

Lovell JT, Grogan K, Sharbel TF, McKay JK. 2014. Mating system and environmental variation drive patterns of adaptation in *Boechera spatifolia* (Brassicaceae). *Molecular Ecology*. 23(18):4486–4497

Makita A. 1998. The significance of the mode of clonal growth in the life history of bamboos. *Plant Species Biology*. 13:85–92

Markov AA. 1906. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. *Proceedings of the Society of Physics and Mathematics at the University of Kazan*. 15(2):135–156

Mark Welch DB, Meselson M. 2000. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science*. 288(5469):1211–1215

Marriage TN, Orive ME. 2012. Mutation-selection balance and mixed mating with asexual reproduction. *Journal of Theoretical Biology*. 308:25–35

Marshall DR, Brown AHD. 1981. The evolution of apomixis. *Heredity*. 47(1):1–15

Marshall DR, Weir BS. 1979. Maintenance of genetic variation in apomictic plant populations. *Heredity*. 42(2):159–172

Masel J, Lyttle DN. 2011. The consequences of rare sexual reproduction by means of selfing in an otherwise clonally reproducing species. *Theoretical Population Biology*. 80(4):317–322

Maynard Smith J. 1978. *The evolution of sex*. CUP Archive

McKey D, Elias M, Pujol B, Duputié A. 2010. The evolutionary ecology of clonally propagated domesticated plants: Tansley review. *New Phytologist*. 186(2):318–332

Mendel G. 1865. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*. 4:3–47

Mendel G. 1869. Über einige aus künstlicher Befruchtung gewonnene *Hieracium*-Bastarde. *Verhandlungen des Naturforschenden Vereines in Brünn*. 8:26–31

Mogie M. 1986. Automixis: Its distribution and status. *Biological Journal of the Linnean Society*. 28:321–329

Muirhead CA, Lande R. 1997. Inbreeding depression under joint selfing, outcrossing, and asexuality. *Evolution*. 51(5):1409–1415

Neaves WB, Baumann P. 2011. Unisexual reproduction among vertebrates. *Trends in Genetics*. 27(3):81–88

Neher RA. 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annual Review of Ecology, Evolution, and Systematics*. 44(1):195–215

Neiman M, Meirmans S, Meirmans PG. 2009. What can asexual lineage age tell us about the maintenance of sex? *Annals of the New York Academy of Sciences*. 1168(1):185–200

Normark BB. 2003. The evolution of alternative genetic systems in insects. *Annual Review of Entomology*. 48(1):397–423

Nougué O, Rode NO, Jabbour-Zahab R, Ségard A, Chevin L-M, et al. 2015. Automixis in *Artemia*: Solving a century-old controversy. *Journal of Evolutionary Biology*. 28(12) :2337-2348

Ohta T, Kimura M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics*. 63(1):229–238

Orive ME. 1993. Effective population size in organisms with complex life-histories. *Theoretical Population Biology*. 44(3):316–340

Otto SP. 2009. The evolutionary enigma of sex. *The American Naturalist*. 174(s1):S1–14

Otto SP, Day T. 2007. *A biologist's guide to mathematical modeling in ecology and evolution*. Princeton University Press

Overath RD, Asmussen MA. 1998. Genetic diversity at a single locus under viability selection and facultative apomixis: Equilibrium structure and deviations from Hardy-Weinberg frequencies. *Genetics*. 148(4):2029–2039

Overath RD, Asmussen MA. 2000a. The cytonuclear effects of facultative apomixis: I. Disequilibrium dynamics in diploid populations. *Theoretical Population Biology*. 58(2):107–121

Overath RD, Asmussen MA. 2000b. The cytonuclear effects of facultative apomixis: II. Definitions and dynamics of disequilibria in tetraploid populations. *Theoretical Population Biology*. 58:123–142

Owen R. 1849. *On parthenogenesis: Or the successive production of procreating individuals from a single ovum; a discourse*. Introduction to the Hunterian Lectures on Generation and Devleopment, for the Year 1849; Delivered at the Royal College of Surgeons of England. London: J. Van Voorst

Pasteur L. 1864. On spontaneous generation. *Revue Des Cours Scientifiques*. I, 1863-64:257–264

Perron O. 1907. Zur Theorie der Matrices. *Mathematische Annalen*. 64(2):248–263

Philip VJ, Nainar SAZ. 1986. Clonal propagation of *Vanilla planifolia* (Salisb.) Ames using tissue culture. *Journal of Plant Physiology*. 122:211–215

Piry S, Luikart G, Cornuet J-M. 1999. BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. *Journal of Heredity*. 90(4):502–503

Prugnolle F, Liu H, de Meeûs T, Balloux F. 2005a. Population genetics of complex life-cycle parasites: An illustration with trematodes. *International Journal for Parasitology*. 35(3):255–263

Prugnolle F, Roze D, Theron A, de Meeûs T. 2005b. F-statistics under alternation of sexual and asexual reproduction: A model and data from schistosomes (platyhelminth parasites). *Molecular Ecology*. 14(5):1355–1365

Raven PH, Evert RF, Eichhorn SE. 2004. *Biology of Plants*. 7th ed. New York: W. H. Freeman.

Reyes-Agüero JA, Aguirre R. JR, Valiente-Banuet A. 2006. Reproductive biology of *Opuntia*: A review. *Journal of Arid Environments*. 64(4):549–85

Richards AJ. 2003. Apomixis in flowering plants: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 358(1434):1085–1093

Rivas E, Eddy SR. 2008. Probabilistic phylogenetic inference with insertions and deletions. *PLoS Comput Biol*. 4(9):e1000172

Roze D. 2009. Diploidy, population structure, and the evolution of recombination. *The American Naturalist*. 174(s1):S79–94

Roze D. 2014. Selection for sex in finite populations. *Journal of Evolutionary Biology*. 27(7):1304–1322

Roze D, Michod RE. 2010. Deleterious mutations and selection for sex in finite diploid populations. *Genetics*. 184(4):1095–1112

Ryndin A, Kirzhner V, Nevo E, Korol A. 2001. Polymorphism maintenance in populations with mixed random mating and apomixis subjected to stabilizing and cyclical selection. *Journal of Theoretical Biology*. 212(2):169–181

Sharbel TF, Voigt M-L, Corral JM, Thiel T, Varshney A, et al. 2009. Molecular signatures of apomictic and sexual ovules in the *Boechera holboellii* complex. *The Plant Journal*. 58(5):870–882

Shaw AJ, Goffinet B, eds. 2000. *Bryophyte Biology*. Cambridge University Press

Signorovitch A, Hur J, Gladyshev E, Meselson M. 2015. Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid rotifers. *Genetics*. 200(2):581–590

Silvertown J. 2008. The evolutionary maintenance of sexual reproduction: Evidence from the ecological distribution of asexual reproduction in clonal plants. *International Journal of Plant Sciences*. 169(1):157–168

Speijer D, Lukeš J, Eliáš M. 2015. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proceedings of the National Academy of Sciences*. 112(29):8827–8834

Stoeckel S, Masson J-P. 2014. The exact distributions of F_{IS} under partial asexuality in small finite populations with mutation. *PLoS ONE*. 9(1):e85228

Taylor JW, Hann-Soden C, Branco S, Sylvain I, Ellison CE. 2015. Clonal reproduction in fungi. *Proceedings of the National Academy of Sciences*. 112(29):8901–8908

Taylor JW, Jacobson DJ, Fisher MC. 1999. The evolution of asexual fungi: Reproduction, speciation and classification. *Annual Review of Phytopathology*. 37(1):197–246

Taylor TN, Taylor EL, Krings M. 2008. *Paleobotany: The biology and evolution of fossil plants*. 2nd ed. Academic Press.

Tibayrenc M, Kjellberg F, Ayala FJ. 1990. A clonal theory of parasitic protozoa: The population structures of *Entamoeba, Giardia, Leishmania, Naegleria, Plasmodium, Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proceedings of the National Academy of Sciences*. 87(7):2414–2418

Trifonov EN. 2011. Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics*. 29(2):259–266

Vallejo-Marín M, Dorken ME, Barrett SCH. 2010. The ecological and evolutionary consequences of clonality for plant mating. *Annual Review of Ecology, Evolution, and Systematics*. 41(1):193–213

Van Dijk P, van Damme J. 2000. Apomixis technology and the paradox of sex. *Trends in Plant Science*. 5(2):81–84

Van Drunen WE, van Kleunen M, Dorken ME. 2015. Consequences of clonality for sexual fitness: Clonal expansion enhances fitness under spatially restricted dispersal. *Proceedings of the National Academy of Sciences*. 112(29):8929–8936

Vanoverbeke J, De Meester L. 2010. Clonal erosion and genetic drift in cyclical parthenogens – the interplay between neutral and selective processes. *Journal of Evolutionary Biology*. 23(5):997–1012

Vitousek PM, Mooney HA, Lubchenco J, Melillo JM. 1997. Human domination of Earth's ecosystems. *Science*. 277(5325):494–499

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annual Review of Genetics*. 47(1):97–120

Von Mises R, Pollaczek-Geiringer H. 1929. Praktische Verfahren der Gleichungsauflösung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*. 9(2):152–164

Wakeley J. 2009. *Coalescent theory: An introduction*. Greenwood Village, Colorado: Roberts & Company Publishers

Wang C-N, Möller M, Cronk QCB. 2004. Altered expression of GFLO , the Gesneriaceae homologue of FLORICAULA/LEAFY , is associated with the transition to bulbil formation in *Titanotrichum oldhamii*. *Development Genes and Evolution*. 214(3):122–127

Weinberg W. 1908. Über den Nachweis der Vererbung beim Menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*. 64:368–382

Weir BS. 1996. *Genetic data analysis 2: Methods for discrete population genetic data*. 2nd ed. Sunderland, Massachusets: Sinauer Associates

Wikipedia contributors. 2015. *Population genetics*. Wikipedia, The Free Encyclopedia. www.wikipedia.org [access 28/10/2015]

Wilkins AS, Holliday R. 2008. The evolution of meiosis from mitosis. *Genetics*. 181(1):3–12

Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KHS. 1997. Viable offspring derived from fetal and adult mammalian cells. *Nature*. 385:810–813

Woese CR. 1987. Bacterial evolution. *Microbiological Reviews*. 51(2):221–271

Wright S. 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. *Genetics*. 6(2):124–143

Yonezawa K, Ishii T, Nagamine T. 2004. The effective size of mixed sexually and asexually reproducing populations. *Genetics*. 166(3):1529–1539

Yu F, Chen Y, Dong M. 2001. Clonal integration enhances survival and performance of *Potentilla anserina*, suffering from partial sand burial on Ordos plateau, China. *Evolutionary Ecology*. 15(4-6):303–318

E Author information

1987	Born in Karl-Marx-Stadt (renamed Chemnitz), Germany
1994-2006	Primary and Secondary School // Rudolfschule / Johannes-Kepler-Gymnasium in Chemnitz, Germany
2006-2012	Diplom – Biologie courses at Uppsala Universitet (Sweden), LMU Munich (Germany) // Dresden University of Technology in Dresden, Germany
2010-2011	Graduate Certificate of Science – Tropical Ecology & Conservation // James-Cook-University in Townsville, Australia
2011-2012	Diplom thesis Population biology of Laserpitium prutenicum (Apiaceae) in Eastern Saxony (Germany) and adjacent regions Senckenberg Museum of Natural History in Görlitz, Germany // Dresden University of Technology in Dresden, Germany
2012-2015	Doctoral candidate Effects of partial asexuality on the dynamics of genotype frequencies in dominantly diploid populations Institut National de la Recherche Agronomique in Le Rheu, France // Agrocampus Ouest in Rennes, France