



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2014

The measure and dynamics of genetic diversity in structured populations

ALCALA Nicolas

ALCALA Nicolas, 2014, The measure and dynamics of genetic diversity in structured populations

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_395E894D2C9C3

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'écologie et évolution

The measure and dynamics of genetic diversity in structured populations

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la Faculté de biologie et de médecine de l'Université de
Lausanne par

NICOLAS ALCALA

Master de l'Institut National des Sciences Appliquées de Lyon

Jury

Président: Prof. Pierre Goloubinof

Directeur: Dr. Séverine Vuilleumier

Co-directeur: Prof. Jérôme Goudet

Expert externe: Prof. Daniel Wegmann

Expert interne: Prof. Laurent Lehmann

Lausanne 2014



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur Prof. Pierre Goloubinoff
Directeur de thèse	Madame Dr Séverine Vuilleumier
Co-directeur de thèse	Monsieur Prof. Jérôme Goudet
Experts	Monsieur Prof. Daniel Wegmann
	Monsieur Prof. Laurent Lehmann

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Nicolas Alcala

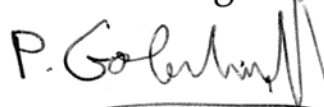
Master de l'INSA de Lyon Lyon, France

intitulée

**The measure and dynamics of genetic
diversity in structured populations**

Lausanne, le 18 juillet 2014

pour La Doyenne
de la Faculté de Biologie et de Médecine



Prof. Pierre Goloubinoff

Contents

Contents	iv
Abstract	vii
Résumé	viii
General introduction	xi
ON THE IMPORTANCE OF GENETIC DIVERSITY	XII
Genetic diversity and the diversity of Life	xii
Genetic diversity and species survival	xiii
Genetic diversity and species diversification	xiv
EVOLUTIONARY PROCESSES AND THE PREDICTED EQUILIBRIUM GENETIC DIVERSITY	XV
Mutation	xv
Natural selection	xvi
Finite population size and the genetic drift.	xvii
Population structure and migration	xviii
HISTORICAL EVENTS AND NON-EQUILIBRIUM GENETIC DIVERSITY	XIX
Demographic changes and the dynamics of genetic diversity	xxi
INFERRING HISTORICAL EVENTS FROM THEIR GENOMIC SIGNATURE	XXII
PLAN OF THE THESIS	XXIII
Chapter 1	
On the Transition of Genetic Diversity Between Isolation and Panmixia: what we can learn from G_{ST} and D	1
ABSTRACT	2
BEHAVIOR OF GENETIC DIFFERENTIATION: FROM ISOLATION TO PANMIXIA	8
Deriving the reference function $f(M)$ from H_s and H_t	9
The effect of mutation, genetic drift and the number of populations on M_T	13
WHAT CAN G_{ST} AND D ACTUALLY MEASURE?	17
USES OF G_{ST} AND D TO ESTIMATE GENETIC DIFFERENTIATION	23
USES OF G_{ST} AND D TO ESTIMATE DEMOGRAPHIC PARAMETERS	26
DISCUSSION	27
Chapter 2	
Peak and Persistent Excess of Genetic Diversity Following an Abrupt Migration Increase	37
ABSTRACT	38
GENETIC DIVERSITY OF POPULATIONS	43
PREDICTING THE DYNAMICS OF GENETIC DIVERSITY.	45
TIME TO REACH GENETIC DIVERSITY EQUILIBRIUM	46
DYNAMICS OF GENETIC DIVERSITY AFTER AN ISOLATION EVENT	48
DYNAMICS OF GENETIC DIVERSITY AFTER A CONNECTION EVENT	51
PEAK OF GENETIC DIVERSITY GENERATED BY A CONNECTION EVENT	53
PEAK OF GENETIC DIVERSITY RESULTING FROM A MIGRATION RATE INCREASE	56
IMPLICATIONS FOR THE INFERENCE OF DEMOGRAPHY AND SELECTION	56
DISCUSSION	62
APPENDIX A DYNAMICS OF GENETIC DIVERSITY	77

Chapter 3**Dynamics of Genetic Diversity Across Multiple Isolation and Connection**

Events	89
ABSTRACT	90
DYNAMICS OF GENETIC DIVERSITY UNDER PERIODIC CONNECTION AND ISOLATION EVENTS.	93
The dynamics of genetic diversities within connection and isolation periods	93
The dynamics of genetic diversities accross connection-isolation cycles	96
The duration of the transient dynamics	97
Determination of domains of the period length	99
The dynamics of the excess of within-population genetic diversity in the different domains	104
DISCUSSION.	106
ACKNOWLEDGEMENTS	110
APPENDIX A THE DYNAMICS OF GENETIC DIVERSITIES ACROSS CYCLES UNDER THE PANMICTIC CONNECTION PERIODS APPROXIMATION.	119
APPENDIX B DYNAMICS OF GENETIC DIVERSITY UNDER STOCHASTIC PERIODS OF ISOLATION AND CONNECTION	120

Chapter 4**The Signature of Past Population Isolation on Gene Genealogies and the Site Frequency Spectrum: Theory and Application to the Evolutionary Dynamics of HIV-1 subtypes**

of HIV-1 subtypes	127
ABSTRACT	128
MODEL.	131
THE SIGNATURE OF PAST ISOLATION ON PAIRWISE COALESCENCE TIMES WITHIN AND BETWEEN POPULATIONS	132
Distribution of pairwise coalescence times	132
Implications for demographic inference from gene genealogies.	136
A METRIC ON THE MINIMUM STRENGTH OF ISOLATION EVENTS	138
On the necessary lengths of isolation events	138
On the necessary strengths of isolation events	140
THE SIGNATURE OF ISOLATION ON THE SITE FREQUENCY SPECTRUM AND ITS DETECTION USING NEUTRALITY TESTS	141
The signature of past isolation events on the local SFS	142
The signature of past isolation events in the total SFS.	144
APPLICATION: DETECTION OF PAST HISTORY OF HIV-1 SUBTYPES IN CHINA	147
DISCUSSION.	150
ACKNOWLEDGMENTS	153
APPENDIX A: DERIVING OPTIMAL TESTS TO DETECT THE SIGNATURE OF ISOLATION AND PAST ISOLATION SCENARIOS FROM THE SFS	161
General discussion	165
MEASURING GENETIC DIFFERENTIATION IN EQUILIBRIUM POPULATIONS	166
THE LARGE IMPACT OF HISTORICAL EVENTS ON GENETIC DIVERSITY	167
Demographic changes and the dynamics of genetic diversity	167
Implications of demographic events on species conservation.	168
Implications of demographic changes on species adaptation.	169
Implications of demographic changes on species diversification	170
INFERRING PAST EVENTS FROM GENOMIC DATA	171

FUTURE DIRECTIONS	172
The dynamics of genetic diversity	172
Genetic diversity and selection	173
Inference of complex scenarios	174
Acknowledgements	181
Appendix A	
Supplementary Files for chapter 1	183
APPENDIX S1: THE RELATIONSHIP BETWEEN $f(M)$ AND THE EFFECTIVE NUMBER OF ALLELES	188
APPENDIX S2: THE USE OF SINGULAR VALUE DECOMPOSITION TO STUDY GE- NETIC DIFFERENTIATION	190
Appendix B	
Supplementary Files for chapter 2	193
FILE S1: GENETIC DIVERSITY EQUILIBRIUM	194
FILE S2: EFFECT OF AN ISOLATION EVENT ON GENETIC DIVERSITY	196
FILE S3: RELAXING THE COMPLETE ISOLATION HYPOTHESIS	197
Appendix C	
Supplementary Files for chapter 4	199

ABSTRACT

Genetic diversity is crucial for species adaptation and promotes species diversification. In addition, genetic diversity patterns are widely used to infer the past history of populations and to guide conservation policies. Population structure, when a set of populations are connected through migration, is known to lead to genetic differentiation and to influence the dynamics of genetic diversity. Nevertheless, how to accurately measure genetic differentiation, and what is the impact of past changes of migration rate on the dynamics of genetic diversity, are still open questions. This thesis has four aims. First, investigate the ability of commonly used measures to describe genetic differentiation. Second, assess the impact of an abrupt change of migration rate on the dynamics of genetic diversity. Third, assess the impact of multiple changes of migration rate. Fourth, investigate the inference of past migration rate changes from genetic diversity patterns. Results show that existing genetic differentiation measures are complementary, as they are accurate in different mutation regimes. In addition, it is shown that a single abrupt increase of migration rate can generate a large persistent excess of genetic diversity. Multiple changes of migration rate can either lead to a large excess of genetic diversity, or a high turnover of genetic diversity, depending on the frequency of the changes. Finally, results show that contrasting genetic diversity patterns from multiple populations allows the distinction of migration changes from other demographic processes. Results of the thesis refine our understanding of genetic diversity in structured populations, and provide guidelines to measure genetic differentiation. Moreover, the processes highlighted could influence the rate and mechanism of adaptation. Finally, results improve the inference of the demographic history of natural populations from genetic data, facilitating their management.

RÉSUMÉ

La diversité génétique est cruciale pour l'adaptation des espèces et favorise leur diversification; les patrons de diversité génétique sont utilisés pour inférer l'histoire des populations et guider les stratégies de conservation. La structure des populations (ensemble de populations connectées par la migration) crée une différenciation génétique et influence la dynamique de la diversité génétique. Néanmoins, deux questions restent en suspens: comment mesurer la différenciation génétique, et quel est l'impact de changements du taux de migration sur la dynamique de la diversité génétique. Cette thèse a quatre objectifs. (i) Etudier la capacité des mesures les plus communes à décrire la différenciation génétique. (ii) Analyser l'effet d'un changement abrupt, et (iii) de multiples changements du taux de migration sur la dynamique de la diversité génétique. (iv) Etudier l'inférence de changements passés de migration à partir des patrons de diversité génétique. Les résultats montrent que les mesures de diversité génétique existantes sont complémentaires, car exactes dans différents régimes de mutation. Une augmentation abrupte de la migration peut générer un large excès de diversité génétique sur une longue durée. De multiples changements du taux de migration peuvent soit entraîner un large excès de diversité génétique, soit un grand renouvellement de la diversité génétique, en fonction de la fréquence des changements. Enfin, comparer les patrons de diversité génétique de plusieurs populations permet de distinguer les changements de migration d'autres processus démographiques. Ces résultats améliorent notre compréhension de la diversité génétique dans les populations structurées et fournissent des recommandations pour mesurer la différenciation génétique. Les processus mis en avant pourraient influencer le taux d'adaptation et ses mécanismes. Enfin, les résultats améliorent l'inférence de l'histoire démographique des populations naturelles à partir des données génétiques, facilitant leur gestion.

"What is the mechanism by which the genetic variability at the molecular level is maintained? This problem has been regarded by some as as the most important problem currently facing population genetics." - Motoo Kimura

General introduction

ON THE IMPORTANCE OF GENETIC DIVERSITY

Genetic diversity and the diversity of Life

ORGANISMS present a tremendous diversity. Recent estimations suggest that there are approximately 8.7 million eukaryotic species on Earth, and hundreds of thousands to millions of prokaryotic species (MORA *et al.* 2011). The smallest organisms are composed of a single cell and measure less than 300 nanometers, while the largest tree, the giant sequoia, has an average height of 80 meters and is composed of hundreds of trillions of cells (BERRY 1943). We also observe an enormous diversity of species' life-spans, life-cycles, mating strategies, and social organizations. In addition, individuals from a same species can display very diverse phenotypes (e.g. flower color variation in Figure 1).

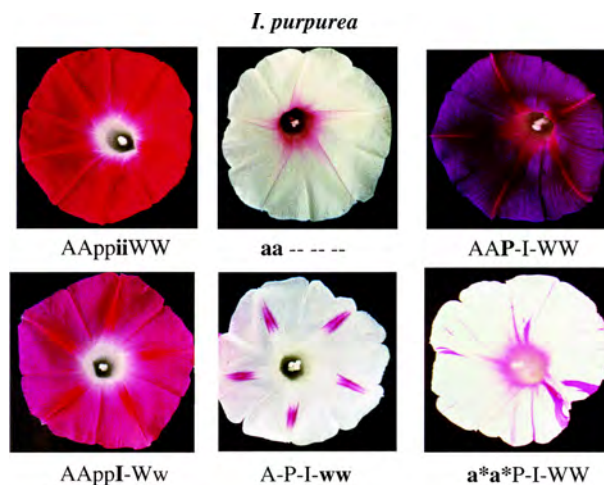


Figure 1 Flower color variation in *I. purpurea*. From CLEGG and DURBIN (2000)

Most of this huge observed diversity is due to genetic diversity. Indeed, most species traits are inherited, rather than acquired during the life-span of the individuals. The discovery that deoxyribonucleic acid (DNA) carries the hereditary information (WATSON and CRICK 1953; MESELSON and STAHL 1958) prompted the era of molecular genetics, which focuses on the variability of DNA molecules. In the fol-

lowing decades, the large diversity of genome size and gene variation was uncovered. Genome sizes can vary from a thousand of base pairs in viruses to hundreds of billions of base pairs in plants (GREGORY *et al.* 2007); similarly, the number of protein-coding genes per genome ranges from a few genes in viruses to almost 90,000 genes in eukaryotes. Within a species, individuals can display very different patterns of genetic diversity: human genes were shown to harbor a very low diversity (less than 0.1% differences in DNA sequences between two individuals from a same population), while *Drosophila* populations have a tenfold higher diversity (LI and SADLER 1991).

Genetic diversity and species survival

Genetic diversity is the raw material for species adaptation. Species with a large genetic diversity are more likely to adapt to new environments (TURNER *et al.* 1993; FEDER *et al.* 2003; COLOSIMO *et al.* 2005; HERNANDEZ *et al.* 2011; JONES *et al.* 2012). Indeed, a large genetic diversity increases the chances that an individual will carry an allele providing larger survival chances in the new environment (GIBSON and DWORKIN 2004). In addition, a large genetic diversity reduces the time to fixation of beneficial alleles (BARRETT and SCHLUTER 2008). For example, in *Drosophila melanogaster*, an allele conferring a resistance to insecticides only recently increased in frequency (during the last century), while evidence suggests it was present in the population for 90,000 years (AMINETZACH *et al.* 2005). This benefit of genetic diversity is even true when the alleles adapted to the new environment were slightly deleterious under the previous environmental conditions (GIBSON and DWORKIN 2004; HERMISSON and PENNING 2005).

Low genetic diversity can lead to population extinction. Indeed, as genetic diversity is low, individuals have an increased probability to be heterozygotes which is known

to reduce fitness by allowing the expression of recessive deleterious alleles (inbreeding depression; GILPIN and SOULE 1986; JIMENEZ *et al.* 1994; HEDRICK and KALINOWSKI 2000). For example, it was shown that inbred captive wolves suffered from reduced reproduction and longevity (LAIKRE and RYMAN 1991). In addition, populations with low genetic diversity often have a reduced effective size, which increases the odds of accumulating deleterious mutations (GILPIN and SOULE 1986; FRANKHAM 1995; LANDE 1998).

Genetic diversity and species diversification

The fast diversification of some taxa have a large contribution to the observed biodiversity. Why do some geographical regions harbor more biodiversity than others is still an open question (ANTONELLI and SANMARTÍN 2011). Mathematical models predict that a high level of genetic diversity promotes allopatric speciation and rapid diversification (GAVRILETS and VOSE 2005; GAVRILETS and LOSOS 2009). In agreement with these results, a large genetic diversity was reported in the ancestral populations of several famous examples of radiations (in cichlid fishes of the great African lakes, SEEHAUSEN *et al.* 2003, in Darwin's finches, FREELAND and BOAG 1999, and Hawaiian crickets, SHAW 2002). Consequently, processes which can generate the large amounts of genetic diversity necessary for radiations are of large interest to understand the origin of biodiversity.

In conclusion, genetic diversity is of tremendous importance for the evolution of species. The large diversity of life raises several questions:

- **How is genetic diversity generated?**
- **Which processes lead to large amounts of genetic diversity?**

EVOLUTIONARY PROCESSES AND THE PREDICTED EQUILIBRIUM GENETIC DIVERSITY

A tentative answer to these questions was first given in the 1920-1930s by Sewall Wright and Ronald Fisher, who provided predictions of the expected long-term genetic diversity (FISHER 1922, 1930; WRIGHT 1931). They showed that observed genetic diversity is the result of the interaction of multiple evolutionary processes, namely mutation, selection, genetic drift and migration. They showed how mathematical modeling enables to predict the long-term level of genetic diversity, given the evolutionary processes at work, and assuming that the environment of the species remains stable under large time-scales.

The following sections briefly describe the main processes influencing the level of genetic diversity and describe their expected impact.

Mutation

Mutations are the primary source of genetic diversity. Mutation results from the random error in DNA replication, which can lead to a new version of a gene, an allele. In population genetics, we are interested in inherited mutations, that is, mutations that affect the germline and are transmitted to the offsprings. The most common conceptual models used in population genetics are presented in Box 1.

Box 1: modeling mutation

In this thesis, we use two mutation models.

1. The infinite alleles model (KIMURA and CROW 1964). Under this model, we assume that two successive mutations have a null probability to lead to the same allele, so the set of possible alleles is infinite. This model can help to quantify genetic diversity, thus they are used in Chapters 1 to 3 to describe the expected level of genetic diversity in equilibrium and non-equilibrium situations.
2. The infinite sites model (KIMURA 1969), where genes are modeled as a sequence of sites, and each mutation is assumed to change the state (nucleotide) of a different site, so the set of possible sites is infinite. This model can predict DNA sequence polymorphism patterns, thus it is used in Chapter 4 to describe the expected genomic signature of particular historical events.

Both models are a good approximation when DNA sequences considered are long.

A largely used alternative mutation model is the stepwise mutation model (OHTA and KIMURA 1973), which assumes that each allele has a given number of repeat units, and mutations either increase or decrease the number of units by 1. This model is relevant for some genetic markers like microsatellites, but not to describe the general polymorphism of the genome, thus it is not used in this thesis.

Natural selection

Natural selection is the process by which some individuals have more offspring that survive and reproduce than others (DARWIN 1859; FISHER 1922). Natural selection arises because individuals in a population have more offsprings than can possibly survive and reproduce, and because survival and reproduction of offsprings is influenced by heritable traits.

Selective processes can have a wide variety of effects on genetic diversity, depending on the fitness of the different alleles. Under directional selection, one allele has

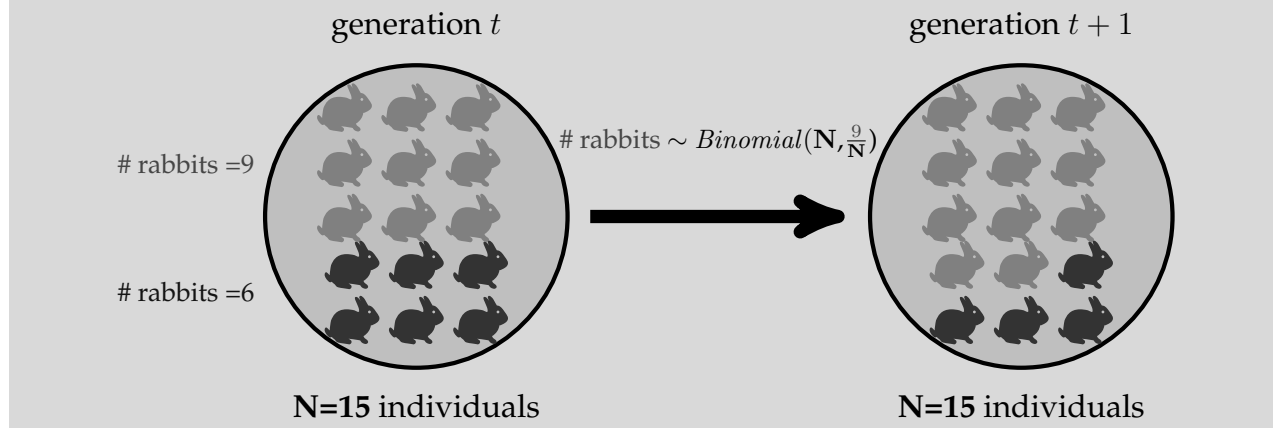
a selective advantage over all others; this results in reduced genetic diversity (FISHER 1922). On the contrary, under balancing selection, several alleles are maintained by natural selection, and thus genetic diversity is maintained (FISHER 1922; KARLIN 1982). It can be due to negative frequency dependent selection, where rare alleles have a selective advantage, or from a selective advantage of heterozygotes.

Finite population size and the genetic drift

Finite population size has an important impact on genetic diversity (WRIGHT 1931). First, the limited number of individuals limits the number of alleles simultaneously present in the population. Second, due to the random sampling of genes from one generation to another (Box 2), we expect genetic diversity to decrease over time in the absence of mutation. This process is called genetic drift. When both mutation and genetic drift are acting, genetic diversity is expected to tend to an equilibrium value, which reflects the number of (selectively neutral) alleles that can be maintained in the population (KIMURA and CROW 1964).

Box 2: modeling finite population size

The Wright-Fisher model (FISHER 1930; WRIGHT 1931) describes the changes of allele frequencies across generations. It assumes that generations are non-overlapping, i.e., that at each generation, all individuals die simultaneously and the new generation consists entirely of their offspring. The alleles of the new generation result from the multinomial sampling of alleles from the current generation.

**Population structure and migration**

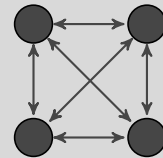
Early population genetics models considered that all individuals have the same probability to mate with any other individual in the population (panmixia). Nevertheless, in many species, populations are structured: individuals mate preferentially with a subpart of the population, called a subpopulation or deme. The mating between different subpopulations requires that an individual leaves its birth deme for another deme; this process is called migration. The most widely used migration models in population genetics are presented in Box 3.

The main consequence of population structure is genetic differentiation, the tendency of individuals from different subpopulations to be more different genetically than individuals from a same population (WRIGHT 1950). Measuring genetic differentiation and understanding its underlying processes is one of the primary goals of

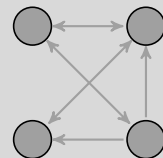
population genetics studies. Recent studies cast doubt on how to best measure genetic differentiation (BALLOUX *et al.* 2000; HEDRICK 2005; JOST 2008; WHITLOCK 2011). Progress is necessary to perfectly understand genetic diversity at equilibrium in structured populations. This is a necessary first step in order to predict and interpret genetic diversity patterns under more complex scenarios, and it will be studied in Chapter 1.

Box 3: modeling population structure

The finite island model assumes that individuals from a subpopulation have the same probability to migrate to any other subpopulation (MAYNARD SMITH 1970). It is the simplest migration model and is thus usually studied as a first approximation to more realistic migration models. Consequently, the thesis focuses on this model, which enables to compare our results to a large body of literature (MARUYAMA 1970; OHTA 1982; WAKELEY 1999; SCHIERUP *et al.* 2000; CHERRY 2003).



Alternative models include the stepping-stone model, where individuals can only migrate to neighboring subpopulations, and migration matrix models, which generalize the migration model to any migration pattern between populations. Nevertheless, due to their mathematical intricacy, they are discussed but no analytical results were derived under these models.



HISTORICAL EVENTS AND NON-EQUILIBRIUM GENETIC DIVERSITY

Nevertheless, the assumption of stable environment is often violated. Abiotic and biotic processes frequently impact species, leading to strong selective and demographic consequences. Climatic events strongly impacted species; the most striking example concerns the glaciation cycles of the Quaternary period, during which temperate and tropical species were repeatedly isolated into refugia and experienced range expan-

sions and secondary contacts (HEWITT 2000, 2004; YOUNG *et al.* 2009). In addition, declines in precipitations affected freshwater, leading to fragmentation and fusion of basins (e.g. in the great African lakes Victoria, Malawi and Tanganyika, GALIS and METZ 1998; STURMBAUER *et al.* 2001). Geological events can also strongly impact species, for example through mountain formation, which isolated numerous populations (e.g. plant species in the American neotropics, HUGHES and EASTWOOD 2006; ANTONELLI *et al.* 2009; ANTONELLI and SANMARTÍN 2011). Nowadays, anthropogenic changes (e.g., due to urbanization and agriculture) strongly impact species distribution and introduce new selective pressures (MILLER and HOBBS 2002; DELANEY *et al.* 2010). Indeed, species can suffer from human-induced sudden habitat reduction and fragmentation (KELLER *et al.* 2004; NOEL *et al.* 2007; VANDERGAST *et al.* 2009). In the meantime, other species experience habitat and population expansion (e.g., sparrows, white-tailed deer, zebra mussels; WAPLES 2010). Demographic changes are not only the consequence of abiotic processes, but they can also be due to spatial and temporal interactions of populations (e.g., secondary contacts; GREEN *et al.* 2010; DOMINGUES *et al.* 2012).

In particular, periods of isolation followed by re-connection of populations are common in natural populations. Our own ancestors, anatomically modern human, and Neanderthal populations were isolated, respectively in Africa and Europe, between 200,000 to 60,000 years ago, and there is recent evidence of subsequent mating between them in Europe (GREEN *et al.* 2010; PATTERSON *et al.* 2012; SANKARARAMAN *et al.* 2012). Similarly, African and non-African populations of *Drosophila melanogaster* were isolated, and subsequently re-connected (POOL *et al.* 2012). Such events also occur between domesticated species of both plants and animals and their wild relatives (e.g. in crops; ELLSTRAND *et al.* 1999). Viral populations are also often temporally iso-

lated within hosts, and transmission events re-connect populations (e.g., in malaria, ROBERT *et al.* 2003). Such re-connections have been reported in the Human Immunodeficiency Virus (HIV), where previously geographically isolated strains are now intermixing (TEBIT and ARTS 2011).

As a result, since the late 1970s, many studies focused on the impact of demographic events on genetic diversity (NAGYLAKI 1977; MARUYAMA and FUERST 1984; WHITLOCK 1992; EXCOFFIER *et al.* 2009). The aims of these studies is to quantify the changes of genetic diversity and the duration of the transient dynamics. The following section briefly describe the main theoretical results and highlight the current gaps.

Demographic changes and the dynamics of genetic diversity

A demographic change can be modeled as a change of some of the model parameters (e.g. the population size or migration rate). Mathematically, this change results in a non-equilibrium state, followed by changes of genetic diversity until it reaches its new equilibrium value expected with the new parameters (NEI and FELDMAN 1972; LATTER 1973; NEI 1973; NAGYLAKI 1974, 1977).

Some particular demographic changes were widely studied and were shown to have strong consequences on genetic diversity. A sudden population size reduction (population bottleneck) was shown to lead to a decrease of genetic diversity, and the time to reach the new equilibrium value depends on the effective population size (NAGYLAKI 1974, 1977). A sudden decrease in migration rates (population fragmentation or isolation) has been shown to to reduce the amount of genetic diversity within populations and allow for population differentiation (LATTER 1973; TAKAHATA and NEI 1985). Quantifying the degree of admixture following a population re-connection has recently received a lot of attention (e.g., PRITCHARD *et al.* 2000; FALUSH *et al.* 2003;

PRICE *et al.* 2009; GRAVEL 2012), in particular in Human populations (ROSENBERG *et al.* 2002; SHRIVER *et al.* 2003). Nevertheless, the dynamics of genetic diversity following past isolation and subsequent population re-connection still lacks theoretical predictions. This is the focus of Chapter 2.

Environmental conditions often change cyclically, so that genetic diversity never reaches an equilibrium value and is constantly maintained in a transient state. For example, serial founder effects occurring during colonization were shown to drastically reduce genetic diversity (AUSTERLITZ *et al.* 1997). Migration fluctuations at small time-scales (a few generations) were shown to moderately affect genetic diversity, so that the fluctuating migration rate can be approximated by an effective migration rate (NAGYLAKI 1979; WHITLOCK 1992; RICE and PAPADOPOULOS 2009; SHPAK *et al.* 2010). Nevertheless, intermediate and long time-scales fluctuations have still no good theoretical basis. This topic is the focus of Chapter 3.

INFERRING HISTORICAL EVENTS FROM THEIR GENOMIC SIGNATURE

Contrasting the observed DNA polymorphism with expected patterns can be used to detect past evolutionary and demographic events. For example, a deficit of low frequency alleles compared to the neutral expectation (i.e. assuming constant population size, no structure and no selection) is usually interpreted as the signature of a population bottleneck (TAJIMA 1989). On the contrary, an excess of low frequency alleles is usually interpreted as a population expansion (TAJIMA 1989; SLATKIN 1996). An excess of both high and low frequency alleles is expected under directional selection (BARTON 1998), while an excess of intermediate frequency alleles is expected under balancing selection (FAY and WU 2000; BARTON and ETHERIDGE 2004; ZENG *et al.* 2006).

Due to recent advances in statistical genetics, increased genomic data and computa-

tional power, it is now also possible to estimate demographic and selective parameters (e.g. populations size and growth rate, proportion of admixture, selection coefficient; BEAUMONT *et al.* 2002; KIM and STEPHAN 2002; KIM and NIELSEN 2004; NIELSEN *et al.* 2005; PRICE *et al.* 2009; EXCOFFIER *et al.* 2013) and estimate the relative likelihood of different demographic scenarios from genomic data (e.g. population bottleneck and subdivision, PETER *et al.* 2010). Nevertheless, it is often difficult to disentangle the impact of demographic changes and the effects of selection (JENSEN *et al.* 2005; NIELSEN 2005; LI and STEPHAN 2006; KIM and GULISIJA 2010; PAVLIDIS *et al.* 2010). It is also difficult to distinguish between the signatures of different demographic changes such as changes in population size or migration rate (WAKELEY 1999; STRASBURG and RIESEBERG 2011). Thus, it is necessary to better characterize the signature of past demographic changes in order to take full advantage of the genomic revolution. This is addressed in Chapter 4.

PLAN OF THE THESIS

The objectives of the thesis are to describe and summarize genetic diversity and genetic differentiation in an equilibrium structured population, to assess the transient impact of migration changes (isolation and subsequent re-connection of populations) on patterns of genetic diversity, and finally to investigate the signature and detection of past isolation and re-connection of populations from genomic data.

Chapter 1 assesses the ability of commonly used genetic differentiation measures G_{ST} and D to describe the degree of isolation and panmixia of populations. It is shown that both G_{ST} and D can indicate no genetic differentiation even though important features of genetic differentiation (e.g. the proportion of private alleles) are very close to their expected value when populations are isolated. What information is given by each

differentiation measure depending on the mutation regime is discussed, their complementarity is highlighted and guidelines are provided for the use of G_{ST} and D . In addition, G_{ST} and D are shown to be complementary to estimate demographic parameters. This chapter is expected to have an impact on the theoretical concept of genetic differentiation as well as on many empirical genetic studies, both in population genetics and conservation genetics.

Chapter 2 presents an investigation of the impact of a period of isolation or low migration followed by a sudden migration increase on the dynamics of genetic diversity. Results show that a sudden migration increase can lead to a large persistent peak of genetic diversity within-populations, and the impact of such events on evolutionary processes is discussed. In addition, the consequences for genetic data analysis is assessed, as migration increases are shown to lead to spurious signals of population size changes using the most common test statistics (e.g. Tajima's D).

Chapter 3 demonstrates how periodic isolation and connection of populations impacts genetic diversity. The importance of the length of the isolation and connection periods and the initial genetic diversity on the temporal accumulation or decrease of genetic diversity is shown. In particular, an important domain of period length previously overlooked by theoretical studies is highlighted, where periods range from a few hundred to tens of thousands of generations, for which genetic diversities undergo large variations and can display a variety of behaviors (accumulation or decrease). The high relevance of this domain is documented for species adaptation to novel environments based on standing genetic variation, pathogen dynamics, and inference of population history.

In Chapter 4, the signature of past isolation on DNA sequence data is investigated. Theoretical expectations are derived for the patterns of DNA sequence polymorphism, and the possibility to infer past isolation from such data is investigated. Results demonstrate that the comparison of gene genealogies and the frequency of single nucleotide polymorphisms (SNPs) between different sampling schemes (one or numerous populations) are strongly informative on past isolation events, and allow the distinction of isolation events from other demographic processes. The theoretical results are illustrated by the genomic signature of past isolation and successive connection events between the major HIV subtypes in China. Finally, the usefulness of the study for building improved methodology for inferring the parameters of past isolation in natural populations is highlighted.

BIBLIOGRAPHY

- AMINETZACH, Y. T., J. M. MACPHERSON, and D. A. PETROV, 2005 Pesticide resistance via transposition-mediated adaptive gene truncation in drosophila. *Science* **309**: 764–767.
- ANTONELLI, A., J. A. A. NYLANDER, C. PERSSON, and I. SANMARTÍN, 2009 Tracing the impact of the andean uplift on neotropical plant evolution. *Proc Natl Acad Sci U S A* **106**: 9749–9754.
- ANTONELLI, A., and I. SANMARTÍN, 2011 Why are there so many plant species in the neotropics? *Taxon* **60**: 403–414.
- AUSTERLITZ, F., B. JUNG-MULLER, B. GODELLE, and P.-H. GOUYON, 1997 Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology* **51**: 148–164.

- BALLOUX, F., H. BRÜNNER, N. LUGON-MOULIN, J. HAUSSE, and J. GOUDET, 2000
Microsatellites can be misleading: an empirical and simulation study. *Evolution* **54**:
1414–1422.
- BARRETT, R. D. H., and D. SCHLUTER, 2008 Adaptation from standing genetic varia-
tion. *Trends in Ecology & Evolution* **23**: 38–44.
- BARTON, N. H., 1998 The effect of hitch-hiking on neutral genealogies. *Genetics Re-
search* **72**: 123–133.
- BARTON, N. H., and A. M. ETHERIDGE, 2004 The effect of selection on genealogies.
Genetics **166**: 1115–1131.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian com-
putation in population genetics. *Genetics* **162**: 2025–2035.
- BERRY, E. W., 1943 The giant sequoia. *Science* **98**: 586.
- CHERRY, J. L., 2003 Selection in a subdivided population with dominance or local fre-
quency dependence. *Genetics* **163**: 1511–1518.
- CLEGG, M. T., and M. L. DURBIN, 2000 Flower color variation: a model for the exper-
imental study of evolution. *Proc Natl Acad Sci U S A* **97**: 7016–7023.
- COLOSIMO, P. F., K. E. HOSEMAN, S. BALABHADRA, G. VILLARREAL, M. DICKSON,
et al., 2005 Widespread parallel evolution in sticklebacks by repeated fixation of ec-
todysplasin alleles. *Science* **307**: 1928–1933.
- DARWIN, C., 1859 *On the origin of species by means of natural selection, or the preservation
of favoured races in the struggle for life*. London: Murray.

DELANEY, K. S., S. P. D. RILEY, and R. N. FISHER, 2010 A rapid, strong, and convergent genetic response to urban habitat fragmentation in four divergent and widespread vertebrates. *PLoS One* **5**.

DOMINGUES, V. S., Y.-P. POH, B. K. PETERSON, P. S. PENNINGS, J. D. JENSEN, *et al.*, 2012 Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution* **66**: 3209–3223.

ELLSTRAND, N. C., H. C. PRENTICE, and J. F. HANCOCK, 1999 Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **30**: pp. 539–563.

EXCOFFIER, L., I. DUPANLOUP, E. HUERTA-SÁNCHEZ, V. C. SOUSA, and M. FOLL, 2013 Robust demographic inference from genomic and snp data. *PLoS Genet* **9**: e1003905.

EXCOFFIER, L., M. FOLL, and R. J. PETIT, 2009 Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40**: 481–501.

FALUSH, D., M. STEPHENS, and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive darwinian selection. *Genetics* **155**: 1405–1413.

FEDER, J. L., S. H. BERLOCHER, J. B. ROETHELE, H. DAMBROSKI, J. J. SMITH, *et al.*, 2003 Allopatric genetic origins for sympatric host-plant shifts and race formation in *rhagoletis*. *Proc Natl Acad Sci U S A* **100**: 10314–10319.

FISHER, R. A., 1922 Darwinian evolution of mutations. *Eugen Rev* **14**: 31–34.

- FISHER, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press.
- FRANKHAM, R., 1995 Conservation genetics. *Annu Rev Genet* **29**: 305–327.
- FREELAND, J. R., and P. T. BOAG, 1999 The mitochondrial and nuclear genetic homogeneity of the phenotypically diverse darwin's ground finches. *Evolution* **53**: 1553–1563.
- GALIS, F., and J. A. METZ, 1998 Why are there so many cichlid species? *Trends Ecol Evol* **13**: 1–2.
- GAVRILETS, S., and J. B. LOSOS, 2009 Adaptive radiation: contrasting theory with data. *Science* **323**: 732–737.
- GAVRILETS, S., and A. VOSE, 2005 Dynamic patterns of adaptive radiation. *Proc Natl Acad Sci U S A* **102**: 18040–18045.
- GIBSON, G., and I. DWORKIN, 2004 Uncovering cryptic genetic variation. *Nature Reviews Genetics* **5**: 681–U11.
- GILPIN, M., and M. SOULE, 1986 *Conservation Biology: The Science of Scarcity and Diversity*, chapter Minimum Viable Populations: Processes of Species Extinction. Sinauer, Sunderland, Mass, pp. 19–34.
- GRAVEL, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–619.
- GREEN, R. E., J. KRAUSE, A. W. BRIGGS, T. MARICIC, U. STENZEL, *et al.*, 2010 A draft sequence of the neandertal genome. *Science* **328**: 710–722.
- GREGORY, T. R., J. A. NICOL, H. TAMM, B. KULLMAN, K. KULLMAN, *et al.*, 2007 Eukaryotic genome size databases. *Nucleic Acids Res* **35**: D332–D338.

HEDRICK, P. W., 2005 A standardized genetic differentiation measure. *Evolution* **59**: 1633–1638.

HEDRICK, P. W., and S. T. KALINOWSKI, 2000 Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics* **31**: pp. 139–162.

HERMISSON, J., and P. S. PENNINGS, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.

HERNANDEZ, R. D., J. L. KELLEY, E. ELYASHIV, S. C. MELTON, A. AUTON, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.

HEWITT, G., 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–913.

HEWITT, G. M., 2004 Genetic consequences of climatic oscillations in the quaternary. *Philos Trans R Soc Lond B Biol Sci* **359**: 183–95; discussion 195.

HUGHES, C., and R. EASTWOOD, 2006 Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the andes. *Proc Natl Acad Sci U S A* **103**: 10334–10339.

JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO, and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using dna polymorphism data. *Genetics* **170**: 1401–1410.

JIMENEZ, J. A., K. A. HUGHES, G. ALAKS, L. GRAHAM, and R. C. LACY, 1994 An experimental study of inbreeding depression in a natural habitat. *Science* **266**: 271–273.

JONES, F. C., M. G. GRABHERR, Y. F. CHAN, P. RUSSELL, E. MAUCELI, *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.

- JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.
- KARLIN, S., 1982 Classifications of selection migration structures and conditions for a protected polymorphism. *Evolutionary Biology* **14**: 61–204.
- KELLER, I., W. NENTWIG, and C. R. LARGIADER, 2004 Recent habitat fragmentation due to roads can lead to significant genetic differentiation in an abundant flightless ground beetle. *Mol Ecol* **13**: 2983–2994.
- KIM, Y., and D. GULISIJA, 2010 Signatures of recent directional selection under different models of population expansion during colonization of new selective environments. *Genetics* **184**: 571–585.
- KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1513–1524.
- KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765–777.
- KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**: 893–903.
- KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
- LAIKRE, L., and N. RYMAN, 1991 Inbreeding depression in a captive wolf (*canis lupus*) population. *Conservation biology* **5**: 33–40.
- LANDE, R., 1998 Risk of population extinction from fixation of deleterious and reverse mutations. *Genetica* **102-103**: 21–27.

- LATTER, B. D., 1973 The island model of population differentiation: a general solution. *Genetics* **73**: 147–157.
- LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genet* **2**: e166.
- LI, W. H., and L. A. SADLER, 1991 Low nucleotide diversity in man. *Genetics* **129**: 513–523.
- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor Popul Biol* **1**: 273–306.
- MARUYAMA, T., and P. A. FUERST, 1984 Population bottlenecks and nonequilibrium models in population genetics. i. allele numbers when populations evolve from zero variability. *Genetics* **108**: 745–763.
- MAYNARD SMITH, J., 1970 Population size, polymorphism, and the rate of non-darwinian evolution. *The American Naturalist* **104**: 231–237.
- MESELSON, M., and F. W. STAHL, 1958 The replication of dna. *Cold Spring Harb Symp Quant Biol* **23**: 9–12.
- MILLER, J., and R. HOBBS, 2002 Conservation where people live and work. *Conservation Biology* **16**: 330–337.
- MORA, C., D. P. TITTENSOR, S. ADL, A. G. B. SIMPSON, and B. WORM, 2011 How many species are there on earth and in the ocean? *PLoS Biol* **9**: e1001127.
- NAGYLAKI, T., 1974 The decay of genetic variability in geographically structured populations. *Proc Natl Acad Sci U S A* **71**: 2932–2936.
- NAGYLAKI, T., 1977 Decay of genetic variability in geographically structured populations. *Proc Natl Acad Sci U S A* **74**: 2523–2525.

- NAGYLAKI, T., 1979 The island model with stochastic migration. *Genetics* **91**: 163–176.
- NEI, M., 1973 Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**: 3321–3323.
- NEI, M., and M. W. FELDMAN, 1972 Identity of genes by descent within and between populations under mutation and migration pressures. *Theor Popul Biol* **3**: 460–465.
- NIELSEN, R., 2005 Molecular signatures of natural selection. *Annual Review of Genetics* **39**: 197–218.
- NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON, *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.
- NOEL, S., M. OUELLET, P. GALOIS, and F.-J. LAPOINTE, 2007 Impact of urban fragmentation on the genetic structure of the eastern red-backed salamander. *Conservation Genetics* **8**: 599–606.
- OHTA, T., 1982 Linkage disequilibrium with the island model. *Genetics* **101**: 139–155.
- OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res* **22**: 201–204.
- PATTERSON, N., P. MOORJANI, Y. LUO, S. MALLICK, N. ROHLAND, *et al.*, 2012 Ancient admixture in human history. *Genetics* **192**: 1065–1093.
- PAVLIDIS, P., J. D. JENSEN, and W. STEPHAN, 2010 Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics* **185**: 907–922.

- PETER, B. M., D. WEGMANN, and L. EXCOFFIER, 2010 Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. *Mol Ecol* **19**: 4648–4660.
- POOL, J. E., R. B. CORBETT-DETIG, R. P. SUGINO, K. A. STEVENS, C. M. CARDENO, *et al.*, 2012 Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS Genet* **8**: e1003080.
- PRICE, A. L., A. TANDON, N. PATTERSON, K. C. BARNES, N. RAFAELS, *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**: e1000519.
- PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- RICE, S. H., and A. PAPADOPOULOS, 2009 Evolution with stochastic fitness and stochastic migration. *PLoS One* **4**: e7130.
- ROBERT, V., K. MACINTYRE, J. KEATING, J.-F. TRAPE, J.-B. DUCHEMIN, *et al.*, 2003 Malaria transmission in urban sub-saharan africa. *Am J Trop Med Hyg* **68**: 169–176.
- ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD, *et al.*, 2002 Genetic structure of human populations. *Science* **298**: 2381–2385.
- SANKARARAMAN, S., N. PATTERSON, H. LI, S. PÄÄBO, and D. REICH, 2012 The date of interbreeding between neandertals and modern humans. *PLoS Genet* **8**: e1002947.
- SCHIERUP, M. H., X. VEKEMANS, and D. CHARLESWORTH, 2000 The effect of subdivision on variation at multi-allelic loci under balancing selection. *Genet Res* **76**: 51–62.

- SEEHAUSEN, O., E. KOETSIER, M. V. SCHNEIDER, L. J. CHAPMAN, C. A. CHAPMAN, *et al.*, 2003 Nuclear markers reveal unexpected genetic variation and a congolese- nilotic origin of the lake victoria cichlid species flock. *Proc Biol Sci* **270**: 129–137.
- SHAW, K. L., 2002 Conflict between nuclear and mitochondrial dna phylogenies of a recent species radiation: what mtdna reveals and conceals about modes of speciation in hawaiian crickets. *Proc Natl Acad Sci U S A* **99**: 16122–16127.
- SHPAK, M., J. WAKELEY, D. GARRIGAN, and R. C. LEWONTIN, 2010 A structured coalescent process for seasonally fluctuating populations. *Evolution* **64**: 1395–1409.
- SHRIVER, M. D., E. J. PARRA, S. DIOS, C. BONILLA, H. NORTON, *et al.*, 2003 Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet* **112**: 387–399.
- SLATKIN, M., 1996 Gene genealogies within mutant allelic classes. *Genetics* **143**: 579–587.
- STRASBURG, J. L., and L. H. RIESEBERG, 2011 Interpreting the estimated timing of migration events between hybridizing species. *Mol Ecol* **20**: 2353–2366.
- STURMBAUER, C., S. BARIC, W. SALZBURGER, L. RÜBER, and E. VERHEYEN, 2001 Lake level fluctuations synchronize genetic divergences of cichlid fishes in african lakes. *Mol Biol Evol* **18**: 144–154.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**: 585–595.
- TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**: 325–344.

- TEBIT, D. M., and E. J. ARTS, 2011 Tracking a century of global expansion and evolution of hiv to drive understanding and to combat disease. *Lancet Infect Dis* **11**: 45–56.
- TURNER, R. C., J. C. LEVY, and A. CLARK, 1993 Complex genetics of type 2 diabetes: thrifty genes and previously neutral polymorphisms. *Q J Med* **86**: 413–417.
- VANDERGAST, A. G., E. A. LEWALLEN, J. DEAS, A. J. BOHONAK, D. B. WEISSMAN, *et al.*, 2009 Loss of genetic connectivity and diversity in urban microreserves in a southern California endemic Jerusalem cricket (Orthoptera: Stenopelmatidae: *Stenopelmatus* n. sp “santa monica”). *Journal of Insect Conservation* **13**: 329–345.
- WAKELEY, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- WAPLES, R. S., 2010 Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size. *Mol Ecol Resour* **10**: 785–796.
- WATSON, J. D., and F. H. CRICK, 1953 Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.
- WHITLOCK, M. C., 1992 Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution* **46**: pp. 608–615.
- WHITLOCK, M. C., 2011 G'_{ST} and D do not replace F_{ST} . *Molecular Ecology* **20**: 1083–1091.
- WRIGHT, S., 1931 Evolution in mendelian populations. *Genetics* **16**: 97–159.
- WRIGHT, S., 1950 Genetical structure of populations. *Nature* **166**: 247–249.

YOUNG, K. A., J. M. WHITMAN, and G. F. TURNER, 2009 Secondary contact during adaptive radiation: a community matrix for lake malawi cichlids. *J Evol Biol* **22**: 882–889.

ZENG, K., Y.-X. FU, S. SHI, and C.-I. WU, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**: 1431–1439.

Chapter 1 On the Transition of Genetic Diversity Between Isolation and Panmixia: what we can learn from G_{ST} and D

Nicolas Alcalá¹, Jérôme Goudet¹, and Séverine Vuilleumier^{1,2}

¹Department of Ecology and Evolution, Biophore, University of Lausanne, CH-1015 Lausanne, Switzerland

²Institute of Microbiology, University Hospital Center and University of Lausanne (CHUV-UNIL), CH-1011 Lausanne, Switzerland

Genetic differentiation

Genetic diversity

F_{ST}

D

G_{ST}

Published, Theoretical Population Biology May 1, 2014 vol. 93, 75–84

ABSTRACT

Population genetic differentiation characterizes the repartition of alleles among populations. It is commonly thought that genetic differentiation measures, such as G_{ST} and D , should be near zero when allele frequencies are close to their expected value in panmictic populations, and close to one when they are close to their expected value in isolated populations. To analyze those properties, we first derive analytically a reference function f of known parameters that describes how important features of genetic differentiation (e.g. gene diversity, proportion of private alleles, frequency of the most common allele) are close to their expected panmictic and isolation value. We find that the behavior of function f differs according to three distinct mutation regimes defined by the scaled mutation rate and the number of populations. Then, we compare G_{ST} and D to f , and demonstrate that their signal of differentiation strongly depends on the mutation regime. In particular, we show that D captures well the variations of genetic diversity when mutation is weak, otherwise it overestimates it when panmixia is not met. G_{ST} detects population differentiation when mutation is intermediate but has a low sensitivity to the variations of genetic diversity when mutation is weak. When mutation is strong the domain of sensitivity of both measures are altered. Finally, we also point out the importance of the number of populations on genetic differentiation measures, and provide recommendations for the use of G_{ST} and D .

QUANTIFICATION of population genetic differentiation (i.e. how different are allele frequencies between populations) is a long standing issue in population genetics. WRIGHT (1951) first proposed the fixation index F_{ST} , which measures both the proportion of alleles that reach fixation and population allelic differentiation for a bi-allelic locus. Wright's fixation index was then extended to multi-allelic loci, G_{ST} , by NEI (1973) and WEIR and COCKERHAM (1984) - also often referred as F_{ST} (HOLSINGER and WEIR 2009; JAKOBSSON *et al.* 2013). G_{ST} characterizes the ratio of within-population, H_s , to total, H_t , gene diversities (notations are summarized in Table 1.1):

$$G_{ST} = 1 - \frac{H_s}{H_t} \quad (1.1a)$$

where

$$H_s = 1 - \frac{1}{n} \sum_j \sum_i p_{ij}^2 \quad (1.1b)$$

$$H_t = 1 - \sum_i \left(\frac{1}{n} \sum_j p_{ij} \right)^2 \quad (1.1c)$$

p_{ij} is the frequency of allele i in population j and n is the number of populations. Thus H_s and H_t correspond to the probability that two genes randomly chosen, respectively, from the same population and from different populations, at a given locus, are different (NEI 1973).

NEI (1973) considered G_{ST} as a statistics which characterizes the properties of sampled populations, while WEIR and COCKERHAM (1984) considered that G_{ST} is independent of the sampling scheme and represents a parameter of the populations that can be estimated (HOLSINGER and WEIR 2009). In this study, we analyze the actual level of genetic differentiation of the populations, and not the properties of a sample, thus we consider genetic differentiation measures as parameters and not as statistics.

Several issues related to G_{ST} have been raised (CHARLESWORTH 1998; NAGYLAKI

Table 1.1 Summary of notations

GENETIC DIFFERENTIATION MEASURES	
G_{ST}	differentiation measure based on allele fixation (NEI 1973)
G'_{ST}	normalized genetic differentiation measure based on allele fixation (HEDRICK 2005)
D	genetic differentiation measure based on genetic composition (JOST 2008)
D_{ST}	absolute genetic differentiation measure (NEI 1973)
GENETIC DIVERSITY MEASURES	
H_s	within-population gene diversity (NEI 1973)
H_t	total gene diversity (NEI 1973)
Δ_S	within-population effective number of alleles (JOST 2008)
Δ_T	total effective number of alleles (JOST 2008)
SUMMARY STATISTICS OF ALLELE FREQUENCIES	
p_{max}	frequency of most frequent allele in the total population (JAKOBSSON <i>et al.</i> 2013)
p^i_{max}	frequency of most frequent allele in each population (JAKOBSSON <i>et al.</i> 2013)
p_{priv}	proportion of private alleles (SLATKIN 1985)
σ	mean singular value of the allele frequency table (GOLUB and KAHAN 1965)
FUNCTIONS	
$f(M)$	function describing the transition from isolation to panmixia
$f_G(M)$	function describing the behavior of G_{ST} as a function of the scaled migration rate
$f_D(M)$	function describing the behavior of D as a function of the scaled migration rate
M_T	threshold migration value of function $f(M)$
M_G	threshold migration value of function $f_G(M)$
M_D	threshold migration value of function $f_D(M)$

1998; HEDRICK 1999; JOST 2008). Authors showed that values of G_{ST} are constrained by the value of within-population gene diversity H_s ; G_{ST} remains inferior to $1 - H_s$ (see Figure 1 from JOST 2008). Therefore, when H_s is large, the range of the genetic differentiation signal is truncated and G_{ST} can be constrained by the frequency of the most frequent allele (JAKOBSSON *et al.* 2013). This has two main consequences; first, when H_s is high, G_{ST} cannot detect differentiation (as shown for high mutation rate loci in BALLOUX *et al.* 2000); second, when loci have different H_s values, corresponding G_{ST} values cannot be compared as G_{ST} does not rank populations by their degree of differentiation (JOST 2008; HELLER and SIEGISMUND 2009). For example, a set of populations with high H_s that does not share alleles has a lower G_{ST} than a set of population with low H_s but that shares alleles (JOST 2008). In addition, G_{ST} can take a value of 1 when in each population the same allele is fixed (GREGORIUS 2010).

To overcome the limitations of G_{ST} , new measures of differentiation were proposed. First, HEDRICK (2005) derived a normalized value of G_{ST} , G'_{ST} , which ranges from 0 to 1 whatever the level of within-population gene diversity, H_s . Nevertheless, like G_{ST} , G'_{ST} cannot detect differentiation when the same allele is fixed in each population, as G'_{ST} is very close to G_{ST} when H_s is low (HEDRICK 2005). Second, JOST (2008) proposed a new measure of genetic differentiation, D , based on the ratio of the within-population Δ_S to the total Δ_T effective number of alleles (KIMURA and CROW 1964; also called "true diversity" in JOST 2008).

$$D = \frac{n}{n-1} \left(1 - \frac{\Delta_S}{\Delta_T}\right) \quad (1.2)$$

with $\Delta_S = \frac{1}{1-H_s}$ and $\Delta_T = \frac{1}{1-H_t}$.

G'_{ST} and D provide similar estimations of population differentiation (HELLER and SIEGISMUND 2009) when H_s is high. However, both have slow rate of convergence to their equilibrium values after a perturbation when mutation is low (RYMAN and LEIMAR 2008, 2009). Thus, following a demographic change (i.e. change in population size or migration rate), G'_{ST} and D values can depend on the prior population size and gene diversity. Impacts can be strong in situation where the within-population gene diversity H_s before the demographic change is far from the expected H_s after the change. On the contrary, when the mutation rate is high, D can converge faster than G_{ST} . When loci are under different mutation regimes their rate of convergence differ accordingly.

The discrepancies between G_{ST} and D have been analyzed and discussed in numerous recent studies (JOST 2008; RYMAN and LEIMAR 2008, 2009; JOST 2009; GERLACH *et al.* 2010; GREGORIUS 2010; KRONHOLM *et al.* 2010; MEIRMANS and HEDRICK 2011; WHITLOCK 2011; WANG 2012). Those studies conclude that D and G_{ST} do not detect

the same type of genetic differentiation. G_{ST} values could be interpreted as a measure of the level of allele fixation in populations (WHITLOCK 2003), and also reflects the population's demographic properties (e.g. the number of migrants per generation under assumed mutation rate) independently of the analyzed loci. D values could be interpreted as measures of the difference in the genetic composition of populations, reflecting the properties of the analyzed loci and are related to genetic distance (JOST 2009).

Many studies that compared G_{ST} and D values used one of them as a reference value (RYMAN and LEIMAR 2009). To compare G_{ST} and D , HELLER and SIEGISMUND (2009) used D as a reference value, which prevented the detection of any issue related to D . Similarly, WHITLOCK (2011) used the coalescent F_{ST} as a reference value, which is closer to G_{ST} than D when the mutation rate is low. On the contrary, JOST (2008) used a genetic differentiation definition to compare D and G_{ST} values and avoid circular arguments (Table 1 and figure 2 in JOST 2008), however it involved only two populations and considered a restrictive number of illustrative examples. To have a deep understanding of the behavior and the properties of measures of genetic differentiation such as G_{ST} and D , an independent reference is necessary. This approach must provide a description of the behavior of the measures of genetic differentiation given any number of populations considered and across a large range of migration and mutation rates.

Here, we do not propose a new measure of genetic differentiation, instead, we propose an innovative approach to characterize the transition of genetic differentiation from isolation to panmixia and to understand the properties of measures of genetic differentiation, G_{ST} and D . We first introduce a function, $f(M)$, that is implicitly defined and describes the degree to which populations are closer to isolation (i.e. no

mating between individuals from different populations) or panmixia (i.e. random mating between individuals from all populations) and how transition between these two states occurs as a function of the number of migrants per generation. The derivation of $f(M)$ requires a complete understanding of the processes shaping genetic diversity and genetic differentiation and thus requires to specify a model with known parameters. We use the most common population structure model, the finite island model and the infinite allele model of mutation, under equilibrium conditions (MARUYAMA 1970; MAYNARD SMITH 1970). Thus, $f(M)$ is not a model-independent statistics measurable from population data, but is a theoretical reference to which the values of G_{ST} and D can be compared under a known model.

Second, we demonstrate that the function, $f(M)$, captures the behavior of the within-population gene diversity H_s and the total gene diversity, H_t , derived by MARUYAMA (1970) under the finite island model. The function $f(M)$ has a threshold migration value M_T that characterizes the number of migrants that leads to a transition between a behavior of gene diversity close to isolation or close to panmixia. Interestingly, the behavior of gene diversities (H_s and H_t) as a function of the number of migrants differs according to three distinct mutation regimes that are defined by the mutation rate and the number of populations. For each domain a different threshold migration value applies.

Third, we demonstrate that the function $f(M)$ implicitly describes the behavior of the within-population Δ_S and total Δ_T effective number of alleles (JOST 2008), as well as the behavior of four statistics representative of distinctive features of genetic differentiation. The four statistics are: the frequency of the most common allele in the total population p_{max} and in each population $\overline{p_{max}^i}$ (e.g. used in JAKOBSSON *et al.* 2013), the proportion of private alleles p_{priv} (e.g. used in SLATKIN 1985 as indicator of gene

flow) and the mean singular value σ of the singular value decomposition of the allele frequency matrix (GOLUB and KAHAN 1965) which represents the degree of overlap of allele frequencies between populations.

Fourth, using function $f(M)$, we analyze the type of genetic differentiation that G_{ST} and D actually measure. In particular, we show that the behavior of G_{ST} and D depends on the mutation regime and on the number of populations. We find that differentiation measures G_{ST} and D display different signals of genetic differentiation in different, specific and restricted mutation regimes. We define those domains given any migration rate, number of populations and mutation rate. We discuss how informative are G_{ST} and D in those domains and highlight the complementarity of G_{ST} and D . Finally, we provide recommendations on the use of measures of population differentiation.

BEHAVIOR OF GENETIC DIFFERENTIATION: FROM ISOLATION TO PANMIXIA

JOST (2008) showed that a genetic differentiation measure requires at least three properties: (i) the measure is 0 when all populations have the same allele at the same frequencies (panmictic state), (ii) the measure is 1 when all populations only have private alleles (isolation state), and (iii) the measure monotonically decreases as allele frequencies change from the isolation state to the panmictic state. Depending on its construction, a measure of differentiation will be sensitive to changes in allele frequencies from isolation to panmixia in a different domain. Thus, two measures can signal differentiation at different changes of allele frequencies: one could have a strong sensitivity close to isolation and the other close to panmixia. Identifying the domain in which a measure provides a strong or a weak signal when frequencies of alleles in population change is thus crucial and is determinant for its relevant use and interpretation.

To identify such domains, we introduce a reference function, f , for which there is a well defined relationship between the transition from 0 to 1 of the measure and the transition of allele frequencies from isolation to panmixia.

To derive the function f , we consider the within-population gene diversity, H_s , and the total gene diversity, H_t , introduced by NEI (1973). H_s and H_t are commonly used to estimate genetic differentiation, such as in G_{ST} , D and G'_{ST} (NEI 1973; HEDRICK 2005; JOST 2008). Then, we show that the function f describes many important features of the distribution of alleles among populations: frequency of the most common allele in each population and among populations, proportion of private alleles and mean singular value of the allele frequency matrix.

Deriving the reference function $f(M)$ from H_s and H_t

We investigate the specific behavior of H_s and H_t as a function of migration in a parametrized finite island model at equilibrium - in which all parameters are known- and assuming that mutation follows the infinite allele model (IAM) (MARUYAMA 1970; MAYNARD SMITH 1970). In this model, n populations composed of N individuals (N or $2N$ genes per population, for haploid or diploid individuals, respectively) are connected through migration at rate m and mutations occur at rate μ (the scaled migration and mutation rates are $M = 2Nm$ and $\theta = 2N\mu$ for haploid individuals, and $M = 4Nm$ and $\theta = 4N\mu$ for diploid individuals). We derive function f of the scaled migration rate, M , using the equilibrium of gene diversities derived by KIMURA and CROW (1964) and MARUYAMA (1970) under the finite island model. The function $f(M)$ is implicitly defined, and describes the relationship between gene diversities and the scaled migration rate M .

Gene diversities, H_s and H_t , can be rescaled and normalized by their values at isola-

tion (no migration among populations), H_s^{iso} and H_t^{iso} , and at panmixia ($m = 1 - 1/n$), H^{pan} , with $H_s = H_t = H^{pan}$. Under the island model and infinite allele model at equilibrium, we have $H_s^{iso} = \frac{\theta}{1+\theta}$, $H^{pan} = \frac{n\theta}{1+n\theta}$ and $H_t^{iso} = 1 - \frac{1}{n} \frac{1}{1+\theta} = \frac{n-1+n\theta}{n(1+\theta)}$ (from KIMURA and CROW 1964). Using the expressions of H_s and H_t from eqs. 2-3 and 2-4 in MARUYAMA (1970) and assuming small mutation and migration rates (terms in $1/N^2$, μ^2 , m^2 , μ/N , m/N and μm can be removed), we obtain:

$$\begin{cases} \frac{H_s - H^{pan}}{H_s^{iso} - H^{pan}} = f(M) \\ \frac{H_t - H^{pan}}{H_t^{iso} - H^{pan}} = f(M) \end{cases} \quad (1.3)$$

Note that assuming small mutation and migration rates μ and m does not imply assuming small θ and M . Indeed, mutation rates rarely exceeds 10^{-2} and satisfy $\mu \ll 1$, thus, $\theta \gg 1$ requires that $N \gg 1$. Similarly, the assumption of small migration rate does not strongly constrain the values of M .

The two rescaled expressions of within and total gene diversities in eq. 1.3 can be described by a unique function $f(M)$. With this function, the transition of gene diversities from isolation to panmictic state can be described as a function of the scaled migration rate, given the scaled mutation rate and the number of populations:

$$f(M) = \frac{M_T}{M + M_T} \quad (1.4a)$$

with

$$M_T = (n - 1)\theta \frac{1 + \theta}{1 + n\theta} \quad (1.4b)$$

From eq. 1.4b, we can see that the value of the threshold migration M_T depends only on the number of populations n and the scaled mutation rate θ .

The relationship between the scaled migration rate M and M_T has a simple interpretation. M_T represents the strength of the mechanisms leading to genetic differentiation (i.e. gene diversities approach their isolation values); while, M represents the strength of the mechanisms reducing genetic differentiation (i.e. gene diversities approach their panmictic values). Their relative strengths determine whether populations tend to have alleles at the same or at different frequencies. The value of function $f(M)$ decreases monotonically as allele frequencies change from the isolation state ($f(M) = 1$), to the panmictic state ($f(M) = 0$) and describes accurately the changes in allele frequency during the transition. The transition from the isolation state to the panmictic state occurs at value $f(M) = 0.5$.

Interestingly, the function $f(M)$ describes the behavior of H_s as well as H_t . This has a strong consequence: comparing either H_s or H_t with its expected value at panmixia (H^{pan}) and isolation (H_s^{iso} and H_t^{iso} , equation 1.3) leads to the same information on closeness to panmixia or isolation. Thus, $f(M)$ can be based on either H_s or H_t , while commonly-used measures of genetic differentiation rely on comparisons between within-population and total genetic diversities, through diversity ratios for G_{ST} (NEI 1973), or ratios of effective number of alleles for D (JOST 2008). Also, the function $f(M)$ describes the relative influence of H_s^{iso} and H^{pan} (resp. H_t^{iso} to H^{pan}) on the value of H_s (resp. H_t). However, these relationships depend on the differences between values of M and M_T (Figure 1.1). When $M \ll M_T$ gene diversities are close to their equilibrium at isolation values $H_s \simeq H_s^{iso}$ and $H_t \simeq H_t^{iso}$ (light grey area in Figure 1.1a, b). When $M \gg M_T$ gene diversities are close to their equilibrium at panmixia H^{pan} (dark grey area in Figure 1.1a, b). When $M = M_T$, the equilibrium diversity is exactly the mean of the isolation and panmictic equilibria $\frac{H^{iso}+H^{pan}}{2}$ (black solid line in Figure 1.1a, b). For example, when M is 19 fold lower than threshold M_T ($M = 0.05/0.95 \times M_T$),

we have $f(M) = 0.95$ and thus gene diversities become $H_s = 0.05H^{pan} + 0.95H_s^{iso}$ and $H_t = 0.05H^{pan} + 0.95H_t^{iso}$: the gene diversity values are at 5% of their expected panmixia value and at 95% of their isolation value.

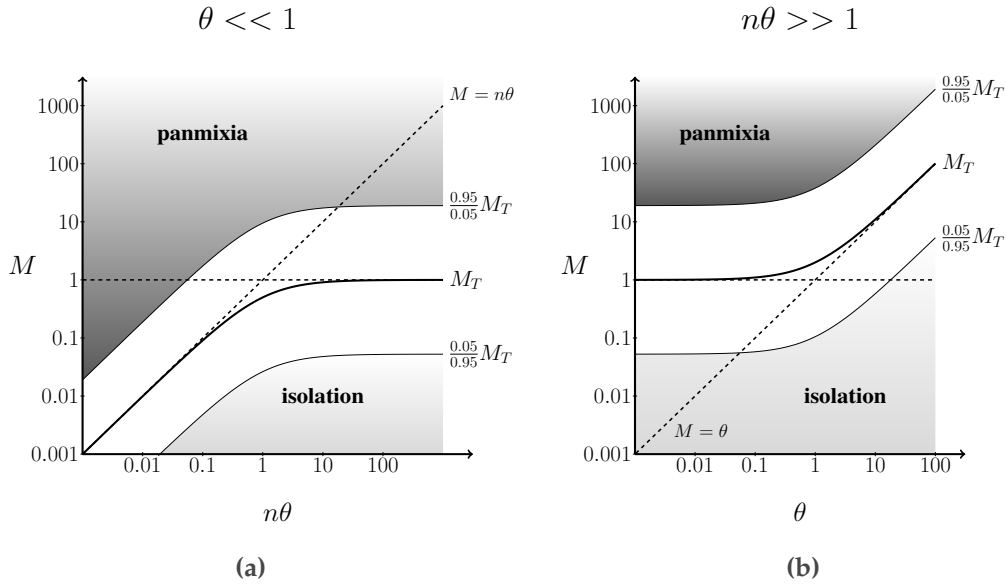


Figure 1.1 Threshold migration value of gene diversities, M_T , as a function of the scaled mutation, θ , and migration rate, M , and the number of populations n , when n is large. The solid line represents the exact value of M_T (eq. 1.4) and the dashed line its linear approximation (eq. 1.5). The domains of panmixia and isolation correspond to the parameter spaces where gene diversity values are at least 95% of their expected panmixia and isolation values, respectively. Gene diversities at 95% of the isolation value, which corresponds to $H_s = 0.05H^{pan} + 0.95H_s^{iso}$ and $H_t = 0.05H^{pan} + 0.95H_t^{iso}$, require that $M = \frac{0.05}{0.95}M_T$. Similarly, gene diversities at 95% of the panmixia value require that $M = \frac{0.95}{0.05}M_T$. (a) When $\theta \ll 1$, M_T is a function of $n\theta$, M_T can be approximated by the line $n\theta$ (dashed line $M = n\theta$) in the weak mutation regime (i.e. when $n\theta < 1$) and by the line 1 (dashed line $M = 1$) in the intermediate mutation regime (i.e. when $\theta < 1 < n\theta$). (b) When $n\theta \gg 1$, M_T is a function of θ and independent of n , M_T can be approximated by the line 1 (dashed line $M = 1$) in the intermediate mutation regime (i.e. when $\theta < 1 < n\theta$) and by the line θ (dashed line $M = \theta$) in the strong mutation regime (i.e. when $1 < \theta$).

Ratios between within and total diversity can provide measures between 0 and 1 (G_{ST} when the mutation rate is low and D rely on this property). Intermediate values between 0 and 1 remain difficult to interpret, for example a value of 0.5 means that the within-population diversity is half the total diversity. However, it does not necessarily mean that main features of allele frequency distributions (e.g. number of private alleles, frequency of the most common allele, etc.) are halfway between isolation and

panmixia. We will show that a value of 0.5 using $f(M)$ means that the main features of allele frequency distributions are halfway between isolation and panmixia.

Finally, it can be shown that $f(M)$ also describes the transition from the isolation state to the panmictic state of the between-population gene diversity, H_d , defined in NEI (1973) and CROW (1986). Indeed, using the relationship $H_d = \frac{n}{n-1}H_t - \frac{H_s}{n-1}$ (NEI 1973) and eq. 1.3, we can show that $f(M) = \frac{H_d - H^{pan}}{H_d^{iso} - H^{pan}}$, where H_d^{iso} is the expected value of H_d in the isolation state.

The effect of mutation, genetic drift and the number of populations on M_T

Now, we investigate how the behavior of M_T is affected by the mutation rate and the number of populations. From eq. 1.4 and using the limit values of M_T when $n\theta \ll 1$, $\theta \ll 1 \ll n\theta$, and $1 \ll \theta$, the threshold M_T can be simplified by piecewise linearisation (see Figure 1.1) and becomes:

$$M_T \simeq \begin{cases} (n-1)\theta & \text{for } n\theta < 1 \\ \frac{n-1}{n} & \text{for } \theta < 1 < n\theta \\ \frac{n-1}{n}\theta & \text{for } 1 < \theta \end{cases} \quad (1.5)$$

Eq. 1.5 decomposes the relationship between gene diversity and migration into the three distinct mutation regimes, which depend on the relative value of the scaled mutation rate θ , the scaled genetic drift, 1, and the number of populations n . The three distinct regimes of mutation are (a) *the weak mutation regime*, (b) *the intermediate mutation regime* and (c) *the strong mutation regime*. In *the weak mutation regime*, the total amount of mutation over n populations is weaker than genetic drift ($n\theta < 1$; Figure 1.1a, left) i.e. there is less than 1 mutant per generation in the total population and the threshold

migration value M_T depends on the mutations in all populations ($n\theta$; dashed line in the left of Figure 1.1a). In *the intermediate mutation regime* genetic drift in a population is stronger than mutation in each population but weaker than mutation occurring over all populations ($\theta < 1 < n\theta$; Figure 1.1a, right, and Figure 1.1b, left) i.e. there is less than 1 mutant per generation in each population but more than 1 mutant per generation in the total population. In this regime, the threshold migration value, which determines whether genetic diversities are closer to their expected isolation or panmictic values, depends mainly on genetic drift (1; dashed line in the right of Figure 1.1a and left of Figure 1.1b). In *the strong mutation regime* mutation is stronger than genetic drift ($1 < \theta$; Figure 1.1b, right) i.e. there are more than 1 mutant per generation in each population and the threshold migration value depends on mutation (θ ; dashed line in the right of Figure 1.1b).

Genetic differentiation can also be estimated based on the effective number of alleles (the within-population Δ_S and total Δ_T effective number of alleles; JOST 2008). Similar to the above analyzed gene diversities, we can describe the degree to which Δ_S and Δ_T , respectively, are close to their isolated and panmictic value. This analysis defines two different functions of the scaled migration rate, one for Δ_S and one for Δ_T (see appendix S1). These functions have different threshold values that differ from the one defined in eq. 1.4. However, as the effective number of alleles and gene diversities are related, $\Delta = 1/(1 - H)$ (from JOST 2008), we can derive a direct relationship between $f(M)$ (eq. 1.4a) and the functions that describe Δ_S and Δ_T (see appendix S1). Thus, $f(M)$ also describes the behavior of Δ_S and Δ_T simultaneously.

Now, we investigate how generally the function, $f(M)$, captures various features of genetic differentiation. We analyze four statistics: the frequency of the most common allele in the total population p_{max} and in each population $\overline{p_{max}^i}$, the proportion of pri-

vate alleles p_{priv} and the mean singular value σ . p_{max} is close to 0 when several alleles are present in all populations, and is 1 when a single allele is fixed in all populations (the same allele). $\overline{p_{max}^i}$ is 0 when many alleles are present in each population and is 1 when one allele is fixed in each population (not necessarily the same allele). p_{priv} is 0 when all alleles are present at least in 2 populations and is close to 1 when most of the alleles are present in only 1 population. Finally, σ represents the degree of overlap of allele frequencies between populations. We can show that $\sigma = \sqrt{F_s/n}$ when the same alleles are present at the same frequency in each population (appendix S2), F_s being the gene identity (within-populations and total gene identities are equal under panmixia). When populations have different alleles, $\sigma = \sum_i \sqrt{F_{s,i}}$ (appendix S2), where $F_{s,i}$ is the within-population gene identity in population i . These four statistics are not commonly used to measure genetic differentiation, however they provide distinctive features of genetic differentiation. Interestingly, as shown in Figure 1.2, the four statistics have a common behavior across scaled migration rates and they follow the function, $f(M)$, in the three distinct regimes of mutation identified eq. 1.5. On the contrary, D and G_{ST} do not have the same behavior as the four statistics for all parameter values (see Figure S1). Figure 1.2 shows that $f(M) \simeq 1$ when all statistics are close to their isolation value and $f(M) \simeq 0$, when all statistics are close to their panmictic value. Figure 1.2 also shows that $f(M)$ and the four statistics are logarithmic functions of the scaled migration rate M . This demonstrates that the difference in order of magnitude of M and M_T determines the value of $f(M)$ and the four statistics and not their absolute differences.

Gene diversities, effective number of alleles and the four statistics describing different features of genetic differentiation follow the function $f(M)$ with threshold migra-

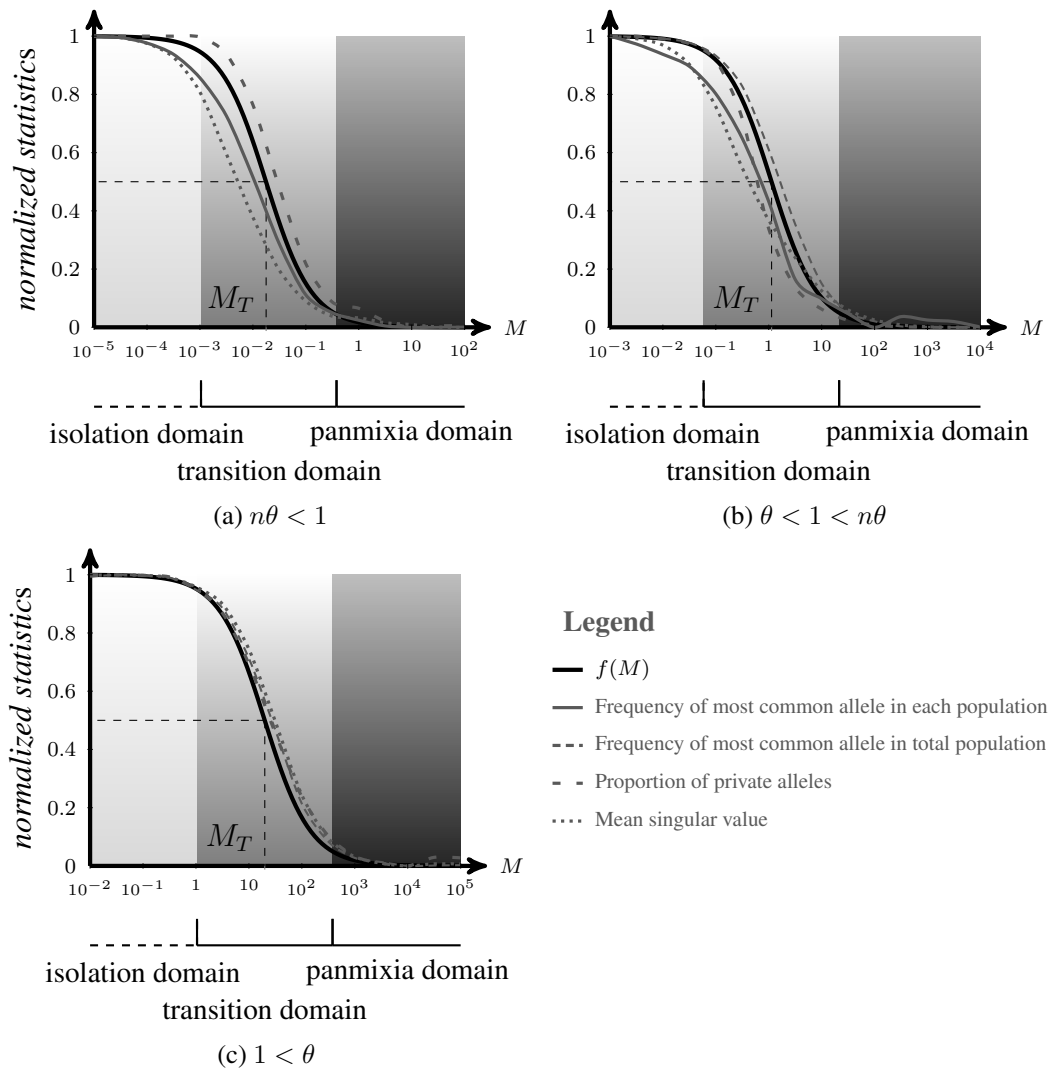


Figure 1.2 Behavior of four normalized statistics of allelic frequencies that describe features of genetic differentiation, as a function of the scaled migration rate M , and comparison with function $f(M)$. Results are presented for (a) weak mutation, (b) intermediate mutation and (c) strong mutation regimes. Statistics are: the frequency of the most common allele in the total population (p_{max}), the frequency of the most common allele in each population ($\overline{p_{max}^i}$), the proportion of private alleles (p_{priv}) and the mean singular value (σ). Migration is represented into three domains depending on how close the expected value is from isolation and panmixia: the isolation domain ($f(M) > 0.95$), the transition domain ($0.05 < f(M) < 0.95$) and the panmixia domain ($f(M) < 0.05$). Simulations are based on an island model (MAYNARD SMITH 1970) in which n populations exchange migrants at a rate m , generations are non-overlapping, and mutation follows an infinite allele model with mutation rate μ .

tion value M_T . Thus, $f(M)$ can be used as a reference function to describe the transition of allele frequencies from the isolation state to the panmictic state. $f(M)$ also allows the determination of the domain in which genetic differentiation measures, in particular G_{ST} and D , have strong signal in relation to changes in allele frequencies. But, the function $f(M)$ does not provide a direct measure of genetic differentiation as it is a reference function to compare commonly used measures of genetic differentiation.

WHAT CAN G_{ST} AND D ACTUALLY MEASURE?

In this section, we first summarize the relationship between gene diversity (the within-population gene diversity, H_s , and the total gene diversity, H_t) and G_{ST} as well as the relationship between the effective number of alleles (the within-population Δ_S and total Δ_T effective number of alleles) and D . Then, we determine the threshold migration value M , for which the behavior of G_{ST} and D changes in the three mutation regimes identified in eq. 1.5.

G_{ST} and D are derived from different normalizations of Nei's "absolute differentiation measure" $D_{ST} = H_t - H_s$ (NEI 1973). As D_{ST} is not bounded by 0 and 1, only the normalized D_{ST} value would allow comparison and ranking of different structured populations. The relationships between measures D_{ST} , G_{ST} and D are as follows:

$$G_{ST} = \frac{D_{ST}}{H_t} \quad (1.6)$$

$$D = \frac{D_{ST}}{\frac{n-1}{n}(1 - H_s)} \quad (1.7)$$

Equations 1.6 and 1.7 show that G_{ST} corresponds to the normalization of D_{ST} by H_t (i.e., its maximum value for a given H_t) and D corresponds to the normalization of D_{ST} by $\frac{n-1}{n}(1 - H_s)$ (i.e., its maximum value for a given H_s). Thus, G_{ST} and D rely on a different conception of differentiation. With G_{ST} , populations are considered completely

differentiated when they have a minimum within-population diversity (i.e., $H_s = 0$), while with D , populations are considered completely differentiated when they have a maximum total gene diversity (the maximum value of H_t given H_s is $H_t = \frac{n-1}{n} + \frac{H_s}{n}$). In consequence, G_{ST} , in opposition to D , does not estimate how close to isolation or panmixia populations are for monomorphic, or almost monomorphic populations.

The normalization of D_{ST} by H_t or by $\frac{n-1}{n}(1 - H_s)$, that leads to G_{ST} and D , determines their relationship with migration. As for the gene diversities (eq. 1.4), the behavior of D_{ST} can be described by a function $f(M)$ of the scaled migration rate with a threshold value M_T . Assuming that population sizes are large and mutation and migration rates are small (removing all term in $1/N^2, \mu^2, m^2, \mu/N, m/N, \mu m$), we have the following relationship between D_{ST} and $f(M)$:

$$D_{ST} = \frac{n-1}{n} \frac{1}{1+\theta} (1 - f(M)) \quad (1.8)$$

Thus, G_{ST} can be described by a function of the scaled migration rate, $f_G(M)$:

$$G_{ST} = \frac{1}{1 + \frac{n}{n-1}\theta} f_G(M) \quad (1.9a)$$

Where

$$f_G(M) = \frac{M_G}{M_G + M} \quad (1.9b)$$

with threshold migration value $M_G = (\frac{n-1}{n})^2 + \frac{n-1}{n}\theta$. Note that G_{ST} is 0 when migration is very strong but is not upper bounded by 1, except under the weak mutation regime ($\theta \ll 1$), where $G_{ST} \simeq f_G(M)$.

D can also be described by a function of the scaled migration rate, $f_D(M)$:

$$\begin{aligned} D &= f_D(M) \\ &= \frac{M_D}{M_D + M} \end{aligned} \tag{1.10}$$

with threshold migration value $M_D = (n - 1)\theta$. Note that eq. 1.10 is equivalent to eq. 17 from JOST (2008), but the expression here does not require $\mu \ll m$ to be valid (only $m \ll 1$ and $\mu \ll 1$). Also D is 0 when migration is very strong, 1 when there is no migration, and it is a monotonic function of M .

The threshold values M_G and M_D of the function $f_G(M)$ and $f_D(M)$ are different and, more importantly, this difference depends on the mutation regime considered. In Table 1.2, the values of the threshold M values for each function that describes the transition from isolation to panmixia of H_s , H_t , D_{ST} , G_{ST} and D in the three mutation regimes identified previously (equations 1.4a, 1.8, 1.9 and 1.10) are presented.

Table 1.2 Summary of threshold values M

Mutation regime	Threshold M value		
	H_s, H_t & D_{ST}	G_{ST}	D
$n\theta < 1$	$(n-1)\theta$	$\left(\frac{n-1}{n}\right)^2$	$(n-1)\theta$
$\theta < 1 < n\theta$	$\frac{n-1}{n}$	$\left(\frac{n-1}{n}\right)^2$	$(n-1)\theta$
$1 < \theta < n\theta$	$\frac{n-1}{n}\theta$	$\frac{n-1}{n}\theta$	$(n-1)\theta$

We now investigate how the differences between threshold values of G_{ST} and D translate into differences in the detection of nearness to panmixia and isolation, and the parameter space for which G_{ST} and D provide an informative signal, in the three mutation regimes.

Under the weak mutation regime ($n\theta < 1$), the within population gene diversity, H_s , is smaller than 0.5 whatever the value of the scaled migration rate, M . Low H_s values indicate that in most populations one allele has a high frequency (frequency $p > 0.5$; see examples Figure S2). However, the value of the total gene diversity, H_t is close to

1 when migration is low but quickly decreases with migration. Thus, in this regime, measure of differentiation, G_{ST} , signals a strong genetic differentiation (Table 1.3), even when migration increases and the total gene diversity, H_t , is very low. Consequently, when only one population has a different genetic composition from many others, G_{ST} has a high value (Figure 1.3d; Figure S2). When H_s is low, G_{ST} has a non-linear signal across changes in H_t (Figure 1.3a). In this situation, G_{ST} is very sensitive to departure from the panmictic state (Figure 1.3d-f; Figure S2). In contrast to G_{ST} , in this regime, D varies almost linearly with H_t and can detect nearness to panmixia and isolation when populations are either monomorphic or polymorphic (Figure 1.3a-f; Figure S2). In this regime, D is monotonic and bounded by 0 and 1. Furthermore, we have $D \simeq f(M)$, thus the domain where D signals genetic differentiation corresponds exactly to the transition of allele frequencies from isolation to panmixia signaled by $f(M)$.

Table 1.3 Genetic differentiation measure values in the 3 mutation regimes

Mutation regime (<i>parameter values</i>)	M	$f(M)$	G_{ST}	D
$\theta (=10^{-3}) < n\theta (=10^{-2}) < 1$	10^{-3}	0.8992	0.9988	0.9000
	10^{-1}	0.0819	0.8892(!)	0.0825
$\theta (=10^{-1}) < 1 < n\theta (=100)$	10^{-1}	0.9158	0.8333	0.9990
	10	0.0981	0.0868	0.9058(!)
$1 < \theta (=10) < n\theta (=1000)$	1	0.9158	0.0825(!)	0.9990
	100	0.0981	0.0088	0.9079(!)

Under the intermediate mutation regime ($\theta < 1 < n\theta$), while the value of the within population gene diversity H_s increases with the migration rate, the total gene diversity H_t is high for any migration value. G_{ST} varies almost linearly with $1 - H_s$, while D changes when $1 - H_s$ is close to 0 (Figure 1.4a). Thus, D signals strong genetic differentiation, even when allele frequencies are close to their expected value under panmixia ($f(M) \simeq 0$, Table 1.3, Figure S3). Also, as shown in Figure 1.4b-d, when populations are polymorphic with similar allele frequencies (weak differentiation), G_{ST} signals no

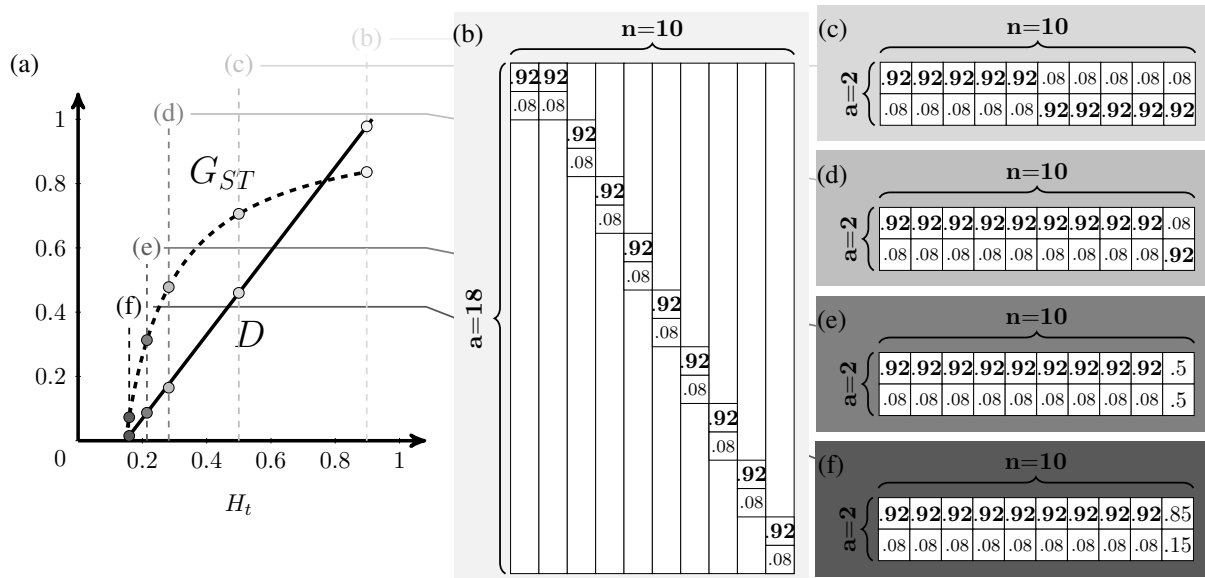


Figure 1.3 Difference in sensitivity of G_{ST} and D to detect genetic differentiation when the within-population gene diversity is low (illustrative value for $H_s = 0.15$). (a) G_{ST} and D as a function of H_t . (b)-(f) Illustrative cases of alleles distributions for different H_t and number of alleles a in $n = 10$ populations, assuming that populations are bi-allelic with one allele at frequency 0.92 and the other at frequency 0.08. The reported allele frequency values match the ones obtained by simulations under the weak mutation regime (Figure S2). Each row represents an allele (frequencies are indicated when the allele is present) and each column a population. (b) Two populations have the same alleles at the same frequencies, the others have only private alleles; $H_t \simeq 0.90$, $G_{ST} \simeq 0.84$ and $D \simeq 0.98$; (c) half of the populations have the same alleles at the same frequencies; $H_t = 0.5$, $G_{ST} \simeq 0.71$ and $D \simeq 0.46$; (d)-(f) All populations have the same alleles at the same frequencies except one. (d) $H_t \simeq 0.28$, $G_{ST} \simeq 0.48$ and $D \simeq 0.17$. (e) $H_t \simeq 0.21$, $G_{ST} \simeq 0.31$ and $D \simeq 0.09$. (f) $H_t \simeq 0.16$, $G_{ST} \simeq 0.07$ and $D \simeq 0.02$. D decreases linearly with the number of genetically identical populations while G_{ST} values are sensitive to small departure from panmixia.

differentiation. In contrast, in this case, D signals strong differentiation (see Figure 1.4). Interestingly, a high D value is observed even when only one population has a different genetic composition than the others (Figure 1.4d, which can occur when migration is strong, see Figure S3). $G_{ST} \simeq f(M)$, thus the domain where G_{ST} signals differentiation corresponds to the transition signaled by $f(M)$.

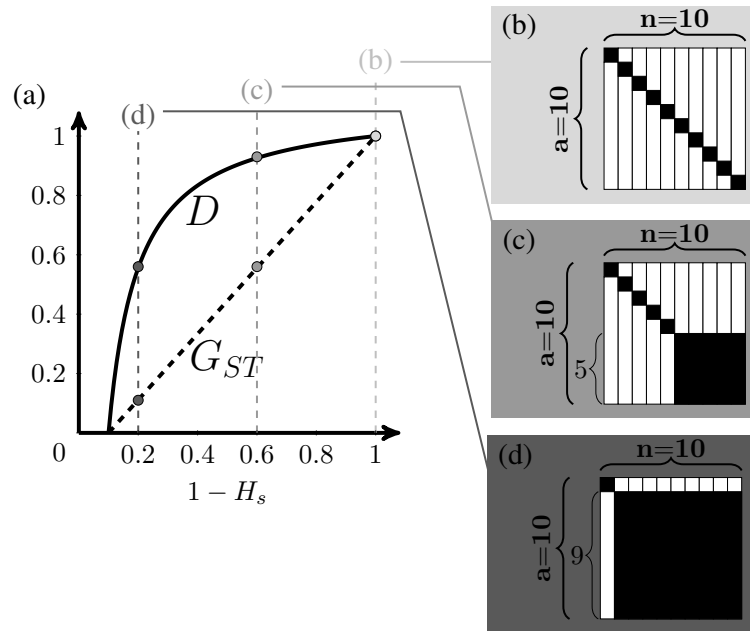


Figure 1.4 Difference in sensitivity of G_{ST} and D to detect genetic differentiation when the total gene diversity is high (i.e. when $H_t \simeq 1$). (a) G_{ST} and D as a function of $1 - H_s$, for a high value of H_t (0.9). (b)-(d) Illustrative cases of alleles distributions for different $1 - H_s$ and number of alleles a in $n = 10$ populations. Few populations can have private alleles while others share alleles at similar frequencies (for illustrative values obtained by simulations see Figure S3). (b) Different alleles are fixed in different populations, $H_s \simeq 0$, $G_{ST} = 1$ and $D = 1$; (c) half of the populations shares five alleles at the same frequency and the other half populations has each one allele fixed, $H_t \simeq 0.4$, $G_{ST} \simeq 0.56$ and $D \simeq 0.93$; (d) all populations share nine alleles at the same frequency, except one with $H_s \simeq 0.8$, $G_{ST} \simeq 0.11$ and $D \simeq 0.56$. G_{ST} decreases with the number of genetically identical populations while D values are high in (b), (c) and (d).

Under the strong mutation regime ($1 < \theta$), the total gene diversity and the within-population gene diversity are very high whatever the value of migration ($H_s \simeq 1$ and $H_t \simeq 1$). G_{ST} signals weak genetic differentiation, as populations are always polymorphic (low fixation level) and is not upper bounded by 1 (fig S4). When $n = 2$, D corresponds exactly to the transition from isolation to panmictic state signalled by

$f(M)$ (as then $f(M) \simeq D$). However, D values increase when the number of populations n is large (see Figure 1.5). Indeed, when $n \gg 1$, we have $f(M) < D$. Thus, D is very sensitive to departures from the expected panmictic values when n is large and reflects that under these conditions (numerous alleles at low frequency, a large number populations), correlations between allele frequencies from different populations might not be expected (fig S4, even under panmixia, allele frequencies might not be identical in all populations). D then signals well how allele frequencies are differentiated between any two demes.

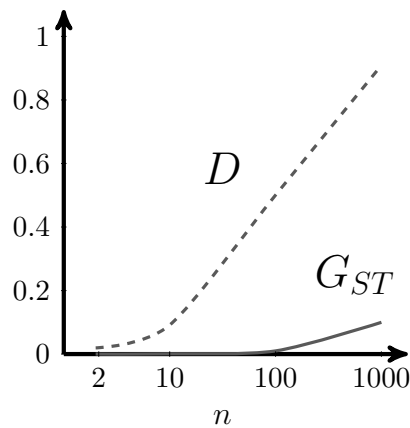


Figure 1.5 Behavior of G_{ST} and D as a function of the number of populations n in panmixia when both the within-population and total gene diversities are high (i.e. H_s and H_t close to 1). D detects genetic differentiation (values superior to 0) when n increases, while G_{ST} detects panmixia (values close to 0). As the number of polymorphic populations increases, the number of alleles present in few copies in few populations increased (as shown Figure S4). As D estimates genes overlap between populations, this results in an overestimation of genetic differentiation by D .

USES OF G_{ST} AND D TO ESTIMATE GENETIC DIFFERENTIATION

Our results have direct implications for the use of differentiation measures. We define a function $f(M)$ that considers genetic differentiation as the degree to which allele frequencies are close to what is expected under isolation or panmixia (see Figure 1.2). Under this definition, we provide in Table 1.4 the measure that is the most appropriate to estimate genetic differentiation given the mutation regime (assuming a finite island

model and the infinite allele model). In empirical studies, only H_s and H_t are commonly available and the values of n and θ are often unknown. Assuming again an island model, an infinite allele model, and equilibrium conditions, it is possible to infer the mutation regime using the values of H_s and H_t . Indeed, we know that H_s and H_t are bounded as follows:

$$\begin{cases} \frac{\theta}{1 + \theta} < H_s < \frac{n\theta}{1 + n\theta} \\ \frac{n\theta}{1 + n\theta} < H_t < 1 - \frac{1}{n(1 + \theta)} \end{cases} \quad (1.11)$$

Table 1.4 Differentiation measures G_{ST} and D uses when parameters n and θ are known

Mutation regime	Measure to use
$n\theta < 1$	D
$\theta < 1 < n\theta$	G_{ST}
$1 < \theta < n\theta$	D^a

^a In this case, D overestimates the level of genetic differentiation; the lower the number of populations n , the more accurate D

Therefore, $H_s < 0.5$ implies that $\frac{\theta}{1+\theta} < 0.5$ and thus $\theta < 1$ which corresponds to the weak or intermediate mutation regime. In a similar way, we can show that $H_s > 0.5$ implies $n\theta > 1$ which corresponds to the intermediate or high mutation regime, and finally $H_t < 0.5$ implies that $n\theta < 1$ which is the weak mutation regime. Consequently, (results summarized Table 1.5). (i) When $H_t < 0.5$, D measures accurately the transition from isolation to panmixia under the weak mutation regime. (ii) When $H_s < 0.5$, either the weak or the intermediate mutation regime apply. When the mutation regime is weak, D measures accurately the transition from isolation to panmixia while G_{ST} is very sensitive to departure from panmictic state (as the threshold migration value of G_{ST} is $1 > M_{T=n\theta}$; Table 1.2). When the mutation regime is intermediate, G_{ST} measures accurately the transition from isolation to panmixia while D is very sensitive to

departure from panmictic state (as the threshold migration value of D is $n\theta > M_T=1$; Table 1.2). Interestingly in those two cases, the lowest value of G_{ST} and D accurately measures the transition. (iii) When $H_s > 0.5$, either the intermediate or the strong mutation regime can apply. When the mutation regime is intermediate, G_{ST} measures accurately the transition and D overestimates it (as its threshold migration value is $n\theta > M_T = 1$; Table 1.2). In the strong mutation regime, G_{ST} weakly detects nearness to the isolation state and D is very sensitive to departure from panmictic state (as the threshold migration value of D is $n\theta > M_T$; Table 1.2). Moreover in this regime, the number of populations n can also rise the values of D .

Table 1.5 Differentiation measures G_{ST} and D uses when parameters n and θ are unknown

Condition	Mutation regime	Measure to use
$H_t < 0.5$	<i>weak</i> mutation ($n\theta < 1$)	D
$H_s < 0.5$	<i>weak</i> or <i>intermediate</i> mutation ($\theta < 1$)	$\min(D, G_{ST})$
$H_s > 0.5$	<i>intermediate</i> or <i>strong</i> mutation ($n\theta > 1$)	G_{ST} (if $\theta < 1$) ^a or D ^b

^a The condition is necessary for G_{ST} to measure correctly genetic differentiation, though θ is usually unknown

^b In this case, D overestimates the level of differentiation; the lower the number of populations n (usually unknown), the more accurate D

Our results imply that assuming an infinite allele model, for low mutation rate markers such as SNPs and allozymes, D should be privileged when there are few populations and G_{ST} when the number of populations is large. For high mutation rate markers, such as microsatellites, D should be favoured, although its value depends on n . In summary, to capture genetic differentiation and to have a complete understanding of genetic differentiation, it is important to estimate the within population gene diversity H_s , the total gene diversity H_t and to use both G_{ST} and D values.

USES OF G_{ST} AND D TO ESTIMATE DEMOGRAPHIC PARAMETERS

The coalescent F_{ST} provides a good estimate of demographic parameters under the island model (WHITLOCK 2011). From our results we can show that G_{ST} and D can also be used to estimate demographic parameters. SLATKIN (1993) already showed that under low mutation rates, G_{ST} performs very well to estimate the amount of migration between populations - the number of migrants per generation in an island model Nm . Besides, WHITLOCK (2011) showed that G_{ST} should be preferred to D to infer demographic parameters as it is less sensitive to the strength of mutations. Our results showed that D and G_{ST} are sensitive to different aspects of genetic diversity among populations that are reflected by the mutation regimes. Together they provide complementary measures to infer demographic parameters of populations. Indeed, assuming that n is known, assuming equilibrium conditions (following assumptions in SLATKIN 1993), and using results in equations 1.9 and 1.10, estimators of the number of migrants per generation, Nm , and the number of mutants per generation, $N\mu$, can be derived as follows:

$$\begin{cases} \widehat{Nm} = \frac{1}{4} \left(\frac{1 - G_{ST}}{G_{ST}} \right) \left(\frac{n-1}{n} \right)^2 \left(\frac{1-D}{1 - \frac{n-1}{n}D} \right) \\ \widehat{N\mu} = \frac{1}{4} \left(\frac{1 - G_{ST}}{G_{ST}} \right) \left(\frac{n-1}{n^2} \right) \left(\frac{D}{1 - \frac{n-1}{n}D} \right) \end{cases} \quad (1.12)$$

The estimator of Nm , \widehat{Nm} in eq. 1.12 corresponds to SLATKIN (1993)'s estimator, corrected by a ratio $\frac{1-D}{1 - \frac{n-1}{n}D}$. Thus, eq. 1.12 corresponds to SLATKIN (1993)'s estimator of \widehat{Nm} when D is close to 0 or n is large. For example, when $n = 100$, the ratio $\frac{1-D}{1 - \frac{n-1}{n}D}$ is larger than 0.95 for all $D < 0.84$. The estimator of $N\mu$, $\widehat{N\mu}$ in eq. 1.12 was also derived in JOST (2008).

DISCUSSION

Here, we present a function f that captures the main features of the transition of allele frequencies from their expected isolation state to their panmictic state, under the finite island model. This function allows a deep analysis of the properties of genetic differentiation measures G_{ST} and D . With this function, we found that G_{ST} and D display accurate signals of genetic differentiation (as defined above) in three different, specific and restricted mutation regimes (Tables 1.4, 1.5). We derive the limits of those domains and discuss the parameter domain for which the use of G_{ST} or D should be favored. The limits of the three mutation regimes identified depend on the number of mutants per generation $\theta/2$ and on the number of populations n . Interestingly, the limit of the weak mutation regime presented here ($n\theta$) corresponds to the threshold value provided by MAYNARD SMITH (1970) that determines the domain in which genetic diversity between different populations is maintained. Indeed, MAYNARD SMITH (1970) found that under the finite island model, when $M \ll n\theta$, genes from different populations are likely to have different alleles, and when $M \gg n\theta$, the genetic diversity is shared between populations, as expected in panmictic populations. This threshold value also corresponds to the threshold value of D (eq. 1.10, and eq. 17 in JOST 2008), which supports the use of D to understand genetic differentiation in the weak mutation regime.

Our description of the transition of allele frequencies from isolation to panmictic state using f reflects the transition of a set of summary statistics (number of private alleles, gene diversities, effective number of alleles, frequency of the most common allele in each population and among populations, mean singular value). Those summary statistics provide a precise meaning to the transition function f . Therefore the use of f allows a deep understanding of differentiation measures by comparing their expected

behavior with f . However, other summary statistics exist and some require new expression for f . For example, using the ratio of within Δ_S and total effective number of alleles Δ_T as summary statistics leads to a function $f_2(M) = D$ (normalizing Δ_S/Δ_T between 0 and 1, as in JOST 2008). Alternatively, normalizing the ratio Δ_T/Δ_S leads to function $f_3(M) = \frac{\frac{n-1}{n}\theta}{\frac{n-1}{n}\theta + M}$. Interestingly, these functions correspond to the function $f(M)$ derived here, under the weak and strong mutation regimes, respectively.

An appropriate measure of genetic differentiation should accurately describe the distribution of alleles among populations. We highlight a complex dependency of genetic differentiation to the mutation rate, population size and migration rate, characterized by function f . This dependency reflects the complex interaction between processes which tend to homogenize (migration) or differentiate (genetic drift, mutation) allele frequencies between populations. While a measure independent of the mutation rate, such as the coalescent F_{ST} , provides insight into important evolutionary forces (WHITLOCK 2011), our results show that it might fail to describe the main features of the distribution of alleles among populations (e.g., Figure 1.3 and S2). For example, the number of private alleles in each population results from the interaction between mutation (generating private alleles) and migration (sharing these alleles).

A large body of literature described the "one migrant per generation rule" (OMPG). This rule states that 1 migrant per generation is the threshold number of migrants to maintain genetic diversity in a structured population, whatever the mutation rate (SPIETH 1974; SLATKIN 1993; MILLS and ALLENDORF 1996). Our results suggest, in agreement with MAYNARD SMITH (1970), that depending on the mutation regime, less or more than one migrant might be enough to reach the highest possible genetic diversity. Our results show that both the threshold identified by MAYNARD SMITH (1970), and the OMPG describe how close populations are to panmixia and isolation states,

but in different mutation regimes. Our results also highlight, as in JOST (2008), the importance of the relative values of $n\theta$ and the number of migrants per generation $M/2$ to characterize the level of differentiation of populations.

When mutation rates are low, in agreement with GREGORIUS (2010), we found that G_{ST} describes accurately the fixation of alleles in populations, but not the transition between panmictic and isolation states (Figure 1.3), even when populations are not monomorphic. G_{ST} values are high regardless of the similarity of allele frequencies between populations. This result is surprising, as G_{ST} is commonly thought as possibly misleading only when H_s is high, and provide reliable estimates when H_s is low (HEDRICK 2005). Although G_{ST} values are not constrained by the value of H_s , when mutation rates are low it might not accurately measure transition from isolation to panmictic states of populations in this domain. We also identify an intermediate mutation regime, where G_{ST} characterizes the transition from isolation state to panmictic state better than D . We show that in the strong mutation regime, both G_{ST} and D can be weakly sensitive to the transition of allele frequencies from isolation state to panmictic state; this result was surprising, as D was developed to solve the unreliability of G_{ST} in this domain (JOST 2008). Our results also show that the total number of populations that belongs to the studied systems, n , (not the number of sampled populations), is important and can impact measures of genetic differentiation. Indeed, we demonstrate (Figure 1.5, Tables 1.2 and 1.3) that population genetic differentiation estimated with D rises when the number of population is large, even under panmixia.

One important point of our study is that D and G_{ST} are complementary, as they are sensitive to different features of the genetic diversity among populations: G_{ST} detects the level of differentiation based on allele fixation and D the difference in genetic composition. Depending on the requested information on genetic differentiation, they

can be used in a complementary manner. This is illustrated in Figures 1.3, 1.4 and 1.5, where, under the weak mutation regime, G_{ST} (but not D) can detect differentiation when a single population is genetically different from the others, while under the intermediate mutation regime, D (but not G_{ST}) can detect differentiation when a single population is genetically different from the others. The complementarity of D and G_{ST} can also be used to infer the mutation and demographic parameters of populations. Indeed, SLATKIN (1993)'s estimator of the number migrants per generation Nm in an island model (assuming an infinite allele model), \widehat{Nm} , which is valid only under weak mutation, can be generalized to any mutation rate. Although parameter Nm is expected to be the same for all loci, \widehat{Nm} must be estimated for each locus independently to correct for differences in mutation rates as they affect differently G_{ST} and D . Estimators of the other model parameters such as the number of mutants per generation $N\mu$ can also be inferred using both D and G_{ST} . Using Estimators in eq. 1.12 in an approximate Bayesian computation framework (BEAUMONT *et al.* 2002), for example, is expected to refine estimates of the model parameters in more complex scenarii as we showed that they are sensitive to different processes.

Results presented here assume equilibrium gene diversities and rely on specific migration and mutation models. The mutation regimes identified and presented here are derived assuming an island model, however, they can be generalized to any migration model which has the isolation and island model as a limit model. Indeed, as long as gene diversities are bounded by the values of H^{iso} and H^{pan} , they depend on the number of populations, n , and on the scaled mutation, θ , thus the three mutation regimes identified apply. Results presented here thus cannot be generalized to stepping-stone models (1-dimension or 2-dimension, finite, circular or infinite) but, to models where the migration has an unconstrained dispersal kernel.

We also assume an infinite allele model. The use of other mutation models, such as the finite allele model, or the stepwise mutation model is expected to lead to different results. Indeed, under those mutation models a lower genetic diversity will be achieved, even under strong mutation (H_s and H_t could be lower than 1 whatever the mutation rate). Thus, under the intermediate and strong mutation regimes, results are expected to differ from our predictions. Other measures of genetic differentiation based on different mutation models could provide additional insight into genetic differentiation under the strong mutation regime (e.g. ϕ_{ST} for haplotype data; EXCOFFIER *et al.* 1992; R_{ST} for microsatellite data SLATKIN 1995).

G_{ST} converges faster than D to its equilibrium values when mutation rate is low, and the opposite is true when the mutation rate is high (RYMAN and LEIMAR 2008). Thus, D might better reflected the level of genetic differentiation in non-equilibrium populations (JOST 2009). However, given the complexity and the diversity of processes that can lead to transient dynamics of genetic diversity (e.g. population size changes, migration changes and selection event), further investigations are needed to disentangle for each of these events the expected impact on measures of genetic differentiation. Moreover the advantage or disadvantage of a slow or fast rate of convergence depends on the questions investigated.

To have an accurate picture of the actual distribution of alleles among populations, we suggest that population genetics studies should investigate the two genetic diversity measures H_s , H_t , and the two genetic differentiation measures D and G_{ST} . Together, they provide a better characterization of the partition of genetic diversity among populations than any single measure.

ACKNOWLEDGEMENT

We thank John Pannell, Lou Jost and two anonymous reviewers for their comments and suggestions that strongly improved our manuscript. This project was funded by the Swiss National Science Foundation (SNF) grants #PZ00P3_139421/1 and #31003A-130065.

BIBLIOGRAPHY

- BALLOUX, F., H. BRÜNNER, N. LUGON-MOULIN, J. HAUSSER, and J. GOUDET, 2000
Microsatellites can be misleading: an empirical and simulation study. *Evolution* **54**:
1414–1422.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian com-
putation in population genetics. *Genetics* **162**: 2025–2035.
- CHARLESWORTH, B., 1998 Measures of divergence between populations and the effect
of forces that reduce variability. *Molecular Biology and Evolution* **15**: 538–543.
- CROW, J. F., 1986 *Basic concepts in population, quantitative, and evolutionary genetics..* WH
Freeman and Company.
- EXCOFFIER, L., P. E. SMOUSE, and J. M. QUATTRO, 1992 Analysis of molecular vari-
ance inferred from metric distances among dna haplotypes: application to human
mitochondrial dna restriction data. *Genetics* **131**: 479–491.
- GERLACH, G., A. JUETERBOCK, P. KRAEMER, J. DEPPERMANN, and P. HARMAND,
2010 Calculations of population differentiation based on G_{ST} and D : forget G_{ST} but
not all of statistics! *Molecular Ecology* **19**: 3845–3852.

GOLUB, G., and W. KAHAN, 1965 Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* **2**: 205–224.

GREGORIUS, H.-R., 2010 Linking diversity and differentiation. *Diversity* **2**: 370–394.

HEDRICK, P. W., 1999 Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313–318.

HEDRICK, P. W., 2005 A standardized genetic differentiation measure. *Evolution* **59**: 1633–1638.

HELLER, R., and H. R. SIEGISMUND, 2009 Relationship between three measures of genetic differentiation G_{ST} , D_{EST} and G'_{ST} : how wrong have we been? *Molecular Ecology* .

HOLSINGER, K. E., and B. S. WEIR, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} . *Nature Reviews Genetics* **10**: 639–650.

JAKOBSSON, M., M. D. EDGE, and N. A. ROSENBERG, 2013 The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics* **193**: 515–528.

JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.

JOST, L., 2009 D vs. G_{ST} : Response to Heller and Siegismund (2009) and Ryman and Leimar (2009). *Molecular Ecology* **18**: 2088–2091.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.

- KRONHOLM, I., O. LOUDET, and J. DE MEAUX, 2010 Influence of mutation rate on estimators of genetic differentiation—lessons from *Arabidopsis thaliana*. *BMC Genetics* **11**: 33.
- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theoretical Population Biology* **1**: 273–306.
- MAYNARD SMITH, J., 1970 Population size, polymorphism, and the rate of non-darwinian evolution. *The American Naturalist* **104**: 231–237.
- MEIRMANS, P. G., and P. W. HEDRICK, 2011 Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources* **11**: 5–18.
- MILLS, L. S., and F. W. ALLENDORF, 1996 The one-migrant-per-generation rule in conservation and management. *Conservation Biology* **10**: 1509–1518.
- NAGYLAKI, T., 1998 Fixation indices in subdivided populations. *Genetics* **148**: 1325–1332.
- NEI, M., 1973 Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences U.S.A.* **70**: 3321–3323.
- RYMAN, N., and O. LEIMAR, 2008 Effect of mutation on genetic differentiation among nonequilibrium populations. *Evolution* **62**: 2250–2259.
- RYMAN, N., and O. LEIMAR, 2009 G_{ST} is still a useful measure of genetic differentiation - a comment on Jost's D . *Molecular Ecology* **18**: 2084–7; discussion 2088–91.
- SLATKIN, M., 1985 Rare alleles as indicators of gene flow. *Evolution* **39**: pp. 53–65.
- SLATKIN, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**: pp. 264–279.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.

SPIETH, P. T., 1974 Gene flow and genetic differentiation. *Genetics* **78**: 961–965.

WANG, J., 2012 On the measurements of genetic differentiation among populations. *Genetics research* **94**: 275–289.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.

WHITLOCK, M. C., 2003 Fixation probability and time in subdivided populations. *Genetics* **164**: 767–779.

WHITLOCK, M. C., 2011 G'_{ST} and D do not replace F_{ST} . *Molecular Ecology* **20**: 1083–1091.

WRIGHT, S., 1951 The genetical structure of populations. *Annals of Eugenics* **15**: 323–354.

Chapter 2 Peak and Persistent Excess of Genetic Diversity Following an Abrupt Migration Increase

Nicolas Alcala, Daniela Streit, Jérôme Goudet, and Séverine Vuilleumier
Department of Ecology and Evolution, Biophore, University of Lausanne, CH-1015
Lausanne, Switzerland

Genetic diversity Migration Population structure Transient dynamics Standing genetic variation

*Published, **Genetics** March 1, 2013 vol. 193 no. 3 953-971*

ABSTRACT

Genetic diversity is essential for population survival and adaptation to changing environments. Demographic processes (e.g., bottleneck and expansion) and spatial structure (e.g., migration, number and size of populations) are known to shape the patterns of the genetic diversity of populations. However, the impact of temporal changes in migration on genetic diversity has seldom been considered, although such events might be the norm. Indeed, during the millions of years of a species' lifetime, repeated isolation and reconnection of populations occur. Geological and climatic events alternately isolate and reconnect habitats. We analytically document the dynamics of genetic diversity after an abrupt change in migration given the mutation rate and the number and sizes of the populations. We demonstrate that during transient dynamics, genetic diversity can reach unexpectedly high values that can be maintained over thousands of generations. We discuss the consequences of such processes for the evolution of species based on standing genetic variation and how they can affect the reconstruction of a population's demographic and evolutionary history from genetic data. Our results also provide guidelines for the use of genetic data for the conservation of natural populations.

GENETIC diversity in a population of constant size results from the balance between the occurrence of new mutations and the loss of alleles by genetic drift (FISHER 1922; WRIGHT 1931; KIMURA and CROW 1964). The expected population genetic diversity can thus be estimated from the effective population size and the mutation rate in the population. In subdivided populations this estimate should further account for the strength of migration (MARUYAMA 1970; MAYNARD SMITH 1970; NEI 1973): limited migration allows for strong differentiation between populations, while strong migration tends to homogenize genetic diversity between populations. Genetic diversity is also known to be impacted by population demographic changes; following bottlenecks and founder events, a loss of genetic diversity is expected to occur (NEI *et al.* 1975). Recently, spatial population expansions were shown to lead to increased differentiation between populations and to generate a low level of genetic diversity at the front of the expansion (EXCOFFIER *et al.* 2009).

Although theoretical studies on the dynamics of genetic diversity in subdivided populations started appearing in the 1970s (NEI and FELDMAN 1972; LATTER 1973; NEI 1973; NAGYLAKI 1974, 1977), the transient dynamics and non-equilibrium states of genetic diversity still do not have a good theoretical basis. Early authors characterized the ultimate rate of change of genetic diversity after a perturbation (either a change in population size or gene flow; NEI and FELDMAN 1972; LATTER 1973; NEI 1973; NAGYLAKI 1974, 1977). They found that changes in genetic diversity are related to the total effective population size, which results in a slow dynamics of genetic diversity change. They thus first highlighted that non-equilibrium states and transient dynamics are expected to act on very large temporal scales. In particular, they showed that decreases in migration rates (population fragmentation or isolation) have long-

term effects on genetic diversity: they reduce the amount of genetic diversity within populations and allow for population differentiation (LATTER 1973; TAKAHATA and NEI 1985). Additionally, it has been shown that short timescale random fluctuations in migration increase population differentiation (NAGYLAKI 1979; WHITLOCK 1992; RICE and PAPADOPOULOS 2009) while cyclic fluctuations of gene flow (such as seasonal fluctuations) mainly impact genetic diversity within populations (KARLIN 1982; SHPAK *et al.* 2010). Although the genetic consequences of migration events (admixture) have recently received much attention (e.g., PRITCHARD *et al.* 2000; FALUSH *et al.* 2003; PRICE *et al.* 2009; GRAVEL 2012), their impact on genetic diversity and more particularly the expected induced transient dynamics have not received much attention.

Genetic diversity has a crucial importance in estimating populations at risk of extinction and species' adaptive potential. Current genetic diversity characterizes species at risk of extinction through inbreeding depression, loss of genetic diversity, and accumulation of deleterious mutations (GILPIN and SOULE 1986; JIMENEZ *et al.* 1994; FRANKHAM 1995; HEDRICK and KALINOWSKI 2000). The current level of genetic diversity (or standing genetic variation) is now widely recognized as a determinant for the adaptation of a population to a novel environment (TURNER *et al.* 1993; FEDER *et al.* 2003; PELZ *et al.* 2005; COLOSIMO *et al.* 2005; HERMISSON and PENNINGS 2005; MYLES *et al.* 2005; HERNANDEZ *et al.* 2011; JONES *et al.* 2012). First, under new selective pressures, the adaptive value of a pre-existing allele can switch from neutral or deleterious to beneficial (GIBSON and DWORKIN 2004; HERMISSON and PENNINGS 2005). Second, alleles from the standing genetic variation are present at higher frequencies in the population than any newly arisen (*de novo*) mutation are, thus they have higher fixation probabilities and lower times to fixation (BARRETT and SCHLUTER 2008). Finally,

these alleles have already passed successive selective filters and are consequently more likely to be compatible with the background genome (ORR and BETANCOURT 2001; SCHLUTER *et al.* 2004; BARRETT and SCHLUTER 2008).

Measures of genetic diversity are widely used to understand and infer the demographic and evolutionary history of populations. Indeed, statistical tests using polymorphism data can detect departure from neutrality and infer demographic or selective processes (e.g. EWENS 1972; WATTERSON 1978; TAJIMA 1983; FU and LI 1993; FAY and WU 2000, see review in KREITMAN 2000). Furthermore, due to recent modeling advances in coalescent theory and increased genomic data and computational power, it is now possible to distinguish different demographic scenarios (e.g. population bottleneck and subdivision, PETER *et al.* 2010) and estimate demographic and selective parameters (e.g. populations size and growth rate, proportion of admixture, selection coefficient) using polymorphism data (BEAUMONT *et al.* 2002; KIM and STEPHAN 2002; KIM and NIELSEN 2004; NIELSEN *et al.* 2005; PRICE *et al.* 2009). Nevertheless, it is often difficult to distinguish between the transient effects of demographic changes and the effects of selection on polymorphism data (JENSEN *et al.* 2005; NIELSEN 2005; LI and STEPHAN 2006; KIM and GULISIJA 2010; PAVLIDIS *et al.* 2010). It is also difficult to distinguish between the signatures of different demographic changes such as changes in population size, number or migration rate (WAKELEY 1999). A better understanding of the impact on genetic data of transient dynamics during demographic changes is necessary to disentangle these processes.

Interestingly, although the impact of population subdivision and short timescale population demographic changes on genetic diversity have received a lot of atten-

tion, other processes, such as long-term isolation and subsequent population reconnection, have received little attention. Such events have, without a doubt, occurred several times in the past, at long and short timescales. Repeated environmental changes have modified habitats and species distribution and created isolation and reconnection of populations. For example, during the climatic oscillations of the Quaternary period, temperate and tropical species were successively isolated into refugia and experienced habitat and population expansion, allowing for population reconnection (HEWITT 2000, 2004; ZHANG *et al.* 2008; YOUNG *et al.* 2009). At the same time, the reduction of sea levels (120 m lower than present, LAMBECK *et al.* 2004) allowed the formation of land bridges that connected isolated lands in several parts of the world (HEWITT 2000). Repeated changes in water level resulted in fragmentation and fusion of basins within continents (as in the Great African Lakes, GALIS and METZ 1998; STURMBAUER *et al.* 2001). Similarly, geological events such as volcanic eruptions induced periodic isolation and reconnection of islands (COOK 2008), while tectonic processes such as the formation of mountains isolated populations and reconnected others (HUGHES and EASTWOOD 2006; ANTONELLI *et al.* 2009; ANTONELLI and SANMARTÍN 2011). More recently, climatic, environmental and anthropogenic changes (e.g., global warming, urbanization and agriculture) have also played important roles in modifying the connectivity pattern between populations (MILLER and HOBBS 2002; DELANEY *et al.* 2010). Consequently, some species are currently subdivided into poorly connected or completely isolated populations; for examples ground beetles (KELLER *et al.* 2004), salamanders (NOEL *et al.* 2007) and crickets (VANDERGAST *et al.* 2009). In the meantime, other species experience habitat and population expansion (e.g., sparrows, white-tailed deer, zebra mussels; WAPLES 2010). Isolation and reconnection of populations not only reflect abiotic processes, but they can also represent spatial and temporal in-

teractions of populations (e.g., secondary contacts; GREEN *et al.* 2010; DOMINGUES *et al.* 2012). Consequently, transient states of genetic diversity are expected to be the norm, and deserve much more attention.

In this study, we analytically characterized the dynamics of genetic diversity following a change in migration rate between populations, given any migration rate, mutation rate, population size and degree of fragmentation. We first analyzed how genetic diversity is affected by an event of isolation of populations, and by an event of reconnection of populations. We then generalized our results for situations where the migration rate between populations displays strong variation. We demonstrate that temporal changes of migration generate periods where genetic diversity reaches unexpectedly high values that can be maintained over thousands of generations. We also show that migration changes can produce a signature on summary statistics such as Tajima's D and Ewens-Watterson's statistics that cannot be differentiated from a signature of population size change or from the signature of selection. Finally, we discuss how such processes can impact observed macro-evolutionary patterns of species diversity and how they can affect the reconstruction of populations' demographic and evolutionary history from genetic data.

GENETIC DIVERSITY OF POPULATIONS

To study the dynamics of genetic diversity after connectivity changes, we consider diploid individuals in a finite island model composed of n random mating populations of size N , so that the total population size is nN . The populations exchange migrants at a rate m . The mutations follow the infinite allele model (each mutation produces a new allele; KIMURA and CROW 1964) and occur at a rate μ . The generations are non-overlapping (Wright-Fisher model; FISHER 1930; WRIGHT 1931).

Genetic diversity, H , is estimated using the identity by descent F between pairs of alleles, through the relationship provided by NEI and FELDMAN (1972):

$$H = 1 - F \tag{2.1}$$

Further, we characterize within-population genetic diversity H_s and between-population genetic diversity H_b , using within- and between-population genetic identities, respectively. Within-population genetic identity, F_s , corresponds to the probability that two genes randomly chosen from the same population are identical by descent. Between-population genetic identity, F_b , corresponds to the probability that two genes randomly chosen from different populations are identical by descent. Considering that within- and between-population genetic identities F_s and F_b at a given time t are, respectively, $F_{s,t}$ and $F_{b,t}$, their values at the next generation (forward in time), respectively, $F_{s,t+1}$ and $F_{b,t+1}$, will follow (MAYNARD SMITH 1970; MARUYAMA 1970; LATTER 1973):

$$\begin{cases} F_{s,t+1} = [a(c + (1 - c)F_{s,t}) + (1 - a)F_{b,t}](1 - \mu)^2 \\ F_{b,t+1} = [b(c + (1 - c)F_{s,t}) + (1 - b)F_{b,t}](1 - \mu)^2 \end{cases} \tag{2.2a}$$

where the parameters are:

$$a = (1 - m)^2 + \frac{m^2}{n - 1} \tag{2.2b}$$

$$b = \frac{1 - a}{(n - 1)} \tag{2.2c}$$

$$c = \frac{1}{2N} \tag{2.2d}$$

System of equations 2.2 can be expressed under matrix notation as:

$$\mathbf{F}_{t+1} = \mathbf{A}\mathbf{F}_t + \mathbf{B} \quad (2.3a)$$

Where:

$$\mathbf{A} = (1 - \mu)^2 \begin{pmatrix} a(1 - c) & 1 - a \\ b(1 - c) & 1 - b \end{pmatrix} \quad (2.3b)$$

$$\mathbf{B} = (1 - \mu)^2 \begin{pmatrix} ac \\ bc \end{pmatrix} \quad (2.3c)$$

$$\mathbf{F}_t = \begin{pmatrix} F_{s,t} \\ F_{b,t} \end{pmatrix} \quad (2.3d)$$

Parameters have the following interpretation: a is the probability that two genes at the same location before migration are still at the same location after migration (either both migrate to the same location or both do not migrate); b is the probability that two genes that were at the same location before migration migrated to different locations; c is the probability that two genes within a population are copies of the same gene; $(1 - \mu)^2$ is the probability that neither of the two randomly chosen genes mutated.

PREDICTING THE DYNAMICS OF GENETIC DIVERSITY

To characterize the impact of connectivity changes on genetic diversity, we analyzed the trajectories of within- and between-population genetic diversity from any initial genetic identity state. Using the equation (2.2), MAYNARD SMITH (1970) and MARUYAMA (1970) showed that genetic identities converge toward an equilibrium value \mathbf{F}^{eq} (value given in Supporting Information File S1). Extending the results obtained by NEI and FELDMAN (1972) for $n=2$ populations, we show that the temporal dynamics of genetic

diversity follow (see appendix A for more details):

$$\mathbf{F}_t = \mathbf{C}_1 \lambda_1^t + \mathbf{C}_2 \lambda_2^t + \mathbf{F}^{eq} \quad (2.4a)$$

Where

$$\mathbf{F}^{eq} = \begin{pmatrix} F_s^{eq} \\ F_b^{eq} \end{pmatrix} \quad (2.4b)$$

λ_1 and λ_2 are respectively the largest and smallest eigenvalues of matrix \mathbf{A} , and they follow:

$$\begin{cases} \lambda_1 = \frac{(1-\mu)^2}{2} [a(1-c) + 1 - b + \sqrt{(1-a(1-c)+b)^2 - 4bc}] \\ \lambda_2 = \frac{(1-\mu)^2}{2} [a(1-c) + 1 - b - \sqrt{(1-a(1-c)+b)^2 - 4bc}] \end{cases} \quad (2.5)$$

\mathbf{C}_1 and \mathbf{C}_2 are column vectors of dimension 2 composed of constant values which depend on the parameters of the model (m , μ , n and N) and on the initial genetic identity \mathbf{F}_0 (appendix A).

In the next section, we provide from eq. 2.4a and 2.5 the temporal change of genetic diversity and derive the corresponding time to reach genetic diversity equilibrium after a connectivity change.

TIME TO REACH GENETIC DIVERSITY EQUILIBRIUM

The change of genetic diversity can be decomposed in two main temporal dynamics: a long-term and a short-term dynamics. Indeed, the temporal change of genetic diversity depends on two components: $|\mathbf{C}_1 \lambda_1^t|$ and $|\mathbf{C}_2 \lambda_2^t|$ (eq. 2.4a). They both follow an exponential decay and their rate of change depends on $r_1 = \ln(\lambda_1)$ and $r_2 = \ln(\lambda_2)$, respectively (appendix A). As $1 > \lambda_1 > \lambda_2 > 0$, $|\mathbf{C}_2 \lambda_2^t|$ decays more rapidly than $|\mathbf{C}_1 \lambda_1^t|$ (see Supporting Information File S1). Thus r_1 determines the ultimate (or long-term) change of genetic diversity and r_2 determines the transient (or short-term) change of

genetic diversity.

When migration and mutation rates are small (i.e. $m \ll 1$ and $\mu \ll 1$) and local population sizes are large (i.e. $N \gg 1$), the decay constants r_1 and r_2 follow:

$$\begin{cases} r_1 = -2\mu - \frac{1}{2N_e} \\ r_2 = -2\mu - 2m \frac{n}{n-1} - \frac{1}{2N} + \frac{1}{2N_e} \end{cases} \quad (2.6a)$$

With

$$N_e = nN \left(1 + \frac{(n-1)}{nM} \right) \quad (2.6b)$$

where $M = 4Nm$ is the scaled migration rate and N_e is the effective population size of the total population (inbreeding, eigenvalue, variance and mutation effective size are equivalent in the finite island model, WHITLOCK and BARTON 1997). As expected from theory (WHITLOCK and BARTON 1997; WAKELEY 1999), in the strong migration limit ($M \gg 1$), the effective size is equal to the total population size nN , while in the weak migration limit ($M \ll 1$), the effective size is higher than the total population size.

We can estimate the durations of the ultimate and transient changes of genetic diversity, denoted t_1 and t_2 , respectively. Formally, we define t_1 and t_2 as the times needed for λ_1^t and λ_2^t to be reduced to a number α , where $\alpha \in]0; 1]$:

$$\begin{cases} \lambda_1^{t_1} = \alpha \\ \lambda_2^{t_2} = \alpha \end{cases} \quad (2.7)$$

Assuming that migration and mutation rates are small and population sizes are large,

t_1 and t_2 simplify to (appendix A):

$$\begin{cases} t_1 = \frac{-\ln(\alpha)}{2\mu + \frac{1}{2N_e}} \\ t_2 = \frac{-\ln(\alpha)}{2\mu + 2m\frac{n}{n-1} + \frac{1}{2N} - \frac{1}{2N_e}} \end{cases} \quad (2.8)$$

The genetic diversity changes as follows (Figure 2.1): (i) a convergence of duration t_2 from the initial genetic diversity value to a transient genetic diversity value and then (ii) a convergence of duration t_1 to the genetic diversity equilibrium H^{eq} . The time to reach genetic diversity equilibrium, t_1 , depends only on two terms: the mutation rate (term 2μ) and the genetic drift at the total population level (term $\frac{1}{2N_e}$). The duration of the transient dynamics, t_2 , depends on four terms: the mutation rate (term 2μ), the migration rate (term $2m$), the genetic drift in each population (term $\frac{1}{2N}$) and the genetic drift at the total population level (term $\frac{1}{2N_e}$). The convergence to the transient and equilibrium values of genetic diversity can occur on separated timescales (i.e. $t_1 \gg t_2$) depending on the parameter values. The timescales t_1 and t_2 can differ from several orders of magnitude. When $n > 14$, differences are the highest ($t_1 \gg t_2$), in the domain where $\mu \ll \frac{1}{2N}$ and also when $\mu \gg \frac{1}{2N}$ and $m > \mu$. When $n \leq 14$, the same conditions apply for $t_1 \gg t_2$ except in a restricted domain where $m \simeq \frac{1}{2N}$ (see appendix A). For example, the duration of the transient dynamics is $t_2 \simeq 134$ and the time to reach equilibrium is $t_1 \simeq 1.5 \times 10^5$ (with $\alpha = 5\%$), when 10 populations of size 2,500 with a mutation rate of 10^{-6} are connected with a migration rate of 0.01.

DYNAMICS OF GENETIC DIVERSITY AFTER AN ISOLATION EVENT

We analyzed the dynamics of genetic diversity after an isolation event, starting with a situation where populations are connected and at their equilibrium value, i.e., within- and between-population genetic diversity H_s and H_b are at the expected connection

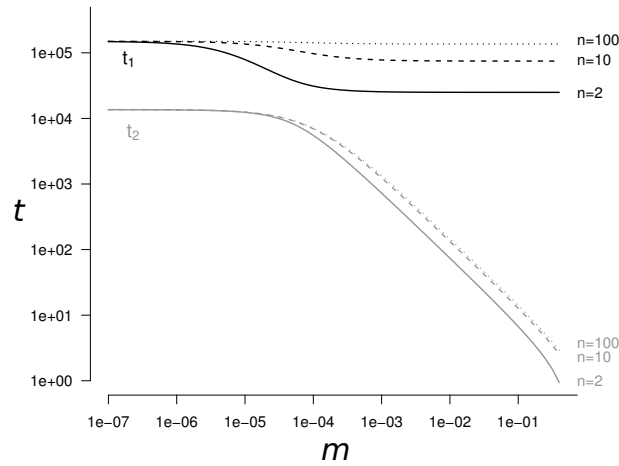


Figure 2.1 The time (in number of generations) t_1 to reach genetic diversity equilibrium and the length of the transient dynamics period t_2 as a function of the migration rate m . The solid line corresponds to $n = 2$ populations, the dashed line to $n = 10$ and the dotted line to $n = 100$. t_1 is always at least one order of magnitude higher than t_2 . This separation of the two periods becomes even greater when $m > \frac{1}{4N} = 10^{-4}$. Parameter values are $N = 2500$, $\mu = 10^{-5}$, $\alpha = 5\%$.

equilibrium values $H_{s,con}^{eq}$ and $H_{b,con}^{eq}$ (see MARUYAMA 1970; MAYNARD SMITH 1970 and Supporting Information File S1).

We observe (Figure 2.2) that immediately after an isolation event, within-population genetic diversity decreases due to genetic drift to the point where it reaches the mutation-drift equilibrium of an isolated population $H_{s,iso}^{eq}$ (see Supporting Information File S1 and KIMURA and CROW 1964), at a rate determined by r_2 (from eq. 2.6). Meanwhile, between-population genetic diversity slowly increases due to the differentiation of populations induced by mutations (at a rate determined by r_1 from eq. 2.6). Populations ultimately reach complete differentiation (equilibrium value of $H_{b,iso}^{eq} = 1$). We can show from eq. 2.2 that following an isolation event, within- and between-population diversities change independently and reach their equilibrium value in t_2 and t_1 generations, respectively (Supporting Information File S2).

The decrease of population genetic diversity (within) can occur quickly relative to

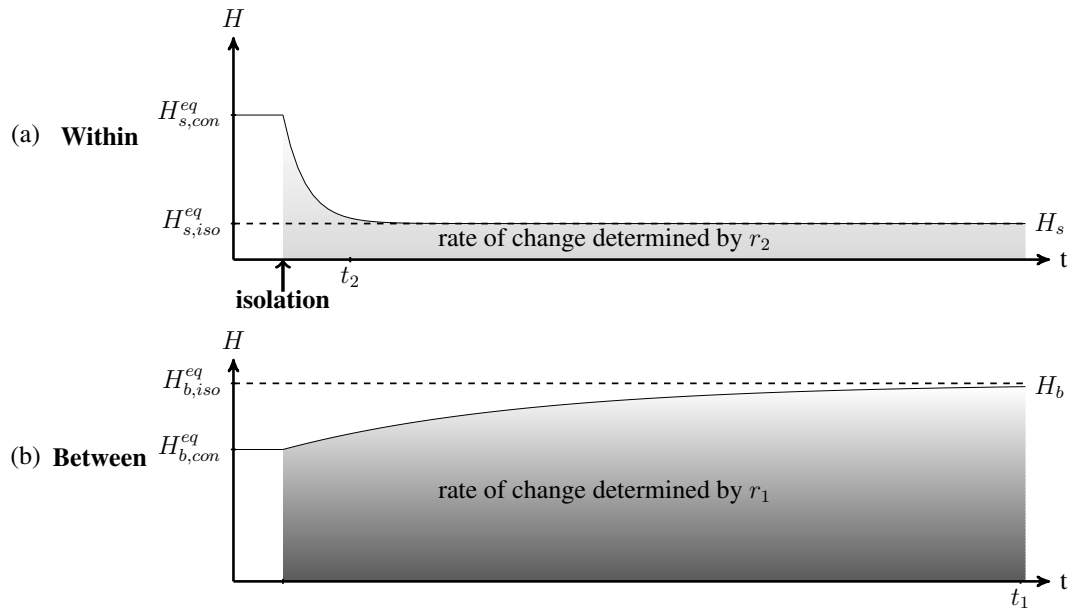


Figure 2.2 Dynamics of (a) within-population H_s and (b) between-population H_b genetic diversity after an isolation event. Within- and between-population diversity (solid lines) were previously at their respective connection equilibrium $H_{s,con}^{eq}$ and $H_{b,con}^{eq}$. After the isolation event, within- and between-population diversities reach their isolation equilibrium $H_{s,iso}^{eq}$ and $H_{b,iso}^{eq}$ (dashed lines) at rates determined by r_2 and r_1 (eq. 2.6). t_2 and t_1 estimate the time to reach the within- and between-population genetic diversity equilibrium, respectively (eq. 2.8). Under the effect of genetic drift, within-population diversity reaches its equilibrium value faster than between-population genetic diversity. Parameters are $n = 10$, $N = 2,500$, $m = 10^{-4}$ before isolation and $m = 0$ afterwards, and $\mu = 10^{-5}$. $t_1 \simeq 149,000$ generations and $t_2 \simeq 13,600$ generations (for $\alpha=5\%$).

population differentiation (between-population genetic diversity, see Figure 2.1). After an isolation event, within-population genetic diversity (H_s) remains above its expected equilibrium $H_{s,iso}^{eq}$, while between-population genetic diversity (H_b) remains below its expected equilibrium $H_{b,iso}^{eq}$. However, the timescales of these non-equilibrium periods differ. When populations are isolated ($m=0$), H_s reaches a value close to its equilibrium value in $t_2 \sim \frac{1}{2\mu + \frac{1}{2N}}$ generations, while H_b reaches a value close to its equilibrium value in $t_1 \sim \frac{1}{2\mu}$ generations (eq. 2.8). Therefore, when $\theta \ll 1$, H_s converges much more quickly than H_b , and when $\theta \gg 1$, both converge in approximately the same amount of time. For example, assuming $\mu = 10^{-5}$ and $N = 1,000$, H_b is significantly lower than the equilibrium value for approximately $t_1 \simeq 150,000$ generations while H_s is significantly higher than the equilibrium value for approximately $t_2 \simeq 6,000$ generations (given $\alpha=5\%$).

DYNAMICS OF GENETIC DIVERSITY AFTER A CONNECTION EVENT

We analyzed the dynamics of genetic diversity after a connection event, starting with a situation where populations are isolated and at their equilibrium value $H_{s,iso}^{eq}$ and $H_{b,iso}^{eq}$ (see MAYNARD SMITH 1970; MARUYAMA 1970 and Supporting Information File S1). After a connection event (Figure 2.3), the genetic diversity accumulated in each population during the isolation period is quickly spread to all populations (Figure 2.3, fast dynamics in light gray, at a rate determined by r_2 from eq. 2.6). Consequently, within-population genetic diversity quickly increases and reaches a high value that is above its expected connected equilibrium value ($H_{s,con}^{eq}$ Figure 2.3, see Supporting Information File S1). This process creates a peak of within-population genetic diversity ΔH_s and a transient excess of genetic diversity between populations ΔH_b (see Figure 2.3). Then, due to genetic drift, both the within- and the between-population diversities decrease (slow dynamics in dark gray in Figure 2.3, at a rate determined by r_1

from eq. 2.6), to the point where the diversities reach the expected value of mutation-migration-drift equilibrium ($H_{s,con}^{eq}$ and $H_{b,con}^{eq}$ from eq. 2.4b; KIMURA and CROW 1964).

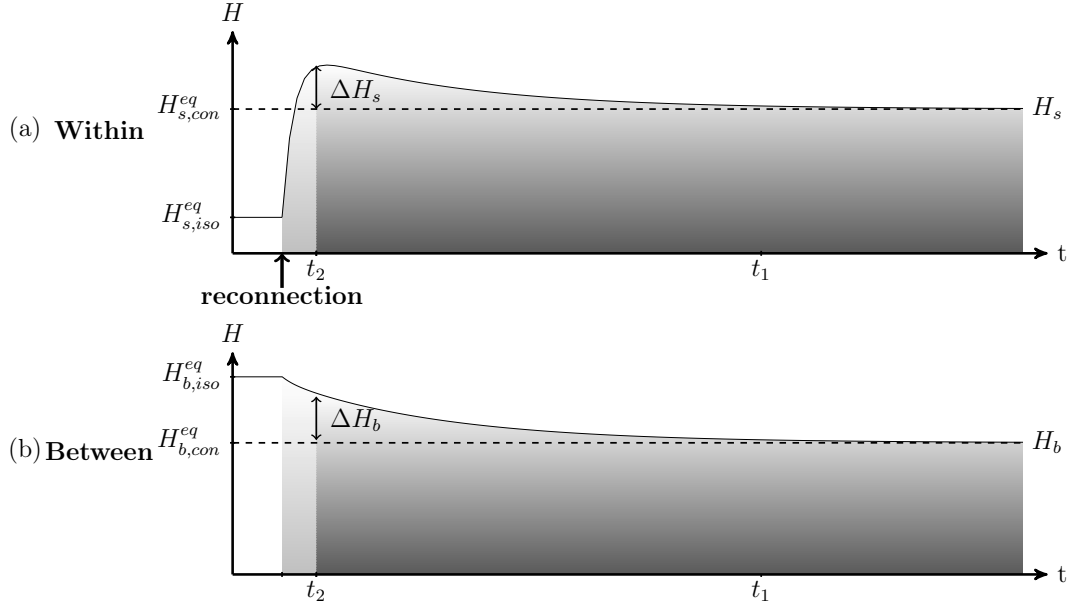


Figure 2.3 Dynamics of (a) within-population genetic diversity H_s and (b) between-population genetic diversity H_b after a reconnection event. Within- and between-population diversities were previously at their respective isolation equilibria $H_{s,iso}^{eq}$ and $H_{b,iso}^{eq}$ (eq. 2.4b). After the reconnection event, within- and between-population diversities reach their respective connection equilibria $H_{s,con}^{eq}$ and $H_{b,con}^{eq}$ (dashed lines). As shown in eq. 2.8, the time to reach genetic diversity equilibrium t_1 and the length of the transient period t_2 are well separated. The two periods are: (1) Fast convergence at a rate determined by r_2 (eq. 2.6) that is driven by the spread of diversity that had accumulated within populations during isolation, which creates the peak of within-population diversity (ΔH_s) and the excess of between-population diversity (ΔH_b). (2) Slow dynamics at a rate determined by r_1 (eq. 2.6) that is caused by the gradual loss of genetic diversity. A large number of generations is needed to reach equilibrium. When $n = 10$, $N = 2,500$, $m = 10^{-4}$ after reconnection, and $\mu = 10^{-5}$, $t_1 \simeq 97,000$ generations, $t_2 \simeq 6,900$ generations (for $\alpha=5\%$) and $\Delta H_s \simeq \Delta H_b \simeq 0.11$.

Within- and between-population diversities change successively according to two timescales: first, a fast transient dynamics, followed by a slow asymptotic dynamics (separation of timescales is derived in appendix A and illustrated in Figure 2.3). Because the transient dynamics can be shorter than the asymptotic dynamics, the excess of genetic diversity (ΔH_s and ΔH_b) can be maintained for a very long period (from Figure 2.1, t_1 is longer than 10,000 generations).

PEAK OF GENETIC DIVERSITY GENERATED BY A CONNECTION EVENT

In this section, we characterize the peak of within-population genetic diversity, ΔH_s , and the excess of between-population genetic diversity, ΔH_b , observed after a connection event as a function of the mutation rate, the genetic drift, the number of populations and the migration rate after connection. The exact value of the within-population genetic diversity peak is represented in Figure 2.4. Assuming that migration and mutation rates are small, we can show that good approximations of the values of ΔH_s and ΔH_b are (see derivations in appendix B):

$$\begin{cases} \Delta H_s = (1 - (1 - F_{s,iso}^{eq}) \frac{M \frac{n}{n-1}}{1 + M \frac{n}{n-1}}) \cdot F_{b,con}^{eq} \frac{M}{1 + M \frac{n}{n-1} - \frac{2N}{N_e}} 0.05^{\frac{t_2}{t_1}} \\ \Delta H_b = (1 - (1 - F_{s,iso}^{eq}) \frac{M \frac{n}{n-1}}{1 + M \frac{n}{n-1}}) \cdot F_{b,con}^{eq} \frac{1 + M - \frac{N}{N_e}}{1 + M \frac{n}{n-1} - \frac{2N}{N_e}} 0.05^{\frac{t_2}{t_1}} \end{cases} \quad (2.9a)$$

Where

$$F_{s,iso}^{eq} = \frac{1}{1 + \theta} \quad (2.9b)$$

is the expected equilibrium identity within an isolated population (KIMURA and CROW 1964), and

$$F_{b,con}^{eq} = \frac{M}{M + (n - 1)\theta(1 + \theta + \frac{n}{n-1}M)} \quad (2.9c)$$

is the expected equilibrium identity between connected populations with the scaled migration rate M , the number of populations n and the scaled mutation rate $\theta = 4N\mu$ (MARUYAMA 1970; MAYNARD SMITH 1970). These approximations lead to the largest absolute error when a small number of populations ($n = 2$) is combined with weak mutation ($\theta < 1$) and intermediate migration ($M \simeq 5$). Nevertheless, this error is small (error smaller than 0.025 for ΔH_s and smaller than 0.08 for ΔH_b), so equation 2.9 provides a good approximation of ΔH_s and ΔH_b for all n , M and θ values (see

appendix B for more details about the validity of approximation 2.9).

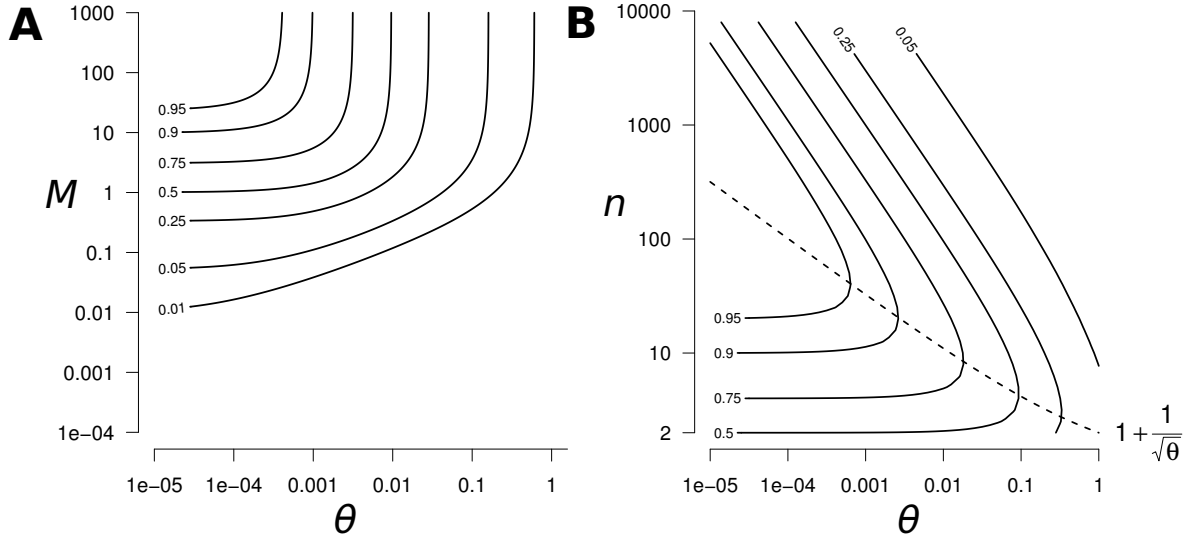


Figure 2.4 Peak of within-population genetic diversity ΔH_s generated by a reconnection event. (A) Contour plot of ΔH_s as a function of θ and M , for $n = 100$. We can clearly see the highest peak of diversity in the high M and low θ region. (B) Contour plot of the peak of genetic diversity after a reconnection event as a function of θ and n , for $M \gg 1$ (high M region identified in (A)). In the high M region, the within-population diversity peak ΔH_s and the between-population diversity excess ΔH_b are equal. The dashed line represents the number of populations which maximizes the peak of diversity $n^* = 1 + 1/\sqrt{\theta}$.

The peak of genetic diversity increases with the difference between the two timescales (ΔH_s and ΔH_b increase with $\frac{t_1}{t_2}$) as genetic diversity increases during the transient phase and decreases during the asymptotic phase. Indeed, when those two phases are separated there is no loss of genetic diversity caused by the asymptotic decay during the transient phase (terms $0.05^{\frac{t_2}{t_1}} \simeq 1$ in eq. 2.9).

In the domain where the peak is the largest ($M \gg 1$ and $\theta \ll 1$), ΔH_s and ΔH_b reach the same value:

$$\Delta H^{max} = \frac{n-1}{n} \frac{1}{1+n\theta} \quad (2.10)$$

In this domain, ΔH^{max} is maximized when the number of populations is (dashed line in Figure 2.4B):

$$n^* = 1 + 1/\sqrt{\theta} \quad (2.11)$$

The corresponding peak of genetic diversity, reached at n^* , is $\Delta H^{max}|_{n^*} = \frac{1}{1+2\sqrt{\theta}}$. Interestingly, the number of populations and the peak of diversity have a non-monotonous relationship. The peak of genetic diversity decreases when the number of populations approaches 2 and when it tends to infinity, while an intermediate number of populations n^* maximizes the peak of genetic diversity. This can be easily explained by the following processes. During isolation, a small number of populations accumulates less between-population genetic diversity, thus, once reconnected, they share a smaller amount of diversity. In contrast, a large number of populations accumulates a higher level of genetic diversity but also has a higher connection equilibrium value; thus, once reconnected, diversity reaches its expected equilibrium and no peak of diversity is observed.

In summary, high peaks of genetic diversity ($\Delta H_s > 0.25$ in Figure 2.4) can occur for a large range of the parameter space: when mutation is weak ($\theta < 0.05$) and migration is moderate to strong ($M > 0.5$). Under these conditions, drastic genetic diversity changes can be observed (ΔH_s values larger than 0.95, Figure 2.4B for $M \geq 50$ and $\theta < 5 \times 10^{-4}$). The number of populations that maximize the peak of diversity, n^* , ranges from a few populations when $\theta \simeq 1$, up to a few hundred populations when $\theta = 10^{-6}$ (values of $\theta < 10^{-6}$ are expected to be very rare, they would require a mutation rate of $2.5 \times 10^{-12}/bp$ for a 1 kb gene and a population size of 100). Interestingly, a significant peak of genetic diversity is also observed when only two populations reconnect ($\Delta H^{max}|_{n=2} = 0.5$ Figure 2.4B).

PEAK OF GENETIC DIVERSITY RESULTING FROM A MIGRATION RATE INCREASE

Complete isolation of populations is not required to generate peaks of genetic diversity. Indeed, an abrupt increase of migration can generate the peak of genetic diversity characterized in the previous sections. In Supporting Information File S3, we determined that if migration crosses a threshold value M_T , peaks of genetic diversity can occur. The value of the threshold M_T , assuming that $m \ll 1$ and $\mu \ll 1$, is:

$$M_T = (n - 1)\theta \frac{1 + \theta}{1 + n\theta} \quad (2.12)$$

Consequently, an increase of migration from M_0 to M crossing the threshold value M_T (i.e. $M_0 \ll M_T$) generates a peak of genetic diversity that can be approximated by eq. 2.9 (see Supporting Information File S3). For example, in a subdivided population of $n = 10$ and $\theta = 0.1$, an increase in migration from $M_0 = 0.01$ to $M = 10$ (which crosses the migration threshold $M_T = 0.495$, eq. 2.12), generates a peak of within-population diversity of 0.350, while a reconnection event in a similar situation would generate a peak of similar intensity (0.358).

IMPLICATIONS FOR THE INFERENCE OF DEMOGRAPHY AND SELECTION

To describe the impact of migration changes on the inference of demography and selection from genetic data, we described the dynamics of two broadly used summary statistics: the Ewens-Watterson statistics (Watterson 1978) and Tajima's D (Tajima 1989). Both the Ewens-Watterson statistics and Tajima's D are known to detect an excess (resp. deficit) of rare alleles, which induces negative (resp. positive) values of the statistics, compared with the expected neutral equilibrium (constant size population

without selection). Usually, an excess of rare alleles is interpreted either as the signature of balancing selection or population expansion, and a deficit of rare alleles is interpreted as the signature of directional selection or as a population bottleneck. We used the Ewens-Watterson statistics, which we denote H_{EW} , and follows (WATTERSON 1978):

$$H_{EW} = H_s - H_A \quad (2.13)$$

where H_s is the genetic diversity, and H_A is the expected genetic diversity given the observed number of alleles K . We also used Tajima's D , which we denote D_T , and follows (Tajima 1989):

$$D_T = \frac{\pi - S/a_1}{\sqrt{\text{Var}(\pi - S/a_1)}} \quad (2.14)$$

where $a_1 = \sum_{i=1}^m \frac{1}{i}$, m is the sample size, π is the average number of pairwise nucleotide differences, and S is the number of segregating sites.

We simulated samples of 50 sequences of 1 kb, with a per nucleotide mutation rate of 2.10^{-8} , in 4 populations of size 2500, and ran 5,000 replicate simulations. We simulated an isolation event, where the migration rate changed from 0.002 to 0, and a reconnection event in which the migration rate changed from 0 to 0.002. The simulations were performed with the software *fastsimcoal* (EXCOFFIER and FOLL 2011), and the data analysis was performed with *Arlequin* (EXCOFFIER and LISCHER 2010). We simulated samples from the same population (with the parameter values that were used, the sampling scheme had a very weak impact; see CHIKHI *et al.* 2010 for a discussion of how the sampling scheme affects the values of Tajima's D). To allow for convergence of the coalescent algorithm, we always assumed that the populations were connected prior to the isolation phase. In the reconnection event simulations, we set the duration of

the isolated phase to $10N$, which allowed the genetic diversity values to reach their equilibrium value.

We followed the dynamics of the statistics and estimated their distribution as a function of time. The results in Figure 2.5 show that an isolation event produces the same signature on the Ewens-Watterson statistics (Figure 2.5A) and Tajima's D (Figure 2.5B), as expected from a bottleneck event and from directional selection. Indeed, following an isolation event, genetic drift first causes the elimination of rare alleles and then eliminates more common alleles. Consequently, the number of alleles K decreases more quickly than genetic diversity H_s (Figure 2.5C). Similarly, the number of segregating sites S decreases faster than the number of pairwise differences π (Figure 2.5D). Therefore, D_T and H_{EW} are skewed toward positive values, as expected after a bottleneck or under the effect of directional selection. Moreover, the statistics remain skewed for a long period of time (less than 10,000 generations in our simulations, see Figure 2.5).

Results in figure 2.6 show that a reconnection event can successively produce the same signature as expected from a population expansion or from a bottleneck event on H_{EW} (Figure 2.6A) and D_T (Figure 2.6B). Indeed, following a reconnection event, migrants first create an excess of rare variants. The number of alleles K increases more quickly than the genetic diversity H_s (Figure 2.6C), and the number of segregating sites S increases faster than the number of pairwise differences π (Figure 2.6D), which skews H_{EW} and D_T toward negative values. Second, new alleles brought by migrants increase in frequency, creating an excess of common variants. Consequently K increases more than H_s , and π increases more than S , which skews D_T and H_{EW} toward

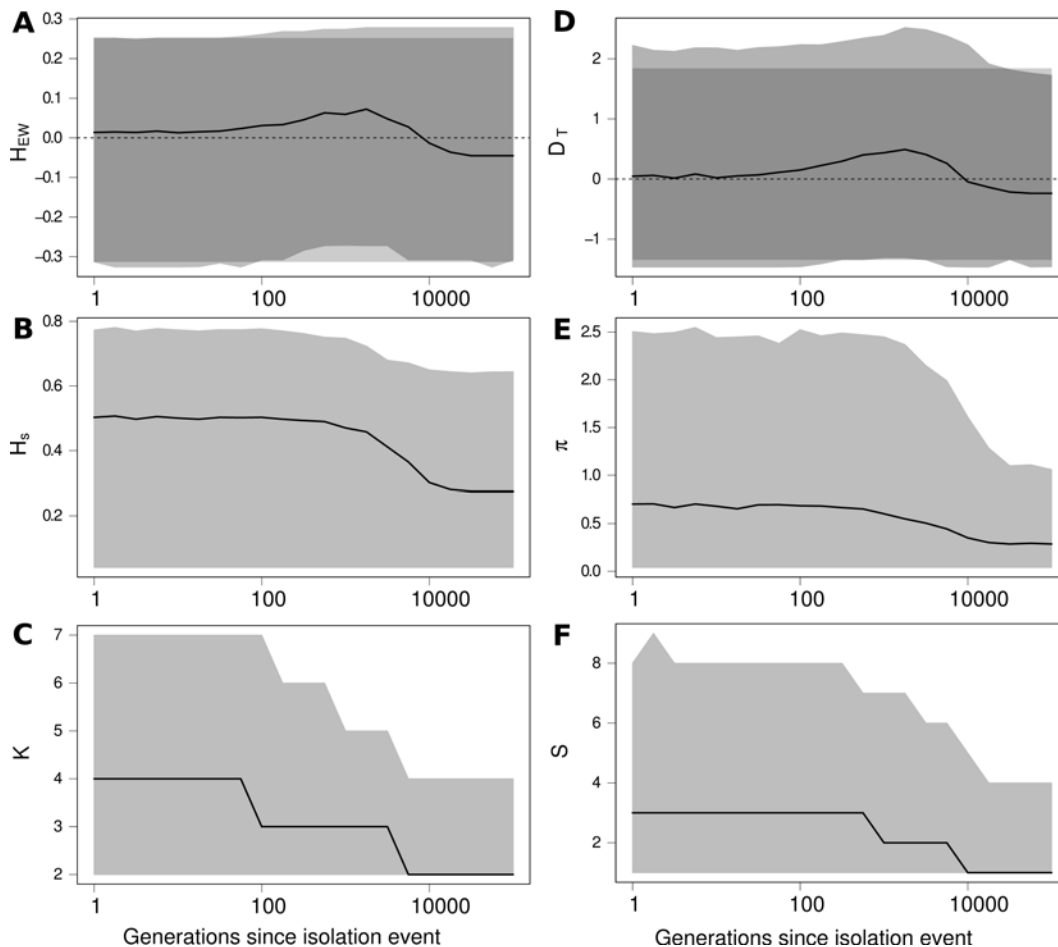


Figure 2.5 Effect of an isolation event on Ewens-Watterson and Tajima's D neutrality tests and on related summary statistics. (A) Ewens-Watterson statistics (H_{EW}), (B) genetic diversity (H_s), (C) number of alleles (K), (D) Tajima's D (D_T), (E) number of pairwise differences (π), and (F) number of segregating sites (S). For each statistics, the black line represents the median of the distribution and the light gray surface represents the 97.5% and 2.5% quantiles of the distribution as a function of the number of generations t after the isolation event. Gray boxes in (A) and (B) represent the expected distribution of the statistics in an isolated population at equilibrium. Values of H_{EW} and D after an isolation event are skewed toward positive values (signature of a bottleneck or directional selection), while there was no change in the size of the population. K and S decrease more quickly than H_s and π , because rare alleles are eliminated by genetic drift more quickly than common alleles. Coalescence simulations of a 1 kb locus with a mutation rate of $2 \cdot 10^{-8}$ per bp, where 4 populations of size 2,500 are isolated; 5,000 replicates.

positive values.

Interestingly, the observed duration of the periods where both statistics are skewed are similar to the expected duration of the dynamics of genetic diversity (from eq. 2.8). After an isolation event, we observe, in Figure 2.5, that all statistics reach their equilibrium value within approximately 10,000 generations ($4N$ generations). This duration corresponds to the value of the time required to reach within-population genetic diversity equilibrium after an isolation event, $t_2 \simeq 12,000$ generations ($4.8N$ generations, estimated from eq. 2.8 with $\alpha=5\%$). In this example, genetic drift is stronger than mutation ($2\mu \ll 1/2N$) and thus $t_2 \sim 2N$. t_2 corresponds to the duration of the period where the deficit of rare alleles skews the distribution of H_{EW} and D_T . After a reconnection event, we observe Figure 2.6 that H_{EW} and D_T reach a "peak" within approximately 600 generations ($0.24N$ generations). This duration corresponds to the value of the duration of the transient dynamics following a reconnection event, $t_2 \simeq 540$ generations ($0.216N$ generations, estimated from eq. 2.8 with $\alpha=5\%$). In this example, migration is stronger than genetic drift and mutation ($m \gg \mu$ and $m \gg 1/2N$), and thus, $t_2 \sim 1/2m$. t_2 corresponds to the period during which the distribution of H_{EW} and D_T is skewed. Subsequently, H_{EW} and D_T reach their equilibrium value in approximately 80,000 generations ($32N$ generations); this duration corresponds to the time required to reach the genetic diversity equilibrium value after a reconnection event, $t_1 \simeq 75,000$ generations ($30N$ generations, estimated from eq. 2.8 with $\alpha=5\%$). t_1 corresponds to the period during which the deficit of rare alleles is eliminated.

In conclusion, both an isolation and a reconnection event induce changes in the proportion of rare alleles, which skews the values of H_{EW} and D_T , thus producing a

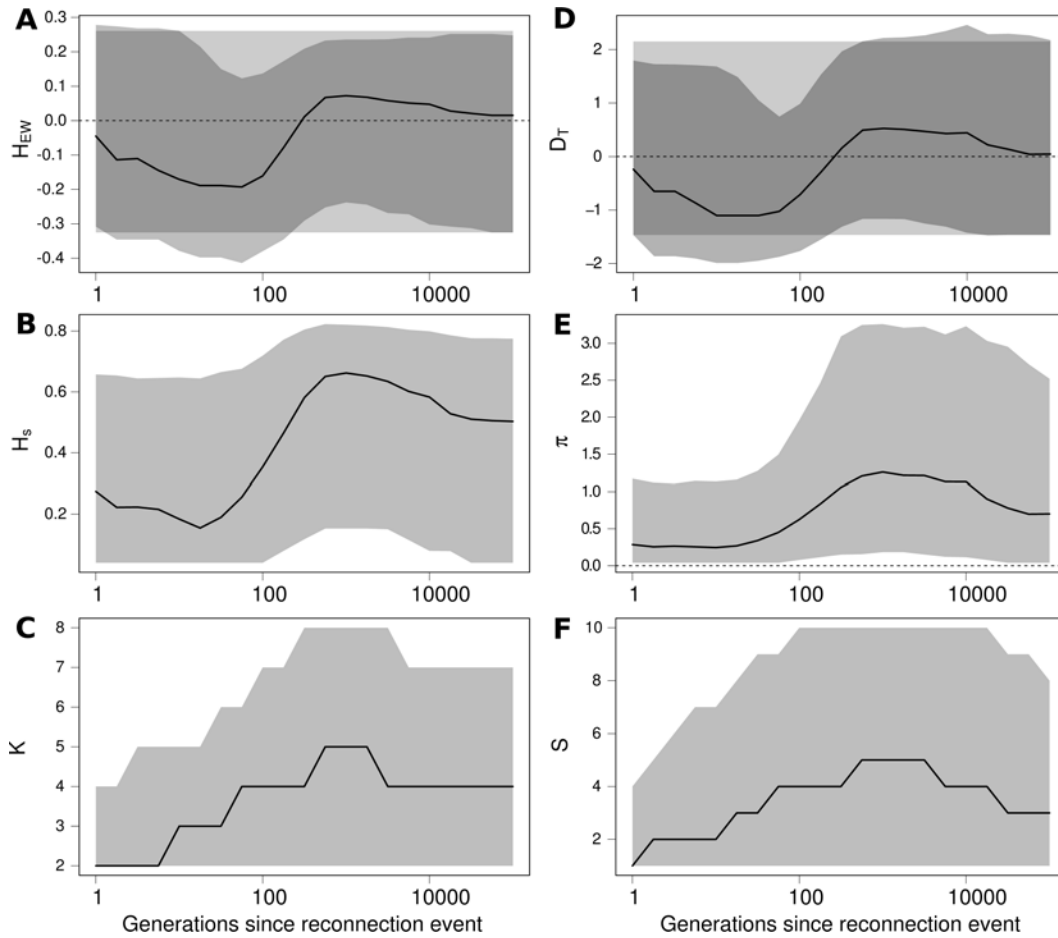


Figure 2.6 Effect of a reconnection event on Ewens-Watterson and Tajima's D neutrality tests and on related summary statistics. (A) Ewens-Watterson statistics (H_{EW}), (B) genetic diversity (H_s), (C) number of alleles (K), (D) Tajima's D (D_T), (E) number of pairwise differences (π), and (F) number of segregating sites (S). For each statistics, the black line represents the median of the distribution, and the light gray surface represents the 97.5% and 2.5% quantiles of the distribution, as a function of the number of generations t after the isolation event. Gray boxes in (A) and (B) represent the expected distribution of the statistics in an isolated equilibrium population. Values of H_{EW} and D after a reconnection event are first skewed toward negative values (signature of a population expansion or balancing selection) and then toward positive values (signature of a bottleneck or directional selection), while there was no change in the size of the population. K and S first increase more quickly than H_s and π because immigrants bring rare alleles, and then H_s and π reach a higher value because immigrant alleles increase in frequency. Finally, alleles are eliminated by genetic drift until the statistics reach their expected equilibrium value when populations are connected. Coalescence simulations of a 1 kb locus with a mutation rate of $2 \cdot 10^{-8}$ per bp, where 4 populations of size 2,500 isolated during 25,000 generations are reconnected with a migration rate $m = 0.002$; 5,000 replicates.

signature that can not be differentiated from the signature of past demographic events or of selection.

DISCUSSION

We documented a simple neutral mechanism, which creates long-term peaks of genetic diversity. This peak of genetic diversity appears shortly after an abrupt increase in migration and is conserved for a long time. We also demonstrated that such genetic diversity peaks can occur for a large and plausible range of population sizes, migration rates, mutation rates and numbers of populations. Subsequent to the genetic diversity peak, the rate of decay of genetic diversity was slow. Consequently, the mechanisms described here leave a strong and long-term footprint on genetic diversity that affects the Ewens-Watterson statistics and Tajima's D that are commonly used to infer the history of populations from genetic data.

The peak of genetic diversity is due to the spread of the genetic diversity accumulated during (partial) isolation. Therefore, the migration model that is assumed (island model of migration) is a leading factor in determining the strength of the observed genetic diversity peak. Assuming isolation by distance, the within-population genetic diversity is expected to have locally lower peaks. At the same time, under this assumption, the between-population genetic diversity is expected to be higher. Indeed, once populations are connected, each population shares with its neighboring populations alleles accumulated during isolation; thus, differentiation between distant populations will be maintained. Additionally, the amount of genetic diversity accumulated during isolation determines the size of the peak of genetic diversity. The maximum value is reached when populations are completely differentiated, i.e., when no alleles are shared between populations. Our results are robust to the relaxation of the complete

isolation and complete differentiation assumptions: when isolation is not complete (because of small migration or non-equilibrium genetic diversity), we show that a genetic diversity peak is still observed (see Supporting Information File S3).

A connection event that occurs after an isolation period might play an important role on species diversification. Indeed, we have demonstrated that such events create an excess of genetic diversity. A high level of genetic diversity has often been hypothesized as being a key factor for species diversification. First, evolution from standing genetic variation might be stronger than from *de novo* mutation (GIBSON and DWORKIN 2004; HERMISSON and PENNINGS 2005; MYLES *et al.* 2005; BARRETT and SCHLUTER 2008). Second, both theoretical (GAVRILETS 2003; GAVRILETS and LOSOS 2009) and empirical work suggest that a high level of pre-existing genetic diversity in a population increases its rate of diversification (HARMON *et al.* 2003; SEEHAUSEN 2004; BARRETT and SCHLUTER 2008). Interestingly, in several cases of adaptive radiation, a high genetic diversity of founder populations has been documented (e.g., BARRIER *et al.* 1999; BEZAULT *et al.* 2011). Several authors argued that the connection of populations after a period of isolation might have played an important role in many adaptive radiations (HUGHES and EASTWOOD 2006; ANTONELLI and SANMARTÍN 2011; BEZAULT *et al.* 2011; JOYCE *et al.* 2011). Therefore, species that experienced population isolation followed by reconnection events could have benefited from a temporary genetic diversity peak which has promoted the diversification of that species. Numerous species are known to have experienced such connectivity changes in the past and show remarkable levels of genetic and species diversity (ARNEGARD *et al.* 1999). For example, cichlid fishes in the great African lakes experienced periods of habitat fragmentation and reconnection due to lake water level fluctuations (ARNEGARD *et al.* 1999); there is

some evidence that these processes might have played a role in the explosive radiation of the species (OWEN *et al.* 1990; YOUNG *et al.* 2009). Additionally, a high rate of speciation is correlated with the timeframe surrounding the uplift of the Northern Andes (SEDANO and BURNS 2010). The mechanisms described here are thus expected to considerably impact the ability of species to adapt to novel environmental conditions and to diversify over a very long period of time.

Statistics on allelic frequencies such as the Ewens-Watterson statistics (WATTERSON 1978) and Tajima's D (TAJIMA 1989) allow the inference of either selection or population demographic changes. Here, we demonstrate that migration changes can lead to signatures that cannot be differentiated from a selection process or a population size change when using the Ewens-Watterson test and Tajima's D. Therefore, past migration changes must be considered more carefully and should be viewed as an alternative explanation of bias in neutrality tests and bottleneck or expansion signals. Recently, authors have shown that population structure can bias neutrality tests and produce false bottleneck signals (LEBLOIS *et al.* 2006; STÄDLER *et al.* 2009; CHIKHI *et al.* 2010) and that shortly after an isolation event departure from the neutrality can be incorrectly inferred (as shown with simulations by BROQUET *et al.* 2010 and discussed in WAPLES 2010). The proper interpretation of genetic signatures is crucial for the understanding of the evolutionary history of populations. An interesting extension of this work would be to analyze in more detail the molecular signature of the mechanisms described here and to provide methods that allow the differentiation of such events from selection or demographic changes. Moreover, our results are also relevant for the study of genealogies. Indeed, genetic identities as considered here are commonly used to describe coalescence time distributions (SLATKIN 1991; ROUSSET 1996; WAKELEY

1999). Future investigations should also investigate the consequences of isolation and connection events on phylogenetic tree reconstruction. Statistical tools that are available to estimate demographic parameters classically focus on a priori specific scenarios (e.g., population bottleneck, expansion, population with constant migration, population split with subsequent migration; see review in KUHNER 2009). Given the strong impact of migration changes on genetic diversity, accounting for such scenarios is necessary. Recent methods allowing a larger range of population demographic scenarios, such as Approximate Bayesian Computation (BEAUMONT *et al.* 2002; BEAUMONT 2010), may be powerful tools to disentangle the signature of demographic processes from the observed genetic diversity.

One of the major goals of conservation genetics is to maintain genetic diversity, decrease extinction risks, avoid inbreeding depression, maintain species evolutionary potential and decrease species vulnerability to environmental change (GILPIN and SOULE 1986; NEWMAN and PILSON 1997; JUMP *et al.* 2009). In this context, conservationists need to estimate the genetic diversity of a population and its effective size. Such measures are commonly obtained from genetic data and are estimated with standard statistics (WRIGHT 1950; JORDE and RYMAN 2007). Although new approaches are emerging that consider populations at a non-equilibrium state, to estimate population size changes and instantaneous migration rates (e.g., HEY and NIELSEN 2004), the expected level of genetic diversity is still commonly estimated assuming that populations are at an equilibrium. As shown here, genetic diversity is more likely to be in a transient state. We have demonstrated that reconnecting isolated populations increases genetic diversity above the expected equilibrium value, while isolating populations induces a slow decrease of genetic diversity. Consequently, any estimate inferred from

data collected from a population that underwent strong migration changes will not reflect the demographic situation of the population (e.g., census size, genetic diversity). This can have drastic consequences on the selection of conservation strategies and for the management of species (PEARSE and CRANDALL 2004; CABALLERO *et al.* 2010).

ACKNOWLEDGMENTS

This project was funded by the Swiss National Research Foundation (SNRF) grant #PZ00P3 – 121702, #PZ00P3_139421/1 and #31003A – 130065.

BIBLIOGRAPHY

ANTONELLI, A., J. A. A. NYLANDER, C. PERSSON, and I. SANMARTÍN, 2009 Tracing the impact of the andean uplift on neotropical plant evolution. *Proc Natl Acad Sci U S A* **106**: 9749–9754.

ANTONELLI, A., and I. SANMARTÍN, 2011 Why are there so many plant species in the neotropics? *Taxon* **60**: 403–414.

ARNEGARD, M. E., J. A. MARKERT, P. D. DANLEY, J. R. STAUFFER, A. J. AMBALI, *et al.*, 1999 Population structure and colour variation of the cichlid fishes *labeotropheus fuelleborni* ahl along a recently formed archipelago of rocky habitat patches in southern lake malawi. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**: 119–130.

BARRETT, R. D. H., and D. SCHLUTER, 2008 Adaptation from standing genetic variation. *Trends in Ecology & Evolution* **23**: 38–44.

BARRIER, M., B. G. BALDWIN, R. H. ROBICHAUX, and M. D. PURUGGANAN, 1999 Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the

hawaiian silversword alliance (asteraceae) inferred from floral homeotic gene duplications. *Mol Biol Evol* **16**: 1105–1113.

BEAUMONT, M. A., 2010 Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**: 379–406.

BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.

BEZAULT, E., S. MWAIKO, and O. SEEHAUSEN, 2011 Population genomic tests of models of adaptive radiation in lake victoria region cichlid fish. *Evolution* **65**: 3381–3397.

BROQUET, T., S. ANGELONE, J. JAQUIERY, P. JOLY, J.-P. LENA, *et al.*, 2010 Genetic bottlenecks driven by population disconnection. *Conserv Biol* **24**: 1596–1605.

CABALLERO, A., S. T. RODRIGUEZ-RAMILO, V. AVILA, and J. FERNANDEZ, 2010 Management of genetic diversity of subdivided populations in conservation programmes. *Conservation Genetics* **11**: 409–419.

CHIKHI, L., V. C. SOUSA, P. LUISI, B. GOOSSENS, and M. A. BEAUMONT, 2010 The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**: 983–995.

COLOSIMO, P. F., K. E. HOSEMAN, S. BALABHADRA, G. VILLARREAL, M. DICKSON, *et al.*, 2005 Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**: 1928–1933.

COOK, L. M., 2008 Species richness in Madeiran land snails, and its causes. *Journal of Biogeography* **35**: 647–653.

- DELANEY, K. S., S. P. D. RILEY, and R. N. FISHER, 2010 A rapid, strong, and convergent genetic response to urban habitat fragmentation in four divergent and widespread vertebrates. *PLoS One* **5**.
- DOMINGUES, V. S., Y.-P. POH, B. K. PETERSON, P. S. PENNINGS, J. D. JENSEN, *et al.*, 2012 Evidence of adaptation from ancestral variation in young populations of beach mice. *Evolution* **66**: 3209–3223.
- EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3**: 87–112.
- EXCOFFIER, L., and M. FOLL, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**: 1332–1334.
- EXCOFFIER, L., M. FOLL, and R. J. PETIT, 2009 Genetic consequences of range expansions. *Annual Review of Ecology, Evolution, and Systematics* **40**: 481–501.
- EXCOFFIER, L., and H. E. L. LISCHER, 2010 Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under linux and windows. *Mol Ecol Resour* **10**: 564–567.
- FALUSH, D., M. STEPHENS, and J. K. PRITCHARD, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
- FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive darwinian selection. *Genetics* **155**: 1405–1413.

- FEDER, J. L., S. H. BERLOCHER, J. B. ROETHELE, H. DAMBROSKI, J. J. SMITH, *et al.*, 2003 Allopatric genetic origins for sympatric host-plant shifts and race formation in *rhagoletis*. *Proc Natl Acad Sci U S A* **100**: 10314–10319.
- FISHER, R. A., 1922 Darwinian evolution of mutations. *Eugen Rev* **14**: 31–34.
- FISHER, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press.
- FRANKHAM, R., 1995 Conservation genetics. *Annu Rev Genet* **29**: 305–327.
- FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GALIS, F., and J. A. METZ, 1998 Why are there so many cichlid species? *Trends Ecol Evol* **13**: 1–2.
- GAVRILETS, S., 2003 Perspective: models of speciation: what have we learned in 40 years? *Evolution* **57**: 2197–2215.
- GAVRILETS, S., and J. B. LOSOS, 2009 Adaptive radiation: contrasting theory with data. *Science* **323**: 732–737.
- GIBSON, G., and I. DWORKIN, 2004 Uncovering cryptic genetic variation. *Nature Reviews Genetics* **5**: 681–U11.
- GILPIN, M., and M. SOULE, 1986 *Conservation Biology: The Science of Scarcity and Diversity*, chapter Minimum Viable Populations: Processes of Species Extinction. Sinauer, Sunderland, Mass, pp. 19–34.
- GRAVEL, S., 2012 Population genetics models of local ancestry. *Genetics* **191**: 607–619.
- GREEN, R. E., J. KRAUSE, A. W. BRIGGS, T. MARICIC, U. STENZEL, *et al.*, 2010 A draft sequence of the neandertal genome. *Science* **328**: 710–722.

- HARMON, L. J., J. A. SCHULTE, A. LARSON, and J. B. LOSOS, 2003 Tempo and mode of evolutionary radiation in iguanian lizards. *Science* **301**: 961–964.
- HEDRICK, P. W., and S. T. KALINOWSKI, 2000 Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics* **31**: pp. 139–162.
- HERMISSON, J., and P. S. PENNINGS, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.
- HERNANDEZ, R. D., J. L. KELLEY, E. ELYASHIV, S. C. MELTON, A. AUTON, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.
- HEWITT, G., 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–913.
- HEWITT, G. M., 2004 Genetic consequences of climatic oscillations in the quaternary. *Philos Trans R Soc Lond B Biol Sci* **359**: 183–95; discussion 195.
- HEY, J., and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HUGHES, C., and R. EASTWOOD, 2006 Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. *Proc Natl Acad Sci U S A* **103**: 10334–10339.
- JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO, and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**: 1401–1410.
- JIMENEZ, J. A., K. A. HUGHES, G. ALAKS, L. GRAHAM, and R. C. LACY, 1994 An experimental study of inbreeding depression in a natural habitat. *Science* **266**: 271–273.

- JONES, F. C., M. G. GRABHERR, Y. F. CHAN, P. RUSSELL, E. MAUCELI, *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.
- JORDE, P. E., and N. RYMAN, 2007 Unbiased estimator for genetic drift and effective population size. *Genetics* **177**: 927–935.
- JOYCE, D. A., D. H. LUNT, M. J. GENNER, G. F. TURNER, R. BILLS, *et al.*, 2011 Repeated colonization and hybridization in lake malawi cichlids. *Curr Biol* **21**: R108–R109.
- JUMP, A. S., R. MARCHANT, and J. PEÑUELAS, 2009 Environmental change and the option value of genetic diversity. *Trends Plant Sci* **14**: 51–58.
- KARLIN, S., 1982 Classifications of selection migration structures and conditions for a protected polymorphism. *Evolutionary Biology* **14**: 61–204.
- KELLER, I., W. NENTWIG, and C. R. LARGIADER, 2004 Recent habitat fragmentation due to roads can lead to significant genetic differentiation in an abundant flightless ground beetle. *Mol Ecol* **13**: 2983–2994.
- KIM, Y., and D. GULISIJA, 2010 Signatures of recent directional selection under different models of population expansion during colonization of new selective environments. *Genetics* **184**: 571–585.
- KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1513–1524.
- KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765–777.
- KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.

- KREITMAN, M., 2000 Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet* **1**: 539–559.
- KUHNER, M. K., 2009 Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* **24**: 86–93.
- LAMBECK, K., F. ANTONIOLI, A. PURCELL, and S. SILENZI, 2004 Sea-level change along the italian coast for the past 10,000 yr. *Quaternary Science Reviews* **23**: 1567 – 1598.
- LATTER, B. D., 1973 The island model of population differentiation: a general solution. *Genetics* **73**: 147–157.
- LEBLOIS, R., A. ESTOUP, and R. STREIFF, 2006 Genetics of recent habitat contraction and reduction in population size: does isolation by distance matter? *Mol Ecol* **15**: 3601–3615.
- LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in drosophila. *PLoS Genet* **2**: e166.
- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor Popul Biol* **1**: 273–306.
- MAYNARD SMITH, J., 1970 Population size, polymorphism, and the rate of non-darwinian evolution. *The American Naturalist* **104**: 231–237.
- MILLER, J., and R. HOBBS, 2002 Conservation where people live and work. *Conservation Biology* **16**: 330–337.
- MYLES, S., N. BOUZEKRI, E. HAVERFIELD, M. CHERKAOUI, J.-M. DUGOUJON, *et al.*, 2005 Genetic evidence in support of a shared eurasian-north african dairying origin. *Hum Genet* **117**: 34–42.

NAGYLAKI, T., 1974 The decay of genetic variability in geographically structured populations. *Proc Natl Acad Sci U S A* **71**: 2932–2936.

NAGYLAKI, T., 1977 Decay of genetic variability in geographically structured populations. *Proc Natl Acad Sci U S A* **74**: 2523–2525.

NAGYLAKI, T., 1979 The island model with stochastic migration. *Genetics* **91**: 163–176.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**: 3321–3323.

NEI, M., and M. W. FELDMAN, 1972 Identity of genes by descent within and between populations under mutation and migration pressures. *Theor Popul Biol* **3**: 460–465.

NEI, M., T. MARUYAMA, and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. *Evolution* **29**: pp. 1–10.

NEWMAN, D., and D. PILSON, 1997 Increased probability of extinction due to decreased genetic effective population size: Experimental populations of *Clarkia pulchella*. *Evolution* **51**: 354–362.

NIELSEN, R., 2005 Molecular signatures of natural selection. *Annual Review of Genetics* **39**: 197–218.

NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON, *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.

NOEL, S., M. OUELLET, P. GALOIS, and F.-J. LAPOINTE, 2007 Impact of urban fragmentation on the genetic structure of the eastern red-backed salamander. *Conservation Genetics* **8**: 599–606.

- ORR, H. A., and A. J. BETANCOURT, 2001 Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157**: 875–884.
- OWEN, R. B., R. CROSSLEY, T. C. JOHNSON, D. TWEDDLE, I. KORNFIELD, *et al.*, 1990 Major low levels of lake malawi and their implications for speciation rates in cichlid fishes. *Proceedings of the Royal Society of London. B. Biological Sciences* **240**: 519–553.
- PAVLIDIS, P., J. D. JENSEN, and W. STEPHAN, 2010 Searching for footprints of positive selection in whole-genome snp data from nonequilibrium populations. *Genetics* **185**: 907–922.
- PEARSE, D. E., and K. A. CRANDALL, 2004 Beyond fst: Analysis of population genetic data for conservation. *Conservation Genetics* **5**: 585–602. [10.1007/s10592-004-1863-z](https://doi.org/10.1007/s10592-004-1863-z).
- PELZ, H.-J., S. ROST, M. HÜNERBERG, A. FREGIN, A.-C. HEIBERG, *et al.*, 2005 The genetic basis of resistance to anticoagulants in rodents. *Genetics* **170**: 1839–1847.
- PETER, B. M., D. WEGMANN, and L. EXCOFFIER, 2010 Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. *Mol Ecol* **19**: 4648–4660.
- PRICE, A. L., A. TANDON, N. PATTERSON, K. C. BARNES, N. RAFAELS, *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5**: e1000519.
- PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- RICE, S. H., and A. PAPADOPOULOS, 2009 Evolution with stochastic fitness and stochastic migration. *PLoS One* **4**: e7130.

- ROUSSET, F., 1996 Equilibrium values of measures of population subdivision for step-wise mutation processes. *Genetics* **142**: 1357–1362.
- SCHLUTER, D., E. A. CLIFFORD, M. NEMETHY, and J. S. MCKINNON, 2004 Parallel evolution and inheritance of quantitative traits. *Am Nat* **163**: 809–822.
- SEDANO, R. E., and K. J. BURNS, 2010 Are the northern andes a species pump for neotropical birds? phylogenetics and biogeography of a clade of neotropical tanagers (aves: Thraupini). *Journal of Biogeography* **37**: 325–343.
- SEEHAUSEN, O., 2004 Hybridization and adaptive radiation. *Trends Ecol Evol* **19**: 198–207.
- SHPAK, M., J. WAKELEY, D. GARRIGAN, and R. C. LEWONTIN, 2010 A structured coalescent process for seasonally fluctuating populations. *Evolution* **64**: 1395–1409.
- SLATKIN, M., 1991 Inbreeding coefficients and coalescence times. *Genet Res* **58**: 167–175.
- STURMBAUER, C., S. BARIC, W. SALZBURGER, L. RÜBER, and E. VERHEYEN, 2001 Lake level fluctuations synchronize genetic divergences of cichlid fishes in african lakes. *Mol Biol Evol* **18**: 144–154.
- STÄDLER, T., B. HAUBOLD, C. MERINO, W. STEPHAN, and P. PFAFFELHUBER, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.
- TAJIMA, F., 1983 Evolutionary relationship of dna sequences in finite populations. *Genetics* **105**: 437–460.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**: 585–595.

- TAKAHATA, N., and M. NEI, 1985 Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**: 325–344.
- TURNER, R. C., J. C. LEVY, and A. CLARK, 1993 Complex genetics of type 2 diabetes: thrifty genes and previously neutral polymorphisms. *Q J Med* **86**: 413–417.
- VANDERGAST, A. G., E. A. LEWALLEN, J. DEAS, A. J. BOHONAK, D. B. WEISSMAN, *et al.*, 2009 Loss of genetic connectivity and diversity in urban microreserves in a southern California endemic Jerusalem cricket (Orthoptera: Stenopelmatidae: *Stenopelmatus* n. sp. “santa monica”). *Journal of Insect Conservation* **13**: 329–345.
- WAKELEY, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- WAPLES, R. S., 2010 Spatial-temporal stratifications in natural populations and how they affect understanding and estimation of effective population size. *Mol Ecol Resour* **10**: 785–796.
- WATTERSON, G. A., 1978 The homozygosity test of neutrality. *Genetics* **88**: 405–417.
- WHITLOCK, M. C., 1992 Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution* **46**: pp. 608–615.
- WHITLOCK, M. C., and N. H. BARTON, 1997 The effective size of a subdivided population. *Genetics* **146**: 427–441.
- WRIGHT, S., 1931 Evolution in mendelian populations. *Genetics* **16**: 97–159.
- WRIGHT, S., 1950 Genetical structure of populations. *Nature* **166**: 247–249.
- YOUNG, K. A., J. M. WHITMAN, and G. F. TURNER, 2009 Secondary contact during adaptive radiation: a community matrix for lake malawi cichlids. *J Evol Biol* **22**: 882–889.

ZHANG, H., J. YAN, G. ZHANG, and K. ZHOU, 2008 Phylogeography and demographic history of chinese black-spotted frog populations (*Pelophylax nigromaculata*): Evidence for independent refugia expansion and secondary contact. *BMC Evolutionary Biology* 8: 21.

APPENDIX A | DYNAMICS OF GENETIC DIVERSITY

In this appendix, we describe the temporal change of genetic diversity (derivation of eq. 2.4a, 2.8 and the separation of the dynamics of genetic diversity into two timescales).

Temporal change of genetic diversity: The solution to eq. 2.3 is:

$$\mathbf{F}_t = \mathbf{A}^t(\mathbf{F}_0 - \mathbf{F}^{eq}) + \mathbf{F}^{eq} \quad (2-A.1)$$

Denoting \mathbf{P} the transformation matrix with eigenvector \mathbf{U}_1 (associated with λ_1) as first column, and eigenvector \mathbf{U}_2 (associated with λ_2) as second column:

$$\mathbf{P} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix} \quad (2-A.2)$$

and denoting

$$\begin{pmatrix} y_{10} \\ y_{20} \end{pmatrix} = \mathbf{P}^{-1}(\mathbf{F}_0 - \mathbf{F}^{eq}) \quad (2-A.3)$$

Eq. 2-A.1 becomes:

$$\begin{aligned}
 \mathbf{F}_t &= \mathbf{P} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^t \mathbf{P}^{-1}(\mathbf{F}_0 - \mathbf{F}^{eq}) + \mathbf{F}^{eq} \\
 &= \mathbf{P} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^t \begin{pmatrix} y_{10} \\ y_{20} \end{pmatrix} + \mathbf{F}^{eq} \\
 &= y_{10} \mathbf{U}_1 \lambda_1^t + y_{20} \mathbf{U}_2 \lambda_2^t + \mathbf{F}^{eq}
 \end{aligned} \tag{2-A.4}$$

Further denoting $\mathbf{C}_1 = y_{10} \mathbf{U}_1$ and $\mathbf{C}_2 = y_{20} \mathbf{U}_2$ leads to eq. 2.4a.

\mathbf{F}_t changes according to two exponential decay functions, λ_1^t and λ_2^t . Their rate of change are $\frac{d\lambda_1^t}{dt} = \ln(\lambda_1) \cdot e^{\ln(\lambda_1)t}$ and $\frac{d\lambda_2^t}{dt} = \ln(\lambda_2) \cdot e^{\ln(\lambda_2)t}$, respectively. Thus $r_1 = \ln(\lambda_1)$ and $r_2 = \ln(\lambda_2)$ are the decay constants which determine the rate of change of functions λ_1^t and λ_2^t .

Therefore, the eigenvalues of matrix \mathbf{A} can be used to compute the rates of change of genetic diversity. As $\lambda_1 > \lambda_2$ and both eigenvalues are less than 1, we have $|r_2| > |r_1|$, and thus $\mathbf{C}_2 \lambda_2^t$ tends to 0 faster than $\mathbf{C}_1 \lambda_1^t$. r_2 determines the transient rate of change of genetic diversity, while r_1 determines the asymptotic rate of change of genetic diversity.

We now want to simplify the expression of the rates of change of genetic diversity.

To do so, we can rewrite eq. 2.5 as:

$$\begin{cases} \lambda_1 = \frac{(1-\mu)^2}{2} [a(1-c) + 1 - b + (1 - a(1-c) + b) \sqrt{1 - \frac{4bc}{(1 - a(1-c) + b)^2}}] \\ \lambda_2 = \frac{(1-\mu)^2}{2} [a(1-c) + 1 - b - (1 - a(1-c) + b) \sqrt{1 - \frac{4bc}{(1 - a(1-c) + b)^2}}] \end{cases}$$

With $4bc \ll (1 - a(1-c) + b)^2$ (as $m \ll 1$ and $N \gg 1$), and given that $\sqrt{1-x} =$

$1 - \frac{1}{2}x + o(x)$, eq. 2.5 further simplifies to:

$$\begin{cases} \lambda_1 = (1 - \mu)^2 \left(1 - \frac{bc}{1 - a(1 - c) + b}\right) \\ \lambda_2 = (1 - \mu)^2 \left(a(1 - c) - b + \frac{bc}{1 - a(1 - c) + b}\right) \end{cases}$$

Considering that migration rates and mutation rates are small, we can neglect terms in m^2 , $\frac{1}{N^2}$, μ^2 , m/N , $m\mu$ and $\frac{\mu}{N}$, which leads to:

$$\begin{cases} \lambda_1 = 1 - 2\mu - \frac{1}{2nN(1 + \frac{n-1}{nM})} \\ \lambda_2 = 1 - 2\mu - 2m\frac{n}{n-1} - \frac{1}{2N} + \frac{1}{2nN(1 + \frac{n-1}{nM})} \end{cases} \quad (2-A.5)$$

and thus

$$\begin{cases} r_1 = \ln\left(1 - 2\mu - \frac{1}{2nN(1 + \frac{n-1}{nM})}\right) \\ r_2 = \ln\left(1 - 2\mu - 2m\frac{n}{n-1} - \frac{1}{2N} + \frac{1}{2nN(1 + \frac{n-1}{nM})}\right) \end{cases} \quad (2-A.6)$$

Assuming that migration and mutation rates are small, and that local population sizes are large ($N \gg 1$), eq. 2-A.6 simplifies to eq. 2.6 as $\ln(1 - x) = -x + o(x)$.

Respective length of the asymptotic and transient dynamics periods:

We denote t_1 and t_2 the times needed for λ_1^t and λ_2^t to be reduced to a value α , where $\alpha \in]0; 1]$:

$$\begin{cases} \lambda_1^{t_1} = \alpha \\ \lambda_2^{t_2} = \alpha \end{cases} \quad (2-A.7)$$

Which leads to:

$$\begin{cases} t_1 = \frac{\ln(\alpha)}{r_1} \\ t_2 = \frac{\ln(\alpha)}{r_2} \end{cases} \quad (2-A.8)$$

Assuming small migration and mutation rates and large local population sizes, we

can replace the expressions of r_1 and r_2 from eq. 2.6 into eq. 2-A.8, and we obtain eq. 2.8. F_t approximately follows:

$$F_t \simeq \begin{cases} F^{eq} + C_1\lambda_1^t + C_2\lambda_2^t & \text{for } t < t_2 \\ F^{eq} + C_1\lambda_1^t & \text{for } t_2 < t < t_1 \\ F^{eq} & \text{for } t_1 < t \end{cases} \quad (2-A.9)$$

Timescales separation:

This section presents the conditions for $t_1 \gg t_2$. When $t_1 \gg t_2$, $\lambda_2^{t_2} \simeq 0$ and $\lambda_1^{t_2} \simeq 1$, thus eq. 2-A.9 simplifies to:

$$F_t \simeq \begin{cases} F^{eq} + C_1 + C_2\lambda_2^t & \text{for } t < t_2 \\ F^{eq} + C_1\lambda_1^t & \text{for } t_2 < t < t_1 \\ F^{eq} & \text{for } t_1 < t \end{cases} \quad (2-A.10)$$

Eq. 2-A.10 decomposes the dynamics of F_t into two timescales: a *transient period* of length t_2 and an *asymptotic period* of length $t_1 - t_2 \simeq t_1$. For $t > t_1$, the genetic identity is close to its equilibrium value F^{eq} , so t_1 can be interpreted as the duration of the disequilibrium period.

Demonstration: Equation 2-A.10 is true if there exists a t such that $\lambda_1^t \simeq 1$ and $\lambda_2^t \simeq 0$. For simplicity, we consider that $\lambda_1^t \simeq 1$ if $\lambda_1^t \geq 0.95$, and that $\lambda_2^t \simeq 0$ if $\lambda_2^t \leq 0.05$. Thus, eq. 2-A.10 is a good approximation if there exists a t such that:

$$\begin{cases} \lambda_1^t \geq 0.95 \\ \lambda_2^t \leq 0.05 \end{cases} \Rightarrow \begin{cases} t \leq \frac{\ln(0.95)}{\ln(\lambda_1)} \\ \lambda_2^t \leq 0.05 \end{cases} \quad (2-A.11)$$

Thus, showing that

$$\lambda_2^{\frac{\ln(0.95)}{\ln(\lambda_1)}} \leq 0.05 \quad (2-A.12)$$

demonstrates that for $t = \frac{\ln(0.95)}{\ln(\lambda_1)}$, we have $\lambda_1^t \geq 0.95$ and $\lambda_2^t \leq 0.05$. This provides a sufficient proof of proposition 2-A.11.

We can demonstrate that proposition 2-A.12 depends only on the ratio $\frac{t_1}{t_2} = \frac{\ln(\lambda_2)}{\ln(\lambda_1)}$ (see definitions of t_1 and t_2 eq. 2.7). Proposition 2-A.12 leads to:

$$\begin{aligned} \lambda_2^{\frac{\ln(0.95)}{\ln(\lambda_1)}} &\leq 0.05 \\ \Leftrightarrow e^{\ln(0.95) \frac{\ln(\lambda_2)}{\ln(\lambda_1)}} &\leq 0.05 \\ \Leftrightarrow \ln(0.95) \frac{t_1}{t_2} &\leq \ln(0.05) \\ \Leftrightarrow \frac{t_1}{t_2} &\geq \frac{\ln(0.05)}{\ln(0.95)} \end{aligned}$$

Therefore, the condition $\frac{t_1}{t_2} \geq \frac{\ln(0.05)}{\ln(0.95)}$ is necessary and sufficient to prove propositions 2-A.12, 2-A.11 and the validity of equation 2-A.10. As $\frac{\ln(0.05)}{\ln(0.95)} \simeq 58.4$, equation 2-A.10 is valid when $\frac{t_1}{t_2} > 58.4$. Considering that m and μ are small, and that N is large, this ratio is approximately equal to (from eq. 2.8):

$$\frac{t_1}{t_2} \simeq \frac{2\mu + 2m \frac{n}{n-1} + \frac{1}{2N} - \frac{1}{2N_e}}{2\mu + \frac{1}{2N_e}} \quad (2-A.13)$$

From eq. 2-A.13 we can derive the conditions of the timescales separation of the dynamics of genetic diversity (i.e. the parameter values for which $t_1 \gg t_2$). When $n > 14$, differences are the highest ($t_1 \gg t_2$), in the domain where $\mu \ll \frac{1}{2N}$ and also when $\mu \gg \frac{1}{2N}$ and $m > \mu$. When $n \leq 14$, the same conditions apply for $t_1 \gg t_2$ except in a restricted domain where $m \simeq \frac{1}{2N}$ ($m \gg \frac{n-1}{2nN}$ or $m \ll \frac{n-1}{2nN}$ are required for

$t_1 \gg t_2$, see Figure 2-A.1). Indeed, denoting $A = \frac{\ln(0.05)}{\ln(0.95)}$, $\frac{t_1}{t_2} > A$ implies that:

$$\begin{aligned} M &> \frac{A + 1 - 2n + (A - 1)n\theta + \sqrt{D}}{2\frac{n^2}{n-1}} \\ \text{or } M &< \frac{A + 1 - 2n + (A - 1)n\theta - \sqrt{D}}{2\frac{n^2}{n-1}} \end{aligned} \quad (2-A.14a)$$

Where

$$D = (A + 1)(A + 1 - 4n) + 2n\theta(A^2 - 1) + n^2(A - 1)^2\theta^2 \quad (2-A.14b)$$

For $\theta \gg 1$, we can neglect terms that do not contain θ , and conditions 2-A.14 simplify to:

$$M > (A - 1)\frac{n - 1}{n}\theta \quad (2-A.15)$$

Which simplifies to $M \gg \theta$.

For $\theta \ll 1$, terms that contain θ can be neglected in eq. 2-A.14, which yields the following conditions:

$$\begin{aligned} M &> \frac{A + 1 - 2n + \sqrt{(A + 1)(A + 1 - 4n)}}{2\frac{n^2}{n-1}} \\ \text{or } M &< \frac{A + 1 - 2n - \sqrt{(A + 1)(A + 1 - 4n)}}{2\frac{n^2}{n-1}} \end{aligned} \quad (2-A.16)$$

Therefore, eq. 2-A.10 is not valid when M is between $\frac{A+1-2n+\sqrt{(A+1)(A+1-4n)}}{2\frac{n^2}{n-1}}$ and $\frac{A+1-2n+\sqrt{(A+1)(A+1-4n)}}{2\frac{n^2}{n-1}}$

This domain is centered around $M = \frac{n-1}{n}$, as $\frac{A+1-2n+\sqrt{(A+1)(A+1-4n)}}{2\frac{n^2}{n-1}}$ and $\frac{A+1-2n+\sqrt{(A+1)(A+1-4n)}}{2\frac{n^2}{n-1}}$

equalize to $M = \frac{n-1}{n}$, for $A=4n-1$. The size of this domain decreases when n increases

(Figure 2-A.1), and eq. 2-A.10 is valid for any value of M in the domain $\theta \ll 1$ when

$n > 14$ (as conditions 2-A.16 are relaxed when $4n > A+1$, with $A = \frac{\ln(0.05)}{\ln(0.95)}$).

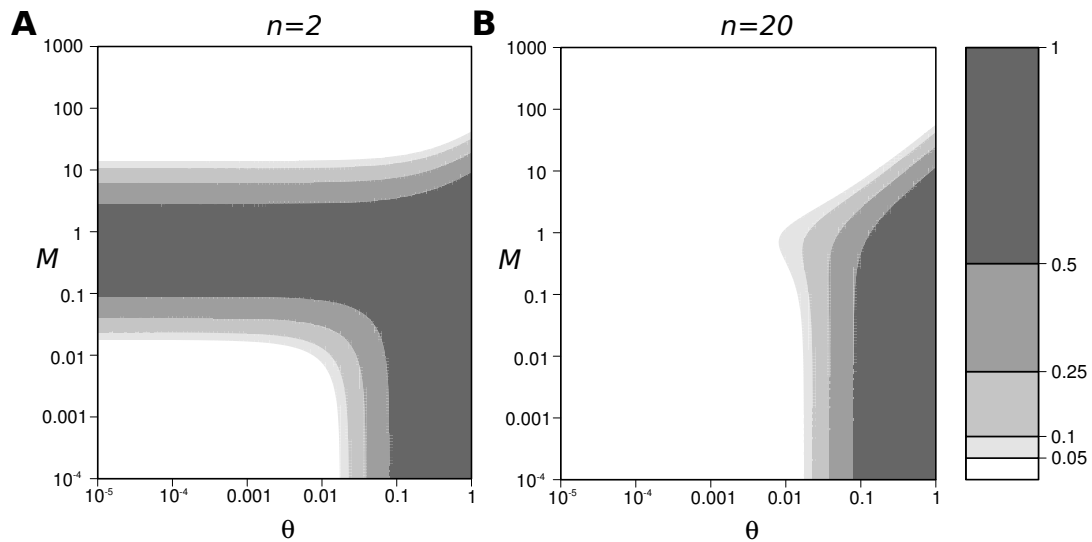


Figure 2-A.1 Domain of validity of equation 2-A.10 (white contours), as a function of the strength of migration (M) and mutation (θ), (A) $n = 2$, (B) $n = 20$. The dark gray filled contours represent the domains where the validity of eq. 2-A.10 is poor (i.e. the exact value of $\lambda_2^{\ln(0.95)/\ln(\lambda_1)} > 0.05$; from eq. 5).

APPENDIX B | PEAK OF GENETIC DIVERSITY GENERATED BY A RE-CONNECTION EVENT

We derive first the value of the peak of within-population genetic diversity, ΔH_s , and the transient excess of between-population genetic diversity, ΔH_b , generated by a re-connection event. Second, we characterize their dependency on the migration rate, mutation rate, size and number of populations.

Derivation of ΔH :

We denote the vector of genetic diversity excess at the end of the transient dynamics phase as $\Delta H = \begin{pmatrix} \Delta H_s \\ \Delta H_b \end{pmatrix}$.

The genetic diversity at the end of the transient dynamics phase is approximately

$\mathbf{H}_{t_2} = \mathbf{1} - (\mathbf{F}^{eq} + \mathbf{C}_1 \lambda_1^{t_2})$ (from eq. 2-A.9 when $t \simeq t_2$). Thus $\Delta \mathbf{H}$ is approximately:

$$\begin{aligned} \Delta \mathbf{H} &= \mathbf{H}_{t_2} - \mathbf{H}^{eq} \\ &= \mathbf{1} - (\mathbf{F}^{eq} + \mathbf{C}_1 \lambda_1^{t_2}) - (\mathbf{1} - \mathbf{F}^{eq}) \\ &= -\mathbf{C}_1 \lambda_1^{t_2} \end{aligned} \quad (2-B.1)$$

The peak of genetic diversity depends on $\mathbf{C}_1 = y_{10} \mathbf{U}_1$ (from eq. 2-A.2 and 2-A.3).

To derive the value of \mathbf{C}_1 , we compute \mathbf{U}_1 and y_{10} . The eigenvector $\mathbf{U}_1 = \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}$ is associated with λ_1 and follows:

$$\mathbf{A} \mathbf{U}_1 = \lambda_1 \mathbf{U}_1 \quad (2-B.2)$$

Eq. 2-B.2 leads to the condition that $u_{21} = -\frac{(a(1-c)-(1-b)-\sqrt{\Delta_{red}})}{2(1-a)} u_{11}$, where $\Delta_{red} = (1 - a(1-c) + b)^2 - 4bc$. We set $u_{11} = 2(1-a)$, leading to:

$$\mathbf{U}_1 = \begin{pmatrix} 2(1-a) \\ 1 - a(1-c) - b + \sqrt{(1 - a(1-c) + b)^2 - 4bc} \end{pmatrix}$$

Similarly, we can determine the eigenvector associated with λ_2 :

$$\mathbf{U}_2 = \begin{pmatrix} 2(1-a) \\ 1 - a(1-c) - b - \sqrt{(1 - a(1-c) + b)^2 - 4bc} \end{pmatrix}$$

Denoting the initial value of genetic identity $\mathbf{F}_0 = \begin{pmatrix} F_{s,0} \\ F_{b,0} \end{pmatrix}$ and replacing the expression of \mathbf{U}_1 and \mathbf{U}_2 in eq. 2-A.3 leads to:

$$\mathbf{Y}_0 = \begin{pmatrix} -\frac{(1-a(1-c)-b-\sqrt{\Delta_{red}})}{4(1-a)\sqrt{\Delta_{red}}} & \frac{1}{2\sqrt{\Delta_{red}}} \\ \frac{1-a(1-c)-b+\sqrt{\Delta_{red}}}{4(1-a)\sqrt{\Delta_{red}}} & -\frac{1}{2\sqrt{\Delta_{red}}} \end{pmatrix} \begin{pmatrix} F_{s,0} - F_{s,con}^{eq} \\ F_{b,0} - F_{b,con}^{eq} \end{pmatrix}$$

Assuming isolation equilibrium for the initial identity leads to $F_{s,0} = F_{s,iso}^{eq}$ and $F_{b,0} = 0$, thus ΔH simplifies to:

$$\Delta H = [(F_{s,iso}^{eq} - F_{s,con}^{eq}) \frac{1 - a(1 - c) - b - \sqrt{\Delta_{red}}}{4(1 - a)\sqrt{\Delta_{red}}} + \frac{F_{b,con}^{eq}}{2\sqrt{\Delta_{red}}}] U_1 \lambda_1^{t_2} \quad (2-B.3)$$

With $4bc \ll (1 - a(1 - c) + b)^2$ (as $m \ll 1$ and $N \gg 1$), and given that $\sqrt{1 - x} = 1 - \frac{1}{2}x + o(x)$, ΔH further becomes:

$$\Delta H = -\frac{1}{2(1 - a(1 - c) + b - \frac{2bc}{1 - a(1 - c) + b})} \left[\frac{(F_{s,iso}^{eq} - F_{s,con}^{eq})}{n - 1} \frac{1 - a(1 - c) + b - c}{1 - a(1 - c) + b} - F_{b,con}^{eq} \right] U_1 \lambda_1^{t_2} \quad (2-B.4)$$

Assuming that migration and mutation rates are always small, we can neglect terms in m^2 , μ^2 , $\frac{1}{N^2}$, $m\mu$, $\frac{\mu}{N}$ and $\frac{m}{N}$ and eq. 2-B.4 simplifies to:

$$\Delta H = -\left[\frac{(F_{s,iso}^{eq} - F_{s,con}^{eq})}{n - 1} \frac{M \frac{n}{n-1}}{1 + M \frac{n}{n-1}} - F_{b,con}^{eq} \right] \left(\frac{\frac{M}{1 + \frac{n}{n-1} M - \frac{2N}{N_e}}}{\frac{1 + M - \frac{N}{N_e}}{1 + \frac{n}{n-1} M - \frac{2N}{N_e}}} \right) \lambda_1^{t_2} \quad (2-B.5)$$

By replacing the term $\lambda_1^{t_2}$ (eq. 2.8 with $\alpha = 0.05$) in eq. 2-B.5, we obtain eq. 2.9.

Eq. 2.9 provides a good approximation of the size of the peak of within-population genetic diversity ΔH_s and of the transient excess of between-population genetic diversity ΔH_b in the entire parameter domain. Figure 2-B.1(A) and (C), and Figure 2-B.2(A) and (C) represent the exact (solid line) and approximate (dashed line; from eq. 2.9) values of ΔH_s and ΔH_b , respectively, as a function of θ and M , for $n = 2$ and $n = 20$; we can see that the true and approximate values are very close. Figure 2-B.1(B) and (D), and Figure 2-B.2(B) and (D), represent the absolute error resulting from the use of eq. 2.9 as an approximation of ΔH_s and ΔH_b , respectively, instead of its exact value, for $n = 2$ and $n = 20$. Discrepancies between eq. 2.9 and the true values of ΔH_s and ΔH_b can first come from the assumption that $m \ll 1$, $\mu \ll 1$ and $N \gg 1$, and second

from the assumption of the existence of t_2 such that $|C_2\lambda_2^{t_2}| \ll |C_1\lambda_1^{t_2}|$.

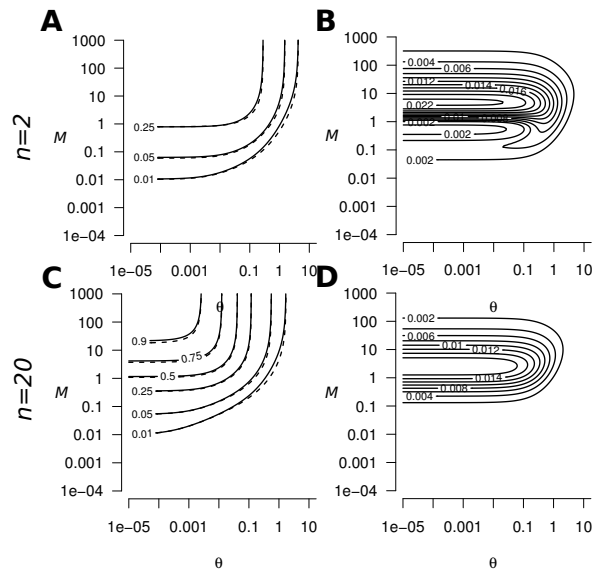


Figure 2-B.1 (A), (C) Exact (solid lines) and approximate (dashed lines, from eq. 2.9) values of the peak of genetic diversity ΔH_s , as a function of the strength of migration (M) and mutation (θ). In both (A) and (C), the exact and approximate values of ΔH_s are very close. (B), (D) Absolute error when using eq. 2.9 to approximate ΔH_s , as a function of M and θ . The maximum absolute error is reached when $M \simeq 5$ and $\theta < 1$ in both (B) and (D). The error decreases when $M \gg 5$ or $M \ll 5$. The absolute error increases when n decreases, but remains weak: (B) the maximum absolute error is 0.025 for $n = 2$, and (D) 0.018 for $n = 20$. Consequently, eq. 2.9 is a good approximation for the peak of genetic diversity whatever the parameter values of θ , M and n considered.

Maximum peak of diversity after a connection event:

The peak of genetic diversity increases monotonously with M ($\frac{d(\Delta H)}{dM} > 0$ for any value of m , μ , n and N), and decreases monotonously with θ ($\frac{d(\Delta H)}{d\theta} < 0$ for any value of m , μ , n and N). The peak of diversity is maximized for intermediate values of n , and reached when:

$$\frac{d(\Delta H)}{dn} = 0 \quad (2-B.6)$$

If we neglect terms in μ^2 , m^2 , $\frac{1}{N^2}$, μm , $\frac{\mu}{N}$ and $\frac{m}{N}$, and assuming small θ and high M , solving eq. 2-B.6 using equation 2.10 for the peak of genetic diversity leads to equation 2.11. When $n = n^*$, eq. 2.10 leads to a peak of genetic diversity of $\Delta H^{max}|_{n^*} = \frac{1}{1+2\sqrt{\theta}}$.

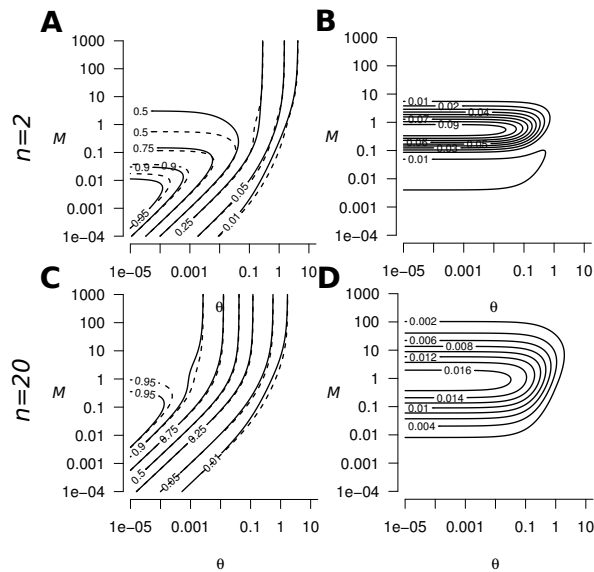


Figure 2-B.2 (A), (C) Exact (solid lines) and approximate (dashed lines, from eq. 2.9) values of the transient excess of between-population genetic diversity ΔH_b , as a function of the strength of migration (M) and mutation (θ). In both (A) and (C), the exact and approximate values of ΔH_b are very close. (B), (D) Absolute error when using equation 2-A.9 to approximate ΔH_b , as a function of M and θ . The maximum absolute error is reached when $M \simeq 1$ and $\theta < 1$, the absolute error is 0.09 for $n = 2$ (B), and 0.016 for $n = 20$ (D). Eq. 2.9 is a good approximation for ΔH_b whatever the parameter values of θ , M and n considered.

Chapter 3 Dynamics of Genetic Diversity Across Multiple Isolation and Connection Events

Nicolas Alcalá¹ and Séverine Vuilleumier^{1,2}

¹ Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

² Institute of Microbiology, Lausanne University Hospital and University of Lausanne, Switzerland

Migration

Population structure

Secondary contact

Model

Stochastic

In prep.

ABSTRACT

Major environmental and geological events (e.g. sea and continental water level fluctuations or islands formation) but also population history (secondary contacts) have generated multiple periods of populations isolation and connection at long and short time-scales. Recently, numerous empirical and theoretical studies suggest that fast evolutionary processes might be triggered by such events, as commonly illustrated in ecology by the adaptive radiation of cichlid fishes (isolation and reconnection of lakes and watersheds) and in epidemiology by the fast adaptation of the influenza virus (isolation and reconnection within hosts). We test the hypothesis that isolation and reconnection events provide the raw material (standing genetic variation) for species adaptation. Our analytical results confirm that population isolation and reconnection can provide, to populations, high levels of genetic diversity (higher than expected at equilibrium) over very large time-periods and that this excess is either cyclic (high allele turnover) or cumulates with time. The limit between those two behaviors, the amplitude of variation of genetic diversity as well as its excess across multiple cycles are defined by the duration of isolation and connection phases and the scaled mutation rate. We demonstrate that our results are robust to stochasticity and discuss the consequences of our results to understand the mechanisms of adaptation.

SUCCESSIVE environmental changes have often modified species habitat in the past, repeatedly isolating and connecting populations, successively suppressing and allowing migration between them. During the glaciations of the quaternary period, many species experienced repeated long-periods of isolation into refugia followed by population reconnection (HEWITT 2000, 2004, 2011; ZHANG *et al.* 2008). At least five wet-dry periods occurred within the African continent (Szabo *et al.* 1995) in which water level fluctuations have successively isolated and connected ecosystems (Sturmbauer *et al.*, 2001). Such dynamics are often associated with rapid species and population diversification, for example, allopatric differentiations of African buffalo populations (Van Hooft *et al.* 2002, Heller *et al.* 2008, 2012), giraffe (Brown *et al.* 2007) and their associated predators (e.g. lions, Barnett *et al.* 2006). Similarly, repeated hybridizations within and between watersheds are expected to have played an important role for Cichlid adaptive radiation (Owen *et al.* 1990, Seehausen 2002, 2004, Seehausen *et al.* 2003, Schwarzer *et al.* 2012, Nevado *et al.* 2013).

Isolation and reconnection events are also common features of virus history. Their dependency on the host cell and their life cycle isolate viral population in-between transmission events. Events of reconnection of populations occur through superinfections (co-infections) where reassortment of different virus strains can be observed. Such isolation and reconnection events have been widely observed and constitute a major cause of pandemics and failures in virus control. For example, reassortment between an avian and human flu virus caused the pandemic flu in 1957 and 1968 and a mix of swine, avian and human influenza was responsible for the 2009 swine flu outbreak (Hsieh *et al.* 2006, Garten *et al.* 2009, Flahault and Zylberman 2010).

Previous studies showed that population genetic diversity is slightly impacted by migration fluctuation and that its influence can be approximated with an effective mi-

gration rate (NAGYLAKI 1979; WHITLOCK 1992; RICE and PAPADOPOULOS 2009; SHPAK *et al.* 2010). In contrast, high transient values of genetic diversity can be observed following a single (ALCALA *et al.* 2013) or multiples (JESUS *et al.* 2006) connection events. Those results raised several questions. In particular, would the excess of genetic diversity observed in JESUS *et al.* (2006) and ALCALA *et al.* (2013) accumulate, be maintained or decrease under periodic isolation and connection events? The answer to this question is of theoretical interest but also has strong implications for the understanding of mechanisms of diversification. Indeed, theoretical and empirical works on speciation and adaptive radiation suggest that a high level of pre-existing neutral genetic diversity in founder populations is determinant to have a high rate of speciation (BARRIER *et al.* 1999; GAVRILETS 2003; HARMON *et al.* 2003; SEEHAUSEN 2004; BARRETT and SCHLUTER 2008; GAVRILETS and LOSOS 2009; BEZAULT *et al.* 2011). It could provide the raw material for evolution from the standing genetic variation (TURNER *et al.* 1993; FEDER *et al.* 2003; PELZ *et al.* 2005; COLOSIMO *et al.* 2005; HERMISSON and PENNINGS 2005; MYLES *et al.* 2005; HERNANDEZ *et al.* 2011; JONES *et al.* 2012). This view is supported by simulation studies showing that alternation between allopatry and sympatry can trigger species diversification (AGUILEE *et al.* 2011; AGUILÉE *et al.* 2013).

While a single (ALCALA *et al.* 2013) or multiple (JESUS *et al.* 2006) events of populations connection can generate a large excess of neutral genetic diversity within populations, the conditions of existence and maintenance of such genetic diversity excesses remain to be determined. To investigate this question, we analytically describe the dynamics of genetic diversity under successive periods of isolation and connection of populations (migration periodically takes the value of 0 and m) of different durations. Assuming that all alleles are selectively neutral, we disentangle the relative importance of the length of the isolation and connection periods in regard to the mutation

rate, the migration rate, the number of populations and the initial population genetic diversity. Our investigations first consider periods of isolation and connection of deterministic lengths. From this, we characterize equilibrium trajectories and transient values of genetic diversity toward equilibrium. We determine four domains in which genetic diversity has a determined behaviour in regards to genetic diversity accumulation and turnover. Robustness of our results to stochasticity is then evaluated. Finally, we discuss the implication of our results for species adaptation to novel environment, pathogen control, and inference of population history.

DYNAMICS OF GENETIC DIVERSITY UNDER PERIODIC CONNECTION AND ISOLATION EVENTS

To study the dynamics of genetic diversity, we consider n populations of size N (diploid individuals) that are periodically isolated and connected (Figure 3.1). Connection and isolation periods are of regular lengths of P generations each, and K cycles of isolation and connection events are considered. Under this model we determine the recursion equations describing the dynamics of genetic diversities within and between isolation and connection periods and the long-term equilibrium trajectory of genetic diversity. We assume a finite island model, non-overlapping generations (Wright-Fisher model; FISHER 1930; WRIGHT 1931) and consider that mutations (rate μ) follow the infinite allele model (KIMURA and CROW 1964).

The dynamics of genetic diversities within connection and isolation periods

Following MARUYAMA (1970), we decompose the genetic diversity $H(t)$ at generation t into the within-population genetic diversity, $h_s(t)$, and the between-population genetic

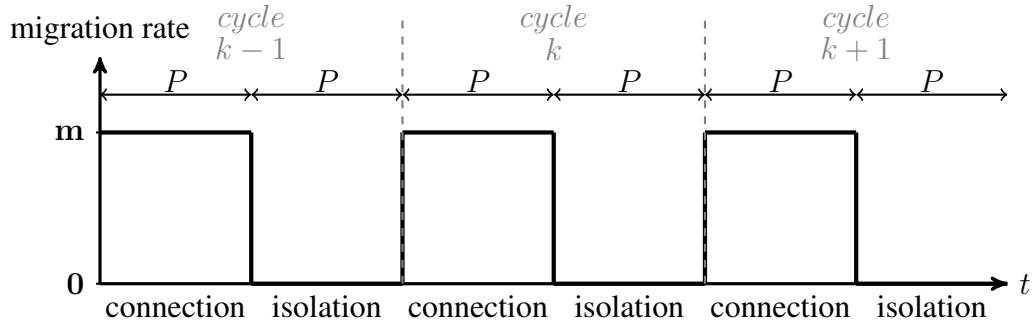


Figure 3.1 Scheme of periodic isolation and connection events. The migration rate m periodically switches from m (connection) to 0 (isolation). Each connection or isolation period lasts P generations and each connection/isolation cycle lasts $2P$ generations. In grey are represented the cycle indexes $k - 1$, k and $k + 1$. We arbitrarily chose to begin cycles with a connection period.

diversity, $h_b(t)$:

$$\mathbf{H}(t) = \begin{pmatrix} h_s(t) \\ h_b(t) \end{pmatrix} \quad (3.1)$$

Where $h_s(t)$ and $h_b(t)$ correspond to the probability that two genes, randomly sampled from a same and from different populations, respectively, are different (NEI and FELDMAN 1972).

We can derive the recursion equations for the change in genetic diversities after P generations during a connection period and during an isolation period, $\mathbf{H}_c(P)$ and $\mathbf{H}_i(P)$, respectively (from MARUYAMA 1970; NEI and FELDMAN 1972):

$$\begin{cases} \mathbf{H}_c(P) = \mathbf{A}_c^P(\mathbf{H}_c(0) - \hat{\mathbf{H}}_c) + \hat{\mathbf{H}}_c \\ \mathbf{H}_i(P) = \mathbf{A}_i^P(\mathbf{H}_i(0) - \hat{\mathbf{H}}_i) + \hat{\mathbf{H}}_i \end{cases} \quad (3.2)$$

Where $\mathbf{H}_c(0)$ and $\mathbf{H}_i(0)$ are the initial genetic diversities during the connection and the isolation periods and $\hat{\mathbf{H}}_c$ and $\hat{\mathbf{H}}_i$ are their expected value at equilibrium (from KIMURA and CROW 1964 and MARUYAMA 1970). Matrix \mathbf{A}_c and \mathbf{A}_i are transition matrices which determine the probability of identity of two genes in a same or differ-

ent populations, given their previous identity in a same or different populations, they follow (MARUYAMA 1970; NEI and FELDMAN 1972):

$$\mathbf{A}_c = (1 - \mu)^2 \begin{pmatrix} a(1 - c) & 1 - a \\ b(1 - c) & 1 - b \end{pmatrix} \quad (3.3a)$$

with parameters:

$$a = (1 - m)^2 + \frac{m^2}{n - 1} \quad (3.3b)$$

$$b = \frac{1 - a}{(n - 1)} \quad (3.3c)$$

$$c = \frac{1}{2N} \quad (3.3d)$$

$$\mathbf{A}_i = (1 - \mu)^2 \begin{pmatrix} 1 - c & 0 \\ 0 & 1 \end{pmatrix} \quad (3.4)$$

In equation 3.3, a (resp. b) corresponds to the probability that two genes sampled in a same population (resp. different populations) were in the same population in the previous generation, and c corresponds to the probability that two genes sampled in a same population are copies of the same gene; $(1 - \mu)^2$ corresponds to the probability that neither of the two genes mutated. Thus $(1 - \mu)^2 a(1 - c)$ (resp. $(1 - \mu)^2 b(1 - c)$) is the probability that two genes that were identical and in a same population (resp. different populations) at a given generation are still identical and in the same population at the next generation; $(1 - \mu)^2(1 - a)$ (resp. $(1 - \mu)^2(1 - b)$) is the probability that two genes that were identical and in a same population (resp. different populations) at a given generation are still identical and in different populations at the next generation.

The dynamics of genetic diversities across connection-isolation cycles

From equation 3.2, we derive the recurrence equations describing the dynamics of genetic diversity across multiple cycles k of connection-isolation periods. We obtain the genetic diversities, $\mathbf{H}_i^{(k+1)}$, at the end of any cycle $k + 1$ (end of isolation period), given genetic diversities, $\mathbf{H}_c^{(k+1)}$, at the end of the connection period of the same cycle $k + 1$. Similarly, $\mathbf{H}_c^{(k+1)}$ depends on $\mathbf{H}_i^{(k)}$, the genetic diversities at end of the previous cycle k :

$$\begin{cases} \mathbf{H}_c^{(k+1)} = \mathbf{A}_c^P (\mathbf{H}_i^{(k)} - \widehat{\mathbf{H}}_c) + \widehat{\mathbf{H}}_c \\ \mathbf{H}_i^{(k+1)} = \mathbf{A}_i^P (\mathbf{H}_c^{(k+1)} - \widehat{\mathbf{H}}_i) + \widehat{\mathbf{H}}_i \end{cases} \quad (3.5)$$

These equations provide a full description of the transient dynamics of genetic diversities across any number of cycles k of duration P . Equation 3.5 can be rewritten as a simple linear equation:

$$\mathbf{H}_c^{(k)} = \Gamma_c^k (\mathbf{H}_c^{(0)} - \mathbf{H}_c^*) + \mathbf{H}_c^* \quad (3.6a)$$

where \mathbf{H}_c^* is the equilibrium value, and where:

$$\Gamma_c = A_c^P A_i^P \quad (3.6b)$$

Consequently, the dynamics of genetic diversities at the end of each cycle k are determined by matrix Γ_c . Similarly, the matrix $\Gamma_i = A_i^P A_c^P$ determines the dynamics of $\mathbf{H}_i^{(k)}$. As Γ_i and Γ_c have the same eigenvalues, the behaviour of genetic diversity studied here is independent of order of the sequence of connection and isolation event.

Equilibrium trajectory of genetic diversities under periodic connection-isolation events

The equilibrium trajectory of genetic diversities is reached when two consecutive isolation events, with a connection event in-between, have the same genetic diversity values, i.e. when:

$$\mathbf{H}_i^{(k+1)} = \mathbf{H}_i^{(k)} \quad (3.7)$$

Denoting \mathbf{H}_c^* and \mathbf{H}_i^* the equilibrium values of $\mathbf{H}_c^{(k)}$ and $\mathbf{H}_i^{(k)}$, respectively, equations 3.5 and 3.7 lead to:

$$\begin{cases} \mathbf{H}_c^* = \mathbf{A}_c^P (\mathbf{H}_i^* - \widehat{\mathbf{H}}_c) + \widehat{\mathbf{H}}_c \\ \mathbf{H}_i^* = \mathbf{A}_i^P (\mathbf{H}_c^* - \widehat{\mathbf{H}}_i) + \widehat{\mathbf{H}}_i \end{cases} \quad (3.8)$$

By solving equation 3.8 we have:

$$\begin{cases} \mathbf{H}_c^* = \widehat{\mathbf{H}}_i + (\mathbf{I} - \mathbf{A}_c^P \mathbf{A}_i^P)^{-1} (\mathbf{I} - \mathbf{A}_c^P) (\widehat{\mathbf{H}}_c - \widehat{\mathbf{H}}_i) \\ \mathbf{H}_i^* = \widehat{\mathbf{H}}_c + (\mathbf{I} - \mathbf{A}_i^P \mathbf{A}_c^P)^{-1} (\mathbf{I} - \mathbf{A}_i^P) (\widehat{\mathbf{H}}_i - \widehat{\mathbf{H}}_c) \end{cases} \quad (3.9)$$

From equation 3.9 it can be seen that the equilibrium trajectory depends only on the parameters of the model (M, θ, n, N and P), and are independent from initial value of genetic diversities.

The duration of the transient dynamics

In this section, we determine the duration of the transient dynamics (relaxation time or time to equilibrium values) for the within- and between-population genetic diversities during isolation and connection period. The relaxation times can be derived from the eigenvalues of the transition matrix \mathbf{A}_c and \mathbf{A}_i considering the dynamics of convergence to equilibrium values. Denoting δ a small number ($0 < \delta < 1$) that characterizes

the convergence to equilibrium value, we can obtain the relaxation time, τ^δ , as the time until $\lambda^{\tau^\delta} = \delta$, where λ is the eigenvalue determining the dynamics of genetic diversity. This yields that $\tau^\delta = \frac{\ln(\delta)}{\ln(\lambda)}$. In our case, the relaxation times of h_s and h_b during a connection and an isolation period are thus determined by the two eigenvalues of matrix \mathbf{A}_c , τ_{1c} and τ_{2c} , and the two eigenvalues of matrix \mathbf{A}_i , τ_{1i} and τ_{2i} (from highest to lowest). From eqs. 3.3 and 3.4, we have:

$$\left\{ \begin{array}{l} \tau_{1c}^\delta = \frac{\ln(\delta)}{\ln\left(\frac{(1-\mu)^2}{2} [1 + a(1-c) - b + \sqrt{(1-a(1-c)+b)^2 - 4bc}]\right)} \\ \tau_{2c}^\delta = \frac{\ln(\delta)}{\ln\left(\frac{(1-\mu)^2}{2} [1 + a(1-c) - b - \sqrt{(1-a(1-c)+b)^2 - 4bc}]\right)} \\ \tau_{1i}^\delta = \frac{\ln(\delta)}{\ln((1-\mu)^2)} \\ \tau_{2i}^\delta = \frac{\ln(\delta)}{\ln((1-\mu)^2(1-c))} \end{array} \right. \quad (3.10)$$

Where relaxation time τ_{1c}^δ corresponds to the time for both h_s and h_b to reach their respective equilibrium value during the connection periods; τ_{1i}^δ and τ_{2i}^δ correspond to the time to reach the equilibrium value of h_b and h_s , respectively, during the isolation periods. Assuming small migration and mutation rates (i.e., $m \ll 1$ and $\mu \ll 1$) and large local population sizes (i.e., $N \gg 1$), and as $\ln(1-X) = -X + o(X)$, eqs. 3.10 simplify to:

$$\left\{ \begin{array}{l} \tau_{1c}^\delta = \frac{-\ln(\delta)}{2\mu + 1/2N_e} \\ \tau_{2c}^\delta = \frac{-\ln(\delta)}{2\mu + 2m\frac{n}{n-1} + 1/2N - 1/2N_e} \\ \tau_{1i}^\delta = \frac{-\ln(\delta)}{2\mu} \\ \tau_{2i}^\delta = \frac{-\ln(\delta)}{2\mu + 1/2N} \end{array} \right. \quad (3.11)$$

The expressions of τ_{1c}^δ , τ_{1i}^δ and τ_{2i}^δ in eq. 3.10 have also a direct interpretation in the coalescence framework (the "coalescent with killings"; DURRETT 2002). The de-

nominator in each expression corresponds to the log probability that lineages did not mutate or coalesce within a generation. For example, the denominator in the expression of τ_{1c}^δ in eq. 3.10 is the log probability that neither a coalescent event (probability $1-1/2N_e$, where $N_e=nN(1+(n-1)/4nNm)$) nor a mutation (probability $(1-\mu)^2$) occurred within a generation when populations are connected (log probability $-\log((1-1/2N_e)(1-\mu)^2) \simeq 2\mu+1/2N_e$). Similarly, the denominator in the expression of τ_{1i} is the log probability that no mutation occurred in isolated population ($-\log((1-\mu)^2) \simeq 2\mu$).

Determination of domains of the period length

In this section, we determine domains of the period P (number of generations) of isolation and connection in which genetic diversities remain or not in a transient state. These domains depend on the relationship between P and the relaxation times, τ_{1c}^δ , τ_{1i}^δ , τ_{2c}^δ and τ_{2i}^δ defined in eq. 3.11.

First, there is the trivial situation where the periods P between connection changes are very large, then both within- and between-population genetic diversities reach successively their equilibrium values within a timeframe shorter than P (cf. Figure 3.2(A)). Therefore, after each event, H_c^* and H_i^* tend to their respective equilibrium, \widehat{H}_c and \widehat{H}_i .

From equation 3.5, this implies that the highest relaxation time from eq. 3.11 is smaller than P . As τ_{1i} is the highest relaxation time for all parameter values, the minimum period duration (in number of generations), P_I , for which genetic diversities reach their equilibrium during isolation and connection periods is:

$$P_I = -\ln(\delta) \frac{1}{2\mu} \quad (3.12)$$

The second situation is when P is smaller than P_I , so H_i^* does not reach the isolation

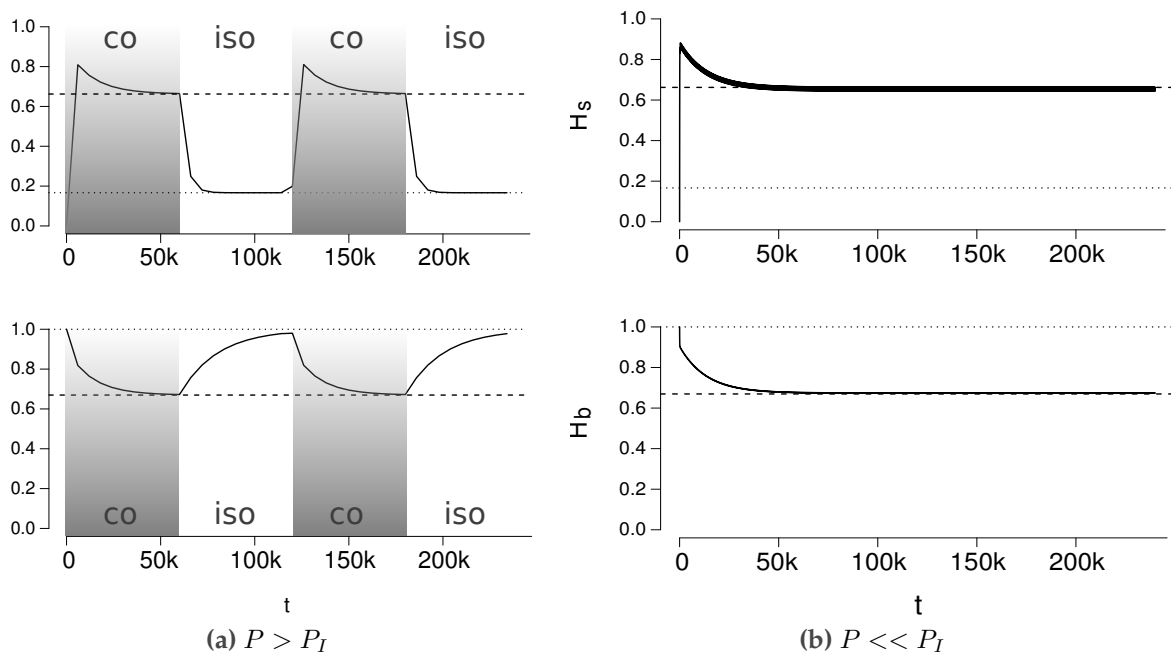


Figure 3.2 Trajectories of within- (H_s) and between-population (H_b) genetic diversities under periodic isolation and connection events (a) in the short-period domain $P < P_W$ and (b) in the long-period domain $P > P_I$. The dashed and dotted lines represent the expected equilibrium value when populations are connected and isolated, respectively. In (a) both H_s and H_b tend to their equilibrium value of connection (dashed line) with very small fluctuations. In (b) both H_s and H_b reach their expected equilibrium value at the end of each connection period and isolation period. Parameters are $M = 40$, $n = 10$, $N = 2,000$, $\mu = 2.5 \times 10^{-5}$. (a) $P = 60,000$, (b) $P = 150$.

equilibrium value within P , and \mathbf{H}_c^* tends to its equilibrium value within P . From equation 3.5, this implies that the relaxation time during connection, τ_{1c} (from eq. 3.11), is smaller than P , thus we can determine a period duration P_C such that:

$$P_C = -\ln(\delta) \frac{1}{2\mu + 1/2N_e} \quad (3.13)$$

When the length of the period P is smaller than P_I but larger than P_C , genetic diversities at the beginning of the connection period has a transient value corresponding to the one observed at the end of the isolation period.

In the third situation, the relaxation times during connection and isolation are all larger than P . The trajectory of genetic diversities to their equilibrium values presents strong fluctuations (Figure 3.3), caused by the succession of peaks of genetic diversity generated by connection events. The trajectories of the peaks alone are monotonous toward their equilibrium values (eqs. 3.9; Appendix A). However, their direction, increasing or decreasing, depends on the initial value of between-population genetic diversity $h_b(0)$. When $h_b(0)$ is below its equilibrium value, successive connection events generate peaks of genetic diversity of increasing value (Figure 3.3(A)). When $h_b(0)$ is above its equilibrium value, successive connection events generate peaks of genetic diversity with decreasing value (Figure 3.3(B)).

Finally we also consider the situation where periods of isolation P are so short that they do not impact values of genetic diversity (no differentiation during isolation). In this situation, the value of genetic diversities at the end of the isolation period is close to the one at the end of the connection period of a same cycle, k , and we have $\mathbf{H}_c^{(k)} \simeq \mathbf{H}_i^{(k)}$. Considering a small value δ' that represents the absolute difference between these two

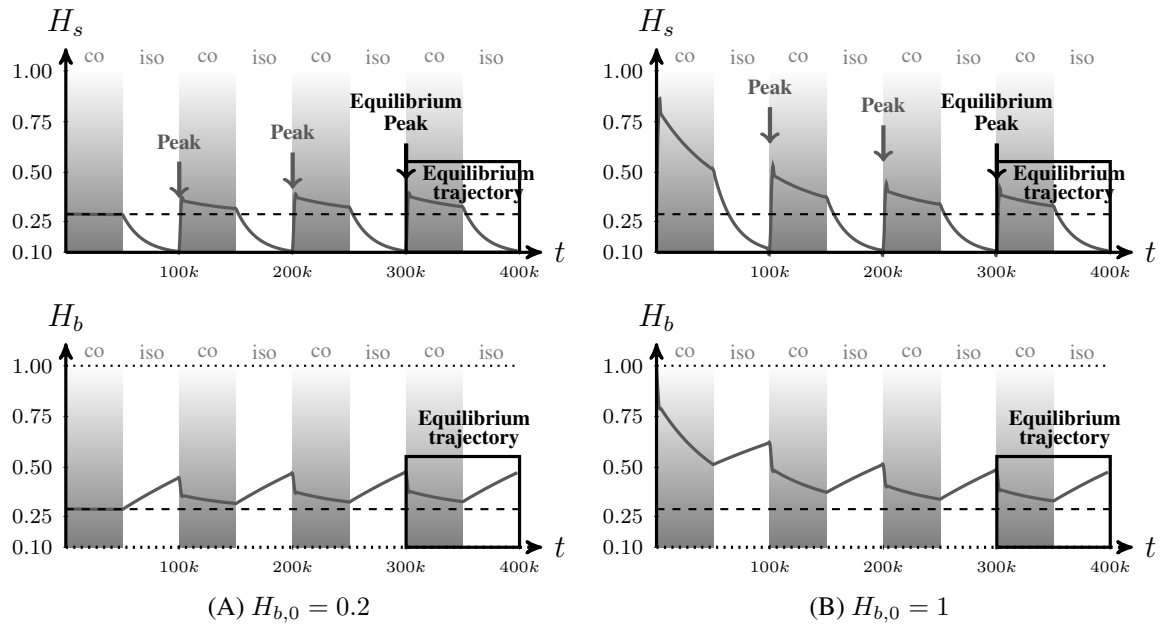


Figure 3.3 Trajectories of within- (H_s) and between-population (H_b) genetic diversities under periodic isolation and connection events in the intermediate-period domain $P_W < P < P_I$ when initial genetic diversity $H_b(0)$ is (a) low and (b) high (i.e. below or above the connection equilibrium). In (A) when $H_b(0)$ is low the successive peaks of within-population genetic diversity (indicated by gray arrows) increase in size, while in (B) they decrease in size. In both (A) and (B), the equilibrium trajectory of genetic diversity (black solid line framed by a rectangle), reached after several events of isolation and connection, is the same. This equilibrium trajectory does not depend on the initial conditions, but on the parameters of the model. In this simulation, parameters are $M = 400$, $\theta = 0.1$, $n = 4$, $N = 10,000$, $P = 50,000$ generations.

values, we have:

$$\begin{cases} |H_{c,s}^{(k)} - H_{i,s}^{(k)}| \leq 2\delta' \\ |H_{c,b}^{(k)} - H_{i,b}^{(k)}| \leq 2\delta' \end{cases} \quad (3.14)$$

From equation 3.5, this implies that the highest relaxation time of the isolation period is lower than P . As τ_{2i} is the highest relaxation time, $P_W = \tau_{2i}^{1-\delta_W}$ is the maximum number of generations for which genetic diversities are not impacted by successive events of isolation and connection:

$$P_W = -\ln(1 - \delta') \frac{1}{1/2N + 2\mu} \quad (3.15)$$

The dynamics of genetic diversities when the period is smaller than P_W are illus-

trated in Figure 3.2(b): the first event of connection generates a large peak of genetic diversity, due to the quick distribution among populations of mutation accumulated during isolation. Following this major change in genetic diversity, the successive isolation and connection events generate only very small fluctuations of the genetic diversity values in their trajectories to their equilibrium. The time to reach the expected equilibrium values is close to what is expected after a single connection event (See Appendix A for demonstration in the case where $m=(n - 1)/n$).

In summary, we have found four domains of the period P of isolation and connection events, where periodic isolation and connection events have a different impact on the trajectories of genetic diversity; these domains depend on the mutation rate μ and the population size N (Figure 3.4).

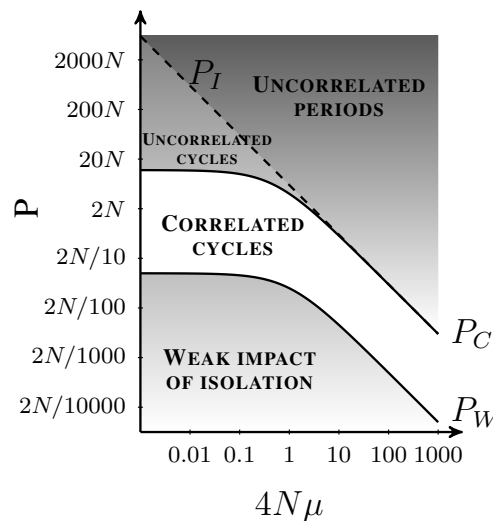


Figure 3.4 Domains of period length where the trajectories of genetic diversity have a different qualitative behavior, as a function of the mutation rate μ scaled by the population size N , $4N\mu$. $N_e = nN(1+(n-1)/4nNm)$ corresponds to the effective population size during the connection period, taking into account the migration rate between populations, m .

Those results are robust to stochasticity in the per generation probability of event occurrence and to stochasticity in the duration of the isolation and connection periods as long as the variance in the duration is smaller than a determined threshold (see

Appendix B).

The dynamics of the excess of within-population genetic diversity in the different domains

We now describe the dynamics of the excess of within-population genetic diversities under periodic connection and isolation events, relative to their expected equilibrium values (\hat{H}_i and \hat{H}_c), as a function of the period length P and the model parameters.

We focus on two measures that are representative of two important processes: genetic diversity turnover and accumulation. Genetic diversity turnover is measured as the amplitude of the variations of the excess of within-population genetic diversity within an isolation and reconnection cycle (i.e., the difference between maximum and minimum value of h_s); this quantity provides information on whether excess of h_s is fluctuating or stationary. Genetic diversity accumulation is measured as the cumulative genetic diversity excess of h_s during a given number of generations. These values are computed numerically from eqs. 3.2 and 3.5 and represented in Figure 3.5.

The highest rate of genetic diversity turnover induced by isolation and connection events is observed when the durations of the isolation and connection period are large $P > 2N$ (figure Figure 3.5 (a)) and when the mutation rate is moderate ($4N\mu < 10$). When period are large, multiple peak of excess of genetic diversity are observed allowing for a high turnover of the excess of genetic diversity. When mutation rate is high, the presence of isolation and reconnection events does not impact the value of genetic diversity in population as the genetic diversity is at its maximum value (close to 1), populations are saturated by mutations (as observed observed when $4N\mu > \theta_{sat} = 9$ in figure Figure 3.5 (a)-(b)).

The highest amount of cumulated genetic diversity is observed when isolation and

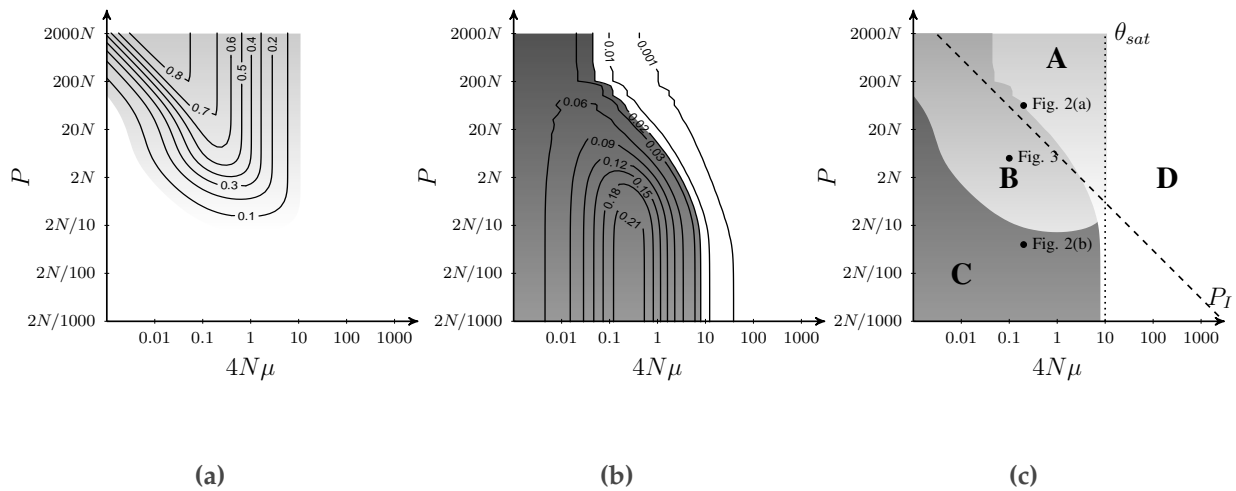


Figure 3.5 Domains of the period P of connection and isolation events, and the scaled mutation rate $4N\mu$, where within-population genetic diversity h_s follows specific dynamics. (a) Amplitude of the variations of within-population genetic diversity h_s within an equilibrium cycle. (b) Mean cumulative excess of h_s compared to its expected equilibrium value; initial genetic diversities correspond to the expected equilibrium values in isolated populations; the mean is computed over 200,000 generations. (c) Schematic representation of the domains presented in (a) and (b). In domain A, there are strong variations of h_s within cycles, but a low excess of genetic diversity is maintained across cycles. In domain B, h_s undergoes small variations within cycles and genetic diversity can accumulate or decrease across cycles, depending on the initial value of h_s . In domain C, genetic diversity is maintained across cycles, and there are almost no variations within cycles. In domain D, h_s saturates and all h_s values are larger than 0.9 during the cycles, regardless of P and $4N\mu$ values. Parameters are $n = 10$ populations, $m = 0.01$ ($4Nm \gg 1$), $2N = 1000$ for each population.

reconnection periods are smaller than $2N$. In this case, relaxation times are larger than the periods. Consequently, genetic diversities remain in a transient state and can cumulate thought time. As the dynamics of h_s and h_b are constrained by the value of $4N\mu$, for all values of $4N\mu$ larger than a threshold $\theta_{sat} = 9$, we have a saturation of h_s ($0.9 \leq h_s \leq h_b$) from eqs. 3.2 and 3.5. In that case almost no variations within or across cycles of isolation and reconnection can occur.

From the estimation of the value of genetic turnover and accumulation, we can define four domains in which genetic diversities have a specific behaviour in regard to turnover and accumulation (Figure 3.5(b)). In domain A (**A** domain Figure 3.5(c)) genetic diversity turnover is observed and no accumulation of genetic diversity occurs. Domain A is defined by $P > P_I$ (**A** domain Figure 3.5(c)) and $4N\mu < \theta_{sat}$. In domain B (**B** domain Figure 3.5(c)) both accumulation and turnover of genetic diversity is observed. Domain B is defined by approximately $P_I/60 < P < P_I$, and $4N\mu < \theta_{sat}$. In domain C (**C** domain fig 3.5(c)), genetic diversity cumulates across cycles and there is almost no variation of genetic diversities within-cycles. Domain C is defined by $P \ll P_I$ and $4N\mu < \theta_{sat}$. Finally, in domain D (domain **D** fig 3.5(c)) genetic diversity is saturated, $4N\mu = \theta_{sat}$, any demographic change will not impact its value.

DISCUSSION

Fluctuations of migration were known to slightly impact the value of genetic diversity (NAGYLAKI 1979; KARLIN 1982; WHITLOCK 1992; SHPAK *et al.* 2010). Our results demonstrates that such results also apply in the presence of periods of isolation and reconnection smaller than P_W (equation 4.9) but only following the first connection event. In this domain, our results are in agreement with NAGYLAKI (1979), LATTER and SVED (1981), KARLIN (1982), WHITLOCK (1992) and SHPAK *et al.* (2010): migration fluctuations of small periods generate small fluctuations of genetic diversity within

populations and slightly increase the mean genetic differentiation and have almost no effect on between-population genetic diversity (Figure 3.3(A)). However, while under stochastic migration studied in NAGYLAKI (1979) and WHITLOCK (1992) the effective migration rate have been shown to depend and decreases with the variance of migration rate, under periodic isolation and connection events (with $P < P_W$ and after the first connection events) the effective migration rate is well described by the highest value of migration rate.

When P is larger than P_I and $4N\mu < \theta_{sat}$ (domain A in Figure 3.5 (c)), we observe as expected from ALCALA *et al.* (2013), successive peaks of genetic diversity that do not overlap. Genetic diversities reach their expected equilibrium values before the occurrence of a new event. Interestingly, it is in this domain that genetic diversities undergo the largest variations allowing for a strong turnover of genetic diversity. The lower limits of this domains is very large, for example, considering a mutation rate of 10^{-5} , the limit between the intermediate and large-period domains is $P_I \simeq 150,000$ generations.

When P is larger than P_W but smaller than P_I and $4N\mu < \theta_{sat}$, the dynamics of genetics diversity within each event of isolation and connection interfere (domain B in Figure 3.5 (c)). The behaviour of the genetic diversities depends on their initial value. Genetic diversity is cumulated by successive events of isolation and connection when the genetic diversity is initially low, and is reduced when the genetic diversity is initially high. The time-scale of this domain is large, for example, considering a mutation rate of 10^{-5} and a population of size 5,000 diploid individuals, the lower limit of this domains is approximately $P_W \simeq 430$ generations.

Interestingly, the time-frame of the domains A and B correspond for many species to the time-frame of the well described major environmental events that isolated and

reconnected populations in the past and that have been suggested to play a major role in speciation events. For example the time-frame of drought events in Africa (SZABO *et al.* 1995; DRAKE and BRISTOW 2006; MASLIN and CHRISTENSEN 2007), similar to that described in SEEHAUSEN (2002) and ELMER *et al.* (2009), is often associated with rapid species diversification (for example for the cichlids; ARNEGARD *et al.* 1999; OWEN *et al.* 1990; SEEHAUSEN 2004; SEDANO and BURNS 2010; YOUNG *et al.* 2009; NEVADO *et al.* 2013). Our results predict that populations might have unexpected large genetic diversity in such fluctuating environments. Such large genetic diversity might then have favored adaptation from standing genetic variation (HERMISSON and PENNINGS 2005; BARRETT and SCHLUTER 2008). However, we also show that successive isolation and connection periods can also lead to a decrease in genetic diversity when the initial level of genetic diversity is high. Those two possible outcomes could explain why populations which underwent similar isolation and connection event in their history can display different genetic signatures such as found for the meadow grasshopper (0.7-0.9% sequence divergence) and for the hedgehog populations (6-12% divergence) in Europe (HEWITT 2000).

Our result also have implications for the understanding of genome polymorphism. Several independent loci might display incongruous signals even when they experienced the same history of isolations. Indeed, polymorphism can be strongly heterogeneous along the genome (e.g. the human genome; ABECASIS *et al.* 2010), thus when periodic migration changes occur on, for example two unlinked loci *A* and *B* with high and low initial level of genetic diversity, respectively, the genetic diversity will decrease at the *A* locus but increase at *B*. Heterogeneous mutation rates throughout the genome would further increase this effect. This is of particular interest for the inference of past population history (demography and selection) from multiple loci (e.g.

using approximate bayesian methods, BEAUMONT *et al.* 2002, or methods using SNP data, GUTENKUNST *et al.* 2009), when inconsistent signals are detected throughout the genome.

The study of the demographic changes on genetic diversity in structured populations could greatly benefit from the theoretical ecology literature where periodic, aperiodic, deterministic or stochastic perturbation on demography have been widely studied (COHEN 1976, 1979; TULJAPURKAR 1982, 1989). Although such developments consider short-period fluctuations (e.g. seasonal fluctuations), the work of JESUS *et al.* (2006) and our own study show that similar methods (e.g. periodic matrix products) could be used to model fluctuations in demographic parameters at longer time-scales. Interestingly, such models can introduce temporal autocorrelation between demographic parameters (such as in our study), which was not investigated in previous studies on stochastic migration (NAGYLAKI 1979; LATTER and SVED 1981). An interesting extension of our work could be to consider gradual transition from isolation to connection, for example using the framework of autoregressive-moving-average (ARMA) models (BOX and JENKINS 1970, e.g. used in ecology in POOLE 1978; TULJAPURKAR 1989) or the techniques of parameter estimation from time-series used in ecology (IVES *et al.* 2010) when repeated (periodic) demographic fluctuations occur (WAHL *et al.* 2002). Also, accounting for stochasticity in the event of isolation and connection does not change the qualitative behavior of genetic diversity. Results presented here could then ease analytical tractability of the dynamic of genetic diversity when complex demographic scenarios are considered (e.g. LAVAL and EXCOFFIER 2004; NEUENSCHWANDER *et al.* 2008; RAY *et al.* 2010) or provide support for the set assumptions on the parameter values.

Short-time scale environmental changes have benefit from a large amount of theo-

retical investigations. Although there is increasing evidence of the importance of large-scale environmental changes for the generation and maintenance of genetic diversity, they received only recently attention (JESUS *et al.* 2006; LEFFLER *et al.* 2012; AGUILÉE *et al.* 2013; ALCALA *et al.* 2013). Consequently, we have a poor understanding of processes that occur in dynamic environments. The study of the impact of large-scale variability of the environment might also contribute to a better understanding of the high level of genetic diversity observed in domesticated organisms despite strong selection (e.g. in rice, CAICEDO *et al.* 2007; ZHAO *et al.* 2010, and cats, LECIS *et al.* 2006; LIPINSKI *et al.* 2008), successive admixture of long-term isolated populations might have strongly promoted the maintenance of genetic diversity. Similarly, pathogens extraordinary genetic diversity (LAWRENCE 2005; HEMELAAR *et al.* 2011), could also be a results of successive secondary contacts occurring after large periods of isolation within hosts (RENZETTE *et al.* 2013), between hosts (CONWAY *et al.* 1999) or populations (BURKE 1997; WARD *et al.* 2013; ALCALA *et al.* In review). How this diversity is related to their past demography, their structure or the environment they encountered remain to be explored in more details. Moreover, similar patterns are expected to occur in the future. Indeed, with nowadays increase ease to travel worldwide, organisms (invasive species, local adapted population, virus, and other pathogens) have increased opportunities to encounter previously isolated related populations, dynamic of genetic diversity within those events might then provide important information to predict species evolutionary potential.

ACKNOWLEDGEMENTS

We thank Thomas Lenormand, Laurent Lehmann and two reviewers for their improving comments. This project was funded by the Swiss National Science Foundation (SNSF) grants #PZ00P3_139421/1 and #31003A – 130065 and the University of Lau-

sanne.

BIBLIOGRAPHY

- ABECASIS, G., D. ALTSHULER, A. AUTON, L. BROOKS, R. DURBIN, *et al.*, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
- AGUILEE, R., A. LAMBERT, and D. CLAESSEN, 2011 Ecological speciation in dynamic landscapes. *J Evol Biol* **24**: 2663–2677.
- AGUILÉE, R., D. CLAESSEN, and A. LAMBERT, 2013 Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics. *Evolution* **67**: 1291–1306.
- ALCALA, N., J. D. JENSEN, and S. VUILLEUMIER, In review The signature of past isolation on gene genealogies and the site frequency spectrum. *Genetics* .
- ALCALA, N., D. STREIT, J. GOUDET, and S. VUILLEUMIER, 2013 Peak and persistent excess of genetic diversity following an abrupt migration increase. *Genetics* **193**: 953–971.
- ARNEGARD, M. E., J. A. MARKERT, P. D. DANLEY, J. R. STAUFFER, A. J. AMBALI, *et al.*, 1999 Population structure and colour variation of the cichlid fishes *labeotropheus fuelleborni* ahl along a recently formed archipelago of rocky habitat patches in southern lake malawi. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**: 119–130.
- BARRETT, R. D. H., and D. SCHLUTER, 2008 Adaptation from standing genetic variation. *Trends in Ecology & Evolution* **23**: 38–44.
- BARRIER, M., B. G. BALDWIN, R. H. ROBICHAUX, and M. D. PURUGGANAN, 1999 Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the

- hawaiian silversword alliance (asteraceae) inferred from floral homeotic gene duplications. *Mol Biol Evol* **16**: 1105–1113.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BEZAULT, E., S. MWAIKO, and O. SEEHAUSEN, 2011 Population genomic tests of models of adaptive radiation in lake victoria region cichlid fish. *Evolution* **65**: 3381–3397.
- BOX, G. E., and G. M. JENKINS, 1970 Time series analysis: Forecasting and control. Holden-D. iv, San Francisco .
- BURKE, D. S., 1997 Recombination in hiv: an important viral evolutionary strategy. *Emerging infectious diseases* **3**: 253.
- CAICEDO, A. L., S. H. WILLIAMSON, R. D. HERNANDEZ, A. BOYKO, A. FLEDELALON, *et al.*, 2007 Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS genetics* **3**: e163.
- COHEN, J. E., 1976 Ergodicity of age structure in populations with markovian vital rates, i: countable states. *Journal of the American Statistical Association* **71**: 335–339.
- COHEN, J. E., 1979 Comparative statics and stochastic dynamics of age-structured populations. *Theoretical population biology* **16**: 159–171.
- COLOSIMO, P. F., K. E. HOSEMAN, S. BALABHADRA, G. VILLARREAL, M. DICKSON, *et al.*, 2005 Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. *Science* **307**: 1928–1933.
- CONWAY, D. J., C. ROPER, A. M. ODUOLA, D. E. ARNOT, P. G. KREMSNER, *et al.*, 1999 High recombination rate in natural populations of plasmodium falciparum. *Proceedings of the National Academy of Sciences* **96**: 4506–4511.

DRAKE, N., and C. BRISTOW, 2006 Shorelines in the sahara: geomorphological evidence for an enhanced monsoon from palaeolake megachad. *The Holocene* **16**: 901–911.

DURRETT, R., 2002 *Probability models for DNA sequence evolution*. Springer.

ELMER, K. R., C. REGGIO, T. WIRTH, E. VERHEYEN, W. SALZBURGER, *et al.*, 2009 Pleistocene desiccation in east africa bottlenecked but did not extirpate the adaptive radiation of lake victoria haplochromine cichlid fishes. *Proc Natl Acad Sci U S A* **106**: 13404–13409.

FEDER, J. L., S. H. BERLOCHER, J. B. ROETHELE, H. DAMBROSKI, J. J. SMITH, *et al.*, 2003 Allopatric genetic origins for sympatric host-plant shifts and race formation in *rhagoletis*. *Proc Natl Acad Sci U S A* **100**: 10314–10319.

FISHER, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press.

GAVRILETS, S., 2003 Perspective: models of speciation: what have we learned in 40 years? *Evolution* **57**: 2197–2215.

GAVRILETS, S., and J. B. LOSOS, 2009 Adaptive radiation: contrasting theory with data. *Science* **323**: 732–737.

GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON, and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics* **5**: e1000695.

HARMON, L. J., J. A. SCHULTE, A. LARSON, and J. B. LOSOS, 2003 Tempo and mode of evolutionary radiation in iguanian lizards. *Science* **301**: 961–964.

HEMELAAR, J., E. GOUWS, P. D. GHYS, S. OSMANOV, *et al.*, 2011 Global trends in molecular epidemiology of hiv-1 during 2000–2007. *Aids* **25**: 679–689.

- HERMISSON, J., and P. S. PENNINGS, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**: 2335–2352.
- HERNANDEZ, R. D., J. L. KELLEY, E. ELYASHIV, S. C. MELTON, A. AUTON, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science* **331**: 920–924.
- HEWITT, G., 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–913.
- HEWITT, G. M., 2004 Genetic consequences of climatic oscillations in the quaternary. *Philos Trans R Soc Lond B Biol Sci* **359**: 183–195.
- HEWITT, G. M., 2011 Quaternary phylogeography: the roots of hybrid zones. *Genetica* **139**: 617–638.
- IVES, A. R., K. C. ABBOTT, and N. L. ZIEBARTH, 2010 Analysis of ecological time series with arma (p, q) models. *Ecology* **91**: 858–871.
- JESUS, F., J. WILKINS, V. SOLFERINI, and J. WAKELEY, 2006 Expected coalescence times and segregating sites in a model of glacial cycles. *Genet. Mol. Res* **5**: 466–474.
- JONES, F. C., M. G. GRABHERR, Y. F. CHAN, P. RUSSELL, E. MAUCELI, *et al.*, 2012 The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**: 55–61.
- KARLIN, S., 1982 Classifications of selection migration structures and conditions for a protected polymorphism. *Evolutionary Biology* **14**: 61–204.
- KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
- LATTER, B., and J. SVED, 1981 Migration and mutation in stochastic models of gene frequency change. ii. stochastic migration with a finite number of islands. *Journal of Mathematical Biology* **13**: 95–104.

- LAVAL, G., and L. EXCOFFIER, 2004 Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20**: 2485–2487.
- LAWRENCE, J. G., 2005 Common themes in the genome strategies of pathogens. *Current opinion in genetics & development* **15**: 584–588.
- LECIS, R., M. PIERPAOLI, Z. BIRO, L. SZEMETHY, B. RAGNI, *et al.*, 2006 Bayesian analyses of admixture in wild and domestic cats (*felis silvestris*) using linked microsatellite loci. *Molecular Ecology* **15**: 119–131.
- LEFFLER, E. M., K. BULLAUGHEY, D. R. MATUTE, W. K. MEYER, L. SEGUREL, *et al.*, 2012 Revisiting an old riddle: What determines genetic diversity levels within species? *PLoS biology* **10**: e1001388.
- LIPINSKI, M. J., L. FROENICKE, K. C. BAYSAC, N. C. BILLINGS, C. M. LEUTENEGGER, *et al.*, 2008 The ascent of cat breeds: genetic evaluations of breeds and worldwide random-bred populations. *Genomics* **91**: 12–21.
- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor Popul Biol* **1**: 273–306.
- MASLIN, M. A., and B. CHRISTENSEN, 2007 Tectonics, orbital forcing, global climate change, and human evolution in africa: introduction to the african paleoclimate special volume. *Journal of Human Evolution* **53**: 443–464.
- MYLES, S., N. BOUZEKRI, E. HAVERFIELD, M. CHERKAOUI, J.-M. DUGOUJON, *et al.*, 2005 Genetic evidence in support of a shared eurasian-north african dairying origin. *Hum Genet* **117**: 34–42.
- NAGYLAKI, T., 1979 The island model with stochastic migration. *Genetics* **91**: 163–176.

- NEI, M., and M. W. FELDMAN, 1972 Identity of genes by descent within and between populations under mutation and migration pressures. *Theor Popul Biol* **3**: 460–465.
- NEUENSCHWANDER, S., F. GUILLAUME, J. GOUDET, *et al.*, 2008 quantinemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation. *Bioinformatics* **24**: 1552–1553.
- NEVADO, B., S. MAUTNER, C. STURMBAUER, and E. VERHEYEN, 2013 Water-level fluctuations and metapopulation dynamics as drivers of genetic diversity in populations of three tanganyikan cichlid fish species. *Mol Ecol* **22**: 3933–3948.
- OWEN, R. B., R. CROSSLEY, T. C. JOHNSON, D. TWEDDLE, I. KORNFIELD, *et al.*, 1990 Major low levels of lake malawi and their implications for speciation rates in cichlid fishes. *Proceedings of the Royal Society of London. B. Biological Sciences* **240**: 519–553.
- PELZ, H.-J., S. ROST, M. HÜNERBERG, A. FREGIN, A.-C. HEIBERG, *et al.*, 2005 The genetic basis of resistance to anticoagulants in rodents. *Genetics* **170**: 1839–1847.
- POOLE, R. W., 1978 The statistical prediction of population fluctuations. *Annual Review of Ecology and Systematics* **9**: 427–448.
- RAY, N., M. CURRAT, M. FOLL, and L. EXCOFFIER, 2010 Splat2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. *Bioinformatics* **26**: 2993–2994.
- RENZETTE, N., L. GIBSON, B. BHATTACHARJEE, D. FISHER, M. R. SCHLEISS, *et al.*, 2013 Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS genetics* **9**: e1003735.

- RICE, S. H., and A. PAPADOPOULOS, 2009 Evolution with stochastic fitness and stochastic migration. *PLoS One* **4**: e7130.
- SEDANO, R. E., and K. J. BURNS, 2010 Are the northern andes a species pump for neotropical birds? phylogenetics and biogeography of a clade of neotropical tanagers (aves: Thraupini). *Journal of Biogeography* **37**: 325–343.
- SEEHAUSEN, O., 2002 Patterns in fish radiation are compatible with pleistocene desiccation of lake victoria and 14,600 year history for its cichlid species flock. *Proc Biol Sci* **269**: 491–497.
- SEEHAUSEN, O., 2004 Hybridization and adaptive radiation. *Trends Ecol Evol* **19**: 198–207.
- SHPAK, M., J. WAKELEY, D. GARRIGAN, and R. C. LEWONTIN, 2010 A structured coalescent process for seasonally fluctuating populations. *Evolution* **64**: 1395–1409.
- SZABO, B., C. HAYNES JR, and T. A. MAXWELL, 1995 Ages of quaternary pluvial episodes determined by uranium-series and radiocarbon dating of lacustrine deposits of eastern sahara. *Palaeogeography, Palaeoclimatology, Palaeoecology* **113**: 227–242.
- TULJAPURKAR, S., 1982 Population dynamics in variable environments. ii. correlated environments, sensitivity analysis and dynamics. *Theoretical Population Biology* **21**: 114–140.
- TULJAPURKAR, S., 1989 An uncertain life: demography in random environments. *Theoretical population biology* **35**: 227–294.
- TURNER, R. C., J. C. LEVY, and A. CLARK, 1993 Complex genetics of type 2 diabetes: thrifty genes and previously neutral polymorphisms. *Q J Med* **86**: 413–417.

- WAHL, L. M., P. J. GERRISH, and I. SAIKA-VOIVOD, 2002 Evaluating the impact of population bottlenecks in experimental evolution. *Genetics* **162**: 961–971.
- WARD, M. J., S. J. LYCETT, M. L. KALISH, A. RAMBAUT, and A. J. L. BROWN, 2013 Estimating the rate of intersubtype recombination in early hiv-1 group m strains. *Journal of virology* **87**: 1967–1973.
- WHITLOCK, M. C., 1992 Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution* **46**: pp. 608–615.
- WRIGHT, S., 1931 Evolution in mendelian populations. *Genetics* **16**: 97–159.
- YOUNG, K. A., J. M. WHITMAN, and G. F. TURNER, 2009 Secondary contact during adaptive radiation: a community matrix for lake malawi cichlids. *J Evol Biol* **22**: 882–889.
- ZHANG, H., J. YAN, G. ZHANG, and K. ZHOU, 2008 Phylogeography and demographic history of chinese black-spotted frog populations (*pelophylax nigromaculata*): Evidence for independent refugia expansion and secondary contact. *BMC Evolutionary Biology* **8**: 21.
- ZHAO, K., M. WRIGHT, J. KIMBALL, G. EIZENGA, A. MCCLUNG, *et al.*, 2010 Genomic diversity and introgression in *o. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS One* **5**: e10780.

APPENDIX A | THE DYNAMICS OF GENETIC DIVERSITIES ACROSS CYCLES UNDER THE PANMICTIC CONNECTION PERIODS APPROXIMATION

In this section, we derive simplified dynamics of the genetic diversities under periodic connection and isolation events in the case where populations are panmictic during the connection periods (i.e., $m = \frac{n-1}{n}$).

In this case, matrix \mathbf{A}_c simplifies to:

$$\mathbf{A}_c = (1 - \mu)^2 \begin{pmatrix} \frac{1}{n}(1 - c) & \frac{n-1}{n} \\ \frac{1}{n}(1 - c) & \frac{n-1}{n} \end{pmatrix} \quad (3-A.1)$$

and its eigenvalues simplify to:

$$\begin{aligned} \lambda_1 &= (1 - \mu)^2(1 - c') \\ \lambda_2 &= 0 \end{aligned} \quad (3-A.2)$$

where $c' = 1/2nN$ is the rate of genetic drift in a panmictic population of size nN .

Consequently,

$$\mathbf{\Gamma}_c = (1 - \mu)^{4P}(1 - c')^{P-1} \begin{pmatrix} \frac{1}{n}(1 - c)^{P+1} & \frac{n-1}{n} \\ \frac{1}{n}(1 - c)^{P+1} & \frac{n-1}{n} \end{pmatrix} \quad (3-A.3)$$

with a first eigenvalue $\lambda_c = (1 - \mu)^{4P}(1 - c')^{P-1}(\frac{1}{n}(1 - c)^{P+1} + \frac{n-1}{n})$, and a second eigenvalue which is null.

So we have, for $k \geq 1$:

$$h_{c,s}^{(k)} = h_{c,b}^{(k)} = \lambda_c^k \left(\frac{\frac{1}{n}(1 - c)^{P+1}h_s^{(0)} + \frac{n-1}{n}h_b^{(0)}}{\frac{1}{n}(1 - c)^{P+1} + \frac{n-1}{n}} - h_c^* \right) + h_c^* \quad (3-A.4)$$

where h_c^* is the equilibrium value of the cycles both within- and between-population.

Interestingly, when P is small, $\log(\delta)/\log(\lambda_c)$ tends to τ_{1c}^δ , thus when periods are very short, the dynamics of genetic diversities approximately follow that under constant connection. On the contrary, when P is very large, $\log(\delta)/\log(\lambda_c) \simeq 2$; nevertheless, in that domain both values are expected to be lower than the period of the cycle P , thus despite this long time to reach equilibrium, the equilibrium cycles dynamics will be reached after a single cycle.

APPENDIX B | DYNAMICS OF GENETIC DIVERSITY UNDER STOCHASTIC PERIODS OF ISOLATION AND CONNECTION

We showed that values of P , P_W and P_I determine the behavior of genetic diversity under periodic isolation and connection. In this section, we consider that each period P is a random variable. Thus, we study the two following quantities: the probabilities that the isolation period is shorter than P_W , $\mathbb{P}(P < P_W)$, and the probability that it is longer than P_I , $\mathbb{P}(P > P_I)$. Two scenarios are considered. Scenario **A** assumes that the probability of an event of isolation (resp. connection) is the same for each generation (i.e. independent of the generation t) but follows a geometric distribution. Scenario **B** assumes that the length of isolation and connection periods are regular but have a Gaussian noise, generating variance around the mean period P (i.e. dependent of the generation t). Thus, Scenario A considers time-homogeneous stochastic changes and Scenario B considers time-inhomogeneous stochastic changes (following TULJAPURKAR 1989).

Dynamics of genetic diversity under scenario A

Under scenario A, we assume that the probability p to switch from isolation state to connection state at a given generation is independent of the current generation t .

Thus, the sequence of isolation and connection events is modelled as a two states time-homogeneous Markov process. Under such a scenario, the duration of each period (corresponding to the waiting time until state switch), P , follows a geometric distribution of parameter p , p being the probability of the occurrence of the isolation or connection event (so the mean period is $\bar{P} = 1/p$). Thus we have:

$$\begin{cases} \mathbb{P}(P < P_W) = 1 - (1 - p)^{P_W} \\ \mathbb{P}(P > P_I) = (1 - p)^{P_I} \end{cases} \quad (3-B.1)$$

Which yields

$$\begin{cases} \mathbb{P}(P < P_W) > 1 - \epsilon_W \Leftrightarrow p > 1 - \epsilon_W^{1/P_W} \\ \mathbb{P}(P > P_I) > 1 - \epsilon_I \Leftrightarrow p < 1 - (1 - \epsilon_I)^{1/P_I} \end{cases} \quad (3-B.2)$$

Where ϵ_W and ϵ_I correspond to the probability that a random period P is larger than P_W and lower than P_I , respectively. Values of ϵ_W and ϵ_I close to 0 lead to a behavior of genetic diversities that follows what is expected under the short-period and long-period domains, respectively.

Using the expression of P_W and P_I from equations 4.9 and 3.13 and given that $e^X = 1 + X + o(X^2)$, we obtain the following approximation for conditions (3-B.2):

$$\begin{cases} \mathbb{P}(P < P_W) > 1 - \epsilon_W \Leftrightarrow p > (1/2N + 2\mu) \frac{\log(\epsilon_W)}{\log(1 - \alpha)} \\ \mathbb{P}(P > P_I) > 1 - \epsilon_I \Leftrightarrow p < 2\mu \frac{\log(1 - \epsilon_I)}{\log(\alpha)} \end{cases} \quad (3-B.3)$$

Where α is a value that determines the difference between genetic diversity during the isolation period and the expected genetic diversity in equilibrium isolated populations (by default, we use $\alpha = 0.05$).

From equation 3-B.3, it is interesting to see that the values of probability p which determine the shape of the equilibrium trajectory are approximately linear functions

of $1/2N$ and μ . Thus, as population size, N , increases, the probability that the trajectories of genetic diversity belong to the short-period domain increases. Similarly, as mutation decreases, μ , the probability that the trajectories of genetic diversity belong to the short-period domain increase and the probability that they belong to the large-period domain decreases. Numerical simulations confirm that when conditions from equation 3-B.3 are met, under scenario A, the genetic diversity reaches the equilibrium trajectory predicted under deterministic periods of isolation and connection, even though periods are stochastic (see Figure 3-B.1).

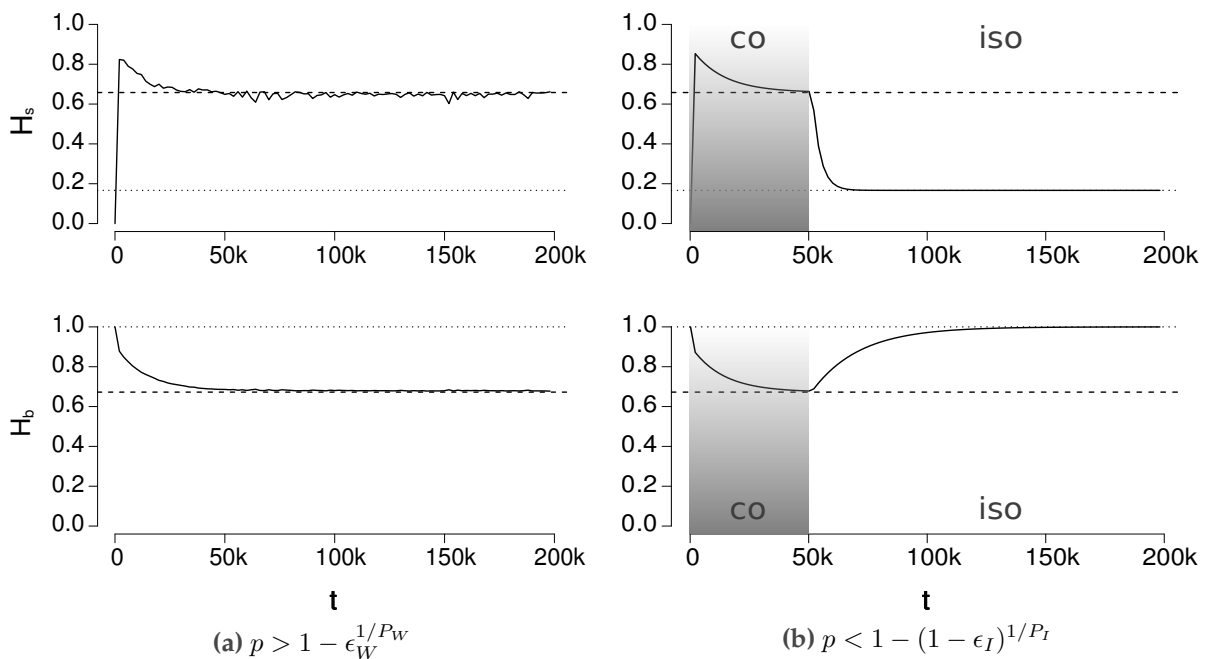


Figure 3-B.1 Illustration of the impact of stochastic period length (scenario A, geometric distribution of parameter p) on the trajectories of within- (H_s) and between-population (H_b) genetic diversities during periodic events of isolation and connection considering (a) short and (b) long expected periods $\mathbb{E}[P] = 1/p$. The dashed and dotted lines represent the expected equilibrium value when populations are connected and isolated, respectively. In (a), $p > 1 - \epsilon_W^{1/P_W}$ (expected short-periods) and both H_s and H_b tend to the connection equilibrium (dashed line). In (b), $p < 1 - (1 - \epsilon_I)^{1/P_I}$ (expected long-periods) and genetic diversities reach their expected equilibrium value at the end of each connection period and isolation period. Parameters are $M = 40$, $n = 10$, $N = 2,000$, $\mu = 2.5 \times 10^{-5}$, $\epsilon_W = \epsilon_I = 0.05$. (a) $p = 0.02$ ($\mathbb{E}[P] = 50$), (b) $p = 5.10^{-7}$ ($\mathbb{E}[P] = 2.10^6$).

Dynamics of genetic diversity under scenario B

Under scenario B, the period P follows a truncated normal distribution (between 0 and ∞). The period of the fluctuations P (i.e. the waiting time until a switch from one state to another) follows a discretized normal distribution of mean \bar{P} and variance σ^2 :

$$f(P = t) = \begin{cases} \frac{K}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\bar{P})^2}{2\sigma^2}} & \text{if } t > 0 \\ 0 & \text{else} \end{cases} \quad (3-B.4)$$

where $K = \sum_{t=0}^{+\infty} f(P=t)$ is a normalization constant taking into account the truncation and discretization of the distribution of P .

We can show that this distribution of waiting time corresponds to a time-inhomogeneous Markov process with two states (isolation and connection), and derive the corresponding transition probabilities $p(t)$ as a function of time t . Indeed, starting from a given state (either connection or isolation), the distribution of P is linked to the transition probabilities $p(t)$ through the relation:

$$f(P = t) = (1 - p(1))(1 - p(2)) \dots (1 - p(t - 1))p(t) \quad (3-B.5)$$

From equation 3-B.4 and 3-B.5, we have:

$$p(0) = f(P = 0) = \frac{K}{\sigma\sqrt{2\pi}} e^{-\frac{\bar{P}^2}{2\sigma^2}} \quad (3-B.6)$$

In addition, from equations 3-B.4 and 3-B.5, we have:

$$p(t + 1) \frac{1 - p(t)}{p(t)} = \frac{f(P = t + 1)}{f(P = t)} = e^{-\frac{2(t-\bar{P})-1}{2\sigma^2}} \quad (3-B.7)$$

which yields

$$p(t+1) = \frac{p(t)}{1-p(t)} e^{\frac{-2(t-\bar{P})-1}{2\sigma^2}} \quad (3-B.8)$$

Using equation 3-B.8 recursively, starting from the expression of $p(0)$ given in equation 3-B.6, leads to all values of $p(t)$.

Assuming that the probabilities, ϵ_W and ϵ_I , that the period P is lower than P_W and larger than P_I are small, respectively (i.e., $\epsilon_W = \epsilon_I = 0.05$) and using the 5% and 95% quantiles of a normal distribution of parameters \bar{P} and σ ($\bar{P} - 1.64\sigma$ and $\bar{P} + 1.64\sigma$, respectively) we have:

$$\begin{cases} \mathbb{P}(P < P_W) > 1 - \epsilon_W \Leftrightarrow \bar{P} + 1.64\sigma < P_W \\ \mathbb{P}(P > P_I) > 1 - \epsilon_I \Leftrightarrow \bar{P} - 1.64\sigma > P_I \end{cases} \quad (3-B.9)$$

Interestingly, we can see from equation 3-B.9 that for a given \bar{P} , increasing σ decreases the probability to reach the short-period domain and the intermediate period domain, as then $\bar{P} + 1.64\sigma$ (resp. $\bar{P} - 1.64\sigma$) becomes closer and possibly larger than P_W (resp. smaller than P_I). The gaussian noise does not change the qualitative behavior of genetic diversity through the periodic isolation and connection events (including the equilibrium trajectory) when conditions from equation 3-B.9 are met (see Figure 3-B.2).

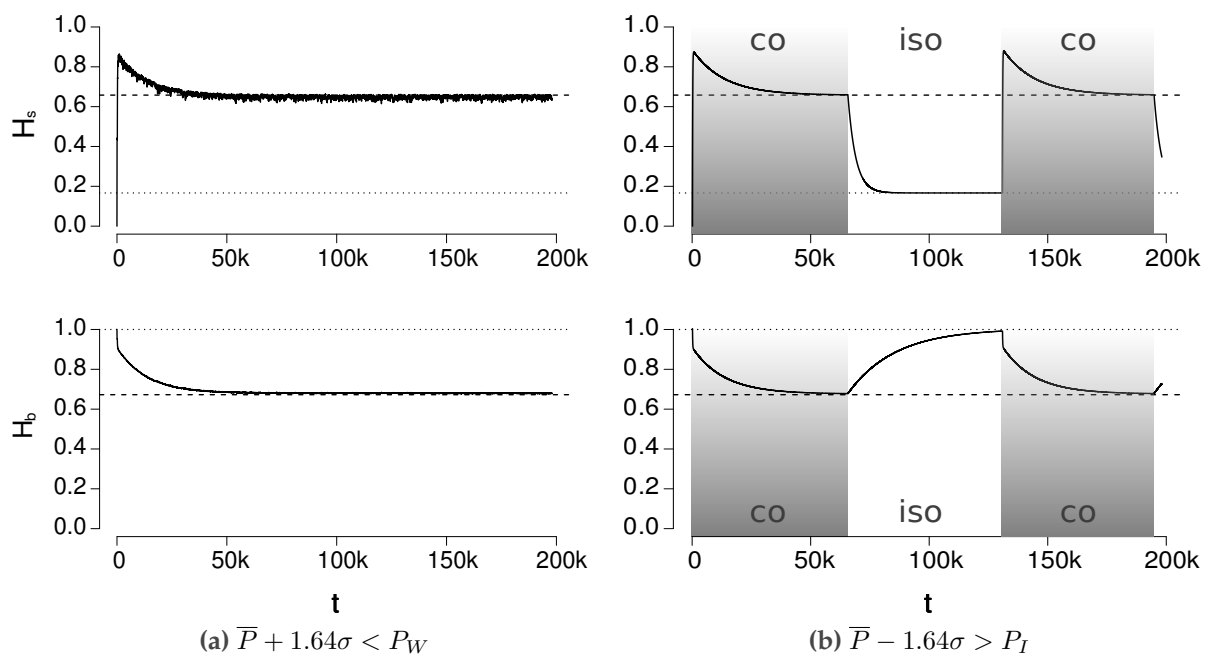


Figure 3-B.2 Illustration of the impact of stochastic period length (scenario B, duration following a normal distribution of mean \bar{P} and standard deviation σ) on the within- (H_s) and between-population (H_b) genetic diversities during periodic events of isolation and connection. The dashed and dotted lines represent the expected value at equilibrium when populations are connected and isolated, respectively. In (a), $\bar{P} + 1.64\sigma < P_W$ (expected short-periods), both H_s and H_b tend to the connection equilibrium (dashed line). In (b), $\bar{P} - 1.64\sigma > P_I$ (expected long-periods), genetic diversities reach their expected equilibrium value at the end of each connection period and isolation period. Parameters are $M = 40$, $n = 10$, $N = 2,000$, $\mu = 2.5 \times 10^{-5}$. (a) $\bar{P} = 100$, $\sigma = 50$ (b) $\bar{P} = 65,000$, $\sigma = 1,000$.

Chapter 4 The Signature of Past Population Isolation on Gene Genealogies and the Site Frequency Spectrum: Theory and Application to the Evolutionary Dynamics of HIV-1 subtypes

Nicolas Alcalá¹, Jeffrey D. Jensen², Amalio Telenti³, Séverine Vuilleumier^{1,3}

¹ Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland

² School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

³ Institute of Microbiology, University Hospital and University of Lausanne, CH-1011 Lausanne, Switzerland

Migration

Population structure

Coalescent

Site frequency spectrum

HIV

In prep.

ABSTRACT

Population genetics has transitioned from the theory-rich data-poor era of Fisher and Wright, to the theory-poor data-rich genomic era – where inference from large-scale polymorphism datasets requires new theoretical developments to understand how past demography and selection have shaped gene genealogies. Despite pervasive evidence in natural populations, past isolation events remain largely unexplored. European mammals and birds were isolated into glacial refugia during the quaternary period and viral populations are isolated within different hosts and between transmission events. Using coalescent theory, we investigate the temporal impact of past population isolation on gene genealogies, the site frequency spectrum, and common summary statistics of DNA polymorphism. We find that past population isolation significantly alters gene genealogies, with coalescent trees displaying an excess of short and long branches - the relative proportion of which is informative for inferring timing of isolation and number of previously isolated populations. We observe in a sample from a single population a transient excess of low and high frequency variants, while we observe in a combined sample from all populations an excess of intermediate frequency variants and a deficit of high frequency variants. We estimate the conditions of detection of these signature. Finally, we illustrate the utility of our results by detecting the history of major HIV-1 subtypes in China. We highlight the importance of detected past isolation events in generating and maintaining an HIV-1 genetic diversity hotspot in China.

PAST isolation events are common in natural populations. Anatomically modern humans are thought to have admixed with Neandertal populations after a period of isolation between the African and European continents (GREEN *et al.* 2010; PATTERSON *et al.* 2012; SANKARARAMAN *et al.* 2012). Similarly, African and non-African populations of *Drosophila melanogaster* also experienced a period of isolation, followed by subsequent admixture (POOL *et al.* 2012). Domesticated species of both plants and animals are isolated from their wild relatives, often to experience subsequent genetic exchange (e.g. in crops; ELLSTRAND *et al.* 1999). Viral populations are temporally isolated within hosts, prior to transmission events (co-infection e.g. in HIV, GROSS *et al.* 2004, and in malaria, ROBERT *et al.* 2003).

Past demographic and evolutionary events leave distinctive signatures on gene genealogies and thus on DNA sequence diversity, patterns commonly described by the *site frequency spectrum* (SFS). For example, a population bottleneck increases the variance of coalescence times and reduces the amount of variants at low frequency in the SFS. Population expansion creates star-shaped genealogies and increases the amount of variants at low frequency in the SFS (TAJIMA 1989; SLATKIN 1996). Population subdivision leads to long internal branches in the genealogy and increases the number of fixed variants (HARPENDING *et al.* 1998). It also results in "structured" genealogies in which the mean and variance of statistics of genealogies (e.g. time to the most recent common ancestor) is large, resulting in an excess of intermediate frequency variants in the SFS (WAKELEY 1999). Balancing selection similarly leads to such structure, and to an excess of variants at intermediate frequency (TAJIMA 1989; FAY and WU 2000; BARTON and ETHERIDGE 2004; ZENG *et al.* 2006). Directional selection leads to genealogies with both star shapes (leading to the common ancestor of the selected lineages) and long branches (due to remaining ancestral lineages and/or recombination with the se-

lected haplotype), and generates an excess of variants at both low and high frequency (BARTON 1998).

Polymorphism data is used to infer past demographic events (TAJIMA 1989; SIMONSEN *et al.* 1995; SCHNEIDER and EXCOFFIER 1999; EXCOFFIER 2004; ZENG *et al.* 2006; GUTENKUNST *et al.* 2009; NADUVILEZHATH *et al.* 2011). Some of the first methods proposed were developed to infer population growth rates (SLATKIN and HUDSON 1991), and recently developed approaches allow for the inference of a wide-range of demographic parameters (e.g., STRIMMER and PYBUS 2001; BEAUMONT *et al.* 2002; DRUMMOND *et al.* 2005; GUTENKUNST *et al.* 2009; EXCOFFIER *et al.* 2013). The isolation of derived populations from a common ancestral population has received particular attention. NIELSEN and WAKELEY (2001) first proposed a method to infer time of divergence, the migration rate between the two populations, as well as the relative sizes of the populations. This approach was recently extended to account for the divergence of multiple populations (HEY 2010). However, inferring the migration rate between diverged populations remains difficult, as models of a recent but complete separation are challenging to distinguish from models of ancient separation with a continued exchange of migrants. Also, parameter inference requires the *a priori* knowledge of a population split (HEY 2010), and the time at which migration event(s) occurred has thus far been difficult to estimate (STRASBURG and RIESEBERG 2011; SOUSA *et al.* 2011). Thus, to develop an accurate inference of past isolation and migration events, it is necessary to estimate parameters that summarize the best these events and thus a full understanding of their impact through time on gene genealogies and polymorphism data is required.

We here analyze the dynamics of signatures of past isolation events on gene genealogies, the SFS and on common polymorphism-based summary statistics. First, us-

ing coalescent theory, we derive analytically the expected pairwise coalescence times after a past population isolation, which determine the shape of coalescent trees. Second, we derive the time since isolation for which a signature on gene genealogies can still be detected, and we describe how the changes in gene genealogies influence methods of demographic inference based on the distribution of coalescence times (i.e., skyline plots). Third, we investigate how different SFS frequency classes are impacted by past isolation events using five estimators of the scaled mutation rate. Fourth, we describe how past isolation can influence commonly used SFS-based test statistics (Tajima's D , Fu and Li's D^* and F^* , Fay and Wu's H , Zeng *et al.*'s E), and discuss the specificity of the signature of past isolation events relative to other demographic and selective processes. Finally, to illustrate the utility of our theoretical results, we analyze the history of HIV-1 subtypes (subtypes B and C and CRF01_AE) in China. We use the signature of the full genome polymorphism data (pairwise nucleotide differences, local, total, and joint SFS) and identify the successive invasion of subtypes B and C and CRF01_AE in China (during the late 1980s, the early 1990s and mid 1990s, respectively; TAKEBE *et al.* 2010; AN *et al.* 2012), and recombinant forms between these subtypes within different risk groups (e.g. AN *et al.* 2012).

MODEL

To investigate the temporal signature of past isolation events on gene genealogies and polymorphism data, we consider d populations previously "connected" that became isolated at time T_{iso} and then "re-connected" at time T_{reco} . We assume that the scaled migration rate M between populations (i.e. twice the number of migrant genes per population per generation) is the same before and after the isolation period. The model assumes that each population has a constant size N and is composed of diploid individuals that reproduce at random, following the Wright-Fisher model (FISHER 1930;

WRIGHT 1931), and that mutations occur at a rate μ and follows the infinite sites model (KIMURA 1969). For analytical simplicity, we consider that times T_{iso} and T_{reco} are scaled in units of $2N$ generations, so $T_{iso} = 1$ corresponds to an isolation event which occurred $2N$ generations ago.

THE SIGNATURE OF PAST ISOLATION ON PAIRWISE COALESCENCE TIMES WITHIN AND BETWEEN POPULATIONS

To analyze how past isolation events impact gene genealogies, we first investigate how it affects the distributions of pairwise within- and between-population coalescence times T_w and T_b . T_w and T_b correspond to the time to the most recent common ancestor of two lineages sampled in the same and in different populations, respectively, scaled by the number of genes per population $2N$. Their distributions provide direct information on the shape of gene genealogies as they reflect the sizes and variability of coalescent tree branches.

Distribution of pairwise coalescence times

The coalescence of two genes under the finite island model can be described by a three states continuous time Markov chain (NOTOHARA 1990): the two genes can be present (1) in the same population, (2) in different populations or (3) be coalesced. The transition probabilities from one state to another depend on the scaled migration rate M between populations and on the number of populations d . Time T is counted in units of $2N$ generations. The Markov chain can be summarized by the following transition

rate matrix (NOTOHARA 1990):

$$\mathbf{Q} = \begin{pmatrix} -M - 1 & M & 1 \\ M/(d - 1) & -M/(d - 1) & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (4.1)$$

The probability that two genes were in state (1), (2) or (3) T generations ago (row vector \mathbf{P}_T), given they are in state (1), (2) or (3) at the current generation is given by $\mathbf{P}_T = \mathbf{P}_0 e^{\mathbf{Q}T}$ (\mathbf{P}_0 at generation 0), where $e^{\mathbf{Q}T}$ is a matrix of dimension $3 \cdot 3$. The elements of matrix $e^{\mathbf{Q}T}$ are denoted $q_{i,j}(T)$ and represent the probability that a gene in state (i) at the current generation was in state (j) T generations ago. Thus, the probability distribution of within-population and between-population pairwise coalescence times are given by $f_{w,coal}(T) = \frac{d}{dT}q_{1,3}(T)$ and $f_{b,coal}(T) = \frac{d}{dT}q_{2,3}(T)$. From that the distributions of within- and between-population pairwise coalescence times, $f_{T_w}(T)$ and $f_{T_b}(T)$, follow:

$$f_{T_w}(T) = \begin{cases} f_{w,coal}(T) & T < T_{reco} \\ q_{1,1}(T_{reco})e^{-(T-T_{reco})} & T_{reco} < T < T_{iso} \\ q_{1,1}(T_{reco})e^{-(T_{iso}-T_{reco})}f_{w,coal}(T - T_{iso}) + q_{1,2}(T_{reco})f_{b,coal}(T - T_{iso}) & T_{iso} < T \end{cases} \quad (4.2a)$$

$$f_{T_b}(T) = \begin{cases} f_{b,coal}(T) & T < T_{reco} \\ q_{2,1}(T_{reco})e^{-(T-T_{reco})} & T_{reco} < T < T_{iso} \\ q_{2,1}(T_{reco})e^{-(T_{iso}-T_{reco})}f_{w,coal}(T - T_{iso}) + q_{2,2}(T_{reco})f_{b,coal}(T - T_{iso}) & T_{iso} < T \end{cases} \quad (4.2b)$$

These distributions have a simple interpretation. The most recent period is the reconnection period, when $T < T_{reco}$, where $f_{T_w}(T)$ and $f_{T_b}(T)$ correspond to the probability of coalescence of 2 lineages under the finite island model, $f_{w,coal}(T)$ and $f_{b,coal}(T)$.

Then, during the isolation period, $T_{reco} < T < T_{iso}$, two lineages can only coalesce if they are in the same population after T_{reco} generations of connection (probabilities $q_{1,1}(T_{reco})$ and $q_{2,1}(T_{reco})$, respectively within and between populations); then their probability of coalescence at generation T in an isolated population is $e^{-(T-T_{reco})}$. During the older connection period, $T > T_{iso}$, two genes can coalesce if they were in the same population during isolation (probabilities $q_{1,1}(T_{reco})$ and $q_{2,1}(T_{reco})$, respectively within and between populations) and they did not coalesce during isolation (probability $e^{-(T_{iso}-T_{reco})}$); then their probability of coalescence at generation T are $f_{w,coal}(T - T_{iso})$. Alternatively, they can coalesce if they were in different populations during isolation (probabilities $q_{1,2}(T_{reco})$ and $q_{2,2}(T_{reco})$); then their probability of coalescence is at generation T is $f_{b,coal}(T - T_{iso})$. Equations 4.2a and 4.2b are used to obtain Figure 4.1.

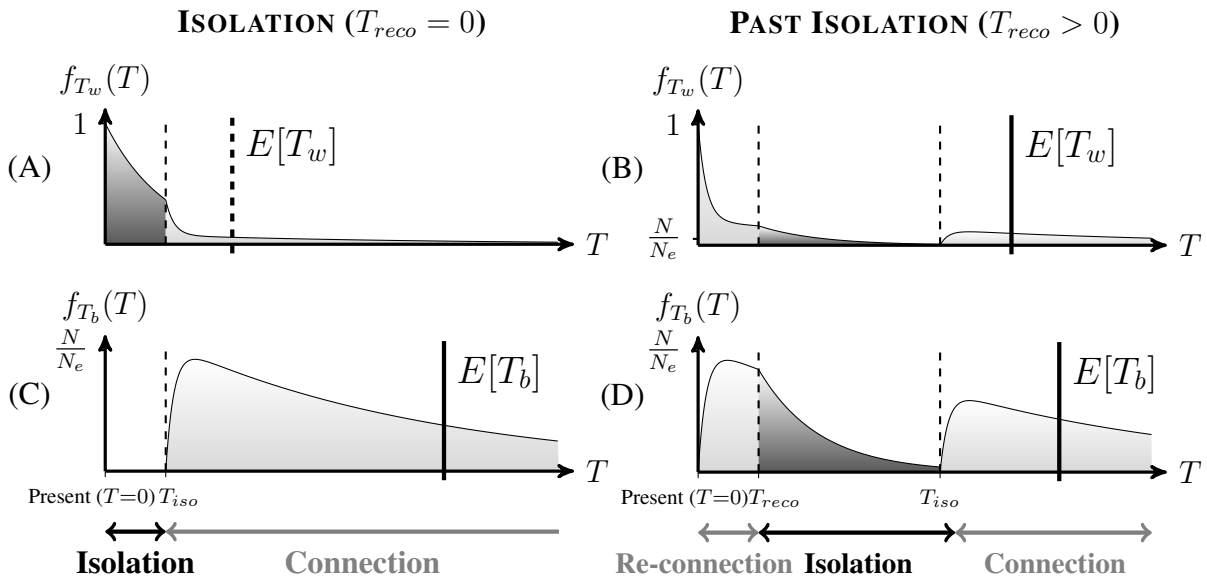


Figure 4.1 Probability distribution of (A)-(B) within-population $P_w(t)$ and (C)-(D) between-population $P_b(t)$ pairwise coalescence times following an isolation event (A, C), and a reconnection event (B, D). Parameters are $M = 1$, $d = 10$, (A) $T_{iso} = 1$, (B) $T_{reco} = 1$, $T_{iso} = 4$.

Thus, shortly after an isolation event (Figure 4.1B, D), coalescent trees will have short terminal branches (reflected by short pairwise coalescence times within-populations), and long internal branches (corresponding to long pairwise coalescent times between-

population). However, the difference of size between short and long branches increases with the duration of the isolation period. Following a reconnection event (Figure 4.1B, D), coalescent trees have a mixture of short and long branches. The distributions of within- and between-population pairwise coalescence times are bimodal. The probability of coalescence decreases monotonically during the isolation period, and increases during the connection period.

From the distribution in eqs. 4.2a and 4.2b, we can derive the expectations of T_w and T_b , $E[T_w] = \int_{T=0}^{+\infty} T f_{T_w}(T) dT$ and $E[T_b] = \int_{T=0}^{+\infty} T f_{T_b}(T) dT$:

$$\left\{ \begin{array}{l} E[T_w] = d \\ E[T_b] = d(1 + \frac{d-1}{dM}) \end{array} \right. \left\{ \begin{array}{l} + \frac{1}{\lambda_2 - \lambda_1} [M(T_{reco} - T_{iso} + 1 - e^{-(T_{iso} - T_{reco})})(e^{\lambda_1 T_{reco}} - e^{\lambda_2 T_{reco}}) \\ + (d-1)(1 - e^{-(T_{iso} - T_{reco})})(\lambda_1 e^{\lambda_1 T_{reco}} - \lambda_2 e^{\lambda_2 T_{reco}})] \\ + \frac{1}{\lambda_2 - \lambda_1} [(T_{iso} - T_{reco})(\lambda_2 e^{\lambda_1 T_{reco}} - \lambda_1 e^{\lambda_2 T_{reco}}) \\ + M(\frac{T_{iso} - T_{reco}}{d-1} + 1 - e^{-(T_{iso} - T_{reco})})(e^{\lambda_1 T_{reco}} - e^{\lambda_2 T_{reco}})] \end{array} \right. \quad (4.3)$$

where λ_1 and λ_2 are the biggest and lowest non-unit eigenvalues of matrix \mathbf{Q} .

Interestingly, we can show that the difference between the mean values of T_w and T_b is informative about the duration of the isolation period when $T_{reco} = 0$ (isolation scenario), but difficult to interpret when $T_{reco} > 0$ (past isolation scenario). In the latter case, the distributions of T_w and T_b are both bimodal, and it is the difference between the position of their respective modes which is most informative about the duration of the isolation period.

Indeed, the difference between the mean within- and between-population pairwise

coalescence times is:

$$E[T_b] - E[T_w] = \frac{d-1}{M} + \frac{1}{\lambda_2 - \lambda_1} [e^{\lambda_1 T_{reco}} ((T_{iso} - T_{reco}) (\frac{dM}{d-1} + \lambda_2) - \lambda_1 (d-1) (1 - e^{-(T_{iso} - T_{reco})})) - e^{\lambda_2 T_{reco}} ((T_{iso} - T_{reco}) (\frac{dM}{d-1} + \lambda_1) - \lambda_2 (d-1) (1 - e^{-(T_{iso} - T_{reco})}))] \quad (4.4)$$

when $T_{reco} > 0$, this value is a function of many parameters, which makes it difficult to interpret. On the contrary, when $T_{reco} = 0$, 4.4 simplifies to:

$$E[T_b] - E[T_w] = (d-1) \left(1 + \frac{1}{M} - e^{-T_{iso}}\right) + T_{iso} \quad (4.5)$$

The value of $E[T_b] - E[T_w]$ is always larger than T_{iso} , as $E[T_b] - E[T_w] \geq \frac{d-1}{M} + T_{iso}$. When the isolation event is recent (i.e. $T_{iso} < 1$), $E[T_b] - E[T_w]$ increases quickly with T_{iso} , at a rate that is given by the derivative $\frac{d(E[T_b] - E[T_w])}{dT_{iso}} \simeq d$. As the isolation event becomes older ($0 < T_{iso} < 3$), the rate of increase $\frac{d(E[T_b] - E[T_w])}{dT_{iso}}$ decreases exponentially until $E[T_b] - E[T_w] \simeq (d-1)(1 + 1/M) + T_{iso}$ when $T_{iso} \simeq 3$. Then, the value of $E[T_b] - E[T_w]$ increases approximately linearly with T_{iso} , at a rate $\frac{d(E[T_b] - E[T_w])}{dT_{iso}} \simeq 1$.

Implications for demographic inference from gene genealogies

The specific signature of past isolation events on gene genealogies have consequences for demographic inference based on coalescence times. A burst of coalescence events is interpreted as a reduction of effective population size by both mismatch distributions-based methods (SCHNEIDER and EXCOFFIER 1999) and skyline plot methods (STRIMMER and PYBUS 2001; DRUMMOND *et al.* 2005). We show in Figure 4.2 that the multimodal coalescence times distributions generated by population isolation produce the same signature as successive population size changes on skyline plot methods (using the method from STRIMMER and PYBUS 2001). An "old" isolation event (i.e. $T_{iso} = 5$,

$T_{reco} = 0$; Figure 4.2B) or a past isolation event ($T_{reco} > 0$; Figure 4.2C) generate skyline plots (Figures 4.2B, C) that detect, first an increase of effective population size (backward in time from the present to $6N$ generations ago) and then a decrease of effective population size (from $6N$ generations ago).

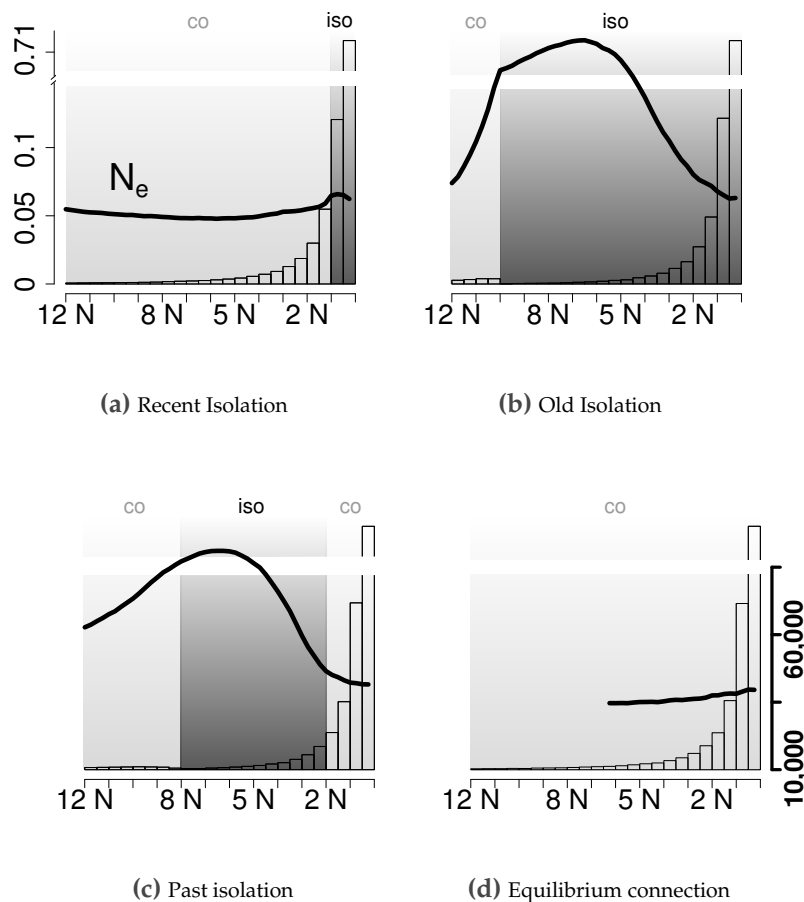


Figure 4.2 Effect of (a), (b) an isolation event and (c) a past isolation event, on the coalescent times distribution (histogram, scale on the left) and the effective population size inferred from coalescence times distribution (thick black line, scale on the right; skyline plots from the method developed by STRIMMER and PYBUS 2001). (d) Coalescence times distribution and skyline plot of an equilibrium connected population. The values are means estimated over 5,000 replicate simulations. Parameters are $d=4$, $N=2,500$, $M=1$ during connection periods, 20 sampled genes of $1kb$ per population, with $\mu=2.10^{-7}$ ($\theta=2$).

A METRIC ON THE MINIMUM STRENGTH OF ISOLATION EVENTS

On the necessary lengths of isolation events

We investigate how long an isolation period must be to leave a signature on gene genealogies (i.e. time to depart from the equilibrium neutral genealogies). We define this time as the duration of the isolation period for which there is a high probability (i.e. probability greater than $1 - \alpha$, where α is small) that a proportion of lineages less than β (with small β) coalesced during this period. The probability ($g_{n,m}(T_{iso})$) that n lineages sampled at present had m ancestral lineages T_{iso} generations in the past is (Equation 6.1 in TAVARÉ 1984):

$$g_{n,m}(T_{iso}) = \begin{cases} 1 - \sum_{k=2}^n \frac{(-1)^k (2k-1) n_{[k]} e^{-\frac{k(k-1)T_{iso}}{2}}}{n_{(k)}} & \text{for } m = 1 \\ \sum_{k=m}^n \frac{(-1)^{k-m} (2k-1) m_{(k-1)} n_{[k]} e^{-\frac{k(k-1)T_{iso}}{2}}}{m!(k-m)! n_{(k)}} & \text{for } m > 1 \end{cases} \quad (4.6)$$

The notation $x_{(y)}$ stands for the rising factorial with y terms, $x_{(y)} = x(x+1)\dots(x+y-1)$, and the notation $x_{[y]}$ stands for the falling factorial with y terms, $x_{[y]} = x(x-1)\dots(x-y+1)$.

Summing $g_{n,m}(T_{iso})$ for all $m < \beta \cdot n$ (rounded to the closest integer) provides the distribution probability that less than a proportion β of lineages coalesced. The quantile $1 - \alpha$ of this distribution provides the duration of the isolation period, $T_{min}^{1-\alpha,\beta}$, for which there is a high probability (i.e. $1 - \alpha$) that no lineages coalesced. When $\beta = 0$, $T_{min}^{1-\alpha,\beta}$ simplifies to:

$$T_{min}^{1-\alpha,0} = \frac{-2\ln(\alpha)}{n(n-1)} \quad (4.7)$$

For $\beta > 0$, this value can be obtained numerically (as in Figure 4.3A for $\beta = 0.05$).

We can see Figure 4.3A and from eq. 4.7 that the probability of detecting recent

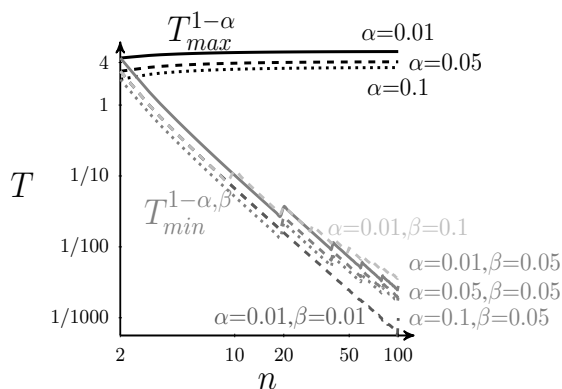
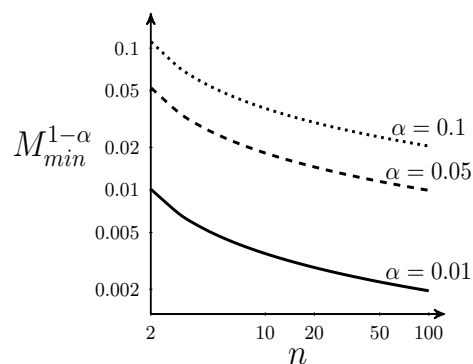
A DURATION OF THE ISOLATION PERIOD

B STRENGTH OF THE ISOLATION EVENT


Figure 4.3 (A) Minimum ($T_{min}^{1-\alpha, \beta}$) and maximum duration ($T_{max}^{1-\alpha}$) and (B) minimum strength of migration $M_{min}^{1-\alpha}$ for which an isolation event is expected not to leave a signature on coalescent trees (with probability $1 - \alpha$), as a function of the sample size in each population n . The term β in $T_{min}^{1-\alpha, \beta}$ denotes the proportion of lineages that coalesced during the isolation period. Note that $T_{max}^{1-\alpha}$ only applies for lineages sampled in a same population; as the signature of an isolation event on genealogies from genes sampled in several isolated populations persists. The minimum strength of migration corresponds to the scaled migration rate during the connection period preceding the isolation period.

isolation events increases with the sample size. Also, $T_{min}^{1-\alpha, \beta}$ does not plateau, but decreases approximately as the square of the sample size when $\beta = 0$. Increasing the sample size by a factor of 10 allows thus theoretically for the detection of isolation periods a hundredfold shorter.

We next investigate how short an isolation event must be to leave a signature on gene genealogies within-populations (the equilibrium coalescent in an isolated population from KINGMAN (1982) is reached). An isolation event of length T_{iso} does not leave a signature if there is a high probability (i.e. $< 1 - \alpha$, where α is small) that all lineages coalesced during that isolation period. This probability can be obtained from Equation 4.6 for $m=1$. As $T_{max} > 1$, ignoring terms in $e^{-\frac{k(k-1)T_{iso}}{2}}$ for $k > 2$ in $g_{n,1}(T_{iso})$ (Equation 4.6) leads to:

$$g_{n,1}(T_{iso}) = 1 - \frac{3(n-1)}{n+1} e^{-T_{iso}} \quad (4.8)$$

Thus, to leave a signature (with probability $1 - \alpha$), the duration of the isolation period, $T_{max}^{1-\alpha}$ (quantile $1 - \alpha$ of distribution $g_{n,1}(T_{iso})$) is a function of the sample size n :

$$T_{max}^{1-\alpha} = \ln(3/\alpha) - \ln\left(\frac{n+1}{n-1}\right) \tag{4.9}$$

Numerical results show that using Eq. 4.9 for T_{max} instead of the true value leads to a relative error smaller than 1.2×10^{-4} .

From Figure 4.3A and eq. 4.9, we can see that, for $n > 20$, $T_{max}^{1-\alpha}$ plateaus for all α values to approximately $\ln(3/\alpha)$, characterizing the maximum duration for which the signature of past demographic changes on gene genealogies can be detected.

On the necessary strengths of isolation events

Signature of isolation cannot be detected if lineages sampled in one population coalesced without having migrated to another population. From STROBECK (1987), the probability that a coalescent event occurred before a migration event, in a sample of size i from a single isolated population, is:

$$p_{coal}^i(M) = \frac{i-1}{i-1+M} \tag{4.10}$$

Thus, we can derive the probability that no migration event occurred during the coalescent of n lineages, as a function of M :

$$p_{coal,n}(M) = \prod_{i=2}^n \frac{i-1}{i-1+M} \tag{4.11}$$

Quantile $1 - \alpha$ of this distribution, gives the limit value of M , denoted $M_{min}^{1-\alpha}$, for which an isolation event has a large probability $1 - \alpha$ to leave no signature. We com-

pute numerically $M_{min}^{1-\alpha}$ for different values of α (Figure 4.3B); we also obtain a good approximation for M_{min} by simplifying the logarithm of eq. 4.11 using the approximations $\ln(i + M) = \ln(i) + M/i + o(M^2)$ and $\sum_{i=0}^n 1/i \sim \ln(n) + \gamma$ as $n \rightarrow \infty$, where γ is Euler's constant:

$$M_{min}^{1-\alpha} \simeq \frac{-\ln(1 - \alpha)}{\ln(n - 2) + \gamma} \quad (4.12)$$

We find numerically that the absolute relative error resulting from using approximation 4.12 is less than 5% for $n > 12$ and decreases when n increases.

We can see Figure 4.3B and from eq. 4.12 that $M_{min}^{1-\alpha}$ decreases when the sample size increases (whatever the α value), but very slowly (inverse of a logarithm decrease). Thus, even with large sample sizes, the power to detect the signature of past isolation from gene genealogies is limited by migration during the connected periods (from Equation 4.10 when probabilities of coalescent events $\frac{i-1}{i-1+M} \simeq 1$ are larger than migration events $\frac{M}{i-1+M} \simeq 0$).

THE SIGNATURE OF ISOLATION ON THE SITE FREQUENCY SPECTRUM AND ITS DETECTION USING NEUTRALITY TESTS

To analyse the temporal signature of population isolation on the SFS, we generate the SFS through time following isolation and reconnection events using the software *fastsimcoal* (EXCOFFIER and FOLL 2011) under two sampling schemes: the *local SFS* and the *total SFS*; also we study the joint SFS between pairs of populations. We then describe the power of five different estimators of the scaled mutation rate θ to detected past isolation: Tajima's D (TAJIMA 1989), Fu and Li's D^* and F^* (FU and LI 1993), Fay and Wu's H (FAY and WU 2000), Zeng et al.'s E (ZENG *et al.* 2006). We also use the framework of FERRETTI *et al.* (2010) and ACHAZ (2009) to analyze the power to detect isolation events of optimal test statistics ($T_{\Omega}^{I,l}$, $T_{\Omega}^{I,t}$, $T_{\Omega}^{II,l}$ and $T_{\Omega}^{II,t}$, Appendix A).

The signature of past isolation events on the local SFS

Immediately after an isolation event, the local SFS displays values expected from populations at "connected" equilibrium, and thus has a strong excess of sites across all frequencies (Figure 4.4A). As time T_{iso} since the isolation event increases, the observed local SFS tends to its expected SFS values at isolation equilibrium (Figure 4.4A). The transient excess of variants across the SFS is the result of the abrupt reduction of population effective size and the relatively slow effect of genetic drift. A characteristic feature of this process is that variants at low frequencies decrease faster than variants at high frequencies (which is detected by θ estimators; Figure S1A). In addition, variants that reach fixation accumulate with time (present in n copies; last class of the local SFS in Figure 4.4A). Following a reconnection event the local SFS shifts from its expected SFS values at isolated equilibrium (Figure 4.4B) to its expected SFS values at "connected" equilibrium (Figure 4.4B). Through this process a "U-shaped" SFS (Figure 4.4B) is formed due to excess of variants at high and low frequency. This pattern is similar to what is expected when variants are under positive selection. Then, we observe an excess of variants at intermediate frequency (Figure 4.4B), similarly to the SFS observed under balancing selection. Exchange of migrants between populations during the reconnection event first brings new variants previously locally fixed (in class $i=n$ of the local SFS; Figure 4.4B). Then, with time, migration increases the number of variants at intermediate frequency (Figure 4.4B). This dynamic is well captured by θ estimators (Figure S1B). This process occurs in two time scales: the initial increase of variants at low frequency is fast and is followed by a slow decrease of variants at low frequency. This translates to the fact that migration has a rapid impact on the SFS while genetic drift acts on a longer time-scale.

All test statistics have a low power to detect an isolation event from the local SFS

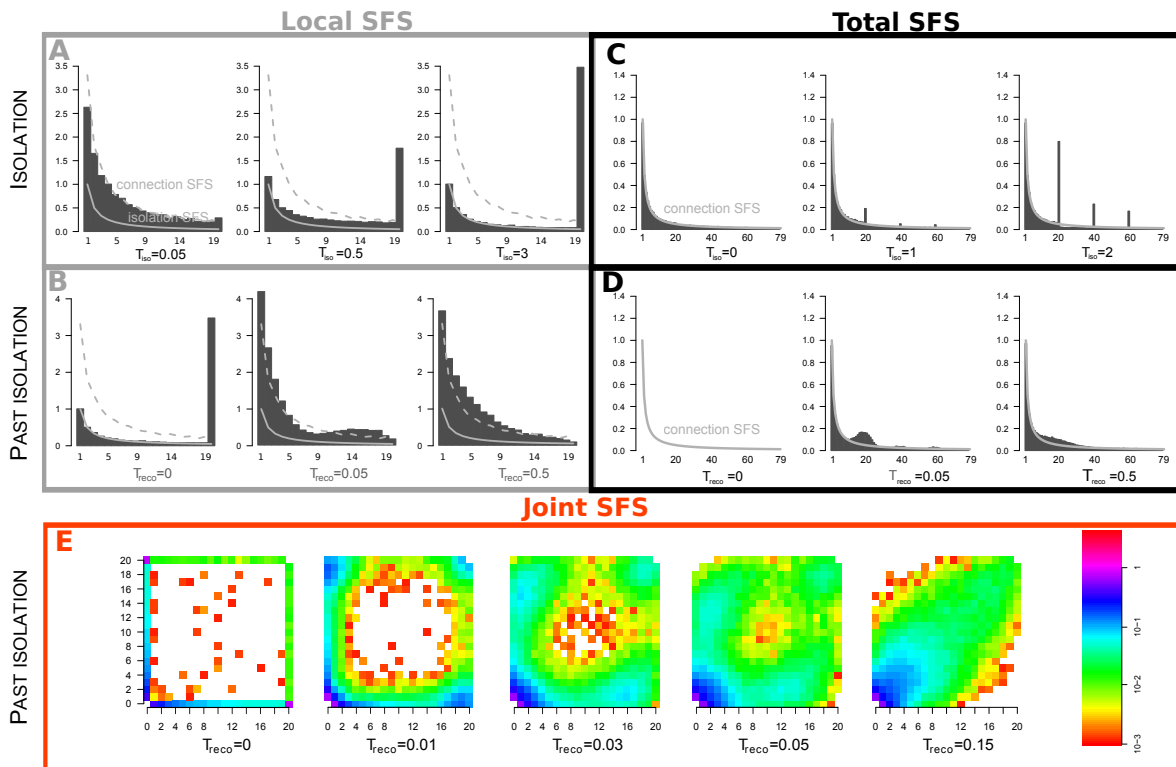


Figure 4.4 Signature of an isolation event (first row) and a past isolation period (second and third rows) on the local (A-B), total (C-D) and joint(E) site frequency spectrum (SFS), as a function of the number of generations since the isolation T_{iso} and connection T_{reco} event. The gray solid and dashed lines represent the expected SFS, in equilibrium isolated and connected populations, respectively. Parameters are $d=4$ populations, $N=2,500$ individuals per population, $n = 20$ sampled genes of $1kb$ per population, with $\mu=2.10^{-7}$ ($\theta=2$). For the past isolation scenario, we consider $T_{iso} = 3$. Means are over 5,000 replicates.

(less than 25%; Figure 4.5A). D , D^* , F^* , E and $T_{\Omega}^{I,l}$ have the largest power as they are sensitive to a deficit of variants at low frequencies. H and $T_{\Omega}^{II,l}$ have the lowest power as they are weakly sensitive to a deficit of low frequencies (statistics are skewed toward positive values; Figure S1). Except for E and $T_{\Omega}^{I,l}$ (Figure S1), the power of test statistics to detect a reconnection event is very low (Figure 4.5B).

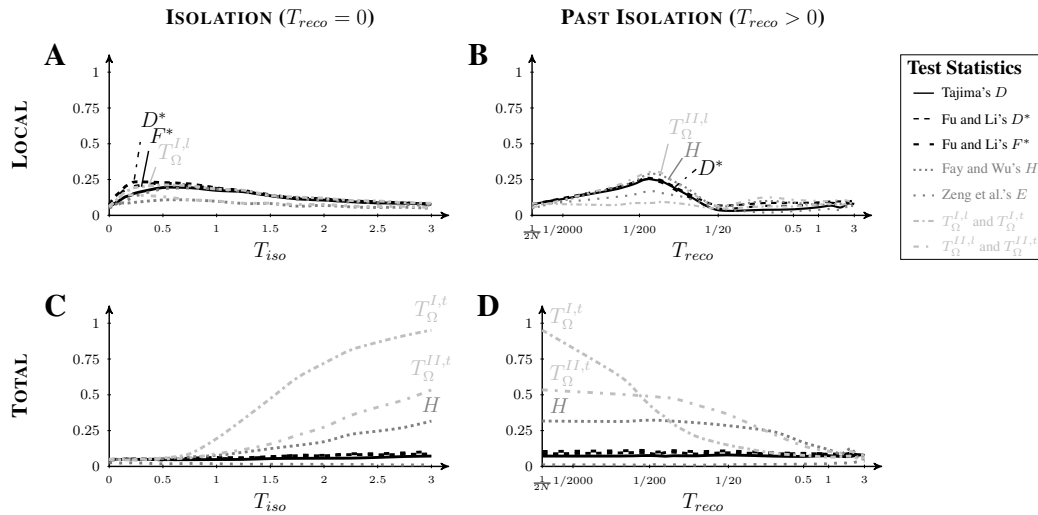


Figure 4.5 Effect of the time since the beginning of an isolation period (left column), and the time since the end of a past isolation period (right column) on the power of test statistics to detect departure from neutrality. Test statistics are either computed from the local site frequency spectrum (SFS; A-B), or estimated from the total SFS (C-D). Test statistics are Tajima's D , Fu and Li's D^* and F^* , Fay and Wu's H , Zeng et al.'s E and optimal test statistics derived using the method from FERRETTI *et al.* (2010) (see Appendix A) that detect the signature of the isolation scenario ($T_{\Omega}^{I,l}$ and $T_{\Omega}^{I,t}$) and II ($T_{\Omega}^{II,l}$ and $T_{\Omega}^{II,t}$) from the local and total SFS, respectively. Parameters are $d=4$, $N=2,500$, $n = 20$ sampled genes of $1kb$ per population, with $\mu=2.10^{-7}$ ($\theta=2$). For the past isolation scenario, we consider $T_{iso} = 3$. Means are over 5, 000 replicate simulations.

The signature of past isolation events in the total SFS

Following an isolation event, the total SFS displays a strongly distinctive pattern. We observe peaks of segregating sites at intermediate frequencies (Figure 4.4C). The frequencies at which peaks are observed depends on the samples sizes n_i , wherer $i = 1, \dots, d$ is the index of the population, and the number of populations d . For example, in Figure 4.4C, the frequencies at which the peaks are observed are 0.25, 0.5, 0.75 and 1 (1 not displayed) as $n_i=n=20$, $d=4$ and $d_i = 1, 2, 3, 4$). The expected size of the peaks

can be derived in the strong migration limit ($M \gg 1$) and when all genes coalesced during isolation. The first peak is composed of two values. First, it is composed of the number of variants that reached fixation during the isolation period which is determined by the number of mutants per generation $\theta/2$ which multiplies the difference between the time since the isolation event T_{iso} and the expected time to the most recent common ancestor in the sample, $E[T_{MRCA}]$. Second, it is composed of the number of variants, ξ_c , which have reached fixation during the past "connection" period. The expected size of the first peak depends thus on the time since isolation and on the rate of coalescence before the isolation event, giving:

$$E[\xi_{n_i}] = d_{n_i}[\theta/2(T_{iso} - E[T_{MRCA}]) + E[\xi_c]] \quad (4.13)$$

Where d_{n_i} is the number of populations that have a sample size of n_i .

Assuming that all genes coalesced within population during isolation, we have $E[T_{MRCA}] = 2(1 - \frac{1}{n_i})$. During the connection period, when $M \gg 1$, the expected number of variants which have fixed ($E[\xi_c]$) corresponds to the number of mutations which occurred on all terminal branches of the coalescent, $d\theta$ (with d lineages), divided by the number of terminal branches, d . Thus, Equation 4.13 simplifies into:

$$\begin{aligned} E[\xi_{n_i}] &= d_{n_i}[\theta/2(T_{iso} - 2(1 - 1/n_i)) + \theta] \\ &= d_{n_i}\theta(T_{iso}/2 + 1/n_i) \end{aligned} \quad (4.14)$$

We can see from Equation 4.14 that for old isolation events, the size of the first peak, ξ_{n_i} , depends almost exclusively on T_{iso} . For example, considering $d = 4$ populations that have been equally sampled, with a sample size of $n_i = 20$, for $T_{iso} = 3$, we expect a peak of size close to 1.55 in class $i = 20$ of the total SFS. Under the same conditions, but with an uneven sampling, where a population has a sample size of 24, a peak of

size close to 0.39 in class $i = 24$ of the total SFS is expected. These estimated values correspond to simulated ones (see Figures 4.4C, Figures S3D).

The size of other peaks, present at higher frequencies (corresponding to variants fixed in several populations), are independent of T_{iso} but dependent on n_i . Their size corresponds to the expected number of fixed variants in a panmictic population of $2dN$ genes (FU 1995):

$$E[\xi_{kn_i}] = d\theta/k, \quad k=2,\dots,d \quad (4.15)$$

Now following a reconnection event, migration distributes and homogenizes variants among populations. The multiple peaks observed during isolation merge to form an excess of variants at intermediate frequencies (Figure 4.4D). This occurs, assuming an island model, at a time frame that is of order of $1/2m$ generations (STROBECK 1987). For example, in the Figure 4.4D, the peaks merge within approximately $N/2$ generations (i.e. $\frac{2.5}{2m}$ generations, as $M = 4Nm = 10$).

θ estimators detect well the excess of variants at intermediate frequency following an isolation event (Figure S1D) and the power of $T_{\Omega}^{I,t}$, $T_{\Omega}^{II,t}$ and H to detect isolation events increase with T_{iso} but decrease with T_{reco} . Other statistics have low power even for long isolation periods (for $T_{iso} = 6N$, D , D^* , F^* and E have less than 10% power).

Past isolation events can also be well identified in the joint SFS (Figure 4.4E). The isolation period induces peaks at the vertices of the joint SFS (classes (20,0), (0,20) in Figure 4.4E), corresponding to private fixed variants in each populations. Following the reconnection, these peaks become progressively wider, decrease in size and move away from the vertices (Figure 4.4E) until merging (Figure 4.4E). This is due migration that distributes variants among populations and homogenize genetic composition of populations. Finally, the joint SFS reaches the expected joint SFS in equilibrium "connected" populations (Figure 4.4E; CHEN 2012) where variants are all shared at the same

frequency in each population and are localized in the diagonal of the joint SFS.

APPLICATION: DETECTION OF PAST HISTORY OF HIV-1 SUBTYPES IN CHINA

We illustrate the utility of our theoretical results by investigating the signature of isolation and past isolation of HIV-1 subtypes in China, using pairwise nucleotide differences between the sequences and the total, local and joint SFS.

We considered the 58 available sequence alignments of whole genome HIV-1 sampled in China in 2009 from the Los Alamos HIV database (<http://www.hiv.lanl.gov/>). Sequences were all identified as HIV-1 subtype B, C, CRF01_AE or recombinant forms between these subtypes (table S1). A sequence from subtype J was used as an outgroup to compute the SFS (the low sensitivity of our results to the chosen outgroup sequence is shown Figure S4 and S5).

Past isolation between the subtypes is determined with clusters using Discriminant Analysis of Principal Components (JOMBART *et al.* 2010). We find a good support for 4 clusters using the Bayesian Information Criterion (ranging from 2 to 100 PCA axes). The first cluster, *B*, contains all sequences from subtype B plus some B/C recombinants; the second cluster, *C*, contains all sequences from subtype C plus some B/C and B/C/CRF01_AE recombinants; the third cluster, CRF01_AE 1, and the fourth cluster, CRF01_AE 2, contain sequences from CRF01_AE, plus one CRF59_01B sequence (a recombinant between subtype B and CRF01_AE). The division of CRF01_AE sequences into two clusters reflects two successive introductions of CRF01_AE in China, in the early 1990s and mid 1990s, respectively (AN *et al.* 2012).

The signature on pairwise nucleotide differences and the total, local and joint SFS suggest successive past isolation events between subtypes. First, a long isolation period between subtypes is suggested by the distance between the modes of the within

and between-cluster distribution of pairwise nucleotide differences (figure S6) and by the large peaks in the total SFS corresponding to variants fixed or almost fixed in each clusters (figure 4.6A). Second, a very recent reconnection event between CRF01_AE and other subtypes is suggested by the small overlap between the within- and between-cluster distribution of pairwise nucleotide differences. This is reflected by the wideness of the peaks in the total SFS (figure 4.6A), by the excess of high frequency variants in the local SFS (figure 4.6B) and by the wide peaks very close to the vertices of the joint SFS (figure 4.6C). Third, an older reconnection event with strong migration between subtype B and C is suggested. This is reflected by the large overlap between the within- and between-cluster distribution of pairwise nucleotide differences, by the wideness of the peaks in the total SFS (figure 4.6A), by the excess of high frequency variants in the local SFS (figure 4.6B) and by the wide peaks close to the diagonal of the joint SFS (figure 4.6C). Although homoplasy is expected to increase the wideness of the peak, simulations show that this does not impact the signature in the SFS (Figure S7).

To gain further insights on past connection events (e.g., origin, timing, duration), we identified the sequences associated with the reconnection event between subtypes, i.e. sequences possessing signals with wide peaks in the joint SFS (e.g., connection of subtypes B and C, and connection between CRF01_AE and other subtypes). Interestingly, we found, in agreement with SU *et al.* 2000; PANG *et al.* 2012; HAN *et al.* 2013, that they are associated with two risk groups: Injecting Drug Users (IDUs) and Men who have Sex with Men (MSM). Our results are, moreover, consistent with the known history of HIV-1 epidemics in China. Subtype B and C are known to have been successively introduced by IDUs (leading to the first HIV outbreak in China in 1989) through the Yunnan province, which was the main entry route of illicit drugs in south-east Asia (YANG *et al.* 2002; LU *et al.* 2008; TAKEBE *et al.* 2010). CRF01_AE was later introduced

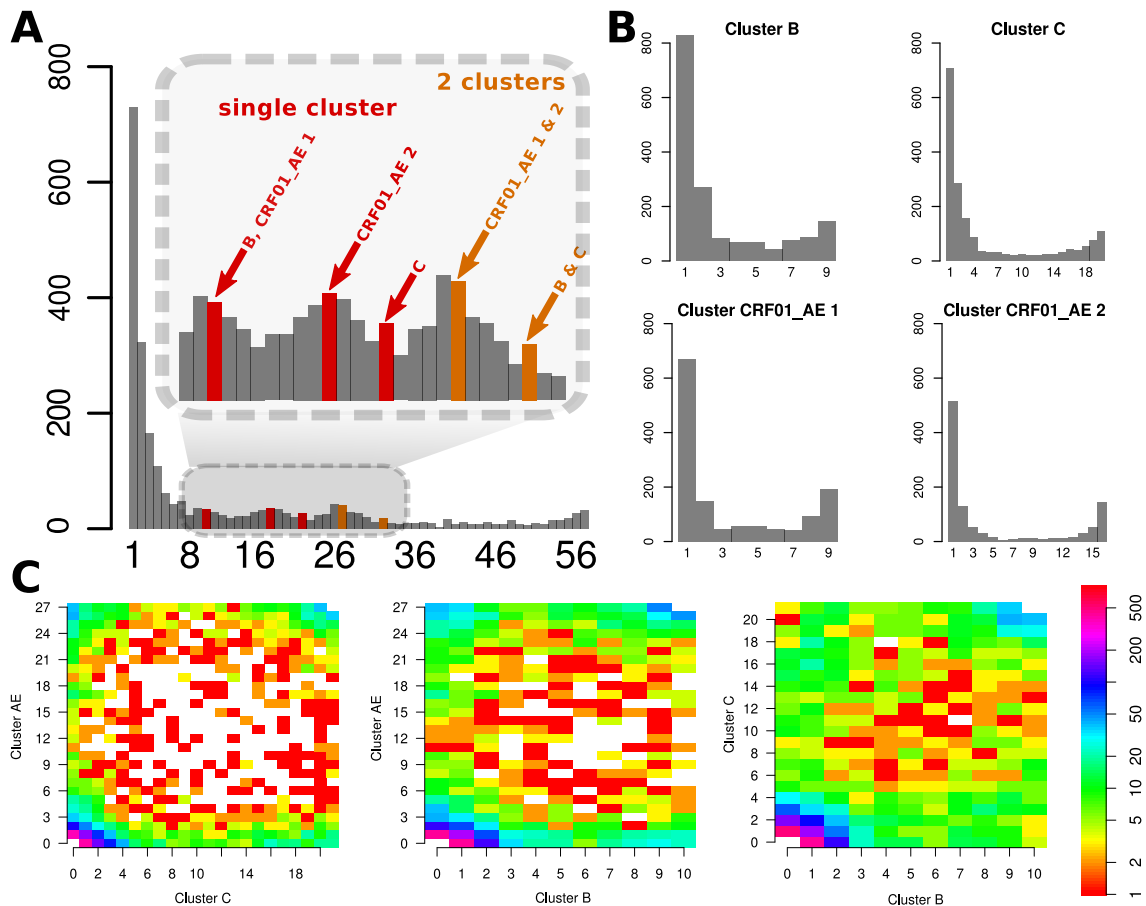


Figure 4.6 (A) Total, (B) local and (C) joint SFS of whole genome HIV-1 sequences sampled in China, from different subtypes. Each cluster is determined with Discriminant Analysis of Principal Components and represents a subtype plus recombinant forms which cluster with this subtype. CRF01_AE is divided into two clusters, CRF01_AE 1 and CRF01_AE 2. (A) Arrows indicate peaks corresponding to variants fixed or almost fixed in a single cluster (red) and in two clusters (orange).

by the MSM risk group (already infected by subtypes B and C) during the 1990s, and is now the most prevalent HIV-1 form in many parts of China (AN *et al.* 2012). This supports the importance of risk groups in generating and maintaining a genetic diversity hotspot in China and is of great importance for managing the treatment and control of HIV (e.g. KORBER *et al.* 2001; GASCHEN *et al.* 2002).

DISCUSSION

Our work highlights features of gene genealogies and the site frequency spectrum that can disentangle past isolation events from other demographic (e.g. bottlenecks) or selective events (e.g. positive, balancing or negative selection). We demonstrate that the difference in pairwise coalescence times within- and between-populations, the changes through time of their distributions, as well as their number of modes, are informative of past isolation events. We determined the necessary duration of an isolation event that allows its detection and the strength of migration that prevents it. We find that increasing the sample size does improve our ability to detect brief isolation periods, but not our ability to detect past isolation when migration is weak.

The analytical work described here is thus of interest to improve inference methods and provides conditions for the existence and detection of a signature of past isolation events. We find that contrasting polymorphism data from different populations helps understand their demographic history. For example, using mismatch distributions as a proxy for pairwise coalescence times can be very informative about the timing of past isolation and the degree of population subdivision. We demonstrate that usual inference methods based on the distribution of coalescence times (tests based on mismatch distributions and skyline plots) do not allow one to distinguish the signature of events of population isolation from other demographic changes, particularly bottleneck and population expansion. However, an increasing distance between the modes

of pairwise within- and between-population coalescence times distributions (and thus between the corresponding mismatch distributions) as identified here, is strongly informative of past isolation event and is not expected from a population size changes (bottleneck or expansion).

Our study suggests that analyses based only on within-population polymorphism data should be avoided and combined samples from several populations is needed to provide an accurate inference of past isolation events. Our conclusions are in agreement with results from STÄDLER *et al.* (2009) and CHIKHI *et al.* (2010) that suggest that the sampling scheme is crucial to detect the signature of past population history. We also suggest the use of multiple test statistics that are representative of different SFS classes of variants (low, intermediate, and high) such as D , D^* , F^* , H , E , and optimal statistics T_Ω (FERRETTI *et al.* 2010) might not capture the full dynamic of the signature of past isolation through time. However, the general statistics developed by ACHAZ (2009), based on the sampling scheme and on variants at frequencies not defined *a priori* provide much high power to detect past isolation than standard statistics.

Empirical studies often analyze population structure (e.g., STRUCTURE analysis; EVANNO *et al.* 2005) and past population size (e.g., RAMBAUT *et al.* 2008 using skyline plot methods) independently. However, when populations are structured inference of past changes of population size is problematic, as migration strongly biases the inference. For example, past migration events could increase the significance of a bottleneck signal (Figure 4.2). Similarly, past bottleneck events followed by continuous migration events will leave a signature similar to an isolation event. The presented signature of past isolation on the total SFS can complement assignment analysis to infer population structure (e.g. using the software STRUCTURE, or principal component analysis-PCA) when they are not informative, and when the joint SFS cannot be computed.

Our results suggest that this might occur when the isolation periods are short or when the time since reconnection is long. In these situations statistical tests do not detect an excess of variants at intermediate frequencies.

We have here explored the temporal dynamics of the gene genealogies and the SFS after isolation events, and the patterns described here would be valuable to analyse demographic change on time series data points (e.g. from ancient DNA; VALDIOSERA *et al.* 2008). Our study, together with others (STÄDLER *et al.* 2009), suggests that non-equilibrium populations can display a large variety of signatures which are difficult to disentangle. Similarly, complex demographic and evolutionary scenarios involve a succession of historical events that disrupt the signature in different manners (e.g. CURRAT *et al.* 2010). Having several time points could be strongly informative and provide more robust estimates for demographic inference. Also, we believe that the full dynamics through time of a signature of a given demographic or selective event on DNA polymorphism should deserve more attention. It provides first the expected duration and conditions under which the signature can be detected and second, this provides crucial information on the statistics that has to be inferred (the one that is the most sensitive to the event under consideration).

The ability to detect past isolation events from DNA polymorphism is of particular relevance to understand the mechanisms of speciation, first because speciation events in sympatry, parapatry, and allopatry have challenged the field for decades (FEDER *et al.* 2003, 2011), and second because models alternating allopatry and sympatry have been proposed to explain adaptive radiations (e.g. AGUILÉE *et al.* 2013).

Our theoretical investigation allowed the description of the past population history of HIV-1 subtypes in China, although HIV-1 might be under strong selective pressures from the immune system (SNOECK *et al.* 2011), undergo frequent bottlenecks followed

by population expansion (during transmission; e.g. KEELE *et al.* 2008) and harbor a high mutation rate (PRESTON *et al.* 1988). Nevertheless, the signature observed (Figure 4.6) only slightly deviates from our theoretical expectations. The robustness of our results to demographic, mutational and selective events can be explained by the fact that such events impact the differentiation between subtypes during the isolation period, but not the signature of a reconnection event, which is driven by migration that acts at short time-scales (see e.g. ALCALA *et al.* 2013). This separation of time-scales allows a robust inference of past population history. Our results also suggest that reconnection following isolation between subtypes is recurrent in the region, and significantly contributes to the evolution of HIV-1 in Asia.

ACKNOWLEDGMENTS

We thank John Wakeley and three anonymous reviewers for their insightful comments, which increased the relevance and outreach of the paper. This project was funded by the Swiss National Science Foundation (SNSF) grants #PZ00P3_139421/1, #31003A-130065 and the University of Lausanne.

BIBLIOGRAPHY

- ACHAZ, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. *Genetics* **183**: 249–258.
- AGUILÉE, R., D. CLAESSEN, and A. LAMBERT, 2013 Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics. *Evolution* **67**: 1291–1306.
- ALCALA, N., D. STREIT, J. GOUDET, and S. VUILLEUMIER, 2013 Peak and persistent excess of genetic diversity following an abrupt migration increase. *Genetics* **193**: 953–971.

- AN, M., X. HAN, J. XU, Z. CHU, M. JIA, *et al.*, 2012 Reconstituting the epidemic history of hiv strain crf01_ae among men who have sex with men (msm) in liaoning, northeastern china: Implications for the expanding epidemic among msm in china. *Journal of virology* **86**: 12402–12406.
- BARTON, N. H., 1998 The effect of hitch-hiking on neutral genealogies. *Genetics Research* **72**: 123–133.
- BARTON, N. H., and A. M. ETHERIDGE, 2004 The effect of selection on genealogies. *Genetics* **166**: 1115–1131.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- CHEN, H., 2012 The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theor Popul Biol* **81**: 179–195.
- CHIKHI, L., V. C. SOUSA, P. LUISI, B. GOOSSENS, and M. A. BEAUMONT, 2010 The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**: 983–995.
- CURRAT, M., E. S. POLONI, and A. SANCHEZ-MAZAS, 2010 Human genetic differentiation across the Strait of Gibraltar. *BMC Evolutionary Biology* **10**.
- DRUMMOND, A. J., A. RAMBAUT, B. SHAPIRO, and O. G. PYBUS, 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**: 1185–1192.

- ELLSTRAND, N. C., H. C. PRENTICE, and J. F. HANCOCK, 1999 Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics* **30**: pp. 539–563.
- EVANNO, G., S. REGNAUT, and J. GOUDET, 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**: 2611–2620.
- EXCOFFIER, L., 2004 Patterns of dna sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol Ecol* **13**: 853–864.
- EXCOFFIER, L., I. DUPANLOUP, E. HUERTA-SÁNCHEZ, V. C. SOUSA, and M. FOLL, 2013 Robust demographic inference from genomic and snp data. *PLoS Genet* **9**: e1003905.
- EXCOFFIER, L., and M. FOLL, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**: 1332–1334.
- FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive darwinian selection. *Genetics* **155**: 1405–1413.
- FEDER, J., S. BERLOCHER, J. ROETHELE, H. DAMBROSKI, J. SMITH, *et al.*, 2003 Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. *Proc Natl Acad Sci U S A* **100**: 10314–10319.
- FEDER, J. L., R. GEJJI, T. H. Q. POWELL, and P. NOSIL, 2011 Adaptive chromosomal divergence driven by mixed geographic mode of evolution. *Evolution* **65**: 2157–2170.
- FERRETTI, L., M. PEREZ-ENCISO, and S. RAMOS-ONSINS, 2010 Optimal neutrality tests based on the frequency spectrum. *Genetics* **186**: 353–365.

- FISHER, R. A., 1930 *The genetical theory of natural selection*. Clarendon Press.
- FU, Y. X., 1995 Statistical properties of segregating sites. *Theor Popul Biol* **48**: 172–197.
- FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GASCHEN, B., J. TAYLOR, K. YUSIM, B. FOLEY, F. GAO, *et al.*, 2002 Diversity considerations in hiv-1 vaccine selection. *Science* **296**: 2354–2360.
- GREEN, R. E., J. KRAUSE, A. W. BRIGGS, T. MARICIC, U. STENZEL, *et al.*, 2010 A draft sequence of the neandertal genome. *Science* **328**: 710–722.
- GROSS, K. L., T. C. PORCO, and R. M. GRANT, 2004 Hiv-1 superinfection and viral diversity. *AIDS* **18**: 1513–1520.
- GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON, and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet* **5**: e1000695.
- HAN, X., M. AN, B. ZHAO, S. DUAN, S. YANG, *et al.*, 2013 High prevalence of hiv-1 intersubtype b /c recombinants among injecting drug users in dehong, china. *PloS one* **8**: e65337.
- HARPENDING, H. C., M. A. BATZER, M. GURVEN, L. B. JORDE, A. R. ROGERS, *et al.*, 1998 Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* **95**: 1961–1967.
- HEY, J., 2010 Isolation with migration models for more than two populations. *Mol Biol Evol* **27**: 905–920.
- JOMBART, T., S. DEVILLARD, and F. BALLOUX, 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics* **11**: 94.

KEELE, B. F., E. E. GIORGI, J. F. SALAZAR-GONZALEZ, J. M. DECKER, K. T. PHAM, *et al.*, 2008 Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. *Proceedings of the National Academy of Sciences* **105**: 7552–7557.

KIMURA, M., 1964 Diffusion models in population genetics. *Journal of Applied Probability* **1**: 177–232.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**: 893–903.

KINGMAN, J. F. C., 1982 On the Genealogy of Large Populations. *J. Appl. Prob.* **19A**: 27–43.

KORBER, B., B. GASCHEN, K. YUSIM, R. THAKALLAPALLY, C. KESMIR, *et al.*, 2001 Evolutionary and immunological implications of contemporary hiv-1 variation. *British medical bulletin* **58**: 19–42.

LU, L., M. JIA, Y. MA, L. YANG, Z. CHEN, *et al.*, 2008 The changing face of hiv in china. *Nature* **455**: 609–611.

NADUVILEZHATH, L., L. E. ROSE, and D. METZLER, 2011 Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Mol Ecol* **20**: 2709–2723.

NIELSEN, R., and J. WAKELEY, 2001 Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics* **158**: 885–896.

NOTOHARA, M., 1990 The coalescent and the genealogical process in geographically structured population. *J Math Biol* **29**: 59–75.

- PANG, W., C. ZHANG, L. DUO, Y.-H. ZHOU, Z.-H. YAO, *et al.*, 2012 Extensive and complex hiv-1 recombination between b', c and crf01_ae among idus in south-east asia. *AIDS* **26**: 1121–1129.
- PATTERSON, N., P. MOORJANI, Y. LUO, S. MALLICK, N. ROHLAND, *et al.*, 2012 Ancient admixture in human history. *Genetics* **192**: 1065–1093.
- POOL, J. E., R. B. CORBETT-DETIG, R. P. SUGINO, K. A. STEVENS, C. M. CARDENO, *et al.*, 2012 Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. *PLoS Genet* **8**: e1003080.
- PRESTON, B. D., B. J. POIESZ, and L. A. LOEB, 1988 Fidelity of hiv-1 reverse transcriptase. *Science* **242**: 1168–1171.
- RAMBAUT, A., O. G. PYBUS, M. I. NELSON, C. VIBOUD, J. K. TAUBENBERGER, *et al.*, 2008 The genomic and epidemiological dynamics of human influenza a virus. *Nature* **453**: 615–619.
- ROBERT, V., K. MACINTYRE, J. KEATING, J.-F. TRAPE, J.-B. DUCHEMIN, *et al.*, 2003 Malaria transmission in urban sub-saharan africa. *Am J Trop Med Hyg* **68**: 169–176.
- SANKARARAMAN, S., N. PATTERSON, H. LI, S. PÄÄBO, and D. REICH, 2012 The date of interbreeding between neandertals and modern humans. *PLoS Genet* **8**: e1002947.
- SCHNEIDER, S., and L. EXCOFFIER, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial dna. *Genetics* **152**: 1079–1089.
- SIMONSEN, K. L., G. A. CHURCHILL, and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for dna polymorphism data. *Genetics* **141**: 413–429.

- SLATKIN, M., 1996 Gene genealogies within mutant allelic classes. *Genetics* **143**: 579–587.
- SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics* **129**: 555–562.
- SNOECK, J., J. FELLAY, I. BARTHA, D. C. DOUEK, and A. TELENTI, 2011 Mapping of positive selection sites in the hiv-1 genome in the context of rna and protein structural constraints. *Retrovirology* **8**: 87.
- SOUSA, V. C., A. GRELAUD, and J. HEY, 2011 On the nonidentifiability of migration time estimates in isolation with migration models. *Mol Ecol* **20**: 3956–3962.
- STRASBURG, J. L., and L. H. RIESEBERG, 2011 Interpreting the estimated timing of migration events between hybridizing species. *Mol Ecol* **20**: 2353–2366.
- STRIMMER, K., and O. G. PYBUS, 2001 Exploring the demographic history of dna sequences using the generalized skyline plot. *Mol Biol Evol* **18**: 2298–2305.
- STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. *Genetics* **117**: 149–153.
- STÄDLER, T., B. HAUBOLD, C. MERINO, W. STEPHAN, and P. PFAFFELHUBER, 2009 The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182**: 205–216.
- SU, L., M. GRAF, Y. ZHANG, H. VON BRIESEN, H. XING, *et al.*, 2000 Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (c/b) recombinant strain in china. *Journal of virology* **74**: 11367–11376.
- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**: 585–595.

- TAKEBE, Y., H. LIAO, S. HASE, R. UENISHI, Y. LI, *et al.*, 2010 Reconstructing the epidemic history of hiv-1 circulating recombinant forms crf07_bc and crf08_bc in east asia: the relevance of genetic diversity and phylodynamics for vaccine strategies. *Vaccine* **28**: B39–B44.
- TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol* **26**: 119–164.
- VALDIOSERA, C. E., J. L. GARCÍA-GARITAGOITIA, N. GARCIA, I. DOADRIO, M. G. THOMAS, *et al.*, 2008 Surprising migration and population size dynamics in ancient iberian brown bears (*ursus arctos*). *Proc Natl Acad Sci U S A* **105**: 5123–5128.
- WAKELEY, J., 1999 Nonequilibrium migration in human history. *Genetics* **153**: 1863–1871.
- WRIGHT, S., 1931 Evolution in mendelian populations. *Genetics* **16**: 97–159.
- YANG, R., X. XIA, S. KUSAGAWA, C. ZHANG, K. BEN, *et al.*, 2002 On-going generation of multiple forms of hiv-1 intersubtype recombinants in the yunnan province of china. *Aids* **16**: 1401–1407.
- ZENG, K., Y.-X. FU, S. SHI, and C.-I. WU, 2006 Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**: 1431–1439.

APPENDIX A: DERIVING OPTIMAL TESTS TO DETECT THE SIGNATURE OF ISOLATION AND PAST ISOLATION SCENARIOS FROM THE SFS

In this appendix, we design, using the method to detect departure from neutrality from FERRETTI *et al.* (2010), four test statistics to detect the signature of the isolation and past isolation scenarios from the local and total SFS. $T_{\Omega}^{I,l}$ and $T_{\Omega}^{I,t}$ can detect the signature of the isolation scenario from the local and total SFS, respectively; $T_{\Omega}^{II,l}$ and $T_{\Omega}^{II,t}$ can detect the signature of the past isolation scenario from the local and total SFS, respectively. We first present the method of FERRETTI *et al.* (2010), which requires the expected SFS under both scenarios. Second, we give two methods to derive the expected SFS under both scenarios. Finally, we present the properties of the resulting test statistics.

Method to derive T_{Ω}^I and T_{Ω}^{II}

ACHAZ (2009) demonstrates that all neutrality tests based on the SFS can be written as:

$$T_{\Omega} = \frac{\sum_{i=1}^{n-1} i\Omega_i\xi_i}{\sqrt{\text{Var}\left(\sum_{j=1}^{n-1} j\Omega_j\xi_j\right)}} \quad (4-A.1)$$

where Ω_i , with $i = 1, 2, \dots, n - 1$ are the weights which uniquely define the test (for example, taking $\Omega_i = 2(n - i)/[n(n - 1)] - 1/(i \sum_j 1/j)$ leads to Tajima's D) and the ξ_i are the classes of the SFS (i.e. the number of variants at frequency i/n). The variance in the denominator can be computed using eq. 9 from ACHAZ (2009). As showed in FERRETTI *et al.* (2010), the optimal test against a given demographic scenario can be built by choosing the Ω_i that maximizes the expected value of T_{Ω} under the given scenario. By denoting $\xi_i^{I,l}$, $\xi_i^{I,t}$, $\xi_i^{II,l}$ and $\xi_i^{II,t}$ the expected values of the classes of the local

and total SFS under the isolation scenario and the local and total SFS under the past isolation scenario, the optimal test statistics $T_{\Omega}^{I,l}$, $T_{\Omega}^{I,t}$, $T_{\Omega}^{II,l}$ and $T_{\Omega}^{II,t}$ have the following weights, denoted $\Omega_i^{I,l}$, $\Omega_i^{I,t}$, $\Omega_i^{II,l}$ and $\Omega_i^{II,t}$, respectively (FERRETTI *et al.* 2010):

$$\begin{aligned}
 \Omega_i^{I,l} &= \frac{\xi_i^{I,l}}{\sum_j \xi_j^{I,l}} - \frac{1}{ia_n} \\
 \Omega_i^{I,t} &= \frac{\xi_i^{I,t}}{\sum_j \xi_j^{I,t}} - \frac{1}{ia_n} \\
 \Omega_i^{II,l} &= \frac{\xi_i^{II,l}}{\sum_j \xi_j^{II,l}} - \frac{1}{ia_n} \\
 \Omega_i^{II,t} &= \frac{\xi_i^{II,t}}{\sum_j \xi_j^{II,t}} - \frac{1}{ia_n}
 \end{aligned} \tag{4-A.2}$$

where $a_n = \sum_{i=0}^{n-1} 1/i$.

$T_{\Omega}^{I,l}$ and $T_{\Omega}^{I,t}$ (resp. $T_{\Omega}^{II,l}$ and $T_{\Omega}^{II,t}$) have positive values under the isolation scenario (resp. the past isolation scenario) and approach 0 under neutrality. Indeed, the positive terms in eq. 4-A.2 correspond to the expected value of the SFS under the isolation scenario or the past isolation scenario, while the negative terms correspond to the expected value under neutrality. Consequently, the statistics should be used in one-sided tests.

Deriving the expected SFS under isolation and past isolation scenarios

Values $\xi_i^{I,l}$, $\xi_i^{I,t}$, $\xi_i^{II,l}$ and $\xi_i^{II,t}$ can be estimated from coalescent simulations under each scenario with fixed parameters. For our study, we used the values presented in Figure 4.4A-B and G, and 4.4C-D, as these values present the most prominent features of the signature of both scenarios on the SFS (large peaks) and have the strongest power to reject neutrality.

Alternatively, values $\xi_i^{I,t}$ and $\xi_i^{II,t}$ can be obtained using formulas approximating the total SFS under both scenarios, assuming that panmixia takes place during the

connected periods and that the isolation periods are long enough ($T_{iso} > 2$). Considering d populations, equal sample size n in each population (total sample size is dn) and scaled mutation rate θ , the total SFS corresponds to the expected SFS in a panmictic population with scaled mutation rate $d\theta$ truncated between 0 and $1/n$, plus excess (peaks) due to the transient dynamics (see Figure 4.4C-D). The size of the peaks is defined by eqs. 4.14 and 4.15 under the isolation scenario, and we can approximate their shape using diffusion approximation under the past isolation scenario:

$$\xi_i^{I,t} = \begin{cases} d\theta/i, & \text{if } i < n \\ d\theta(T_{iso}/2 + 1/n), & \text{if } i = n \\ d\theta/k, & \text{if } i = k \times n, k = 2, \dots, d \\ 0, & \text{otherwise} \end{cases} \quad (4-A.3)$$

$$\xi_i^{II,t} = \begin{cases} d\theta[1/i + (T_{iso}/2 + 1/n)\phi(i/(dn), 1/d, T_{iso}) + \sum_{k=2}^d (1/k)\phi(k/d, T_{iso})], & \text{if } i < n \\ d\theta[(T_{iso}/2 + 1/n)\phi(i/(dn), 1/d, T_{iso}) + \sum_{k=2}^d (1/k)\phi(k/d, T_{iso})], & \text{if } i \geq n \end{cases} \quad (4-A.4)$$

where $\phi(p, p_0, T_{iso})$ corresponds to the solution to the diffusion equation $\frac{\partial \phi}{\partial t} = \frac{1}{4N} \frac{\partial^2}{\partial p^2} (p(1-p)\phi)$, where p is the current frequency of the variant and p_0 the initial frequency of the variant (given in eq. 4.9 in KIMURA 1964).

We found numerically that using only the first 30 terms of the infinite series solution to the diffusion equation provides a good approximation for $T_{iso}/2d > 0.005$. Also, eq. 4-A.4 is accurate when $T_{iso}/2d$ is not large (from numerical simulations, eq. 4-A.4 is a good approximation for $T_{iso}/2d < 0.1$), as we only considered the temporal variations of the peaks and not the one of the full spectrum.

General properties of test statistics $T_{\Omega}^{I,t}$ and $T_{\Omega}^{II,t}$

Weights of the tests of neutrality from the total SFS under the isolation ($\Omega_i^{I,t}$) and the past isolation scenarios ($\Omega_i^{II,t}$) are presented in Figure 4-A.1. As expected from the signature under the isolation scenario, few positive Ω_i^I values exist (correspond to the expected peaks) and they have a strong weight. Under the past isolation scenario, the number of positive and negative $\Omega_i^{II,t}$ values is balanced, as peaks are wider, reflecting the signature under this scenario. Thus, test statistics $T_{\Omega}^{I,t}$ will have a larger variance than $T_{\Omega}^{II,t}$, which might reduce their power to reject neutrality.

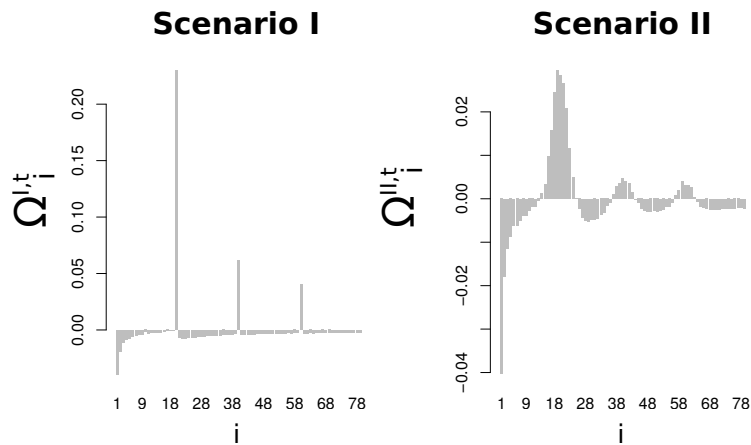


Figure 4-A.1 Weights of the optimal neutrality tests from the total SFS for the isolation ($\Omega_i^{I,t}$) and past isolation scenarios ($\Omega_i^{II,t}$), as a function of the class (i) of the variant in the total SFS. The weights were computed for $d = 4$ populations with sample sizes $n = 20$, with an isolation period of $6N$ generations and a scaled migration rate $M = 10$ during the connection periods.

General discussion

MEASURING GENETIC DIFFERENTIATION IN EQUILIBRIUM POPULATIONS

The results presented in this thesis improve our understanding of genetic differentiation. In particular, we saw that genetic diversity and the main features of allele frequency distributions follow a same transition from isolation to panmixia as a function of the migration rate. Different transition functions from isolation to panmixia were previously proposed (TAKAHATA and NEI 1984; JOST 2008), and their relevance to the theoretical concept of genetic differentiation is controversial (JOST 2008, 2009; WHITLOCK 2011). Nevertheless, they were based on the value of a differentiation measure (G_{ST} or D), and thus relied on the ability of differentiation measures to describe accurately genetic differentiation. This thesis reports for the first time an analytical description of such a transition function directly based on the allele frequency distributions, independent of all differentiation measures. We learn from the study of this transition function that the transition domain between isolation and panmixia is not very wide—approximately 2 orders of magnitude—and this transition domain can be in a range of low migration rates (when mutation is weak) or high migration rates (when mutation is strong). Thus, we predict that there are large domains of migration rates for which genetic diversities and allele frequency distributions are qualitatively the same.

We show that measures commonly used to describe the level of genetic differentiation of populations, G_{ST} and D (WEIR and COCKERHAM 1984; JOST 2008), accurately describe the level of genetic diversity only under specific mutation regimes. This result suggests that a single measure cannot summarize genetic differentiation under all mutation regimes. Instead, we show that using several measures capturing different aspects of genetic differentiation considerably increases the understanding of the genetic differentiation of populations.

Our results emphasize the usefulness of parametrized models to understand the behavior of differentiation measures, and of summary statistics in general. Indeed, previous model-free comparisons such as in HELLER and SIEGISMUND (2009) and JOST (2008) (table 1) did not allow a comprehensive interpretation of the values of G_{ST} and D . On the contrary, the results of this thesis provide a complete description of the behavior of the measures as a function of all model parameters.

THE LARGE IMPACT OF HISTORICAL EVENTS ON GENETIC DIVERSITY

Demographic changes and the dynamics of genetic diversity

The results of the thesis show that demographic changes, and in particular past isolation and re-connection of populations, can have strong genetic consequences. For a wide range of population size, number and migration rate, a population re-connection event (or migration increase) creates a peak and a large excess of genetic diversity within-populations. The temporal decay of this peak is low, which allows the excess to persist over large time-scales. Thus the thesis highlights a demographic event which enables to go beyond the theoretical amount of "alleles that can be maintained in a finite population" predicted by KIMURA and CROW (1964) in an isolated population, and by MARUYAMA (1970) and MAYNARD SMITH (1970) in a structured population. To the extent of my knowledge, no other demographic or selective event can generate such large amounts of genetic diversity within-populations in such a small period of time; indeed, population expansion and balancing selection can lead to large genetic diversity within-populations, but genetic diversity is limited by the rate of mutation, which tends to be low (TAJIMA 1989; SLATKIN 1996).

The results also show that when the demographic changes are periodic (with or without stochasticity in the period length), genetic diversity can have a variety of be-

haviors depending on its initial value and the period length. Periodic isolation and connection of populations can either lead to a large persistent excess of genetic diversity under small periods, lead to large variations of genetic diversity under large periods, or lead to a temporal accumulation or decay of genetic diversity (depending on its initial value) under intermediate periods. These results are in line with the literature on the impact of short time-scale migration fluctuations, which predicts a weak impact on genetic diversity (WHITLOCK 1992). In addition, these results fill a gap on the current literature concerning the dynamics of genetic diversity under intermediate and large-period migration changes. Finally, the results provide a simple interpretation for the limit between these three domains: the limit represents the ability of population to differentiate during the isolation periods (or periods of low migration).

Implications of demographic events on species conservation

Even in stable environments with equilibrium genetic diversity values, the results of the thesis show that it is difficult to assess correctly the genetic state of the populations and to provide recommendations for conservation and management of populations. The "one migrant per generation" (OMPG) rule is a perfect illustration. The OMPG rule is a rule of thumb used in conservation genetics, which states that one migrant per generation between populations is enough to maintain genetic diversity (e.g. reviewed in WANG 2004). The OMPG rule was formulated based on the equilibrium value of the G_{ST} statistics under the finite island model, and is based on the interpretation of G_{ST} as a differentiation measure. The assumptions of the OMPG rule were criticized (MILLS and ALLENDORF 1996; WHITLOCK and MCCAULEY 1999), nevertheless, this thesis brings a much deeper criticism of the rule. Even when the assumptions (finite island model at equilibrium) of the rule are met, genetic diversities can be very close

to what is expected in a panmictic population with much less, or much more than one migrant per generation. Thus we conclude that, while maintaining gene flow between endangered populations is important, no simple rule which does not take into account the mutation processes can be used to maintain or restore a maximum genetic diversity.

In addition, the thesis shows that assessing the genetic state of populations is even more difficult after demographic events with non-equilibrium genetic diversity. The assessment of the current effective size and the detection of recent population declines are of primary importance to design conservation initiatives (GILPIN and SOULE 1986; NEWMAN and PILSON 1997; JUMP *et al.* 2009). Although some estimators of population size are valid in non-equilibrium populations (e.g. HEY and NIELSEN 2004; JORDE and RYMAN 2007), studies often rely on standard measures of genetic diversity under equilibrium assumptions (e.g. θ estimators, gene diversity). We show that past isolation leads to an overestimation of the current effective size using such genetic diversity measures. In addition, past isolation can lead to false signals of recent population bottleneck even when the population size was stable for a long time (using neutrality tests such as Tajima's D , and skyline plot methods). Consequently, conservation studies should imperatively take into account the possible confounding effects of population structure before providing any recommendation.

Implications of demographic changes on species adaptation

Although the results of the thesis concern neutral genetic diversity, there is strong evidence that they are relevant for adaptation. Indeed, as noted in LENORMAND *et al.* (2009), "neutral mutations are in fact often conditionally neutral", and could provide a fitness advantage in another environment or with another genetic background. Similarly, GIBSON and DWORKIN (2004) describe "cryptic genetic variation": previously

neutral genetic variation that can suddenly provide a fitness advantage due to genome-genome interactions (epistasis) or genome-environment interactions (change of environment). In this context, the accumulation of large amounts of "neutral" genetic diversity could provide a tremendous advantage to a species to adapt to future environmental changes, allowing species to be prepared for many different situations.

Implications of demographic changes on species diversification

The large amounts of genetic diversity generated by a sudden connection event or short time-scale periodic isolation and connection events could play a role in species diversification. This hypothesis is supported by empirical evidence concerning the most famous examples of adaptive radiations. Darwin's finches, which were discovered by Charles Darwin during his second voyage of the Beagle (DARWIN 1909) and played a central part in his theory of evolution by natural selection (DARWIN 1859), underwent an adaptive radiation where episodes of hybridization between populations played a role (FREELAND and BOAG 1999; GRANT and GRANT 2008). Another famous example of adaptive radiation concerns the cichlid fishes of the great African lakes, where hundreds of species appeared within a few tens or hundred thousand years (BARRIER *et al.* 1999). There is evidence that connection events before the start of the radiation (BARRIER *et al.* 1999; BEZAULT *et al.* 2011), and periodic events later on (KELLER *et al.* 2013) provided large amounts of neutral and adaptive genetic variation which influenced the sympatric speciation. In other regions which harbor a very large biodiversity, population isolation and re-connection events due to geological and climatic events might have played a role in the diversification process (e.g. in the neotropics, RULL 2005; ANTONELLI *et al.* 2009; ANTONELLI and SANMARTÍN 2011; SEDANO and BURNS 2010).

Results of the thesis show that periodic isolation and connection events of interme-

diate length have a different impact on genetic diversity. Instead of providing a large excess of genetic diversity as a single connection event, they lead to the temporal accumulation or decrease of genetic diversity across cycles. In accordance with our results, AGUILEE *et al.* (2011) and AGUILÉE *et al.* (2013) showed with simulations that genetic diversity increased across isolation and connection cycles. Nevertheless, their results relied on divergent selection to maintain genetic diversity during the isolation periods, while we show that neutral mechanisms alone can lead to high levels of genetic diversity.

The large diversity of genetic consequences of a similar abiotic process might be due to the importance of the life-cycle (e.g. generation time) and the population parameters (e.g. size) of the species. In accordance with this theoretical prediction, it was shown that species with very different life-cycles which underwent the same glaciation cycles display very different genetic diversity levels (in Europe, the Meadow grasshopper shows 0.7-0.9% sequence divergence while hedgehog shows 6-12% divergence; HEWITT 2000).

We provide theoretical improvements to disentangle the mode of speciation using genomic data. The relative importance of sympatric versus allopatric speciation is still under debate (VIA 2001; FEDER *et al.* 2003; BUTLIN *et al.* 2008; FEDER *et al.* 2011). In addition, models with alternating allopatric and sympatric phases have been recently proposed to explain some speciation events (AGUILÉE *et al.* 2013; ROUX *et al.* 2013). How pervasive is this mode of speciation is an open question, and our results provide the first steps to address it.

INFERRING PAST EVENTS FROM GENOMIC DATA

The results of the thesis show that standard neutrality tests are difficult to interpret, which is due to the more general problem of model identifiability in statistical infer-

ence (i.e. the possibility to infer the true value of a parameter after an infinite number of observations WALTER and PRONZATO 1996). Indeed, even when sample sizes tend to infinity, it is unclear whether two different demographic models can be distinguished using genomic data (MYERS *et al.* 2008). When finite sample sizes are considered, we show that past isolation events in structured populations produce the same signature as many different scenarios (population size changes, migration rate changes) on neutrality tests.

Nevertheless, results of the thesis show that contrasting polymorphism patterns from several populations can be used to disentangle possible demographic scenarios in structured populations. Indeed, the distribution of pairwise nucleotide differences within and between populations and the site frequency spectra of several populations are shown to display signatures specific to past isolation events. In addition, results identify salient features of the signature which could be used as summary statistics for the inference of demographic changes, for example in an approximate bayesian computation framework (BEAUMONT *et al.* 2002).

FUTURE DIRECTIONS

The dynamics of genetic diversity

There is a need to better understand the behavior of measures of genetic differentiation under non-equilibrium scenarios. Indeed, the results of the thesis show that genetic diversity can stay in non-equilibrium states for a very long time. Consequently, the equilibrium assumption made in chapter 1 to describe the behavior of measures of genetic differentiation G_{ST} and D will often be violated. An interesting extension of this thesis would be to describe the behavior of genetic differentiation under non-equilibrium scenarios by starting from a panmictic population and simulating an isolation event.

Such a simulation was performed in JOST (2008); nevertheless, only a single example involving two populations was provided, and the impact of the model parameters was not assessed (mutation rate, number and size of populations). Recording the transition of allele frequency distributions and comparing it to that of G_{ST} and D would allow a quantification of their ability to describe allele frequency distributions, across the whole parameter space.

Another possible extension of the results of the thesis would be to investigate the spatial dynamics of genetic diversity after a reconnection event (secondary contact following a range expansion). Such a scenario would for example better represent the secondary contact zones established following the end of the last glaciation in Europe (HEWITT 2000). We expect from our result that the peak of genetic diversity generated by such events would temporally spread from the "epicenter" of the re-connection, creating a traveling wave of genetic diversity. How far would this wave go and what would be its expected spatial and temporal decay as a function of the migration model (dispersal kernel of the species) is of particular interest.

Genetic diversity and selection

Another interesting extension of the results of this thesis would be to investigate the impact of selection on the peak of genetic diversity generated by a re-connection event, either starting (i) before or (ii) after the re-connection event. Scenarios (i) and (ii) would be of strong biological relevance. (i) Isolated populations are often locally adapted to their environment. Such local adaptation can either reduce or increase the probability of fixation of a migrant allele (VUILLEUMIER *et al.* 2010), and is thus expected to strongly affect the size and duration of the peak of genetic diversity. (ii) The biotic or abiotic processes leading to population reconnection often result in new selective pres-

tures (ANTONELLI and SANMARTÍN 2011) or changes of genetic background which change the selective advantage of the alleles (GIBSON and REED 2008; SEEHAUSEN 2004). The exact impact of the strength of local adaptation on the size and duration of the peak in cases (i) and (ii) remain to be investigated.

Inference of complex scenarios

Our results suggest that using both G_{ST} and D enables to estimate the parameters of the finite island model under equilibrium assumptions. This illustrates the importance of having independent measures of genetic differentiation, and in general, of having non-redundant summary statistics to infer model parameters. Using G_{ST} and D as summary statistics in an ABC framework could thus improve the estimation of model parameters. Nevertheless, further investigations are needed to assess their usefulness to infer parameters under non-equilibrium scenarios.

The results of chapter 4 suggest the possibility of deriving F-statistics based on sequence data for non-equilibrium populations which underwent past isolation, allowing to infer the current and past migration rates and duration of the isolation period. HUDSON *et al.* (1992) showed that the F_{ST} derived from sequence data (contrary to G_{ST} , which only considers alleles) could provide estimates of the rate of gene flow between populations, assuming equilibrium demography. In the case of past isolation, we saw that the distribution of pairwise differences were bimodal, which would bias the estimator of gene flow from F_{ST} . Nevertheless, we also showed that the first mode of each distribution was close to its equilibrium distribution; consequently, it might be possible to derive an F statistics based on their first modes, which would provide estimates of the current level of gene flow even when the pairwise differences depart strongly from their equilibrium distribution. Similarly, it might be possible to derive informa-

tive statistics about the duration of the isolation period and the ancestral effective size from the comparison of the position of the modes within- and between-populations. Such statistics would be very useful for the inference of introgression rates from summary statistics, for example improving the ABC framework used in ROUX *et al.* (2013). This would be a promising tool to investigate the genomics of speciation, and identify genomic islands of speciation (FEDER *et al.* 2003).

Finally, the results of the thesis suggest that revisiting the parameter identifiability problem is very important for the future of statistical genomics. On the one hand, the infinite sample size assumptions made in traditional parameter identifiability studies (MYERS *et al.* 2008) prevents from understanding the relationship between sample size and parameter estimation precision. On the other hand, papers introducing a new method usually do quantify the relationship between sample size and parameter estimation (e.g. HEY and NIELSEN 2007; GUTENKUNST *et al.* 2009; EXCOFFIER *et al.* 2013), but their results are only relevant for a specific method. It would be very informative to quantify the theoretical relationship between sample size and parameter identifiability for a given model, independently of the statistical inference method used. Comparing these theoretical predictions with the performance of current methods would highlight their weaknesses, and it could highlight promising improvements for experimental design, as well as for future methodological and technological development.

BIBLIOGRAPHY

AGUILEE, R., A. LAMBERT, and D. CLAESSEN, 2011 Ecological speciation in dynamic landscapes. *J Evol Biol* **24**: 2663–2677.

AGUILÉE, R., D. CLAESSEN, and A. LAMBERT, 2013 Adaptive radiation driven by the interplay of eco-evolutionary and landscape dynamics. *Evolution* **67**: 1291–1306.

- ANTONELLI, A., J. A. A. NYLANDER, C. PERSSON, and I. SANMARTÍN, 2009 Tracing the impact of the andean uplift on neotropical plant evolution. *Proc Natl Acad Sci U S A* **106**: 9749–9754.
- ANTONELLI, A., and I. SANMARTÍN, 2011 Why are there so many plant species in the neotropics? *Taxon* **60**: 403–414.
- BARRIER, M., B. G. BALDWIN, R. H. ROBICHAUX, and M. D. PURUGGANAN, 1999 Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the hawaiian silversword alliance (asteraceae) inferred from floral homeotic gene duplications. *Mol Biol Evol* **16**: 1105–1113.
- BEAUMONT, M. A., W. ZHANG, and D. J. BALDING, 2002 Approximate bayesian computation in population genetics. *Genetics* **162**: 2025–2035.
- BEZAULT, E., S. MWAIKO, and O. SEEHAUSEN, 2011 Population genomic tests of models of adaptive radiation in lake victoria region cichlid fish. *Evolution* **65**: 3381–3397.
- BUTLIN, R. K., J. GALINDO, and J. W. GRAHAME, 2008 Review. sympatric, parapatric or allopatric: the most important way to classify speciation? *Philos Trans R Soc Lond B Biol Sci* **363**: 2997–3007.
- DARWIN, C., 1859 *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: Murray.
- DARWIN, C., 1909 *Voyage of the Beagle*. Courier Dover Publications.
- EXCOFFIER, L., I. DUPANLOUP, E. HUERTA-SÁNCHEZ, V. C. SOUSA, and M. FOLL, 2013 Robust demographic inference from genomic and snp data. *PLoS Genet* **9**: e1003905.

- FEDER, J. L., S. H. BERLOCHER, J. B. ROETHELE, H. DAMBROSKI, J. J. SMITH, *et al.*, 2003 Allopatric genetic origins for sympatric host-plant shifts and race formation in *rhagoletis*. *Proc Natl Acad Sci U S A* **100**: 10314–10319.
- FEDER, J. L., R. GEJJI, T. H. Q. POWELL, and P. NOSIL, 2011 Adaptive chromosomal divergence driven by mixed geographic mode of evolution. *Evolution* **65**: 2157–2170.
- FREELAND, J. R., and P. T. BOAG, 1999 The mitochondrial and nuclear genetic homogeneity of the phenotypically diverse darwin's ground finches. *Evolution* **53**: 1553–1563.
- GIBSON, G., and I. DWORKIN, 2004 Uncovering cryptic genetic variation. *Nature Reviews Genetics* **5**: 681–U11.
- GIBSON, G., and L. K. REED, 2008 Cryptic genetic variation. *Curr Biol* **18**: R989–R990.
- GILPIN, M., and M. SOULE, 1986 *Conservation Biology: The Science of Scarcity and Diversity*, chapter Minimum Viable Populations: Processes of Species Extinction. Sinauer, Sunderland, Mass, pp. 19–34.
- GRANT, B. R., and P. R. GRANT, 2008 Fission and fusion of darwin's finches populations. *Philos Trans R Soc Lond B Biol Sci* **363**: 2821–2829.
- GUTENKUNST, R. N., R. D. HERNANDEZ, S. H. WILLIAMSON, and C. D. BUSTAMANTE, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet* **5**: e1000695.
- HELLER, R., and H. R. SIEGISMUND, 2009 Relationship between three measures of genetic differentiation G_{ST} , D_{EST} and G'_{ST} : how wrong have we been? *Molecular Ecology* .
- HEWITT, G., 2000 The genetic legacy of the quaternary ice ages. *Nature* **405**: 907–913.

- HEY, J., and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**: 747–760.
- HEY, J., and R. NIELSEN, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A* **104**: 2785–2790.
- HUDSON, R. R., M. SLATKIN, and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.
- JORDE, P. E., and N. RYMAN, 2007 Unbiased estimator for genetic drift and effective population size. *Genetics* **177**: 927–935.
- JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.
- JOST, L., 2009 D vs. G_{ST} : Response to Heller and Siegmund (2009) and Ryman and Leimar (2009). *Molecular Ecology* **18**: 2088–2091.
- JUMP, A. S., R. MARCHANT, and J. PEÑUELAS, 2009 Environmental change and the option value of genetic diversity. *Trends Plant Sci* **14**: 51–58.
- KELLER, I., C. E. WAGNER, L. GREUTER, S. MWAIKO, O. M. SELZ, *et al.*, 2013 Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol Ecol* **22**: 2848–2863.
- KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
- LENORMAND, T., D. ROZE, and F. ROUSSET, 2009 Stochasticity in evolution. *Trends Ecol Evol* **24**: 157–165.

- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor Popul Biol* **1**: 273–306.
- MAYNARD SMITH, J., 1970 Population size, polymorphism, and the rate of non-darwinian evolution. *The American Naturalist* **104**: 231–237.
- MILLS, L. S., and F. W. ALLENDORF, 1996 The one-migrant-per-generation rule in conservation and management. *Conservation Biology* **10**: 1509–1518.
- MYERS, S., C. FEFFERMAN, and N. PATTERSON, 2008 Can one learn history from the allelic spectrum? *Theor Popul Biol* **73**: 342–348.
- NEWMAN, D., and D. PILSON, 1997 Increased probability of extinction due to decreased genetic effective population size: Experimental populations of *Clarkia pulchella*. *Evolution* **51**: 354–362.
- ROUX, C., G. TSAGKOGEOGA, N. BIERNE, and N. GALTIER, 2013 Crossing the species barrier: genomic hotspots of introgression between two highly divergent ciona intestinalis species. *Mol Biol Evol* **30**: 1574–1587.
- RULL, V., 2005 Biotic diversification in the guayana highlands: a proposal. *Journal of Biogeography* **32**: 921–927.
- SEDANO, R. E., and K. J. BURNS, 2010 Are the northern andes a species pump for neotropical birds? phylogenetics and biogeography of a clade of neotropical tanagers (aves: Thraupini). *Journal of Biogeography* **37**: 325–343.
- SEEHAUSEN, O., 2004 Hybridization and adaptive radiation. *Trends Ecol Evol* **19**: 198–207.
- SLATKIN, M., 1996 Gene genealogies within mutant allelic classes. *Genetics* **143**: 579–587.

- TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**: 585–595.
- TAKAHATA, N., and M. NEI, 1984 F and g statistics in the finite island model. *Genetics* **107**: 501–504.
- VIA, S., 2001 Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol Evol* **16**: 381–390.
- VUILLEUMIER, S., J. GOUDET, and N. PERRIN, 2010 Evolution in heterogeneous populations: from migration models to fixation probabilities. *Theor Popul Biol* **78**: 250–258.
- WALTER, E., and L. PRONZATO, 1996 On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and Computers in Simulation* **42**: 125–134.
- WANG, J., 2004 Application of the one-migrant-per-generation rule to conservation and management. *Conservation Biology* **18**: 332–343.
- WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating f-statistics for the analysis of population structure. *evolution* : 1358–1370.
- WHITLOCK, M. C., 1992 Temporal fluctuations in demographic parameters and the genetic variance among populations. *Evolution* **46**: pp. 608–615.
- WHITLOCK, M. C., 2011 G'_{ST} and D do not replace F_{ST} . *Molecular Ecology* **20**: 1083–1091.
- WHITLOCK, M. C., and D. E. MCCAULEY, 1999 Indirect measures of gene flow and migration: F_{ST} not equal to $1/(4nm + 1)$. *Heredity (Edinb)* **82 (Pt 2)**: 117–125.

Acknowledgements

First and foremost, I am indebted to my PhD supervisor, Dr. Séverine Vuilleumier. Séverine inspired me to think creatively, without ever sacrificing scientific rigor. Her natural easiness with the grasping of conceptual problems and with synthesis proved invaluable at every step of the formulation, analysis and discussion of our studies. In addition, she shared her enthusiasm for good science, and I could always count on her unfailing support. She has provided someone for me to look up to.

Next, I would like to thank the experts on my thesis, who kindly agreed to evaluate this work –Professors Daniel Wegmann and Laurent Lehmann. I also thank Laurent for his advices in the presentation of mathematical results. I am also grateful to the president of my thesis Pr. Pierre Goloubinoff.

I would like to acknowledge Professor Jérôme Goudet, for his contribution to chapters 1 and 2 as well as for pleasant and insightful discussions about science and other things; attending Jérôme’s lab meetings made me dive into population genomics and helped me understand the current challenges of the field. I also acknowledge Jeffrey D. Jensen and Amalio Telenti for their contribution to chapter 4 and give them my most sincere gratitude for making me feel at home in their respective labs, and for very constructive comments and discussions. I also thank Jeff for his strong support during the thesis and beyond. Finally, I would like to acknowledge the contribution of Daniela Streit for Chapter 2, and for providing results which inspired a large part of the thesis.

I express my gratitude to the Swiss National Science Foundation and the department of Ecology and Evolution of the University of Lausanne for their financial support during my thesis. I also thank the Société Académique Vaudoise and the fondation du 450e anniversaire de l’université de Lausanne for their financial support.

I thank the members of the department of Ecology and Evolution, and in particular my office mates –past and present–, for making DEE the fun and intellectually rich environment it is. I thank my parents Christine and Luis, for always believing in me, and telling me about Switzerland in the first place. I thank my brother Thomas, for his boasting about his little brother since we were children; sometimes I even fall for it myself. I thank my friends, from high school, university and before, from Switzerland, France and elsewhere.

Finally, I thank Karine, who provided the emotional support for this thesis. Among many other things, I would like to thank her for always reminding me of my strengths and flaws, and sometimes making the latter the former. Her infallible optimism inspires me to be better, as a scientist and as a person.

Appendix A Supplementary Files for chapter 1

Nicolas Alcalá¹, Jérôme Goudet¹, and Séverine Vuilleumier^{1,2}

¹Department of Ecology and Evolution, Biophore, University of Lausanne, CH-1015 Lausanne, Switzerland

²Institute of Microbiology, University Hospital Center and University of Lausanne (CHUV-UNIL), CH-1011 Lausanne, Switzerland

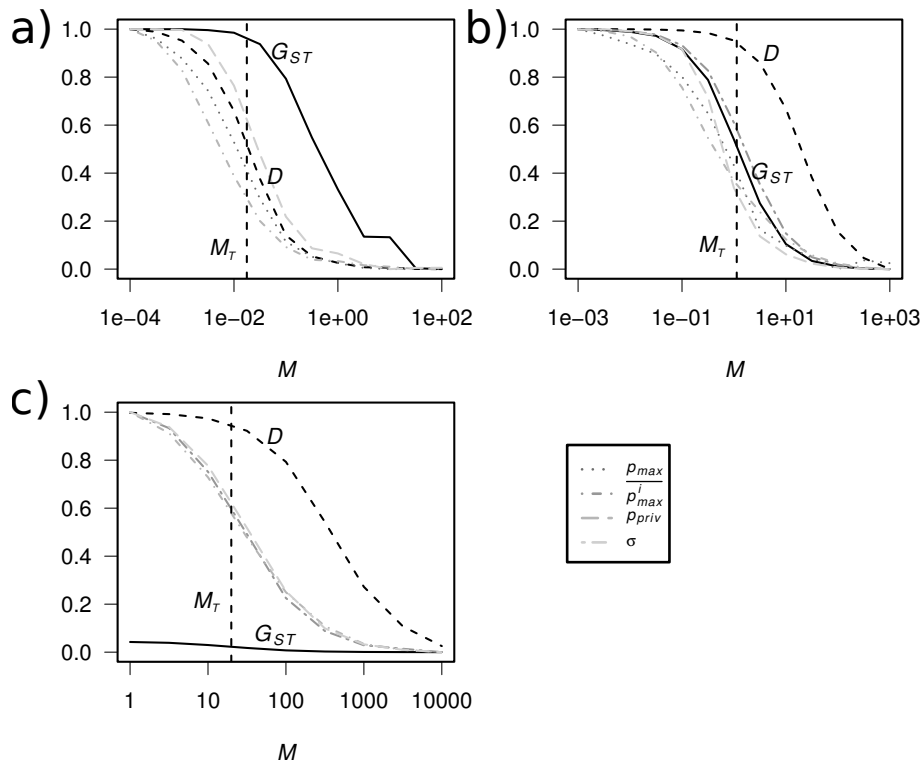


Figure S1 Comparison of the behaviour of G_{ST} and D with four normalized statistics of allelic frequencies (as for f in fig. 2) as a function of the scaled migration rate M . Results are presented for (a) weak mutation regime, (b) intermediate mutation regime and (c) strong mutation regime. We can see that D and G_{ST} follow the behaviour of the normalized statistics, respectively under the weak and intermediate mutation regime. Parameters are the same as in fig. 2.

$M = 0$	<i>pop. 1</i>	<i>pop. 2</i>	<i>pop. 3</i>	<i>pop. 4</i>	<i>pop. 5</i>	<i>pop. 6</i>	<i>pop. 7</i>	<i>pop. 8</i>	<i>pop. 9</i>	<i>pop. 10</i>	$G_{ST} \simeq 0.94$ $D = 1$
	1	0	0	0	0	0	0	0	0	0	
	0	0.994	0	0	0	0	0	0	0	0	
	0	0.006	0	0	0	0	0	0	0	0	
	0	0	0.998	0	0	0	0	0	0	0	
	0	0	0.002	0	0	0	0	0	0	0	
	0	0	0	1	0	0	0	0	0	0	
	0	0	0	0	1	0	0	0	0	0	
	0	0	0	0	0	1	0	0	0	0	
	0	0	0	0	0	0	1	0	0	0	
	0	0	0	0	0	0	0	1	0	0	
	0	0	0	0	0	0	0	0	1	0	
	0	0	0	0	0	0	0	0	0	0.447	
	0	0	0	0	0	0	0	0	0	0.553	
$M = M_T$	<i>pop. 1</i>	<i>pop. 2</i>	<i>pop. 3</i>	<i>pop. 4</i>	<i>pop. 5</i>	<i>pop. 6</i>	<i>pop. 7</i>	<i>pop. 8</i>	<i>pop. 9</i>	<i>pop. 10</i>	$G_{ST} \simeq 0.83$ (!) $D \simeq 0.50$
	1	1	1	1	0.955	0.33	0.181	0	0	0	
	0	0	0	0	0.045	0.67	0.819	1	1	1	
$M = M_G$	<i>pop. 1</i>	<i>pop. 2</i>	<i>pop. 3</i>	<i>pop. 4</i>	<i>pop. 5</i>	<i>pop. 6</i>	<i>pop. 7</i>	<i>pop. 8</i>	<i>pop. 9</i>	<i>pop. 10</i>	$G_{ST} \simeq 0.49$ (!) $D \simeq 0.11$
	1	1	1	1	1	0.999	0.999	0.943	0.797	0.317	
	0	0	0	0	0	0.001	0.001	0.057	0	0.674	
	0	0	0	0	0	0	0	0	0.203	0.009	
$M = 4N \frac{n-1}{n}$	<i>pop. 1</i>	<i>pop. 2</i>	<i>pop. 3</i>	<i>pop. 4</i>	<i>pop. 5</i>	<i>pop. 6</i>	<i>pop. 7</i>	<i>pop. 8</i>	<i>pop. 9</i>	<i>pop. 10</i>	$G_{ST} \simeq 0.00$ $D \simeq 0.00$
	0.98	0.971	0.968	0.964	0.964	0.963	0.956	0.956	0.954	0.95	
	0.02	0.029	0.032	0.036	0.036	0.037	0.044	0.044	0.046	0.05	

Figure S2 Representative examples of simulated allele distributions under the weak mutation regime, for different scaled migration rates M . To ensure the representativeness of the examples, for each value of M , the examples are chosen such that G_{ST} and D values are both close (less than 5% absolute difference) to their expected value given the value of M (from eq. 1.9 and 1.10). $D \simeq 0.5$ when approximately half populations have similar allele frequencies, while $G_{ST} \simeq 0.5$ when almost all populations have similar allele frequencies. Distributions were simulated using coalescent program *msms* (EWING and HERMISSON 2010). Sample sizes are 1000 per population; parameters are $n = 10, \theta = 0.0173$.

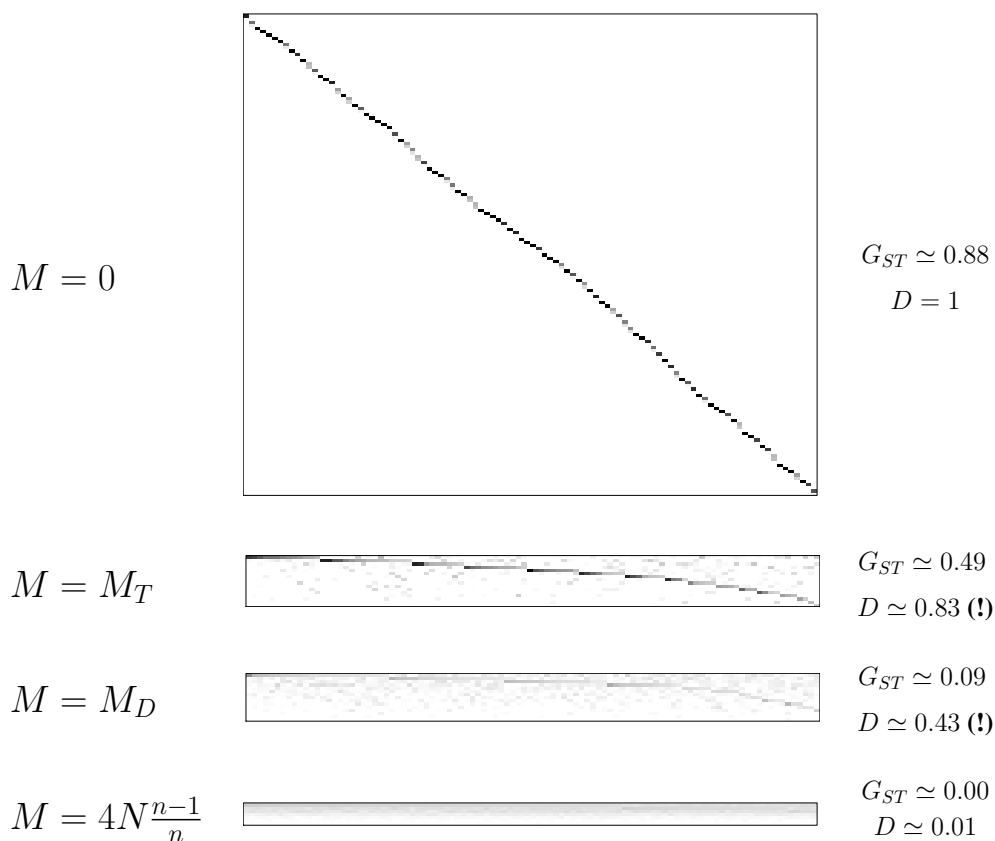


Figure S3 Representative examples of simulated allele distributions under the intermediate mutation regime, for different scaled migration rates M . To ensure the representativeness of the examples, for each value of M , the examples are chosen such that G_{ST} and D values are both close (less than 5% absolute difference) to their expected value given the value of M (from eq. 1.9 and 1.10). $G_{ST} \simeq 0.5$ when alleles are present in several populations (genetic structure is still visible), while $D \simeq 0.5$ when almost all populations have similar allele frequencies (genetic structure is barely visible). Distributions were simulated using coalescent program *msms* (EWING and HERMISSON 2010). Sample sizes are 1000 per population; parameters are $n = 100$, $\theta = 0.1$.

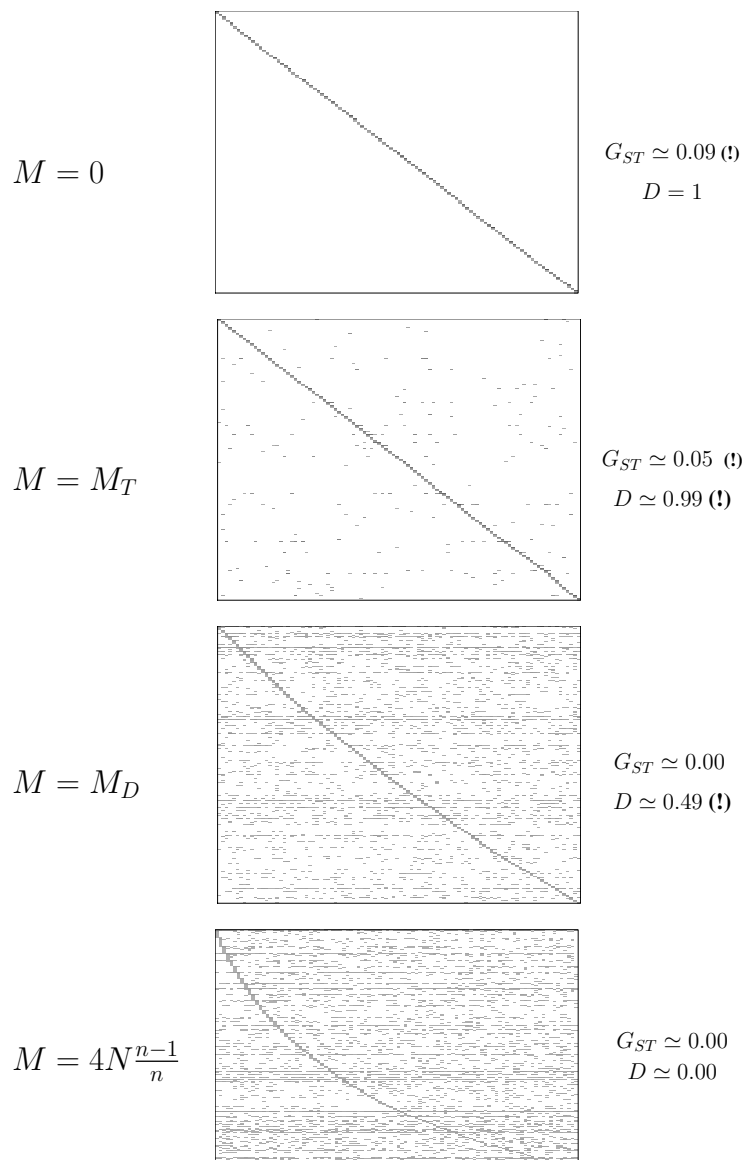


Figure S4 Representative examples of simulated allele distributions under the strong mutation regime, for different scaled migration rates M . To ensure the representativeness of the examples, for each value of M , the examples are chosen such that G_{ST} and D values are both close (less than 5% absolute difference) to their expected value given the value of M (from eq. 1.9 and 1.10). In this regime, $G_{ST} \simeq 0$ for all M values; $D \simeq 0.5$ when most alleles are present in multiple populations (allele frequencies are close to what is expected under panmixia, i.e. when $M = 4N \frac{n-1}{n}$). We can see that even under panmixia (row $M = 4N \frac{n-1}{n}$), allele frequencies are expected to be quite different between populations, as we expect sampling effects (see fig. 1.5) due to low allele frequencies (maximum allele frequency is 0.012). Distributions were simulated using coalescent program *msms* (EWING and HERMISSON 2010). Sample sizes are 1000 per population; parameters are $n = 100$, $\theta = 10$.

APPENDIX S1: The relationship between $f(M)$ and the effective number of alleles

Using the expected gene diversity H in an isolated population, and using the relationship $\Delta = 1/(1 - H)$ (KIMURA and CROW 1964), we can derive the expected equilibrium within-population effective number of alleles in an isolated population $\Delta_S^{iso} = 1 + \theta$ and the expected equilibrium total effective number of alleles of n isolated populations $\Delta_T^{iso} = n(1 + \theta)$. Similarly, in a panmictic population of size nN , the expected within-population effective number of alleles and total effective number of alleles reach the same value $\Delta^{pan} = 1 + n\theta$. We derive the equilibrium within-population and total effective number of alleles, Δ_S and Δ_T , in the finite island model from the results of MARUYAMA (1970). We can then describe the degree to which Δ_S (resp. Δ_T) is close to isolation or panmixia using the function of migration f_{Δ_S} (resp. f_{Δ_T}):

$$\begin{cases} f_{\Delta_S}(M) = \frac{\Delta_S - \Delta^{pan}}{\Delta_S^{iso} - \Delta^{pan}} \\ f_{\Delta_T}(M) = \frac{\Delta_T - \Delta^{pan}}{\Delta_T^{iso} - \Delta^{pan}} \end{cases} \quad (\text{S1.1a})$$

where

$$\begin{cases} f_{\Delta_S}(M) = \frac{M}{M + M'_T} \\ f_{\Delta_T}(M) = \frac{M}{M + M''_T} \end{cases} \quad (\text{S1.1b})$$

are monotonic functions of the migration rate, with two thresholds M'_T and M''_T :

$$\begin{cases} M'_T = (n - 1)\theta \\ M''_T = \frac{(n - 1)\theta}{n} \end{cases} \quad (\text{S1.1c})$$

The relative values of M and M'_T (resp. M''_T) describe the relative influence of Δ_S^{iso} and Δ^{pan} (res. Δ_T^{iso} to H^{pan}) on the value of Δ_S (resp. Δ_T). When $M \ll M'_T$ effective number of alleles are close to their equilibrium at isolation values $\Delta_S \simeq \Delta_S^{iso}$, and when $M \ll M''_T$, $\Delta_T \simeq \Delta_T^{iso}$. When $M \gg M'_T$ (resp. $M \gg M''_T$), Δ_S (resp. Δ_T) is close to its equilibrium at panmixia Δ^{pan} .

Functions $f_{\Delta_S}(M)$ and $f_{\Delta_T}(M)$ have two interesting properties. First, the thresholds M'_T and M''_T are both different from the threshold M_T (eq. 1.4), meaning that the absolute values of the effective number of alleles and gene diversities are not sensitive to the same mechanisms of differentiation. Indeed, from eq. S1.1, Δ_S depends on the relative values of the total number of mutants $(n - 1)\theta$ and the number of migrants M per generation, and Δ_T depends mainly on the relative values of the number of mutants θ and the number of migrants M per generation. In contrast, gene diversities depend on a more complicated relationship between n , θ and M (see eq. 1.4a). Second, the thresholds M'_T and M''_T are also different from one another. Therefore, neither $f_{\Delta_S}(M)$ nor $f_{\Delta_T}(M)$ describe alone the transition between the isolation state to the panmictic state. We then investigate how they are related to $f(M)$. From the definition of $f(M)$ (eq. 1.4a) and using the relationships $\Delta_S = (1 - H_s)^{-1}$ and $\Delta_T = (1 - H_t)^{-1}$ (JOST 2008), we find a direct relationship between both within-population and total effective

number of alleles and function $f(M)$:

$$\begin{aligned} f(M) &= \frac{(\Delta_S - \Delta^{pan})/\Delta_S}{(\Delta_S^{iso} - \Delta^{pan})/\Delta_S^{iso}} \\ &= \frac{(\Delta_T - \Delta^{pan})/\Delta_T}{(\Delta_T^{iso} - \Delta^{pan})/\Delta_T^{iso}} \end{aligned} \tag{S1.2}$$

Thus $f(M)$ also measures the transition of Δ_S and Δ_T values from their expected isolation to panmictic values. Indeed, we have $f(M) = 0$ when $\Delta_S = \Delta_T = \Delta^{pan}$, and $f(M) = 1$ when $\Delta_S = \Delta_S^{iso}$ and $\Delta_T = \Delta_T^{iso}$. The absolute value of the numerator in eq. S1.2, $|\frac{\Delta_S - \Delta^{pan}}{\Delta_S}|$ (resp. $|\frac{\Delta_T - \Delta^{pan}}{\Delta_T}|$), measures the distance of Δ_S (resp. Δ_T) from its expected value under panmixia, Δ^{pan} , scaled by its value Δ_S (resp. Δ_T). The denominator in S1.2, $\frac{\Delta_S^{iso} - \Delta^{pan}}{\Delta_S^{iso}}$ (resp. $|\frac{\Delta_T^{iso} - \Delta^{pan}}{\Delta_T^{iso}}|$), is the maximum absolute value of its numerator, which is reached when $\Delta_S = \Delta_S^{iso}$ (resp. $\Delta_T = \Delta_T^{iso}$).

APPENDIX S2: The use of singular value decomposition to study genetic differentiation

Given a rectangular matrix A of dimension $k \times n$, that represents the populations (column n) and the different alleles in each population (row k), the matrix A can be decompose as follows (singular value decomposition (SVD)):

$$A = U\Sigma V' \quad (\text{S2.1})$$

where Σ is an $k \times n$ rectangular diagonal matrix with positive or null values on the diagonal (called the singular values of A), and U is a $k \times k$ matrix and V' is the transpose of V that is a $n \times n$ matrix. The columns of U and V are the left-singular vectors and the right-singular vectors of A , respectively. The number of non-zero singular values of Σ is the rank of the matrix A . We can also compute the *effective* rank of a matrix, i.e. the number of "approximately independent" columns with the number of "approximately zero" singular values values (up to a rounding error). The latter value is of particular interest as it gives the number of approximately independent columns, which is, in our case, with matrix A , the number of populations that are genetically independent.

Moreover, singular values of A summarize the pairwise gene identities between population pairs. They are equal to the square root of the eigenvalues of matrix AA' , with A' the transpose of A . AA' is the "pairwise identity matrix" of dimension $n \times n$. In the matrix AA' , the row i and column j are the populations and the value ij , with $i \neq j$, are the between-population identity between populations i and j and value ij , with $i = j$, are the within-population identity (on the diagonal).

In summary, it is possible to use singular values of A to determine how independent are the populations and to decompose the pairwise identities in matrix AA' .

A simple informative value can then be derived from singular values: the mean singular value, σ . When population are in panmixia and all the population have the same allele at the same frequencies, σ value is the square root of the within-population gene identity divided by the number of populations, $\sqrt{F_s/n}$ (square root of trace of AA' divided by n). When population are isolated, σ value is the mean of the square roots of the within-population gene identities in each populations $\sum_i \sqrt{F_{s,i}/n}$.

Illustrative examples

Now let us consider the 3 following allele frequency tables:

	Species A		Species B		Species C		
	pop. 1	pop. 2	pop. 1	pop. 2	pop. 1	pop. 2	pop. 3
allele 1	0.5	0.5	0.5	0	0.5	0	0.25
allele 2	0.5	0.5	0.5	0	0.5	0	0.25
allele 3			0	0.5	0	0.5	0.25
allele 4			0	0.5	0	0.5	0.25

For species A, both populations have the same allele frequencies, thus it corresponds to 1 independent population, thus $\sigma_1 = 1$ and $\sigma_2 = 0$. For species B, populations have non-overlapping sets of alleles and they have the same within-population gene identities, we have $\sigma_1 = \sigma_2 = \sqrt{0.5}$. For species C, there are 2 populations with a non-overlapping set of alleles and the third population is a mixture of them, we have $\sigma_1 = \sqrt{0.75}$, $\sigma_2 = \sqrt{0.5}$ and $\sigma_3 = 0$.

BIBLIOGRAPHY

- EWING, G., and J. HERMISSON, 2010 Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064–2065.
- JOST, L., 2008 G_{ST} and its relatives do not measure differentiation. *Molecular Ecology* **17**: 4015–4026.
- KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.
- MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theoretical Population Biology* **1**: 273–306.

Appendix B Supplementary Files for chapter 2

Nicolas Alcalá, Daniela Streit, Jérôme Goudet, and Séverine Vuilleumier
Department of Ecology and Evolution, Biophore, University of Lausanne, CH-1015
Lausanne, Switzerland

FILE S1: GENETIC DIVERSITY EQUILIBRIUM

In this Supporting Information File, we provide for eq. 2.3 the conditions of existence of a genetic diversity equilibrium value, we derive its value and finally determine its stability.

Condition of existence of an equilibrium:

Eq. 2.3 is a non-homogeneous linear matrix recurrence equation, with an additive constant vector \mathbf{B} . If $\mathbf{I} - \mathbf{A}$ is invertible, there is a unique equilibrium value for eq. 2.3:

$$\mathbf{F}^{eq} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \quad (\text{S1.1})$$

and from any initial value \mathbf{F}_0 , genetic identities after t generations follow:

$$\mathbf{F}_t = \mathbf{A}^t(\mathbf{F}_0 - \mathbf{F}^{eq}) + \mathbf{F}^{eq} \quad (\text{S1.2})$$

$\mathbf{I} - \mathbf{A}$ is invertible if and only if the determinant of $\mathbf{I} - \mathbf{A}$ is not null, which is true whenever $\mu \neq 0$. This condition is always met in our model.

Equilibrium value:

We can express the equilibrium value S1.1 as a function of the migration rate m , mutation rate μ , population size N and number of populations n :

$$\begin{aligned} \mathbf{F}^{eq} &= (\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \\ &= \begin{pmatrix} 1 - a(1-c)(1-\mu)^2 & -(1-a)(1-\mu)^2 \\ -b(1-c)(1-\mu)^2 & 1 - (1-b)(1-\mu)^2 \end{pmatrix}^{-1} (1-\mu)^2 \begin{pmatrix} ac \\ bc \end{pmatrix} \\ &= \frac{1}{ac(\frac{1}{(1-\mu)^2} - 1) + bc + (\frac{1}{(1-\mu)^2} - 1)(\frac{1}{(1-\mu)^2} + b - a)} \begin{pmatrix} ac(\frac{1}{(1-\mu)^2} - 1) + bc \\ \frac{bc}{(1-\mu)^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{1 + \frac{(2-\mu)}{c(1-\mu)^2}\mu \frac{(2-\mu)}{a} \frac{(2-\mu)}{(1-\mu)^2}\mu + b} \\ \frac{b}{b(1-\mu)^2 + (2-\mu)\mu(a+c^{-1}(\frac{(2-\mu)}{(1-\mu)^2}\mu + nb))} \end{pmatrix} \end{aligned}$$

When mutation rates and migration rates are small, $2-\mu$ is close to 2, $(1-\mu)^2$ is close to 1, $a = (1-m)^2 + \frac{m^2}{n-1}$ is close to $1-2m$, and $b = \frac{1-a}{n-1}$ is close to $\frac{2m}{n-1}$. Therefore, taking into account that $c = \frac{1}{2N}$, the expression above further simplifies to:

$$\mathbf{F}^{eq} = \begin{pmatrix} \frac{1}{1 + 4N\mu \frac{2\mu + 2m \frac{n}{n-1}}{2\mu + \frac{2m}{n-1}}} \\ \frac{\frac{2m}{n-1}}{\frac{2m}{n-1} + 2\mu(1-2m+2N(2\mu + 2m \frac{n}{n-1}))} \end{pmatrix}$$

Denoting $\theta = 4N\mu$ and $M = 4Nm$, for the scaled mutation and migration rate, respectively, the above expression can be described as:

$$\mathbf{F}^{eq} = \begin{pmatrix} \frac{1}{1 + \theta(1 + \frac{M}{\theta + \frac{M}{n-1}})} \\ \frac{M}{M + (n-1)\theta(1 + \theta + \frac{n}{n-1}M)} \end{pmatrix} \quad (\text{S1.3})$$

which is the equilibrium value that was first derived by MARUYAMA (1970). We can then derive the genetic diversities $\mathbf{H}_t = \begin{pmatrix} H_{s,t} \\ H_{b,t} \end{pmatrix} = \mathbf{1} - \mathbf{F}_t$, and the equilibrium genetic diversities $\mathbf{H}^{eq} = \mathbf{1} - \mathbf{F}^{eq}$.

Stability of the equilibrium:

Equilibrium S1.1 is stable if and only if $\lim_{t \rightarrow \infty} \mathbf{A}^t = 0$, thus if and only if the absolute value of all eigenvalues of matrix \mathbf{A} are below 1.

The eigenvalues of matrix \mathbf{A} are the roots of the characteristic equation $\chi_{\mathbf{A}}(T) = T^2 - tr(\mathbf{A})T + det(\mathbf{A})$. They are presented in eq. 2.5. As $4bc \geq 0$ for all values of m, n and N , $\sqrt{(1 - a(1 - c) + b)^2 - 4bc} \leq 1 - a(1 - c) + b$. Thus from eq. 2.5, $\lambda_1 \leq (1 - \mu)^2$. λ_1 is below 1 for all mutation rates strictly less than 1. As $\lambda_1 < 1$ and $\lambda_2 < \lambda_1$, both eigenvalues are below 1. In addition, λ_1 and λ_2 are positive for any possible migration rate, mutation rate, size and number of populations. Therefore $0 < \lambda_1 < 1$ and $0 < \lambda_2 < 1$, which proves that equilibrium S1.1 is always stable, and is reached whatever the initial identity \mathbf{F}_0 considered.

FILE S2: EFFECT OF AN ISOLATION EVENT ON GENETIC DIVERSITY

We show in this section that following an isolation event, within- and between-population genetic diversities change independently and that their times to reach equilibrium values are t_2 and t_1 , respectively. When populations are isolated ($m = 0$), eq. 2.2 simplifies to:

$$\begin{cases} F_{s,t+1} = (c + (1-c)F_{s,t})(1-\mu)^2 \\ F_{b,t+1} = F_{b,t}(1-\mu)^2 \end{cases} \quad (\text{S2.1})$$

In eq. S2.1, $F_{s,t+1}$ depends only on $F_{s,t}$ but not on $F_{b,t}$; similarly, $F_{b,t+1}$ depends only on $F_{b,t}$ but not on $F_{s,t}$. Thus, $F_{s,t}$ and $F_{b,t}$ both follow a one dimensional recurrence equation:

$$\begin{cases} F_{s,t} = F_s^{eq} + (F_{s,0} - F_s^{eq})((1-c)(1-\mu)^2)^t \\ F_{b,t} = F_{b,0}(1-\mu)^{2t} \end{cases} \quad (\text{S2.2})$$

Therefore, when populations are isolated, within- and between-population genetic diversities change according to $(1-c)(1-\mu)^2$ and $(1-\mu)^2$, respectively. As when $m=0$, $\lambda_1=(1-\mu)^2$ and $\lambda_2=(1-c)(1-\mu)^2$ (from eq. 2.5), we can rewrite S2.2 as:

$$\begin{cases} F_{s,t} = F_s^{eq} + (F_{s,0} - F_s^{eq})\lambda_2^t \\ F_{b,t} = F_{b,0}\lambda_1^t \end{cases} \quad (\text{S2.3})$$

We can conclude that when populations are isolated, F_s and F_b change according to λ_2^t and λ_1^t , respectively. This demonstrates that after isolation, within- and between-population genetic diversities reach their equilibrium value in t_2 and t_1 generations, respectively.

FILE S3: RELAXING THE COMPLETE ISOLATION HYPOTHESIS

Throughout the study, we consider a complete isolation event, where populations that were previously connected suddenly become completely isolated, and a reconnection event where previously completely isolated populations suddenly become connected. Nevertheless, we show in this Supporting Information File that the assumption of complete isolation can be relaxed. Indeed, the results for complete isolation are a very good approximation of results for a strong but incomplete isolation.

Equilibrium genetic identity under incomplete isolation:

Incomplete isolation corresponds to a state where populations are connected through migration at a rate $0 < m \ll 1$. We characterize in this section the threshold value of migration rate under which the complete isolation equilibrium genetic identity is a good approximation of the incomplete isolation equilibrium value.

To do so, we can rewrite eq. S1.3 under the following form (isolating terms in M):

$$\begin{aligned} \mathbf{F}^{eq} &= \begin{pmatrix} \frac{\theta + \frac{M}{n-1}}{\theta + \frac{M}{n-1} + \theta(\theta + \frac{M}{n-1} + M)} \\ \frac{M}{M + (n-1)\theta(1+\theta) + n\theta M} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{1+\theta} + \left(\frac{1}{1+n\theta} - \frac{1}{1+\theta}\right) \frac{M}{M + (n-1)\theta \frac{1+\theta}{1+n\theta}} \\ \left(\frac{1}{1+n\theta}\right) \frac{M}{M + (n-1)\theta \frac{1+\theta}{1+n\theta}} \end{pmatrix} \end{aligned}$$

We know that an isolated population of size N has a within-population genetic identity equilibrium value of $\frac{1}{1+\theta}$ (KIMURA and CROW 1964), and a between-population equilibrium value of 0. Therefore, the genetic identity of isolated populations at equilibrium is $\mathbf{F}^{iso} = \begin{pmatrix} \frac{1}{1+\theta} \\ 0 \end{pmatrix}$. The equilibrium within- and between-population genetic identity of a panmictic population of size nN is $\mathbf{F}^{pan} = \begin{pmatrix} \frac{1}{1+n\theta} \\ \frac{1}{1+n\theta} \end{pmatrix}$ (KIMURA and CROW 1964). Eq. S1.3 can be written:

$$\mathbf{F}^{eq} = \mathbf{F}^{iso} + (\mathbf{F}^{pan} - \mathbf{F}^{iso})f_{n;\theta}(M) \quad (\text{S3.1a})$$

with

$$f_{n;\theta}(M) = \frac{M}{M + (n-1)\theta \frac{1+\theta}{1+n\theta}} \quad (\text{S3.1b})$$

$f_{n;\theta}(M)$ is similar to a Michaelis-Menten function (in the form $f(M) = vM/(M + M_T)$), with a maximum value $v=1$ and a threshold $M_T = (n-1)\theta \frac{1+\theta}{1+n\theta}$ (MICHAELIS and MENTEN 1913). We can predict by analogy to the Michaelis-Menten function that the behavior of the function depends on the relative value of M and M_T , with in our case a threshold value of:

$$M_T = (n-1)\theta \frac{1+\theta}{1+n\theta} \quad (\text{S3.2})$$

If M is much below the threshold, $\mathbf{F}^{eq} \simeq \mathbf{F}^{iso}$: the genetic identity equilibrium is close to the isolation equilibrium. If M is larger than the threshold, $\mathbf{F}^{eq} \simeq \mathbf{F}^{pan}$: the genetic identity equilibrium is close to the panmictic equilibrium. When $M = M_T$, $\mathbf{F}^{eq} = \frac{\mathbf{F}^{iso} + \mathbf{F}^{pan}}{2}$: the genetic identity equilibrium is the mean of the panmictic and isolation equilibria. Fig. S3 illustrates the variations of \mathbf{F}_s^{eq} and \mathbf{F}_b^{eq} as a function of M .

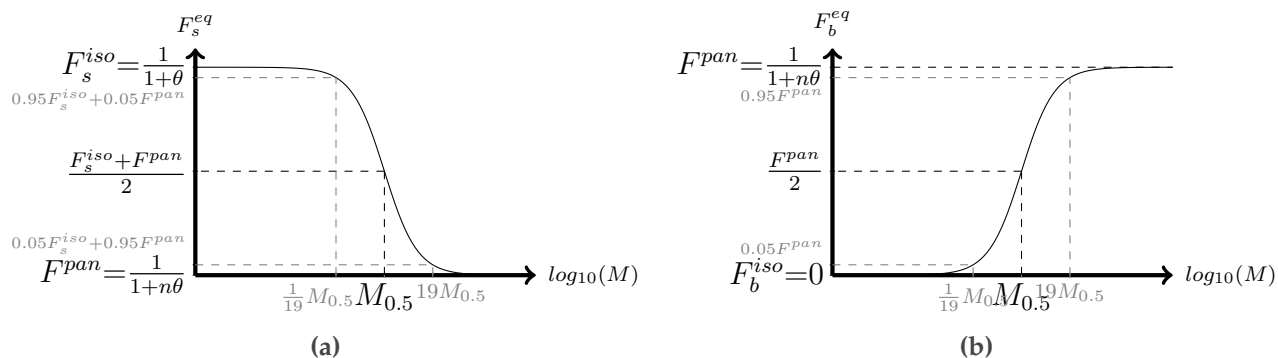


Figure S1 (A) Within-population F_s^{eq} , and (B) between-population F_b^{eq} genetic identity, as a function of the strength of migration M (scaled migration rate). F_s^{iso} and F_b^{pan} are the expected equilibria when all populations are isolated or panmictic, respectively. The inflexion point (transition between panmictic and isolation equilibrium) is $M_T = (n-1)\theta / (1+n\theta)$ for both within- and between-population genetic identity. For $M < M_{0.05} = \frac{0.05}{0.95} M_T$ (M_T is 19 fold $M_{0.05}$), the genetic identity equilibrium is close to the isolation equilibrium.

Peak of genetic identity generated by a migration rate increase:

We now consider an event of abrupt migration rate increase (from scaled rate M_0 to $M > M_0$). In this case, eq. 2-B.5 becomes:

$$\Delta H = - \left[\frac{(F_s^{eq} - F_s^{pan})}{n-1} \frac{M \frac{n}{n-1}}{1 + M \frac{n}{n-1}} - F_b^{pan} \right] (f(M) - f(M_0)) \left(\frac{\frac{M}{1 + \frac{n}{n-1} M - \frac{2N}{N_e}}}{1 + M - \frac{N}{N_e}} \right) 0.05 \frac{t_2}{t_1} \quad (S3.3)$$

The value of the peak of diversity depends on the relative value of M_0 and the threshold M_T . Indeed, from eq. S3.1b and S3.3 we can see that $f(M_0) \simeq 0$ and eq. S3.3 and 2-B.5 equalize when $M_0 \ll M_T$. Thus, an increase in migration rate above the threshold produces the same peak as a reconnection event. This demonstrates that genetic diversity peaks are generated whenever the scaled migration rate increases abruptly and crosses the threshold value M_T . Thus, an increase in migration rate from a small M_0 to M produces approximately the same peak of genetic diversity as an increase from 0 to M ($M_0 < \frac{0.05}{0.95} M_T$ ensures that $f(M_0) < 0.05$).

BIBLIOGRAPHY

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725–738.

MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. *Theor Popul Biol* **1**: 273–306.

MICHAELIS, L., and M. L. MENTEN, 1913 The kinetics of the inversion effect. *Biochemische Zeitschrift* **49**: 333–369.

Appendix C Supplementary Files for chapter 4

Nicolas Alcalá¹, Jeffrey D. Jensen², Amalio Telenti³, Séverine Vuilleumier^{1,3}

¹ Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland

² School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

³ Institute of Microbiology, University Hospital and University of Lausanne, CH-1011 Lausanne, Switzerland

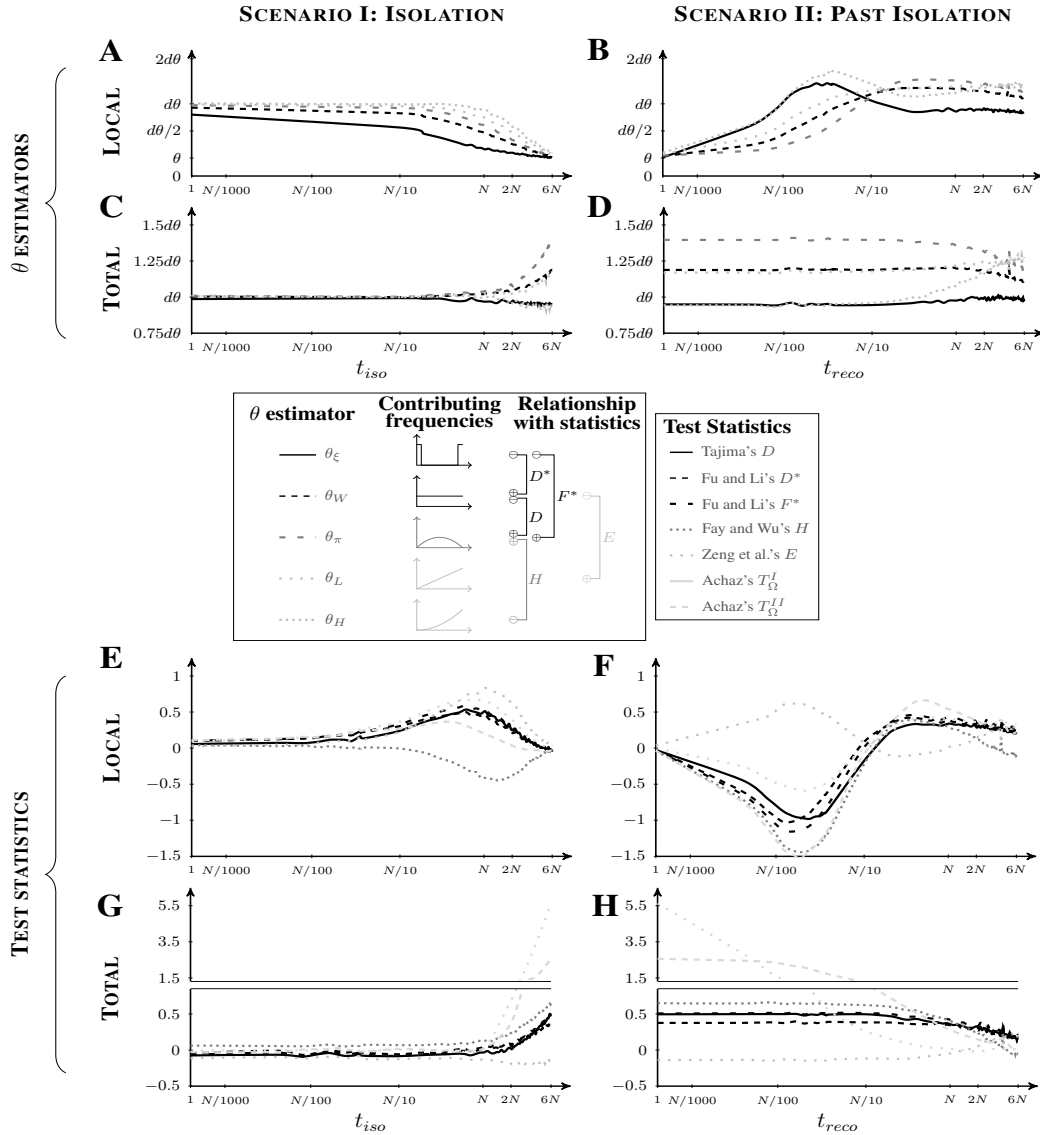


Figure S1 Effect of the time since the beginning of an isolation period, t_{iso} (scenario I; left column), and the time since the end of a past isolation period, t_{reco} (scenario II; right column) on estimators of the scaled mutation rate θ (A)-(D) and on the expected value of test statistics based on estimators of θ (E)-(H). θ estimators and test statistics are either computed from the local site frequency spectrum (SFS; A and B, and E and F), or estimated from the total SFS (C and D, and G and H). The panel in the center presents schematically for each θ the range of the frequencies concerned (column "contributing frequencies") and the related test statistics (column "relationship with statistics"). (A) Estimators based on low and intermediate frequency variants decrease first (θ_ξ , θ_W and θ_π), skewing all test statistics in (E). (B) Estimators based on variants at very low and very high frequencies increase first (θ_ξ and θ_H), then it is the ones based on high frequencies (θ_L) followed by those based on intermediate and low frequencies (θ_W and θ_π). Consequently, test statistics have successively negative and positive values in (F). (C-D) Values of estimators based on intermediate frequency variants are large (θ_W , θ_π and θ_L); this translates into positive values on the related statistics (G-H). Achaz's test statistics are the most skewed in all scenarios, for the local and total SFS (E-H). Parameters are $d=4$, $N=2,500$, 20 sampled genes of 1kb per population, with $\mu=2 \cdot 10^{-7}$ ($\theta=2$). Means are over 5,000 replicate simulations.

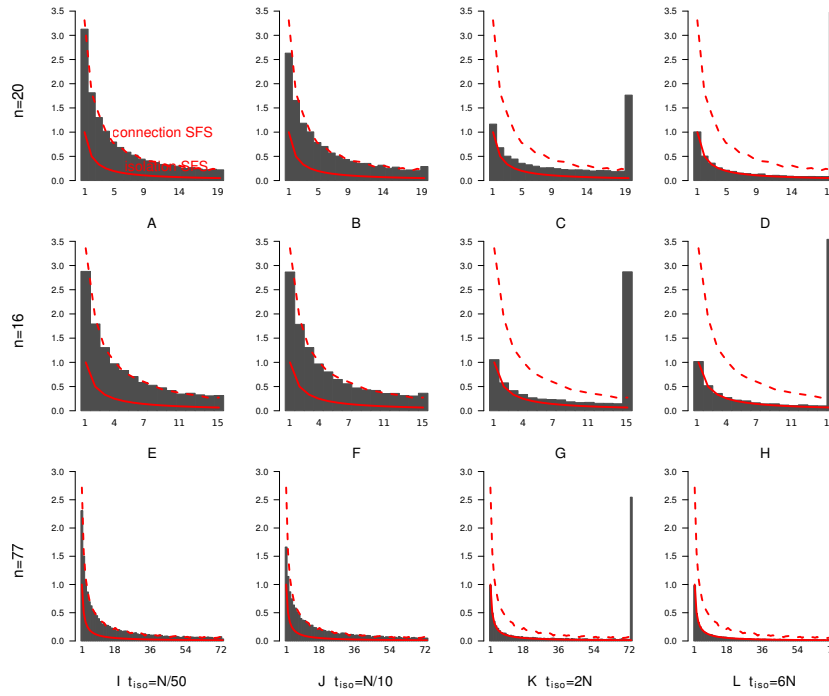


Figure S2 Temporal signature of an isolation event (scenario I; A-D) and a past isolation event (scenario II; E-H) on the local site frequency spectrum (SFS), for different sampling schemes. The red solid line represents the expected SFS in equilibrium connected populations. The total SFS following scenario I and II is represented as a function of the number of generations since the isolation event (t_{iso}) and the reconnection event (t_{reco}), respectively. Parameters are $N=2,500$ individuals per population, with $\mu=2.10^{-7}$ ($\theta=2$). Means are over 5,000 replicate simulations.

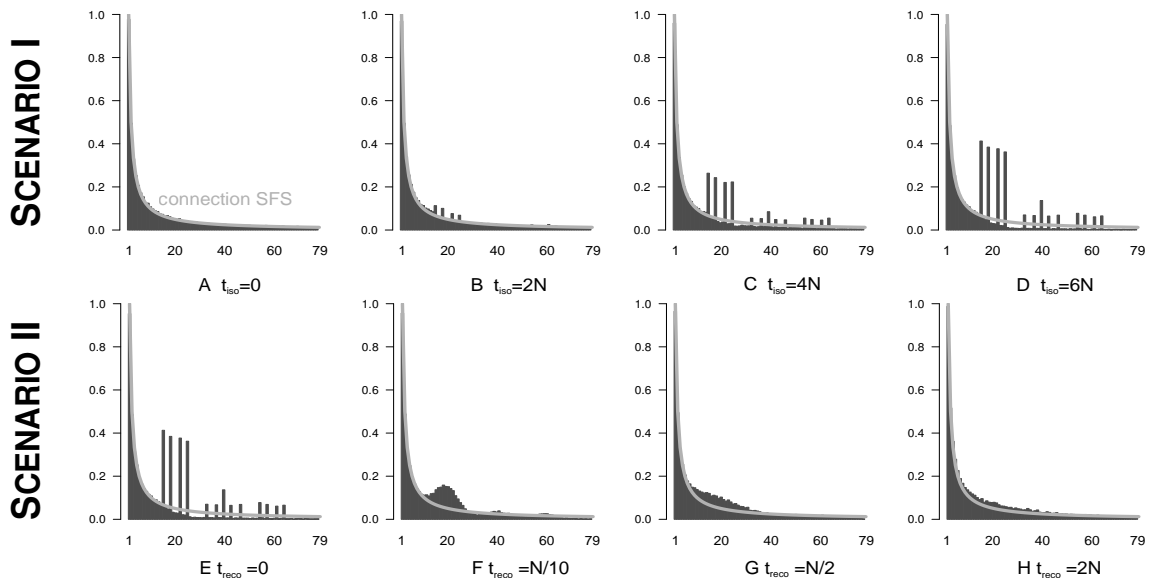


Figure S3 Temporal signature of an isolation event (scenario I; A-D) and a past isolation event (scenario II; E-H) on the total site frequency spectrum (SFS), for an uneven sampling scheme (sample sizes in each of the four populations are 16, 18, 22 and 24). The gray solid line represents the expected SFS in equilibrium connected populations. The total SFS following scenario I and II is represented as a function of the number of generations since the isolation event (t_{iso}) and the reconnection event (t_{reco}), respectively. Results are for $d = 4$ populations. Other parameters are $N=2,500$ individuals per population, with $\mu=2.10^{-7}$ ($\theta=2$). Means are over 5,000 replicate simulations.

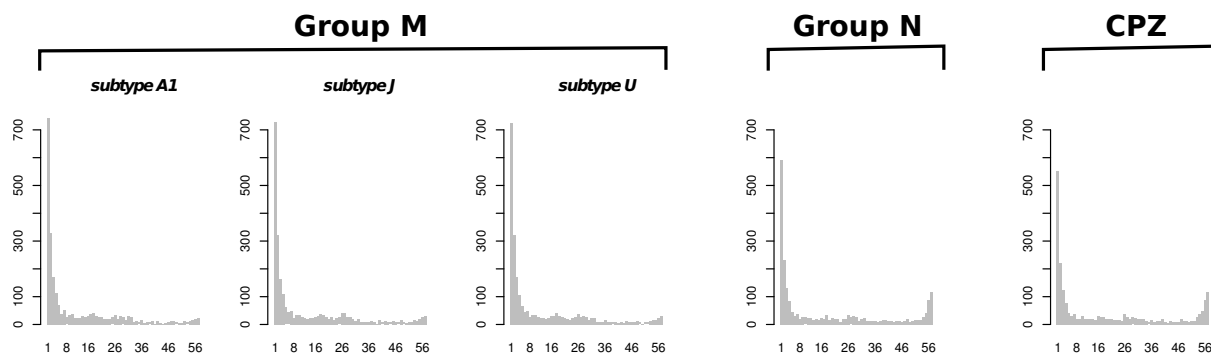


Figure S4 Sensitivity of the total SFS to the outgroup choice for ancestral state for the Chinese subtype HIV-1 polymorphism data used. Each plot corresponds to a different outgroup sequence, from a different groups (M, N, CPZ) or subtype with M group (A1, J, U). All subtypes A1, J and U in Group M have the lowest sequence divergence from our sampled sequences, Group N and CPZ (Chimpanzee Simian Immunodeficiency Virus) have the largest divergence which leads to an important excess of variants at high frequency, as expected from ancestral state misinference (BAUDRY and DEPAULIS 2003). However, the excess of intermediate frequency variants ("peaks") is robust to the chosen outgroup, and is thus not due to ancestral state misinference.

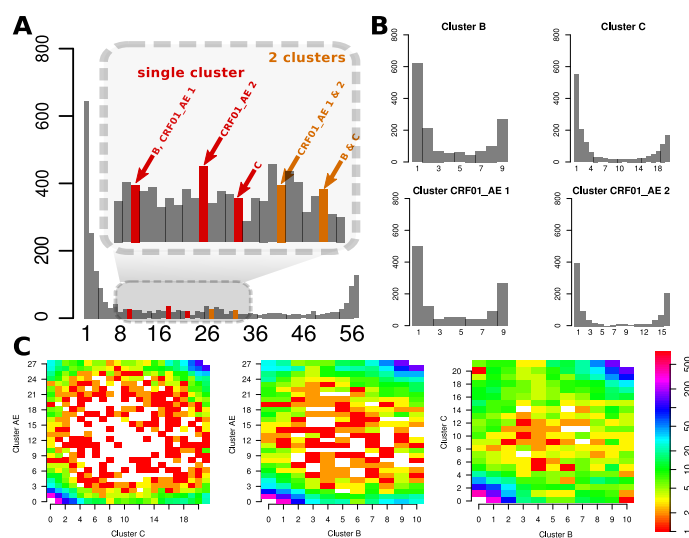


Figure S5 (A) Total, (B) local and (C) joint SFS of whole genome HIV-1 sequences from different subtypes sampled in China, using a sequence of HIV-1 group N as an outgroup for ancestral state inference. Although, the total, local and joint SFS present an excess of sites almost fixed in all population (pattern not observed when subtype J is used as outgroup, fig. 4.6) due to ancestral state misinference (BAUDRY and DEPAULIS 2003), other features of the SFS (e.g. presence and position of peaks are qualitatively the same as in fig. 4.6).

BIBLIOGRAPHY

AN, M., X. HAN, J. XU, Z. CHU, M. JIA, *et al.*, 2012 Reconstituting the epidemic history of hiv strain crf01_ae among men who have sex with men (msm) in liaoning, northeastern china: Implications for the expanding epidemic among msm in china. *Journal of virology* **86**: 12402–12406.

BAUDRY, E., and F. DEPAULIS, 2003 Effect of misoriented sites on neutrality tests with outgroup. *Genetics* **165**: 1619–1622.

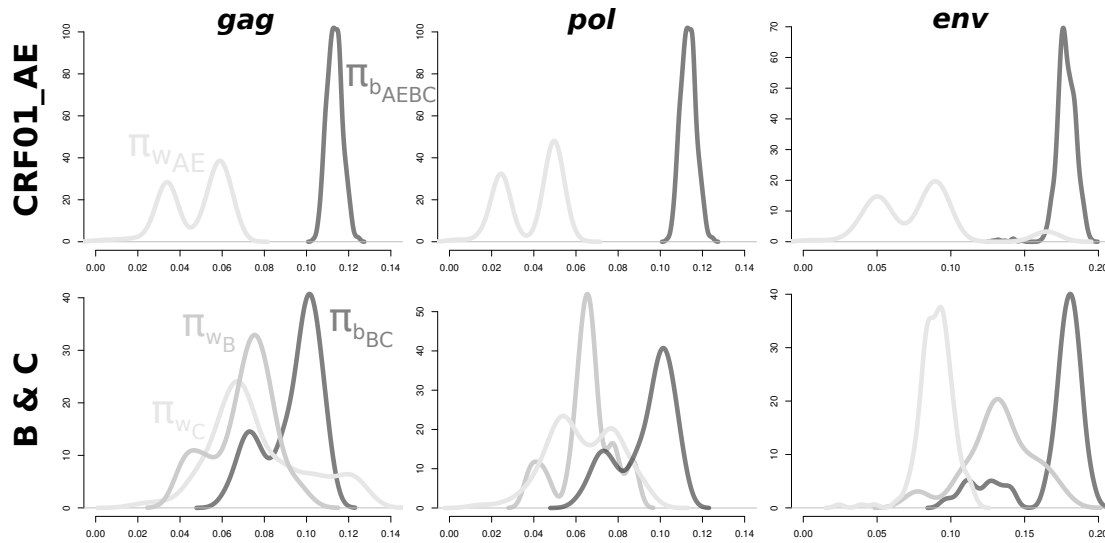


Figure S6 Within- (π_w) and between-cluster (π_b) distributions of pairwise nucleotide differences between HIV-1 sequences sampled in China, for 3 main HIV genes (columns *gag*, *pol* and *env*). Panels in the first and second lines correspond to clusters CRF01_AE and clusters B and C, respectively. The within-CRF01_AE cluster distribution presents two modes at low pairwise differences (at 0.025 – 0.05 and 0.05 – 0.09, respectively) for all three genes signalling substructure in the sample. The presence of two CRF01_AE clusters (from DAPC analysis) reflects the successive introduction of the CRF in China (AN *et al.* 2012). An additional mode at high pairwise differences (0.17), overlapping with the between-cluster distribution, is observed for gene *env* within-CRF01_AE cluster signalling past isolation scenario. Old past isolation is suggested for the three genes by the multimodal distributions, the high variance and the overlap of differences within-C cluster, within-B cluster and between-cluster B and C.

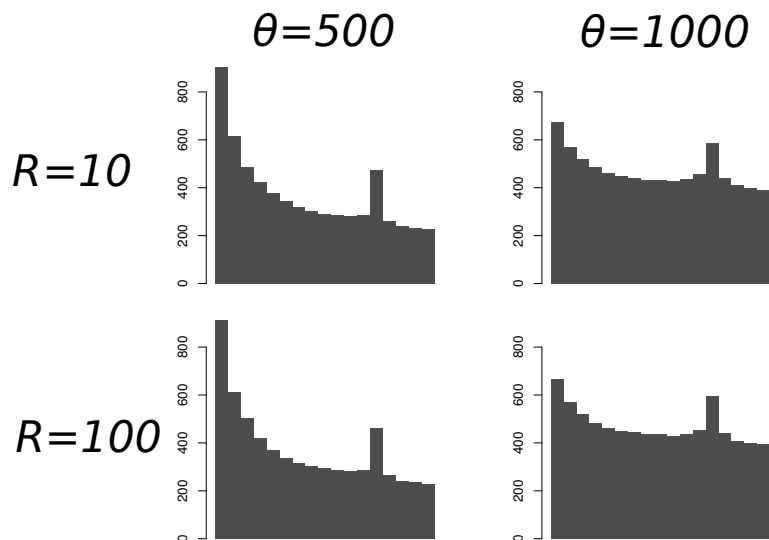


Figure S7 Sensitivity of the total SFS to different scaled mutation and recombination rates (θ and R). SFS are simulated under an isolation scenario and assuming a finite sites model (using software *fastsimcoal*; EXCOFFIER and FOLL 2011) assuming ancestral populations split $2N$ generations ago; sample sizes are $n = 13$ in the first population and $n = 22$ in the second; the number of sites is 9000 per sequence.

Table S1 Counts of HIV-1 subtypes and recombinant forms. To avoid sampling bias, we excluded redundant samples (from a same patient) and samples with identity larger than 98%. We excluded the gaps in sequences alignment and segregating sites with more than 2 states (due to homoplasy), resulting in 3676 segregating sites. Alignments are available upon request.

B	C	CRF 01_AE	CRF 07_BC	CRF 59_01B	URF BC	URF 01BC
8	2	26	2	1	17	2

EXCOFFIER, L., and M. FOLL, 2011 fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* 27: 1332–1334.