Inferences of Malagasy Evolutionary History from Genomic Data

Omar Alva Sánchez

▶ **To cite this version:**

Omar Alva Sánchez. Inferences of Malagasy Evolutionary History from Genomic Data. Social Anthropology and ethnology. Université Paul Sabatier - Toulouse III, 2022. English. NNT : 2022TOU30199 . tel-04190175

HAL Id: tel-04190175
https://theses.hal.science/tel-04190175

Submitted on 29 Aug 2023

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

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

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

OMAR ALVA SÁNCHEZ

Le 28 septembre 2022

Inférence de l'histoire évolutive de la population malgache à partir de données génomiques

---

Ecole doctorale : **BSB - Biologie, Santé, Biotechnologies**

Spécialité : **ANTHROPOBIOLOGIE**

Unité de recherche :
EvolSan - Evolution et Santé Orale

Thèse dirigée par
Denis PIERRON

Jury

M. Michel RAYMOND, Rapporteur
M. Etienne PATIN, Rapporteur
M. Jean-Pierre MAZAT, Rapporteur
Mme Chantal RADIMILAHY, Examinatrice
M. Thierry LETELLIER, Examinateur
M. Denis PIERRON, Directeur de thèse
Mme Monique COURTADE-SAÏDI, Présidente

RÉSUMÉ

AUTEUR : ALVA SANCHEZ Omar

TITRE : Inférence de l'histoire évolutive de la population malgache à partir de données génomiques

Directeur de Thèse : PIERRON Denis

LIEU ET DATE DE SOUTENANCE : Toulouse, le 28 septembre 2022

RESUME en français

Les récentes avancées en anthropologie moléculaire ont permis d'éclairer l'histoire du peuplement de la terre par l'espèce humaine. Confrontées à de nouveaux environnements, les populations humaines se sont adaptées tant génétiquement que culturellement. Au cours de ces processus de peuplement, elles se sont progressivement diversifiées au niveau génomique du fait de mutations, de dérives génétiques et d'effets fondateurs.

Au cours de cette thèse, nous avons exploré les conséquences de l'histoire évolutive humaine sur la diversité génétique et son impact sur la santé des populations actuelles. Nous nous sommes particulièrement intéressés à l'effet des processus d'admixture (mélange de population) et de peuplement de nouveaux milieux sur le partage d'éléments génétiques entre individus. Nous avons de plus étudié les effets de ces processus sur l'efficacité de la sélection naturelle sur les mutations génétiques rendant non fonctionnelle les protéines. Pour cela nous nous sommes focalisés sur la population humaine vivant actuellement à Madagascar et issue d'un brassage génétique qui s'est produit au cours du dernier millénaire, entre des populations africaines de langue bantoue et des populations asiatiques de langue austronésienne.

Dans ce travail, nous avons donc cherché à comprendre comment l'histoire humaine du peuplement de l'île influence la quantité de mutations génétiques délétères des individus vivant actuellement à Madagascar. Dans un premier temps, nous avons réalisé un travail bibliographique pluridisciplinaire (historique, linguistique et génétique) couvrant plus de deux siècles de travaux scientifiques pour lister et identifier différents scénarios plausibles de peuplement de Madagascar.

Dans un second temps, en étudiant les données issues du génotypage de 700 individus malgaches, nous avons testé la vraisemblance de ces différents scénarios avec la diversité génétique des individus vivant sur Madagascar actuellement. Ainsi, grâce aux progrès récents des algorithmes bioinformatiques, nous avons décidé de mettre en œuvre la modélisation génétique de l'histoire évolutive humaine, en lançant des simulations informatiques pour en déduire l'histoire démographique et la migration des populations malgaches. La comparaison des données observées et des données issues de simulations en termes de partage de segments chromosomiques a permis de rejeter plusieurs scénarios couramment répandus à Madagascar qui ne sont pas compatibles avec les données empiriques. A l'inverse, nous avons mis en évidence un phénomène de goulot d'étranglement lors du peuplement qui a conduit le plus probablement à la diversité génétique observée à Madagascar. Ainsi, nous avons déduit que les ancêtres asiatiques de la population malgache ont vécu isolés pendant environ 1000 ans, avec une taille effective de seulement quelques centaines d'individus. Cet isolement ayant pris fin environ 1000 ans avant le présent.

Dans un troisième temps, nous avons analysé 67 génomes complets (WGS) d'individus malgaches afin de rechercher l'impact de cette histoire démographique sur la variation fonctionnelle. Nous avons en particulier étudié l'effet du goulot d'étranglement pour les ancêtres asiatiques sur la fréquence des mutations délétères. Ainsi nous avons produit un catalogue des mutations dérivées présentes à Madagascar au niveau de génome complet, en identifiant des milliers des mutations potentiellement délétères qui pourront être interrogées dans des travaux futurs.

En conclusion, ce travail de doctorat d'anthropologie permet de mieux appréhender comment différents scénarios de peuplement d'un nouveau territoire peuvent influencer de manière différente le génome des populations actuelles de ce territoire. De plus, ce travail souligne la complémentarité et le dialogue nécessaire entre les sciences humaines, les sciences environnementales, les sciences de l'information et les sciences bio-médicales pour comprendre la santé des populations humaines.

Mots-clés : Évolution ; Anthropologie; Admixture; Peuplement; Madagascar; Génétique; genome-wide ; Simulations informatiques

English summary

Recent advances in molecular anthropology have shed light on the history of the peopling of the world by the human species. Faced with new environments, human populations have adapted both genetically and culturally. During these settlement processes, they gradually diversified at the genomic level due to mutations, genetic drift and founder effects. During this thesis, we explored the consequences of human evolutionary history on genetic diversity and its impact on the health of current populations. We were particularly interested in the effects of admixture processes and the arrival to new environments on the sharing of genetic elements between individuals. We further investigated the effects of these processes on the efficacy of natural selection, where genetic mutations disrupt the function of proteins. For these reasons, we focused on the human population currently living in Madagascar. Previous genetic and linguistic analyzes have shown that the Malagasy population is the result of a genetic admixture that has occurred over the last millennium, between Bantu-speaking African populations and Austronesian-speaking Asian populations. During this work, we therefore sought to understand how the human history of the settlement of the island has influenced the amount of deleterious genetic mutations carried by individuals currently living in Madagascar.

First, we conducted a multidisciplinary bibliographic work (historical, linguistic and genetic) covering more than two centuries of scientific work, listing and identifying different plausible scenarios for the settlement of Madagascar. Secondly, by studying data from the genotyping of 700 Malagasy individuals, we tested the similarity of these different scenarios with the genetic diversity of Malagasy individuals. Thus, thanks to recent advances in bioinformatics algorithms, we decided to implement genetic modeling of human evolutionary history, launching computer simulations in order to deduce the demographic history and migration events of Malagasy populations. The comparison of empirical and simulated data, in terms of the sharing of chromosomal segments, made it possible to reject several scenarios previously proposed in Madagascar, as they are not compatible with the empirical data. Conversely, we detected a bottleneck phenomenon happening before the admixture, which most probably led to the genetic diversity observed in Madagascar. Thus, we deduced that the Asian ancestors of the Malagasy population lived in isolation for about 1000 years, with an effective size of only a few hundred

individuals. This isolation ended about 1000 years before the present. Thirdly, we analyzed 67 complete genomes (WGS) of Malagasy individuals in order to investigate the impact of this demographic history on functional variation. In particular, we studied the effect of the bottleneck for Asian ancestors on the frequency of deleterious mutations. We produced a catalog of derived mutations present in Madagascar at the whole genome level, identifying thousands of potentially deleterious mutations that can be interrogated in future works.

In conclusion, this anthropology doctoral thesis allows us to better understand how different settlement scenarios can influence differently the genome of the current populations inhabiting this territory. In addition, this work underlines the complementarity and the necessary dialogue between the human sciences, the environmental sciences, the computational sciences and the biomedical sciences, all in order to better understand the health of current human populations.

Keywords: Evolution; Anthropology; Admixture; Settlement; Madagascar; Genetics; Genome-wide; Computational simulations

---

DISCIPLINE ADMINISTRATIVE : Anthropobiologie

Laboratoire Évolution et santé oral (EVOLSAN).
Université Toulouse III - Paul Sabatier - UFR Odontologie - 3, chemin des Maraîchers, CP 31062, Toulouse, Midi-Pyrénées, France.

This page is intentionally left blank

ACKNOWLEDGEMENTS

I would like to thank all the people in Madagascar, Mexico and France, who made possible the accomplishment of this doctoral thesis.

First of all, I would like to thank the people who have walked by my side in this life, and who with their love and support always encourage me to pursuit in the struggle for life. Thank you for never allowing anyone to extinguish the spirit of freedom that exists in me.

I also thank the people in France who gave me their support and advice. Thanks to them, I was able to recognize and learn how to deal with the challenges that arose during the time that I was in this country.

I would like to thank the scientists and students from the *Musée d'Art et d'Archeologie*, the University of Antananarivo and the Paul Sabatier University, for all their efforts in the comprehension of the history of Madagascar.

Last but not least, I thank the members of the *Médicine Evolutive* team from the EVOL-SAN laboratory, as well as the scientists and students from the Paul Sabatier University, for their time and collaboration in order to achieve and accomplish our professional goals. Thanks to every one of you, I was able to see, learn and unlearn many things.

My doctoral study was financially supported by the *Ministère* de l'*Éducation* Nationale de l'*Enseignement* Supérieur of France. The accomplishment of every step during my Ph.D. was also possible thanks to the academic staff from the Paul Sabatier University and the *École Doctorale Biologie-Santé-Biotechnologies* of Toulouse, whose titles and responsibilities include teaching, non teaching, research, and administrative positions.

This page is intentionally left blank

TABLE OF CONTENTS

INTRODUCTION

One of the most interesting questions in anthropology studies involves knowing how humans settled and populated the world. There are several scientific disciplines that have offered elements of answers and proposed particular scenarios according to the time and the place of interest in human evolutionary history. We have learned about the environmental conditions and the population's configurations (osteological features, effective population size, time and rate of population growth) during the advent of modern humans in Africa, around 200 000 years ago. We know from anatomical and archaeological studies that they had skeletons very similar to those of present-day people, and they lived in Africa and the Middle East until sometime after 60,000 years ago. Genetic studies of human origins based on mitochondrial DNA propose an important split event around 50,000 years ago, where few individuals left Africa and continued to migrate through Europe, Asia, Oceania and America. Interestingly, those same studies suggest that an effective size of no more than 3,000 individuals was involved in the departure event. During these global migrations happening in the course of thousands of years, human individuals have reached new environments, leading to the settlement of new places. Thanks to different scientific disciplines, we know that migrations, settlements, and adaptation of new environments are an important part of human evolutionary history.

Through scientific efforts during the last 60 years, we have expanded the knowledge on biology, chemistry and computational science, and along with technological development, we were able to study the DNA molecule, which is behind the expression and perpetuation of all known life forms on earth. One essential advance happened with the launch of the Human Genome Project in 1990, which offered several tools and strategies to accurately assess almost every base pair of an individual's DNA. During the following decade, scientists from different areas of research (biochemistry, biology, computation, statistics) collaborated internationally to assemble, for the first time, the entire human genome. Interestingly, these efforts culminated in technological advances that collaborated to reduce the cost of sequencing on a 30-fold scale. Almost 20 years later, it became accessible to genotype hundreds of individuals and produce public scientific databases

(such as HapMap and 1000 Genomes) that became essential to contemporaneous research. Although the projects offered many exciting applications, one of the biggest breakthroughs in anthropology has been to effectively assess the evolutionary history of our species, with a better understanding of human genetic diversity. In this manner, and since the last two decades, several research groups have interrogated important and interesting clues regarding which evolutionary forces could act on a population that moves into a new place; and ultimately, what histories and clues can we recover from the genetic information gathered in the individuals' genes. In the search of knowing how human populations settled the world, we learned that Madagascar appears to be one of the last large territories being settled by human populations.

Figure A: Map of Indian Ocean. The origin of the Malagasy population has been debated for several centuries. Studies based on cultural, linguistic and genetic elements show that the Malagasy population share ancestors with populations living across the Indian Ocean. Red and blue zones represent the geographical locations of the reference populations used to study the demographic process in Madagascar (green).

For many centuries, the origin of the Malagasy population remained enigmatic. Historical texts on the question are abundant and can be traced back—at least—to reports by Portuguese sailors arriving to the island in the 16th century. While Madagascar is an

island located <400 km from the East African coast and more than one thousand to several thousand kilometers from any other shoreline (**Figure A: Map of Indian Ocean**), ancient contributions from Asian, Indian, Melanesian, Arabic, Persian and Semitic populations have been suggested. However, by the 20th century, there was no consensus on the origin(s) of the population. More recently, genetic and linguistic analyses have shown that the Malagasy population is the result of a genetic admixture that has occurred over the last millennium, between Bantu-speaking African populations and Austronesian-speaking Asian populations. In this manner and by studying the genetic information carried by individuals currently living in Madagascar, we looked to characterize the human history of the settlement of the island. During this thesis, we therefore sought to propose a settlement model(s) that could explain Malagasy genetic diversity, particularly inspecting the evolutionary history of Malagasy ancestral populations before the admixture.

First, we conducted a multidisciplinary bibliographic work (historical, linguistic and genetic) covering more than two centuries of scientific work, in order to list and identify different plausible scenarios for the settlement of Madagascar, which is essential for the construction of a proper evolutionary model of Madagascar's human settlement. Our first approach was a broad literature research involving the historical questions and the different disciplinary approaches that have studied the settlement of Madagascar. Next, we oriented our scope in order to investigate the genetic ancestors of Malagasy, discussing the genetic methods for studying admixture and the published results obtained from genetic data. Next, we did bibliographic research about the reasoning of computational simulations and their application on human evolutionary studies, as we wanted to evaluate how this strategy could help us to better interpret the results, in addition to a better understanding of the bioinformatics methods used in anthropological studies.

In a second time, we began to build the evolutionary model by implementing computational simulations. Based on the bibliographic work, and on the genomic analysis of a large representative sample of Madagascar's inhabitants and extensive genetic modeling, we tested the similarity of different simulated scenarios with the genetic diversity of Malagasy individuals. Thanks to this strategy, we were able to detect and discard plausible demographic scenarios that could explain the observed genetic diversity in Madagascar, showing that today's Malagasy populations originate from an admixture between two small

ancestral populations around 1,000 years ago. Our analyses revealed that the genetic admixture with an African population ended an intriguing thousand-year period of isolation for the Asian ancestral population. We characterized this event as a long-term bottleneck, where the Asian ancestors of the Malagasy population had an effective size of only a few hundred individuals and lived in isolation for about 1000 years.

Finally, in order to better understand the evolutionary forces acting on Madagascar, we performed whole-genome sequencing on the Malagasy population. As theory predicts that the efficacy of natural selection is reduced in populations who have suffered bottlenecks with low effective population sizes, we studied the effect of the bottleneck for Asian ancestors on the frequency of deleterious mutations. In this manner, we produced a catalogue of derived mutations present in Madagascar at the whole genome level, identifying thousands of potentially deleterious mutations that can be interrogated in future works.

Current genetic diversity observed in human populations is the result of past evolutionary processes, geographic expansions and demographic histories. During this doctoral thesis, we developed a strategy for testing anthropological theories using a simulation framework, which allowed us to test the agreement between these human evolutionary theories and genetic markers. Importantly, we were able to interpret the results with archaeological, ecologic, linguistic and ethnographic data, elucidating how different settlement scenarios can influence differently the genome of the populations inhabiting a territory

CHAPTER I

IDENTIFYING PLAUSIBLE SCENARIOS FOR THE SETTLEMENT OF MADAGASCAR

CHAPTER I

IDENTIFYING PLAUSIBLE SCENARIOS FOR THE SETTLEMENT OF MADAGASCAR

Introduction

Due to lacking historical records, ancient migration, gene flow and settlement events in human evolutionary history cannot be easily addressed. Indeed, most of these events occurred so far in time that precedes written recorded history, they are limited to some geographical regions and they depict events from the last 5,000 years (Gross, 2012), and proper geographical and chronological attribution is needed for the depicted events (Assael et al., 2022). In the cases where the access to historical records is very limited, scholars and researchers might hypothesize numerous conflicting scenarios (or theories) regarding the same events, sometimes based on cultural evidence (such as archeological remains) or linguistic studies (such as loan words). More recently, genetic studies have helped elucidate the context of some past evolutionary events in human history, confirming or discarding diverse hypotheses based on the observed genetic diversity of current-day human populations (Choudhury et al., 2021; Hudjashov et al., 2017; Jacobs et al., 2019; Larena et al., 2021; Nielsen et al., 2017; Schlebusch et al., 2012). Nevertheless, the analysis of genetic data, as well as the interpretation of these results, can be a complex challenge.

It has been shown that different ancestral events might produce a similar genetic effect or a combination of events might lead to complex genetic information (Arenas & Posada, 2014). In the case of Madagascar, this was evidenced by studying the split between the Malagasy and source populations from south Borneo, where two scenarios showed similar likelihoods: (i) a hard split around 2,500 years ago and (ii) a slow divergence with less and less migration between 3,000–2,000 years before present (Pierron et al., 2017). In addition, there exist computational methods that can infer the histories of populations, but little is known about the quality of these inferences or their robustness to deviations from their underlying assumptions (Adrion et al., 2020). For instance, local ancestry inference

methods depend on the available proxies for the source populations, which could potentially affect the power to detect African and Asian ancestry in the Malagasy genomes. A strategy that can help with the above uncertainties involves the application of computer simulations.

Interestingly, along with the rapid development of genotyping and re-sequencing technologies, there were also a wide catalogue of tools for simulating large-scale genomic data under realistic scenarios that include the effects of natural selection, recombination, gene conversion, and complex demographic and environmental factors (Yuan et al., 2012). Thus, computer simulations have been used for understanding current genetic diversity underlying different demographic histories, allowing the study of evolutionary aspects that can affect an entire system, which can also help to validate and compare analytical frameworks (Arenas & Posada, 2014). Up to date, computational simulations do not permit to infer the evolutionary history of populations without realistic priors (particularly for a landscape like Madagascar), but they can help to compare different scenarios, test their plausibility and potentially identify the most probable according to empirical data. In this manner, a multidisciplinary approach (combining archaeology, history, genetics) is needed to provide correct input parameters to the computer in order to launch a model for studying *in silico* the settlement of a territory. Even if there have been attempts on modeling the settlement of Madagascar, these efforts have been limited by the available data, the genetic markers used and the specificity of each tested scenario. For these reasons, we decided to implement computational simulations for the study of Madagascar settlement, testing different hypotheses and inferring the evolutionary parameters involved in this process.

In order to build a realistic simulation of Malagasy genetic diversity, we needed a definition of scenarios corresponding to different plausible hypotheses (or theories) of settlement. Thus, we conducted a bibliographical work assembling the historical and genetic evidence available on the study of Madagascar settlement. We looked for answers in the literature regarding the possible ancestral populations involved in the settlement, as well as the dates and places of possible events involved in the process. Indeed, the origin of the Malagasy population has been a subject of speculation since the 16th century. Contributions of African, Asian, Indian, Melanesian, Arabic and Persian populations have been suggested based on physical and cultural anthropology, oral tradition, linguistics and

later also by archaeology. It is nowadays admitted that Malagasy populations are the result of an original admixture of populations coming from different continents, as genetic studies show that the current Malagasy share ancestors with populations living in Eastern Africa, Southeast Asia and Western Eurasia. Nevertheless, many questions remain unanswered, and there is still much uncertainty regarding when, how and the circumstances under which the ancestors of the modern Malagasy population arrived on the island. Given these reasons, we completed a catalogue of scenarios based on historical texts and genomic results for building a computational model of Malagasy genetic diversity.

At the beginning of this chapter, we present the historical questions and classical methodological approaches regarding the settlement of Madagascar, reviewing the extent to which genetic results have settled historical questions concerning the origin of the Malagasy population. We present an overview of the early literature, a discussion of the genetic results of the 20th and 21th centuries, as well as the latest results in genome-wide analyses. This work entitled “Genetic evidence and historical theories of the Asian and African origins of the present Malagasy population” was published in the Human Molecular Genetics indexed journal. In the second part of this chapter, we expanded the discussion regarding the genetic ancestors of Malagasy, presenting the methods used to discuss genetic admixture and the evidence for the multiple genetic origins of the Malagasy population. We also examine the links between Malagasy and other populations located on different continents, discussing the origin of maternal lineages (based on mitochondrial DNA), and paternal lineages (based on the Y chromosome). Finally, at the end of this chapter we discuss the reasoning and application of computational simulations in human evolutionary studies, and how this strategy can help us to refine the interpretations of results based on genetic data, in addition to a better understanding of the methods already in use for studying the human settlement of Madagascar.

We conclude this chapter by presenting a selection of settlement models that could explain Malagasy genetic diversity, particularly inspecting the evolutionary history of ancestral populations before the admixture.

Article 1

Genetic evidence and historical theories of the Asian and African origins of the present Malagasy population

Margit Heiske, Omar Alva, Veronica Pereda-Loth, Matthew Van Schalkwyk, Chantal Radimilahy, Thierry Letellier, Jean-Aimé Rakotarisoa and Denis Pierron

Human Molecular Genetics, 2021, Vol. 30, No. 2

Human Molecular Genetics, 2021, Vol. 30, No. 2 R72–R78

doi: 10.1093/hmg/ddab018
Advance Access Publication Date: 22 January 2021
Invited Review Article

INVITED REVIEW ARTICLE

Genetic evidence and historical theories of the Asian and African origins of the present Malagasy population

Margit Heiske[1,†], Omar Alva[1,†], Veronica Pereda-Loth[1,†], Matthew Van Schalkwyk[2], Chantal Radimilahy[3], Thierry Letellier[1], Jean-Aimé Rakotarisoa[3] and Denis Pierron[1,*]

[1]1Équipe de Médecine Evolutive, Faculté de Chirurgie Dentaire URU EVOLSAN Université Toulouse III, France, [2]Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology, University of Cambridge, Cambridge, UK and [3]Musée d'Art et d'Archéologie, University of Antananarivo, Antananarivo, Madagascar

*To whom correspondence should be addressed at: Équipe de Médecine Evolutive, Faculté de Chirurgie Dentaire, 31400 Toulouse, France. Tel: +33 (0)5 62172929; Fax: +33(0)561254719; Email: denis.pierron@univ-tlse3.fr

Abstract

The origin of the Malagasy population has been a subject of speculation since the 16th century. Contributions of African, Asian, Indian, Melanesian, Arabic and Persian populations have been suggested based on physical and cultural anthropology, oral tradition, linguistics and later also by archaeology. In the mid-20th century, increased knowledge of heredity rules and technical progress enabled the identification of African and Asian populations as main contributors. Recent access to the genomic landscape of Madagascar demonstrated pronounced regional variability in the relative contributions of these two ancestries, yet with significant presence of both African and Asian components throughout Madagascar. This article reviews the extent to which genetic results have settled historical questions concerning the origin of the Malagasy population. After an overview of the early literature, the genetic results of the 20th and 21th centuries are discussed and then complemented by the latest results in genome-wide analyses. While there is still much uncertainty regarding when, how and the circumstances under which the ancestors of the modern Malagasy population arrived on the island, we propose a scenario based on historical texts and genomic results.

Introduction

For many centuries, the origin of the Malagasy population remained enigmatic. Historical texts on the question are abundant and can be traced back—at least—to reports by Portuguese sailors arriving on the island in the 16th century. While Madagascar is an island located <400 km from the East African coast and more than one thousand to several thousand kilometers from any other shoreline, ancient contributions from Asian, Indian, Melanesian, Arabic, Persian and Semitic populations have been suggested (1–3). However, by the 20th century, there was no consensus on the origin(s) of the population, and one of the most discussed questions was the importance of African ancestry (4).

In this article, we aim to evaluate the extent to which genetic results have settled historical questions concerning the origin of the Malagasy population. First, we examine why so many source populations have been suggested. We then review the genetic

†These authors contributed equally to this study.

Received: September 11, 2020. **Revised:** December 23, 2020. **Accepted:** January 6, 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

results of the 20th century. Finally, we present the most 'current' answers to the elusive question of Malagasy origins based on the latest results (5,6).

Historical Elements and Theories on the Origin

The oldest textual references on the origin of the Malagasy population might be the descriptions of the people of 'Komr' (5) linked to Srivijava (al-Idrisi, 12th century), as well as of individuals related to Chinese people (Ibn Sa'id, 13th century) (6–8). However, there is uncertainty concerning the names used to describe the islands in the Indian Ocean by ancient Greco-Roman, Arabic and Chinese writers (6). As early as 1604, Megiser even questioned whether Madagascar was really known by ancient cosmographers or only 'recently discovered', and consequently whether the island should be classified as an 'Ancient world' or 'New world' (9).

At the turn of the 16th century, the chroniclers reporting the first travels of a Portuguese fleet to the shores of Madagascar, described two populations; one consisting of dark-skinned individuals related to African groups, and a second population residing in seaside cities that comprised of Arab settlers (9–12). In 1559, the sailors of 'Nossa senhora da barca' reported individuals who looked and spoke like the Javanese and concluded that the eastern part of Madagascar might have been settled by a Javanese population (9,13). In 1611, Pyrard de Laval similarly reported that 'it is said' that the island was once settled by Chinese people and that Malagasy faces were similar to those of the Chinese except for their complexion (14). In 1614, the Jesuit priest Luis Mariano described a population originally from Mozambique and Malindi that lived in a part of the north-west coast of the island. However, throughout the interior of the island, the population spoke a language that seemed to him very similar to Malay. Therefore, he concluded that the first inhabitants came from the ports of Malacca (15).

In addition to these accounts, inhabitants of the south of Madagascar reported to Dutch sailors that some of their ancestors came from Portugal, Mangalore and Mecca (9,16). In 1655, based on a study of religious practices, Etienne de Flacourt suggested that a part of the population belonged to an ancient Abraham lineage (17), while Nacquart assumed descendance from Persian Mohammedans. Later on in 1781, Court de Gebelin proposed ancient Phoenician trading posts in Madagascar based on linguistic similarities (9).

These various tales of origin were discussed over the course of the 19th century (18), and in 1908 it was reported that at least 50 different authors had hypothesized different origins (19). At this time, one dominant view was that people from the central highlands were descended from the recent arrival of a Malay-speaking population, and those on the coast with a darker complexion from an old migration of an African Bantu-speaking population. In addition to this, later arrivals and limited contributions by Arabs, Indians and Europeans were assumed. Based on oral tradition, some authors also described the earlier presence of groups of unknown origins, such as the Vazimba/Kimosy, related to either Austronesians or African pygmy/bushmen who could have been replaced by later arrivals (18). In 1927, Dubois identified five independent theories on the origin of the Malagasy people (20). One of these, proposed by Grandidier, was that the coastal populations were not of African descent but rather of Papuan/Melanesian descent (19). This hypothesis turned the question of African versus Austronesian contributions to the current Malagasy population into an animated controversy (16).

In the 20th century, progress in linguistics and archaeology sustained the debate. Despite a lack of resources, archaeological studies have suggested an ancient human presence on Madagascar (21,22). Indeed, cutmarks were identified on bones of elephant birds (*Mullerornis* and *Aepyornis*) dated to around 10 000 BP. However, to date there is no further indication regarding the origin of these occupants. In contrast, linguistic analyses have shown that the Malagasy language is closely related to Ma'anjan, a language spoken in South Borneo (23,24). Dating based on linguistic borrowings suggest a recent split of a proto-Malagasy population from other Indonesian populations within the first millennium of the Christian era (24,25). The African contribution to the modern Malagasy language was considered very limited (25,26).

Estimation of the African and Asian Ancestral Proportions

Since the beginning of the 20th century, genetic heredity laws made it possible to go beyond considerations founded on visual phenotypical characteristics (27–29). Based on the allele frequencies of the B blood group in Madagascar and given their absence in the Papuan/Melanesian population, David *et al.* (27) excluded heavy contribution from these populations (suggested by Grandidier) and proposed instead a putative ancestry from India and Java. Further, by computing the ratio between Bantu and Javanese populations required to obtain the Rh blood groups frequencies observed in the Malagasy haplotypes, Singer *et al.* (30) posited in 1957 that genetically, the Malagasy were approximately two parts African and one part Asian. The presence of the S type of Hemoglobin in the Malagasy population engendered a debate as to whether this presence is due to gene flow from Africa (31), India or the Arabian Peninsula (32). The two latter origins were proposed by Fourquet *et al.* (32) who hypothesized that before settling Madagascar, sailors from Borneo and maybe Celebes would have mixed with the population of South India. Consequently, they suggest that the contribution of African populations to the Malagasy ancestry was limited and recent (32).

In the mid-90s, the first studies with DNA-based evidence were published. Using this new molecular approach, Hewitt et al showed that the Hemoglobin S type present in Madagascar originated from an African Bantu speaking population, thereby contradicting the hypothesis by Fourquet *et al.* (33). They further identified a blood type present in West Africa and Portugal that is also present in Malagasy populations. They linked this phenomenon to the arrival of the European sailors, possibly with West African slaves, in the 16th century. Additionally, the analysis of the non-mutated type suggested a mixed contribution from East Asia/Oceania and Africa. In contrast, two studies identified significant contribution from East Asia but only a minor signal from Africa. Based on the analysis of a 9 bp deletion in mitochondrial DNA, Soodyal *et al.* (34) identified a Polynesian motif in 18% of a sample of 280 Malagasy individuals. Another molecular study, based on HLA class II haplotype analysis and data derived from 55 individuals from a rural community of the central highlands, demonstrated molecular evidence of population affinity between the Malagasy and the Javanese (35).

Several factors can explain the apparently conflicting results regarding the extent of Asian/African contributions to the Malagasy genetic makeup obtained by the different genetic studies of the 20th century. For instance, the influence of founder effects

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

should be taken into consideration as well as the fact that the evolutionary history of a single marker may not be representative of the whole genome. Also, the relevance of the sample composition was rarely discussed, while the frequency of one specific marker can vary substantially depending on the studied population. For example, in Singer *et al.* (36), the majority of the sampled individuals were assigned to a 'tribe' called 'Merina', however, 'Merina' is actually a geographic term that refers to the central part of Madagascar. In fact, assignment to a group is often inappropriate and should be treated with caution since it might be based on particular phenotype(s), a particular ancestor(s), cultural specificities, a putative benefit, birth location or the location that is currently inhabited.

Eventually, at the turn of the 21st century, a series of studies confirmed the presence of both African and Austronesian components (37–40). However, analyses attempting to quantify the relative contribution of both ancestries still showed variable results (41–44). Regueiro *et al.* (44) suggested a 66.3% genetic makeup from Africa based on the profile of 15 autosomal STR loci of 67 individuals. Alternatively, by attributing a continental origin to the mitochondrial DNA (mtDNA) and Y chromosome of Malagasy lineages, several studies estimated the admixture ratio. Importantly, these markers are specific to the female/male lineages and are thus not representative of the entire ancestry. Hurles *et al.* (41) reported 'approximately equal' African/Asian contributions for the Malagasy in the central highlands: 38% (14/37) of the mtDNA lineages and 51% (18/35) of the Y-chromosomal lineages were traced back to Africa. By comparing the populations of the central highlands with populations from the southeast, Tofanelli *et al.* (42) found signals of regional differences and sex-biased admixture: 73% African Y-chromosomal lineages and 38% African mtDNA lineages.

Conversely, a different analysis of a similar data set suggested that the maternal African contribution was very limited i.e. only 7% (43). This was based on an earlier discovery of a specific mtDNA lineage that is frequent in Madagascar (the Malagasy motif) but absent in the putative source population (Borneo) (40). Simulations suggested that this could be the result of the settling of Madagascar by a very limited number of women of 93% Asian descent. However, this model did not account for the other parameters (present admixture ratio, global mtDNA diversity etc.) and did not explore alternative hypotheses (arrival of multiple populations, etc.).

In 2017, a comprehensive study of the genomic landscape of the entire island provided a global view in terms of genetic diversity (45). This was made possible by 10 years of intensive work by the ICMAA Institute in Antananarivo, sampling 300 villages across Madagascar. For the first time, the genomic diversity of a large territory such as Madagascar was sampled using a systematic field work approach that combined genomics, linguistics and ethnography.

The presence of both African and Asian ancestry in all regions of Madagascar was demonstrated by assessing the origins of the genome segments of each individual (local ancestry) as well as by applying a similar computational approach as that of Singer *et al.* (46) on thousands of independent polymorphisms across the genome. On average, African ancestry represented 59%, Asian ancestry 37%, and West-Eurasian ancestry only 4%. Interestingly, the ancestry percentage was not identical over all chromosomes. In fact, the Malagasy genome showed one of the strongest signals of recent selection reported for modern humans. Alleles of African origin have been favored in over one quarter of the chromosomes 1 (~60 million bp). The most probable candidate for this selection is the Duffy null blood group (46,47). This genotype provides resistance to *Plasmodium vivax* (malaria), a parasite absent in Africa. Remarkably, this suggests that this parasite may have been introduced to Madagascar by Asian populations while the protective mutation (Duffy null blood group) was introduced by African populations (46).

Further, correlations between the genetic structure of the population and the geography of the island were detected (46). Some genetic clusters are identical in term of admixture date and percentage. This suggests that external factors like physical geography and the formation of kingdoms on the island have impacted the genomic landscape of Madagascar, probably by influencing population movements and exchanges. Nevertheless, there is striking regional variability in the relative proportion of African and Asian ancestry (Figures 1 and 2), which explains why previous conclusions concerning origin based on localized sampling could not be extrapolated to 'Madagascar' as a whole. Interestingly, populations with the largest African ancestry were found in the northeast, specifically in the region described by Luis Mariano in 1614 as one settled by a population who spoke an African language (in contrast to the rest of Madagascar), while the largest Asian ancestry was found in the central highlands (9).

Identification of Specific Source Populations and Time of Peopling

Beside the proportion of ancestry, DNA-based researches have focused on understanding the scenario that led to the peopling of the island. Genome-wide analyses have determined that Malagasy individuals share a strong genetic bond with Bantu-speaking African populations as well as Austronesian-speaking Asian populations, suggesting a peopling by individuals of either descent (48). These analyses object against hypotheses that propose any large contribution by populations related to those currently living in India, the Horn of Africa or Melanesia. In addition, there is no evidence linking the Malagasy population with African populations that have a hunter-gatherer subsistence strategy. In particular, a distinct origin has been assumed for the Malagasy hunter-gatherer groups living in the Mikea forest. However, a genome-wide analysis showed that they shared the same ancestries as the current Malagasy, suggesting a recent cultural reversion (48). Nevertheless, there are other minor components that have been identified in the Malagasy ancestry such as those from West Eurasia. This is particularly relevant to certain Malagasy groups that seem to incorporate Arab-Islamic aspects in their culture, such as the use of the Arabic script *Sorabe* (49,50). But presumably, these West Eurasian components of Malagasy ancestry might be also derived from a more recent contribution from Europe. However, this needs to be clarified through further analyses.

In accordance with linguistic studies, uniparental and genome-wide analyses have determined that the populations closest to the Malagasy are African Bantu-speaking populations from East Africa and Austronesian-speaking populations from around the Makassar Strait (41,42,45,51). However, with respect to the details of the uniparental markers, there is a conundrum concerning the Asian side. Y chromosome diversity is similar to populations in the west of the Makassar Strait while mtDNA diversity seems closer to the populations further east (presence of the Polynesian motif). This might reflect the migration process (individuals coming from different populations, founder events, etc.) or demographic events in Madagascar (genetic drift). Moreover, variations that may have occurred since the Malagasy

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

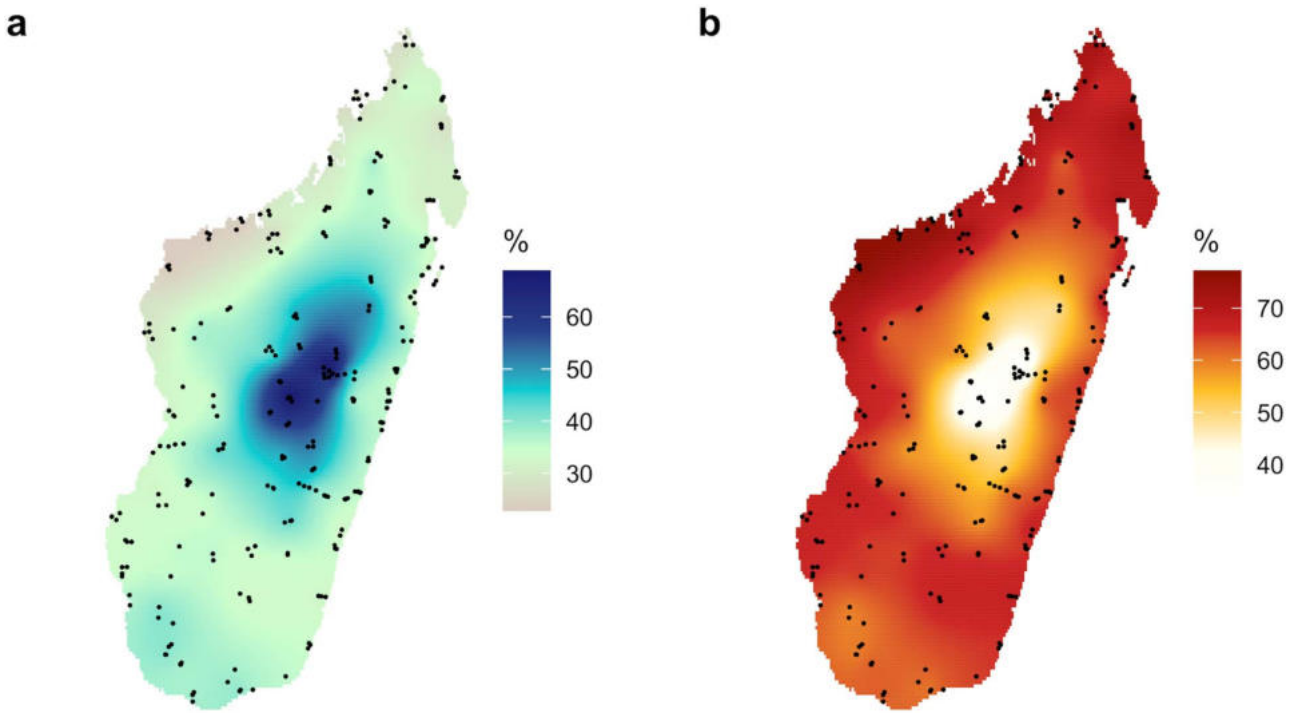


Figure 1. Exponential kriging interpolation of the (**a**) east-Asian and (**b**) African ancestry across Madagascar landscape based on genome-wide data (autosomal data) published in Pierron *et al.* (45).

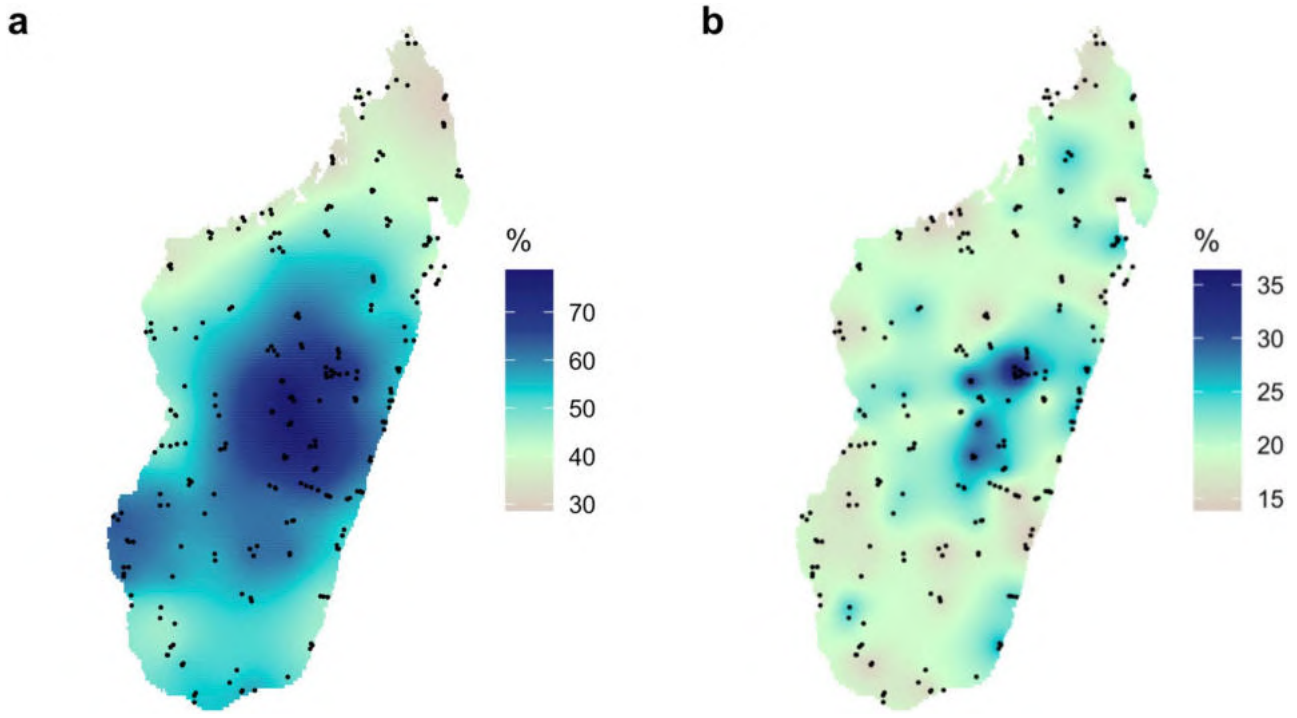


Figure 2. Exponential kriging interpolation of the east-Asian lineages across Madagascar landscape according to their a) mitochondrial lineages (**b**) African lineages based on data published in Pierron *et al.* (45).

ancestors left Indonesia could have played a role. Another reason could be due to insufficient genetic sampling from Indonesia. Analyses of ancient DNA from Indonesia should provide more details about the Asian source population, which could help resolve this enigma.

Remarkably, the relatedness to the Bantu-speaking and Austronesian-speaking populations supports the theory of a recent arrival of the Malagasy ancestors on the island. The arrival of Bantu-speaking populations on the East African coast in particular is a recent phenomenon (52). Through

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

analyses based on uniparental markers (42,43) and whole genome analyses, it was possible to date their admixture with Austronesian-speaking population in the last two millennia (45,48). Evidence of some East-Asian ancestry was also identified in individuals from the Comoros, East Africa and Arabia (53). Interestingly, the admixture that occurred in one island of Comoros (Anjouan) might have been older than in Madagascar. However this result was not replicated when the authors change their computation methods or the set of parameters (54).

Compared to the Asian ancestors of the Malagasy, their African ancestors seemed to have split from their respective source populations much more recently. The overrepresentation of the African paternal lineage in addition to the more recent link between African Bantu-speaking and Malagasy populations led to the hypothesis of a late arrival of Bantu-speaking populations on a territory already settled by Austronesian-speaking populations. In support of this hypothetical scenario are the testimonies of Arabic writers, and Pyrard de Laval's statement that the island was once populated by Asians. Furthermore, a slave testimony reported in 1575 by André Thevet (9) described the combination of a cyclone and a tsunami (linked to an earthquake) that ravaged existing villages and cities all over Madagascar. The only survivors of this catastrophe would have been those who had fled to high ground. Faced with devastated territories and only a few survivors, two African kingdoms (Sofala and Mozambique) might have sent three or 4000 individuals to resettle the island. Christian kingdoms evoked in this narration suggest a date compatible with the genomic-based time estimation of the admixture, Bantu-speaking population movement, and with evidence of a tsunami in the Indian Ocean (55). However, this is the only citation of this scenario that we know of, and given all the existing hypotheses, it is not surprising to find one that matches the genomic results. More studies that combine genomics and archaeology should enable further examination of this scenario. It should also be kept in mind that genetics can only provide information on the direct ancestors of the present population and therefore not on the first settlers of Madagascar.

Conclusions

The quest to understand the contribution of African and Asian ancestors in today's Malagasy population has not been a straight path free of ulterior motives. For example, in 1957, Singer *et al.* criticized other researchers by stating 'There appears to be some stigma about the possibility of an African origin of the population in all the French and Malagasy writers' articles' (30). Subsequent DNA sequencing did not resolve the controversy as rapidly as one would expect. Today, both African and Asian contributions have been proven, in addition to a West-Eurasian component, although in a rather small proportion. Nonetheless, Madagascar's history of settlement is still largely unknown and further investigation is needed to understand the extent of other contributions posited by authors in the past.

Access to human genetic information has caused a disruption in the field since it allows researchers to distinguish between questions on the origins of cultural factors (i.e. language) and of humans (individuals), as well as their diffusion. Today's challenge consists of forming a comprehensive view of the migration phenomena that led to the settlement of Madagascar—one that incorporates cultural and biological aspects of Malagasy history. Over the next decades, studying human, animal or plant DNA based on present or ancient samples should provide unexpected new insights regarding the history of Madagascar (56–59).

Acknowledgements

The authors thank the editors and guest editors for their suggestions for this special issue. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest statement. The authors declare no conflict of interest.

References

1. Radimilahy, C. and Rajaonarimanana, N. (2011) Civilisations des mondes insulaires: Madagascar, îles du canal de Mozambique, Mascareignes, Polynésie, Guyanes. In *Mélanges en l'honneur du professeur Claude Allibert*. Hommes et sociétés, Karthala.
2. Radimilahy, C. (2011) Réflexions sur la production préeuropéenne du textile dans le Nord de Madagascar. In *Études Océan Indien*. Inalco, Paris, pp. 162–176.
3. Beaujard, P. (2011) Les plantes cultivées apportées par les premiers Austronésiens à Madagascar. In *Civilisations des mondes insulaires - (Madagascar, îles du canal de Mozambique, Mascareignes, Polynésie, Guyanes)*, Karthala, pp. 357–385.
4. Ferrand, G. (1909) L'Origine africaine des Malgaches. *Bul. Mém. Soc. Anthropol. Paris*, **10**, 22–35.
5. Allibert, C. (2000) Le mot Ḳomr dans l'Océan indien et l'incidence de son interprétation Sur l'ancienneté du savoir que l'on a de la région. *Topoi*, **10**, 319–334.
6. Martin, N. (2011) Madagascar, une île au carrefour d'influences. In *Études Océan Indien*. Inalco, Paris, pp. 46–47.
7. Vicente, M.A. (2015) Madagascar avant que les Portugais n'y arrivent. In *Temas Insulares*, CLEPUL/Instituto Europeu de Ciências da Cultura Padre Manuel Antunes, Carcavelo.
8. Allibert, C. (1990) *Documents pédagogiques: textes anciens sur la côte est de l'Afrique et l'océan Indien occidental*. CEROI avec l'aide du GRECO-Ocean Indien, Paris.
9. Grandidier, A., Charles-Roux, J., Delhorbe, C., Froidevaux, H. and Grandidier, G. (1903) *Collection des ouvrages anciens concernant Madagascar: Ouvrages ou extraits d'ouvrages portugais, hollandais, anglais, français, allemands, italiens, espagnols et latins relatifs à Madagascar (1500 à 1613)*. Comité de Madagascar, Paris.
10. Lafitau, J.-F. (1733) *Histoire des découvertes et conquêtes des Portugais dans le nouveau monde*. Slatkine, Genève.
11. de Barros, J. (1777) *Da Asia decada Segunda Livro I. Capitulo I.* Na Regia officina typografica, Lisboa, pp. p1–p15.
12. Vicente, M.A. (2015) Madagascar dans les sources portugaises. In *Fontes e Temas Insulares, Centro de Literaturas e Culturas Lusófonas e Europeias da Faculdade de Letras da Universidade de Lisboa*. CLEPUL/Instituto Europeu de Ciências da Cultura Padre Manuel Antunes, Carcavelo.
13. de Couto, D. (1778) *Da Asia decada setima parte primeira Capitulo VII., liv VIII, ch1*. Na Regia officina typografica, Lisbao, pp. 175–179.
14. Pyrard, F. (1611) *Discours du voyage des François aux Indes Orientales*. David Le Clerc, Paris, p. 371.
15. Grandidier, A., Charles-Roux, J., Delhorbe, C., Froidevaux, H. and Grandidier, G. (1904) *Tome II, Collection des ouvrages*

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

anciens concernant Madagascar: 1613–1640. Comité de Madagascar, Paris.
16. Ferrand, G. (1905) Les migrations Muslulmanes et Juives à Madagascar. *Rev. Hist. Relig.*, **52**, 381–417.
17. Flacourt, E. (1661) In Oudot, G. (ed), *Histoire de la grande isle Madagascar, composée par le sieur de Flacourt, avec une relation de ce qui s'est passé ès années 1655, 1656 et 1657N*. Clouzier, Troyes, Paris.
18. de Rialle, G. (1889) La population de Madagascar: d'après des publications récentes. *Rev. Hist. Relig.*, **20**, 180–192.
19. Grandidier, A. and Grandidier, G. (1908) *Histoire physique, naturelle et politique de Madagascar*. Imprimerie Nationale, Paris, Vol. **4**.
20. Dubois, S.J. (1927) Les Origines des Malgaches. *Anthropos*, **22**, 80–124.
21. Douglass, K., Hixon, S., Wright, H.T., Godfrey, L.R., Crowley, B.E., Manjakahery, B., Rasolondrainy, T., Crossland, Z. and Radimilahy, C. (2019) Critical review of radiocarbon dates clarifies the human settlement of Madagascar. *Quat. Sci. Rev.*, **221**, 105878.
22. Davis, D.S., Andriankaja, V., Carnat, T.L., Chrisostome, Z.M., Colombe, C., Fenomanana, F., Hubertine, L., Justome, R., Lahiniriko, F., Léonce, H. *et al.* (2020) Satellite-based remote sensing rapidly reveals extensive record of Holocene coastal settlement on Madagascar. *J. Archaeol. Sci.*, **115**, 105097.
23. Dahl, O.C. (1951) *Malgache et maanjan: Une comparaison linguistique*. Egede-Instituttet, Oslo.
24. Adelaar, K.A. (1989) Les langues austronésiennes et la place du malagasy dans leur ensemble. *Arch. Androl.*, **38**, 25–52.
25. Serva, M. and Pasquini, M. (2020) Dialects of Madagascar. *PLoS One*, **15**, e0240170.
26. Beaujard, P. (2011) The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. *Azania Archaeol. Res. Afr.*, **46**, 169–189.
27. David, R. (1939) Le problème anthropologique malgache. Nouvelles observations chez les Māhaf'aly du Sud-Ouest de Madagascar. *J. Afr.*, **9**, 119–152.
28. Ratsimamanga, A.R. (1940) Tache pigmentaire héréditaire et origines des Malgaches. *Rev. Anthropol.*, **N° 1–3 (Janvier-mars)**, 1–3.
29. Geipel, G. (1957) Die finger und Handleisten der Neger Madagaskars, zugleich ein Beitrag zur Frage ihres Ursprungs. *Z. Für Morphol. Anthropol.*, **48**, 234–253.
30. Singer, R., Budtz-Olsen, O.E., Brain, P. and Saurain, J. (1957) Physical features, sickling and serology of the Malagasy of Madagascar. *Am. J. Phys. Anthropol.*, **15**, 91–124.
31. Buettner-Janusch, J. and Buettner-Janusch, V. (1964) Hemoglobins, haptoglobins, and transferrins in the peoples of Madagascar. *Am. J. Phys. Anthropol.*, **22**, 163–169.
32. Fourquet, R., Sarthou, J., Roux, J. and Acri, K. (1974) Hémoglobine S et Origines du peuplement de Madagascar: nouvelle hypothèse Sur son introduction en Afrique. *Arch. Inst. Pasteur Madagascar*, **N° 43/1**, 185–220.
33. Hewitt, R., Krause, A., Goldman, A., Campbell, G. and Jenkins, T. (1996) Beta-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from bantu-speaking Negroids. *Am. J. Hum. Genet.*, **58**, 1303–1308.
34. Soodyall, H., Jenkins, T. and Stoneking, M. (1995) 'Polynesian' mtDNA in the Malagasy. *Nat. Genet.*, **10**, 377–378.
35. Migot, F., Perichon, B., Danze, P.M., Raharimalala, L., Lepers, J.P., Deloron, P. and Krishnamoorthy, R. (1995) HLA class II haplotype studies bring molecular evidence for population affinity between Madagascans and Javanese. *Tissue Antigens*, **46**, 131–135.
36. Allibert, C. (2008) Austronesian migration and the establishment of the Malagasy civilization: contrasted readings in linguistics, archaeology, genetics and cultural anthropology. *Diogene*, **55**, 7–16.
37. Poetsch, M., Wiegand, A., Harder, M., Blöhm, R., Rakotomavo, N., Freitag-Wolf, S. and von Wurmb-Schwark, N. (2013) Determination of population origin: a comparison of autosomal SNPs, Y-chromosomal and mtDNA haplogroups using a Malagasy population as example. *Eur. J. Hum. Genet.*, **21**, 1423–1428.
38. Rabe, T., Jambou, R., Rabarijaona, L., Raharimalala, L., Rason, M.A., Ariey, F. and Dhermy, D. (2002) South-east Asian ovalocytosis among the population of the highlands of Madagascar: a vestige of the island's settlement. *Trans. R. Soc. Trop. Med. Hyg.*, **96**, 143–144.
39. Chow, R.A., Caeiro, J.L., Chen, S.-J., Garcia-Bertrand, R.L. and Herrera, R.J. (2005) Genetic characterization of four Austronesian-speaking populations. *J. Hum. Genet.*, **50**, 550–559.
40. Razafindrazaka, H., Ricaut, F.-X., Cox, M.P., Mormina, M., Dugoujon, J.-M., Randriamarolaza, L.P., Guitard, E., Tonasso, L., Ludes, B. and Crubézy, E. (2010) Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *Eur. J. Hum. Genet.*, **18**, 575–581.
41. Hurles, M.E., Sykes, B.C., Jobling, M.A. and Forster, P. (2005) The dual origin of the Malagasy in island Southeast Asia and East Africa: evidence from maternal and paternal lineages. *Am. J. Hum. Genet.*, **76**, 894–901.
42. Tofanelli, S., Bertoncini, S., Castrì, L., Luiselli, D., Calafell, F., Donati, G. and Paoli, G. (2009) On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. *Mol. Biol. Evol.*, **26**, 2109–2124.
43. Cox, M.P., Nelson, M.G., Tumonggor, M.K., Ricaut, F.-X. and Sudoyo, H. (2012) A small cohort of island southeast Asian women founded Madagascar. *Proc. Biol. Sci.*, **279**, 2761–2768.
44. Regueiro, M., Mirabal, S., Lacau, H., Caeiro, J.L., Garcia-Bertrand, R.L. and Herrera, R.J. (2008) Austronesian genetic signature in east African Madagascar and Polynesia. *J. Hum. Genet.*, **53**, 106–120.
45. Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L.M., Rakotomalala, M.M., Razafiarivony, M., Rasoarifetra, B. *et al.* (2017) Genomic landscape of human diversity across Madagascar. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 32.
46. Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olaso, R., Deleuze, J.F. *et al.* (2018) Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nat. Commun.*, **9**, 932.
47. Hodgson, J.A., Pickrell, J.K., Pearson, L.N., Quillen, E.E., Prista, A., Rocha, J., Soodyall, H., Shriver, M.D. and Perry, G.H. (2014) Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. *Proc. R. Soc. B Biol. Sci.*, **281**, 20140930.
48. Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.A., Blench, R.M. *et al.* (2014) Genome-wide evidence of Austronesian-bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 936–941.
49. Capredon, M., Brucato, N., Tonasso, L., Choesmel-Cadamuro, V., Ricaut, F.-X., Razafindrazaka, H., Rakotondrabe, A.B., Ratolojanahary, M.A., Randriamarolaza, L.-P., Champion, B.

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

et al. (2013) Tracing Arab-Islamic inheritance in Madagascar: study of the Y-chromosome and mitochondrial DNA in the Antemoro. *PLoS One*, **8**, e80932.
50. Capredon, M., Sanchez-Mazas, A., Guitard, E., Razafindrazaka, H., Chiaroni, J., Champion, B. and Dugoujon, J.-M. (2012) The Arabo-Islamic migrations in Madagascar: first genetic study of the GM system in three Malagasy populations. *Int. J. Immunogenet.*, **39**, 161–169.
51. Kusuma, P., Cox, M.P., Pierron, D., Razafindrazaka, H., Brucato, N., Tonasso, L., Suryadi, H.L., Letellier, T., Sudoyo, H. and Ricaut, F.X. (2015) Mitochondrial DNA and the Y chromosome suggest the settlement of Madagascar by Indonesian sea nomad populations. *BMC Genomics*, **16**, 191.
52. Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., Prista, A., Damasceno, A., Fehn, A.-M., Schlebusch, C. *et al.* (2020) Along the Indian Ocean coast: genomic variation in Mozambique provides new insights into the bantu expansion. *Mol. Biol. Evol.*, **37**, 406–416.
53. Brucato, N., Fernandes, V., Kusuma, P., Černý, V., Mulligan, C.J., Soares, P., Rito, T., Besse, C., Boland, A., Deleuze, J.-F. *et al.* (2019) Evidence of Austronesian genetic lineages in East Africa and south Arabia: complex dispersal from Madagascar and Southeast Asia. *Genome Biol. Evol.*, **11**, 748–758.
54. Brucato, N., Fernandes, V., Mazières, S., Kusuma, P., Cox, M.P., Ng'ang'a, J.W., Omar, M., Simeone-Senelle, M.-C., Frassati, C., Alshamali, F. *et al.* (2018) The Comoros show the earliest Austronesian gene flow into the Swahili corridor. *Am. J. Hum. Genet.*, **102**, 58–68.
55. Maselli, V., Oppo, D., Moore, A.L., Gusman, A.R., Mtelela, C., Iacopini, D., Taviani, M., Mjema, E., Mulaya, E., Che, M. *et al.* (2020) A 1000-yr-old tsunami in the Indian Ocean points to greater risk for East Africa. *Geology*, **48**, 808–813.
56. Herrera, M.B., Thomson, V.A., Wadley, J.J., Piper, P.J., Sulandari, S., Dharmayanthi, A.B., Kraitsek, S., Gongora, J. and Austin, J.J. (2017) East African origins for Madagascan chickens as indicated by mitochondrial DNA. *R. Soc. Open Sci.*, **4**, 160787.
57. Brouat, C., Tollenaere, C., Estoup, A., Loiseau, A., Sommer, S., Soanandrasana, R., Rahalison, L., Rajerison, M., Piry, S., Goodman, S.M. *et al.* (2014) Invasion genetics of a human commensal rodent: the black rat Rattus rattus in Madagascar. *Mol. Ecol.*, **23**, 4153–4167.
58. Linz, B., Vololonantenainab, C.R.R., Seck, A., Carod, J.-F., Dia, D., Garin, B., Ramanampamonjy, R.M., Thiberge, J.-M., Raymond, J. and Breurec, S. (2014) Population genetic structure and isolation by distance of helicobacter pylori in Senegal and Madagascar. *PLoS One*, **9**, e87355.
59. Sauther, M.L., Bertolini, F., Dollar, L.J., Pomerantz, J., Alves, P.C., Gandolfi, B., Kurushima, J.D., Mattucci, F., Randi, E., Rothschild, M.F. *et al.* (2020) Taxonomic identification of Madagascar's free-ranging "forest cats". *Conserv. Genet.*, **21**, 443–451.

Downloaded from https://academic.oup.com/hmg/article/30/R1/R72/6106232 by guest on 05 December 2022

Article 2

The Multiple Sources of the Malagasy Genetic Diversity

Omar Alva, Harilanto Razafindrazaka, Chantal Radimilahy, Jean-Aimé Rakotoarisoa,

Thierry Letellier, Denis Pierron

This chapter will appear on the book Malagasy World, with 03/2023 as preview date for publication according to the publisher (Taylor & Francis Group). Co-edited by:
Zoë Crossland, Dept. of Anthropology, Columbia University
Kristina Douglass, Dept of Anthropology, Penn State University
Chantal Radilmilahy, Museum of Art and Archaeology, University of Antananarivo

The Multiple Sources of the Malagasy Genetic Diversity

Omar Alva[1], Harilanto Razafindrazaka[2], Chantal Radimilahy[3], Jean-Aimé Rakotoarisoa[3], Thierry Letellier[1], Denis Pierron[1*]

Affiliations:

1- Équipe de Médecine Evolutive, EVOLSAN faculté de chirurgie dentaire, Université Toulouse III, Toulouse, France

2- Aix Marseille University, CNRS, EFS, ADES, 13344 Marseille, France

3- Musée d'Art et d'Archéologie, University of Antananarivo, Antananarivo, Madagascar.

[*] Corresponding authors:

Denis Pierron

Équipe de Médecine Evolutive, EVOLSAN

Faculté de Chirurgie Dentaire

Université Toulouse III, 31400 Toulouse, France

denis.pierron@univ-tlse3.fr; Tél. +33 (0)5 62 17 29 29; Fax: +33(0)5 61 25 47 19

Abstract

The origin of the Malagasy population has been debated for several centuries. Along with historical, cultural and archeological studies, genetic studies have also highlighted the extent of the diversity of the inhabitants of the island of Madagascar. It is nowadays admitted that Malagasy populations are the result of an original admixture of populations coming from different continents. Indeed, genetic studies show that the current Malagasy share ancestors with populations living in Eastern Africa, Southeast Asia and Western Eurasia. Nevertheless, many questions remain unanswered. In this chapter, we present and discuss the genetic evidence for this diversity. We begin by presenting the methods used to discuss genetic admixture, and the evidence for the multiple genetic origins of the Malagasy population. Then we examine the links between Malagasy and other populations located on different continents. In particular, we discuss the origin of maternal lineages based on mitochondrial DNA, and paternal lineages based on the Y chromosome study.

Introduction

The multiple origins of the Malagasy population has been debated for several centuries, with numerous historical and archeological studies addressing the origin of this population (Vérin, 1976 ; Crowther et al, 2016; Wood, 2015). Studies based on cultural and linguistic elements show that the Malagasy population share cultural aspects with many populations across the Indian Ocean (Adelaar & Himmelmann, 2005; Beaujard, 2011; Serva, 2012). For example, many musical instruments considered traditional find their origins across different regions of the world: South East Asia, Africa, Western Eurasia, etc (Rakotomalala, 1996 ; Sachs, 1938). For instance, it seems that the language of the Ma'anyan population, inhabiting south Borneo, is the closest language to the Malagasy (Dahl, 1951) with malays and javanese loanwords (Adelaar, 2009). These elements make Madagascar an example of a proto-globalization process that happened across the Indian Ocean before modern globalization, where food and cultural elements were transferred across the Indian Ocean between human societies (Lawler, 2014).

While archeology, linguistic and ethnographic studies follow the transmission of cultural elements; studies based on genetic data will assess the transmission of biological traits. At the beginning, molecular anthropology started with limited elements regarding phenotype studies based on blood group (Mourant, 1952; Singer et al., 1957) . Nevertheless, the rapid progress of molecular biology since 1990 to nowadays has lead to an accumulation of data, rendering possible a global view of the general ascendance of human populations, allowing to specifically characterize the paternal lineage (through the study of chromosome Y) or the maternal lineage (through the study of Mitochondrial DNA). Based on genetic analysis, it was possible to suggest a common origin of human populations, dated around 200,000 years BP in Africa (Cann et al., 1987). DNA technology advances were accompanied by a better understanding of mechanism of heredity at the chromosomal level, as the development of bioinformatics methods and algorithms were able to unravel the information contained in each individual's genome. That was the case of the international HapMap project, which sought to genotype at least one common SNP every 5 kilobases (kb) across the genome of individuals from diverse populations. Through analysing this dataset, it was possible to build recombination maps from diverse populations and to

identify genes that have experienced recent adaptive evolution (The International HapMap Consortium, 2005, 2007; Voight et al., 2006). In this way, the history of populations can be understood and leveraged under the light of biological, cultural and social anthropological studies. Since the first studies focused on few individuals and/or few genetic markers, the number of experiments and questions has increased with technological progress. In the same manner, the number of individuals sampled and the diversity of locations have also increased in Madagascar and across the rest of the world, which have rendered possible building a relatively global understanding of Malagasy genetic diversity. The aim of this article is to review the scientific evidence produced from genetic studies, in order to provide elements of discussion for scientists coming from other disciplines. Specifically, we will present key elements in order to discuss the diversity landscape coming from genetic evidence.

Methods to study genetic admixture

Due to lacking historical records, ancient migration and gene flow events cannot be frequently addressed (Korunes & Goldberg, 2021). However, these past demographic processes have an impact on the patterns of genetic variation across individual genomes, which can be used to better understand the complex dispersal of modern humans from Africa and their population expansion across the globe (Nielsen et al., 2017). Such studies have shown that neighboring populations frequently exchange individuals that contribute to an on-going process of gene flow between them. However, there have also been range expansions or migration events in human evolutionary history that have put in contact previously isolated populations. These events involving the formation of a new population by the genetic exchange from at least two source populations are called admixture (Jobling et al, 2014).

A series of different admixture estimation procedures have been developed based on DNA data, which can detect and describe admixture using different genetic markers: chromosome Y (chrY), mitochondrial DNA (mtDNA), and autosomal DNA. Unlike the autosomes that are inherited from both parents, chromosome Y is passed down only from father to son, making it possible to reconstruct the paternal lineages within populations; on

the other hand, mitochondrial DNA is passed on only from mothers to children, forming a maternally inherited counterpart to the Y chromosome. The different inheritance patterns of these molecules have shaped the genetic diversity and geographical differentiation among populations (Cann et al., 1987; Mitchell & Hammer, 1996; Nielsen et al., 2017). Similarly, by sampling the DNA of individuals of the population of interest (through saliva, blood, etc.), we can assess the maternal lineage (inspecting the genetic variants present in their mtDNA) and their paternal lineage (inspecting their chromosome Y). Identifying the likely origins of all paternal and maternal lineages found in a population (through phylogenetic analysis) can reveal admixture if more than one ancestral population has contributed to the population of interest (Hurles et al., 2005). Interestingly such analyses can reveal that males or females of the ancestral populations may have contributed disproportionately, a phenomenon known as sex-biased admixture is produced.

Beside mitochondrial and Y chromosome data, advances in DNA sequencing have recently led to genome-wide genotype data production for many populations throughout the world, as well as the development of several approaches to simultaneously assess population structure and admixture from such data. One method of reconstructing genetic ancestry is based on principal component analysis (PCA), a general method for summarizing high-dimensional data (Novembre & Stephens, 2008). PCA can be used to assess the clustering of individuals or populations according to allele frequency (dis) similarities: individuals representing admixture of two distinct ancestral populations would be expected to lie on a cline between the clusters, as the exact position of admixed individuals on this cline would be determined by the ancestral contributions of the two parental populations (see Figure 1) (Jobling et al, 2014; Ma & Amos, 2012). Similarly, clustering algorithms attempt to classify individuals into a discrete number of clusters based on associations among their genetic patterns (Wangkumhang et al., 2018). Many of these algorithms determine the proportion of each individual's DNA derived from a given number of clusters (inferred or specified by the researcher). In some cases, the number of clusters inferred can be interpreted as the number of ancestral source groups that potentially intermixed in the past. In that way, individuals' DNA can be assigned to multiple clusters descending from historical admixture events.

Based on the block-like manner in which autosomal DNA is inherited, other methods for detecting admixture have been developed (Wangkumhang & Hellenthal, 2018). In fact, due to differences in break points of recombination and allele frequencies between the ancestral populations, admixed genomes can be explained as the mosaics of chromosome segments from different ancestries (Korunes & Goldberg, 2021). In theory, comparing the individual's genetic variation patterns to that of a set of reference populations (meant to represent the source groups) can help to identify the ancestral origin of each block of DNA inherited from the admixing sources (Thornton & Bermejo, 2014; Wangkumhang & Hellenthal, 2018). The total amount these inferred tracts can help to determine the proportions of ancestry inherited from each admixing source. This process, known as Local Ancestry Inference (LAI), can be assessed through particular loci across the genome, and the average number of blocks inherited from each ancestral population helps to determine the global ancestry of an individual or a population (see Figure 2). Other approaches are based on linkage disequilibrium (LD) profiles, whereby alleles at different loci tend to be co-inherited (Jobling et al., 2014; Loh et al., 2013). As LD at physically linked loci decays more slowly due to recombination events, recently admixed populations should exhibit LD over greater genetic distances than the source populations. As recombination breaks down these associations, leaving a signature of the time elapsed since admixture (Wangkumhang & Hellenthal, 2018). Several methods have been developed to further examine the relation between time since admixture and the extent of LD decay. Admixture estimates should be done considering that all present-day populations of the world are far from homogeneous (there is substantial population structure), with the possibility of subgroups tracing predominant ancestry to different source populations. These approximations have been implemented for elucidating the past evolutionary history of human populations. Based on these recent developments and by studying the DNA of Malagasy's individuals, it is possible to find hints in order to better understand the history of the populations that have settled the island.

Evidence of genetic admixture in Madagascar

Although the Madagascar is situated less than 500 km from the African shore and far away from any other continent, the Malagasy population share not only characteristics with populations from this continent, but also a large amount with the Austronesian populations of South East Asia. The first DNA analyses studying the origin of Malagasy population come from 1995 (Soodyall et al., 1995), showing that Malagasy shares a particular mitochondrial DNA type, defined as the haplogroup B4a1a1a. This sequence is found at high frequency in Polynesia, suggesting a connection between the founders of Madagascar and Polynesian ancestors (Soodyall et al., 1995). In the following decades, the study of autosomal DNA sequences that tend to be hyper variable across individuals, known as autosomal short tandem repeat (STR) loci, allowed to confirmed that the Malagasy population came from an admixture of two ancestral populations: Bantu from Africa and Austronesian from Asia, as the Malagasy gene pool represented 66.3% of African genetic contribution and 33.7% of South East Asia genetic contribution (Regueiro et al., 2008).

More recently, the genetic data from Madagascar have been expanded thanks to the innovations in DNA sequencing technologies. Particularly between 2008 and 2018, the international consortium *Madagascar, Anthropologie, Génétique et Ethno-linguistique* (MAGE) launched an extensive survey of Malagasy genetic diversity across the island. Indeed 2,691 mitochondrial samples were fully sequenced; chrY haplogroups were determined for 1,554 male individuals; and using microarray technology, 700 individuals were genotyped across the 22 autosomes, representing 2.2 millions of Single Nucleotide Polymorphisms (SNPs) (Pierron et al., 2017). Based on genome-wide data, the African genetic contribution was 59.4 ± 0.4%; the Asian component was 36.6 ± 0.4%, whereas the West-Eurasian component was only 3.9 ± 0.1%.

All studied individuals presented a similar pattern associating African and East Asian components, but with considerable variation, as estimates based on population global ancestry show diversity in such proportions across Madagascar. Based on the genetic patterns of each individual, clustering algorithms revealed that the individual's spatial

distribution in the island presented a strong correlation of geography and distinctive genetic clusters. This genetic diversity shows a heterogeneous situation across the island and suggests that the admixture process should be studied locally. By analyzing autosomal linkage-disequilibrium decay and according to the different genetic clusters, the oldest admixture event was dated around 800-900 years BP, and corresponds to the populations who inhabit the south and east coast of Madagascar; the population living in the central region of the island showed an admixture date around 700 years BP. The most recent date of admixture (~ 600 years BP) corresponds to the populations located in the north of the island. When taking into account the 700 genotyped individuals, the date of admixture converges to a recent event happening between 600 and 900 years BP (Pierron et al., 2017).

The analysis of mtDNA and chrY from the MAGE consortium showed that the paternal lineages of African origin were much more frequent in Madagascar than lineages from East Asia. The distribution of paternal lineages (chromosome Y) show a pronounced African ancestry all over the island, with a low frequency of Asian paternal lineages, reaching 30% in the central highland. On the other hand, African and Asian mtDNA proportions were similar (42.4% and 50.1%, respectively). The analysis of the maternal lineages showed that the African genetic contribution is only in the majority on the north of the island; with a higher frequency of Asian maternal lineages in the center and the south of the island. The distribution of African and Asian ancestry across the island reveals that the admixture was probably sex biased and happened heterogeneously across Madagascar, suggesting independent migration in Madagascar from African and Asians populations. Additionally, the whole sequencing of mtDNA revealed the presence of a specific ancient haplogroup (M23) with no member outside Madagascar, which might suggest an old settlement of Madagascar (Razafandrizaka, 2010; Razafindrazaka et al., 2010; Ricaut et al., 2009). Later results showed that while the M23 haplogroup is still only found in Madagascar, all Malagasy share a recent ancestor (1,200 ± 300 years BP, Pierron et al., 2017). As mentioned by the authors (Pierron et al., 2017), M23 diversity does not support an ancient settlement of Madagascar by a population predating the Bantu and Austronesian arrivals.

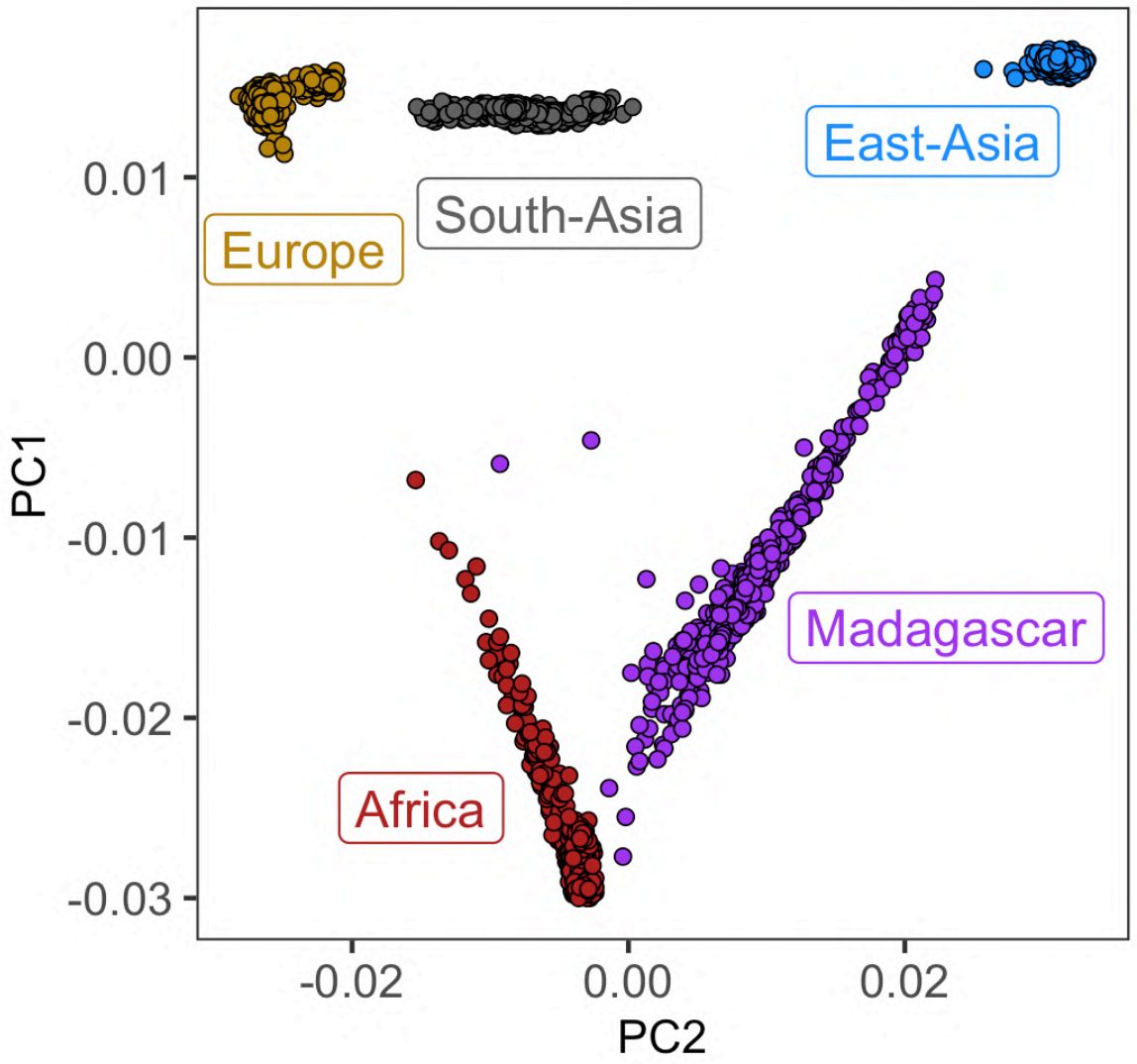


Figure 1. PCA results for 2,851 individuals from Madagascar and 1000 Genomes reference panel (The 1000 Genomes Project Consortium et al., 2015)**.** The colors for individuals from Madagascar, African, European, South Asian and East Asian populations correspond to purple, red, yellow, grey and blue. On the *x*-Axis the value of PC2 is shown, while *y*-axis denotes the value of PC1, with each dot in the figure representing one individual. Principal component analysis was performed with the smartPCA method (Patterson et al., 2006).

South Eastern Africa genetic contribution

Genetically speaking, there is evidence that a large proportion of the ancestors of the present Malagasy population were from African descent, as the African gene flow represents at least 60% of the autosomal genetic background. On average, the proportion of African ancestry is much more represented in the coastal region than in the highland. Particularly, the populations with the highest proportion of African ancestry are located on the north of the island (Pierron et al., 2017). However, it is important to specify that even in people with the smallest African ancestry contribution, near the capital Antananarivo, the level still reach 30%.

Regarding the possible geographic origin of Malagasy African ancestors, it is possible to infer the African ancestral contributions by computing the sharing of chromosome fragments between Malagasy and other populations. These fragments, called identity-by-descent (IBD) segments (Browning & Browning, 2013), have shown that populations from Kenya and the horn of Africa do not seem to share recent ancestors with the Malagasy. Similarly, hunter-gatherer populations from east and South Africa are distantly related to Malagasy (Pierron et al., 2014). Nevertheless, there are hunter-gatherer populations inhabiting Madagascar nowadays. In fact, one study based on the genetic variants of Mikea (hunter-gatherers), Vezo (semi-nomadic fisherman) and Temoro (farmers) populations, the authors tested whether and to what extent the Mikea population share their genetic ancestry with their neighboring Malagasy populations, concluding that the Mikea population have reverted to a hunter-gatherer lifestyle (Pierron et al., 2014), indicating that they originated recently from the admixture of Bantu and Austronesian genetic ancestors. Later, in 2017, it was confirmed that Bantu-speaking populations inhabiting South Eastern Africa shared the highest number of IBD segments. Consistently, ~98% of the African maternal and paternal lineages sampled across the island were associated with Bantu-speaking populations (Pierron et al., 2017). The size of IBD segments shared between Malagasy and Bantu-speaking populations indicate recent connections, pointing to a recent split, around 1,500 years BP, between the African ancestral population of Malagasy and the Bantu-speaking population probably inhabiting the African east coast.

South East Asia genetic contribution

In addition to a major African genetic input, there is at least a 30% Asian input into the Malagasy genetic pool (Pierron et al., 2014; Tofanelli et al., 2009). One key signature of this genetic input has been described thanks to mitochondrial DNA analyses, showing that a variant of the mitochondrial Polynesian motif (B4a1a1a), termed the Malagasy motif (characterized by two specific polymorphisms 1473 and 3423A) is found at an elevated frequency in Madagascar (Razafindrazaka, 2010). The rarity of the Polynesian motif in Indonesia whose frequency reaches only 2% make this key signature very difficult to detect among Indonesian populations (Cox et al, 2010; Kusuma et al, 2015). Another possibility is that this Malagasy motif arose in Madagascar among the first Indonesian settlers (Razafindrazaka et al., 2010). Based on the mitochondrial haplotype's diversity, a very controversial study has suggested a small number of initial Austronesian settlers arriving in Madagascar. Cox *et al*., has proposed that Polynesian motif's high frequency could be explained if the settlement of Madagascar was done by two very small women populations: 27 females from the Austronesian-speaking population and 3 females from the Bantu-speaking population (Cox et al., 2012). However, this scenario having two extreme founder events has not been confirmed by subsequent genetic data representative of Malagasy diversity, where there is a high diversity of mitochondrial lineages, specifically of African origins (Pierron et al., 2017).

Concerning the geographical origins of the Malagasy Asian ancestors, some other studies have attempted to determine the ancestral connections based on associations among contemporaneous populations (Brucato et al., 2016; Kusuma et al., 2016; Pierron et al., 2014). Based on the genetic patterns of Malagasy and Indonesian populations, it was suggested that the Austronesian ancestral component could come from either Java, Borneo, or Sulawesi (Pierron et al., 2014). Later, one study pointed out that the most plausible Asian ancestor of the Malagasy were the Banjar people, a population who lived in Banjarmasin, southeast Borneo (Brucato et al., 2016). Based on the data produced by the MAGE consortium, extending the number of Malagasy individuals analyzed, it was shown that the populations from Indonesia, especially south Borneo, shared the highest number of chromosome fragments. This confirmed that among the populations sampled from outside

Africa, they are the closest link with the Malagasy (Pierron et al., 2017). This dataset also revealed that the Asian contribution was heterogeneous across Madagascar, where individuals presenting the highest proportions of Asian ancestry were located in the central highlands. As mentioned above, half of the maternal lineages analyzed in Madagascar are of Asian origin and they are distributed heterogeneously across the island, with a major representation in the highlands. Similarly, Asian paternal lineages, representing ~20% from all paternal lineages surveyed, are mostly found in populations near the center-highlands of Madagascar. Based on the number of chromosome fragments shared though time, demographic simulations suggested a split between the Malagasy Asian ancestors and source populations from south Borneo, leading to at least two possible scenarios: a hard split 2,500 years BP, and a slow divergence with less and less migration between 3,000–2,000 years BP (Pierron et al., 2017). This implies a complex evolutionary history for Malagasy Asian ancestors, as they seem to have diverged from South Borneo populations between 2000 - 3000 years BP. Future work involving ancient DNA and more genetic data from Island South East Asia will be crucial to continue the study of the past of the Asian genetic ancestors.

West Eurasia genetic contribution

Given the context of connections and trade across the Indian Ocean (Boivin et al., 2013), the possible genetic flux to Madagascar from other populations, rather than African and Asian, has been addressed (Capredon et al., 2012, 2013; Pierron et al., 2017). For instance, their contribution has been described based paternal lineages, where other Y chromosome lineages with an origin different than African or Asian are also present; some of these (R1a, J2, T1, G2) are also present in the Middle East and may reflect the Muslim influence on Madagascar and the Comoros. Haplogroup R1b, characteristic of western Europeans, is present in low frequency (0.9%), suggesting a limited paternal contribution from western Europeans. For instance, archeological evidence and ethnographic studies has linked the Arab contribution to a context of trading posts (Capredon et al., 2012; Vérin, 1967), while the west European contribution occured through colonial migratory events composed mainly of men. Consistently, based on 2,691 mtDNA sequences, there was no evidence of maternal gene flow from Europe or the Middle East (Pierron et al., 2017). By studying the

genetic information shared between Malagasy and other populations through IBD segments, it was detected a minor genetic contribution from West Eurasian populations, allowing to exclude other populations, besides African and Asian, as putative major contributors. The evolutionary history of Malgasy populations is complex, further work extending the available genetic data, with additional reference populations could help to identify such small contributions.

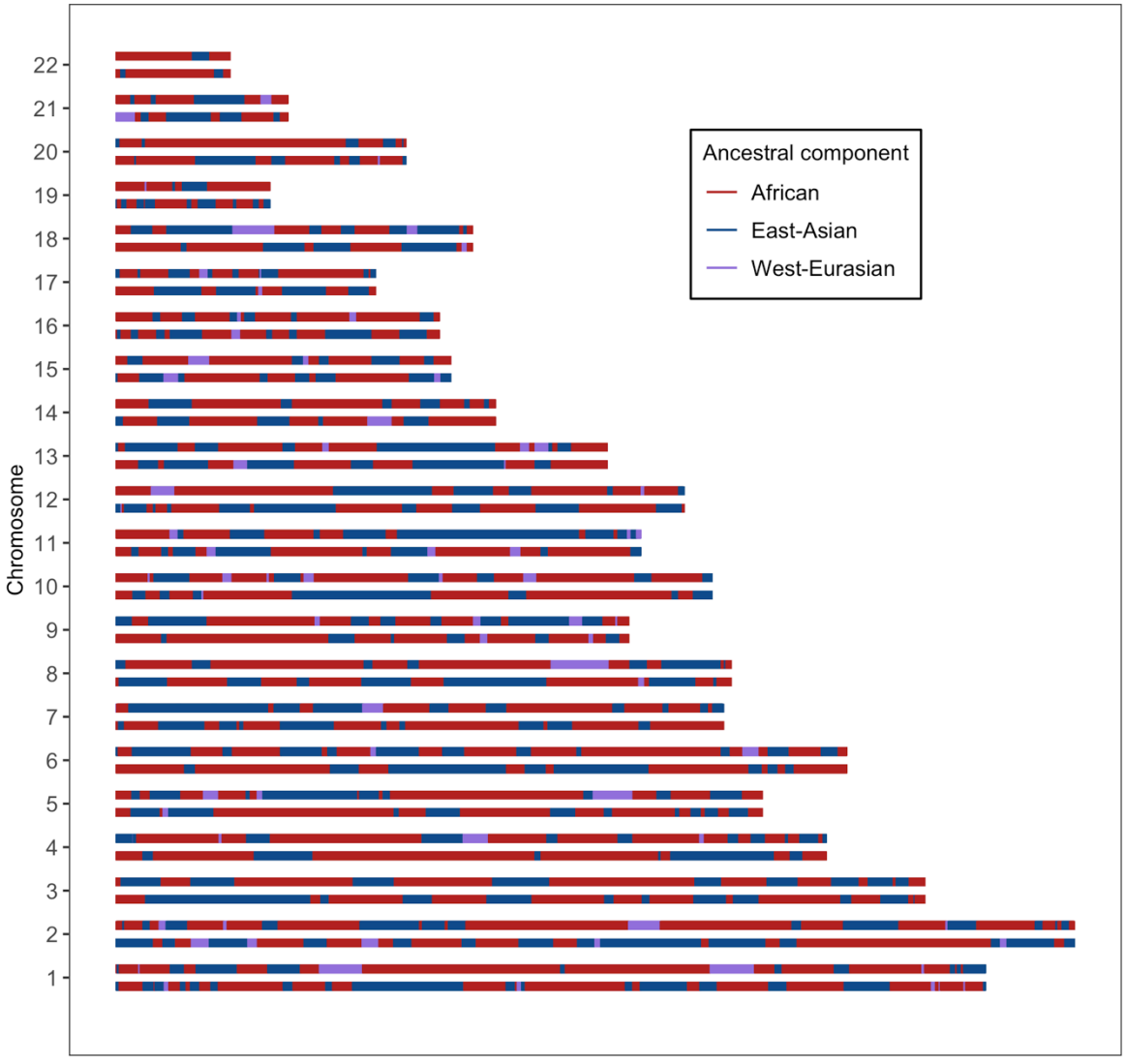


Figure 2. Local Ancestry Inference results for a Malagasy individual using as proxy of the ancestral populations the reference data set of African, European and East Asian individuals from 1000 Genomes reference panel (The 1000 Genomes Project Consortium et al., 2015). The software RFMix (Maples et al., 2013) was used for inferring local ancestry

across the 22 autosomes. We can observe that the majority of the genome was inferred as of African origin (red), with a significant fraction being of East-Asian origin (blue). The West-Eurasian contribution is much less represented, accounting for < 2% of the sites studied.

Conclusions

As it has been described, genetic analyses converge on two main ancestries for the entire actual Malagasy population, namely Bantu from southeast Africa and Austronesians from Indonesia (in particular, south Borneo), with a very limited contribution from Europe and the Middle East). The global ancestry heterogeneity across Madagascar seems to be affected by the settlement of the island, geographic barriers and demography. This has impacted the human genetic diversity across the island, as an elevated Asian contribution is found in the populations from the central highlands, while individuals from the coastal regions present a higher African contribution. Despite the growing research focused on the settlement of the island, there is still little information about the ancestral populations before the admixture. The time when the first Bantu speakers arrived and the manner of their arrival are not known; and the date of arrival of the Austronesian population is even more debated, as this event led this population to move 7000 km from their closest genetic ancestors in South Borneo. While the process of arrival of human populations is still unresolved, recent studies raise also the question of its impact on endemic biodiversity and in particular its role in the extinction of numerous endemic vertebrates weighing >10kg in Madagascar (Crowley, 2010; Douglass et al., 2019; Godfrey et al., 2019). Despite there is no historical record of all the events that have happened during the settlement on the island, genetic, archeological, cultural information allow step by step to reconstruct the past of the Malagasy population.

Bibliography

Adelaar, K. A. (2009). *Towards an integrated theory about the Indonesian migrations to Madagascar. Ancient Human Migrations : a multidisciplinary approach.* Foundations of Archeological Inquiry.

Adelaar, K. A., & Himmelmann, N. (2005). *The Austronesian languages of Asia and Madagascar*. Routledge.

Beaujard, P. (2011). The first migrants to Madagascar and their introduction of plants: Linguistic and ethnological evidence. *Azania: Archaeological Research in Africa*, *46*(2), 169–189. https://doi.org/10.1080/0067270X.2011.580142

Boivin, N., Crowther, A., Helm, R., & Fuller, D. Q. (2013). East Africa and Madagascar in the Indian Ocean world. *Journal of World Prehistory*, *26*(3), 213–281. https://doi.org/10.1007/s10963-013-9067-4

Browning, B. L., & Browning, S. R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. *Genetics*, *194*(2), 459–471. https://doi.org/10.1534/genetics.113.150029

Brucato, N., Kusuma, P., Cox, M. P., Pierron, D., Purnomo, G. A., Adelaar, A., Kivisild, T., Letellier, T., Sudoyo, H., & Ricaut, F.-X. (2016). Malagasy Genetic Ancestry Comes from an Historical Malay Trading Post in Southeast Borneo. *Molecular Biology and Evolution*, *33*(9), 2396–2400. https://doi.org/10.1093/molbev/msw117

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). *Mitochondrial DNA and human evolution*. 6.

Capredon, M., Brucato, N., Tonasso, L., Choesmel-Cadamuro, V., Ricaut, F.-X., Razafindrazaka, H., Rakotondrabe, A. B., Ratolojanahary, M. A., Randriamarolaza, L.-P., Champion, B., & Dugoujon, J.-M. (2013). Tracing Arab-Islamic Inheritance in Madagascar: Study of the Y-chromosome and Mitochondrial DNA in the Antemoro. *PLoS ONE*, *8*(11), e80932. https://doi.org/10.1371/journal.pone.0080932

Capredon, M., Sanchez-Mazas, A., Guitard, E., Razafindrazaka, H., Chiaroni, J., Champion, B., & Dugoujon, J.-M. (2012). The Arabo-Islamic migrations in Madagascar: First genetic study of the GM system in three Malagasy populations: Arabo-Islamic migrations in Madagascar. *International Journal of Immunogenetics*, *39*(2), 161–169. https://doi.org/10.1111/j.1744-313X.2011.01069.x

Cox, M. P., Nelson, M. G., Tumonggor, M. K., Ricaut, F.-X., & Sudoyo, H. (2012). A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1739), 2761–2768. https://doi.org/10.1098/rspb.2012.0012

Crowley, B. E. (2010). A refined chronology of prehistoric Madagascar and the demise of the megafauna. *Quaternary Science Reviews*, *29*(19–20), 2591–2603. https://doi.org/10.1016/j.quascirev.2010.06.030

Dahl OC. 1951. Malgache et maanjan: une comparaison linguistique. Oslo, Norway: Edege-Intituttet.

Douglass, K., Hixon, S., Wright, H. T., Godfrey, L. R., Crowley, B. E., Manjakahery, B., Rasolondrainy, T., Crossland, Z., & Radimilahy, C. (2019). A critical review of radiocarbon dates clarifies the human settlement of Madagascar. *Quaternary Science Reviews*, *221*, 105878. https://doi.org/10.1016/j.quascirev.2019.105878

Godfrey, L. R., Scroxton, N., Crowley, B. E., Burns, S. J., Sutherland, M. R., Pérez, V. R., Faina, P., McGee, D., & Ranivoharimanana, L. (2019). A new interpretation of

Madagascar's megafaunal decline: The "Subsistence Shift Hypothesis". *Journal of Human Evolution*, *130*, 126–140. https://doi.org/10.1016/j.jhevol.2019.03.002

Hurles, M. E., Sykes, B. C., Jobling, M. A., & Forster, P. (2005). The Dual Origin of the Malagasy in Island Southeast Asia and East Africa: Evidence from Maternal and Paternal Lineages. *The American Journal of Human Genetics*, *76*(5), 894–901. https://doi.org/10.1086/430051

Jobling, M. (2014). *Human Evolutionary Genetics, Second Edition*. 690.

Korunes, K. L., & Goldberg, A. (2021). Human genetic admixture. *PLOS Genetics*, *17*(3), e1009374. https://doi.org/10.1371/journal.pgen.1009374

Kusuma, P., Brucato, N., Cox, M. P., Pierron, D., Razafindrazaka, H., Adelaar, A., Sudoyo, H., Letellier, T., & Ricaut, F.-X. (2016). Contrasting Linguistic and Genetic Origins of the Asian Source Populations of Malagasy. *Scientific Reports*, *6*(1), 26066. https://doi.org/10.1038/srep26066

Lawler A. 2014. Sailing Sinbad's seas. Science 344:1440–1445

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, *193*(4), 1233–1254. https://doi.org/10.1534/genetics.112.147330

Ma, J., & Amos, C. I. (2012). Principal Components Analysis of Population Admixture. *PLoS ONE*, *7*(7), e40115. https://doi.org/10.1371/journal.pone.0040115

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Mitchell, R. J., & Hammer, M. F. (1996). Human evolution and the Y chromosome. *Current Opinion in Genetics & Development*, *6*(6), 737–742. https://doi.org/10.1016/S0959-437X(96)80029-3

Mourant, A.E. (1954). The distribution of the human blood groups. Blackwell Scientific Publications; Oxford.

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

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, *40*(5), 646–649. https://doi.org/10.1038/ng.139

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L. M.-A., Rakotomalala, M. M., Razafiarivony, M., Rasoarifetra, B., Raharijesy, M. A., Razafindralambo, L., Ramilisonina, Fanony, F., Lejamble, S., Thomas, O., Mohamed Abdallah, A., Rocher, C., Arachiche, A., … Letellier, T. (2017). Genomic landscape of human diversity across Madagascar. *Proceedings of the National Academy of Sciences*, *114*(32), E6498–E6506. https://doi.org/10.1073/pnas.1704906114

Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.-X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.-A., Blench, R. M., Letellier, T., & Kivisild, T. (2014). Genome-wide evidence of Austronesian-Bantu admixture and cultural

reversion in a hunter-gatherer group of Madagascar. *Proceedings of the National Academy of Sciences*, *111*(3), 936–941. https://doi.org/10.1073/pnas.1321860111

Rakotomalala, M. (1996). *Recherches en musicologie malgache*. Les Cahiers du Cite.

Razafindrazaka, H. (2010). *Le peuplement humain de Madagascar : anthropologie génétique de trois groupes traditionnels* [Doctoral dissertation, Victoria University]. Université Toulouse III.

Razafindrazaka, H., Ricaut, F.-X., Cox, M. P., Mormina, M., Dugoujon, J.-M., Randriamarolaza, L. P., Guitard, E., Tonasso, L., Ludes, B., & Crubézy, E. (2010). Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *European Journal of Human Genetics*, *18*(5), 575–581. https://doi.org/10.1038/ejhg.2009.222

Regueiro, M., Mirabal, S., Lacau, H., Caeiro, J. L., Garcia-Bertrand, R. L., & Herrera, R. J. (2008). Austronesian genetic signature in East African Madagascar and Polynesia. *Journal of Human Genetics*, *53*(2), 106–120. https://doi.org/10.1007/s10038-007-0224-4

Ricaut, F.-X., Razafindrazaka, H., Cox, M. P., Dugoujon, J.-M., Guitard, E., Sambo, C., Mormina, M., Mirazon-Lahr, M., Ludes, B., & Crubézy, E. (2009). A new deep branch of eurasian mtDNA macrohaplogroup M reveals additional complexity regarding the settlement of Madagascar. *BMC Genomics*, *10*(1), 605. https://doi.org/10.1186/1471-2164-10-605

Sachs, C. (1938). Les Instruments de Musique de Madagascar. *Nature* 142, 191.

Serva, M. (2012). The Settlement of Madagascar: What Dialects and Languages Can Tell Us. *PLoS ONE*, *7*(2), e30666. https://doi.org/10.1371/journal.pone.0030666

Singer, R., Budtz-Olsen, O. E., Brain, P., & Saugrain, J. (1957). Physical features, sickling and serology of the Malagasy of Madagascar. *American Journal of Physical Anthropology*, *15*(1), 91–124. https://doi.org/10.1002/ajpa.1330150113

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Thornton, T. A., & Bermejo, J. L. (2014). Local and Global Ancestry Inference and Applications to Genetic Association Analysis for Admixed Populations: Inferring and Accounting for Ancestry Admixture. *Genetic Epidemiology*, *38*(S1), S5–S12. https://doi.org/10.1002/gepi.21819

Tofanelli, S., Bertoncini, S., Castri, L., Luiselli, D., Calafell, F., Donati, G., & Paoli, G. (2009). On the Origins and Admixture of Malagasy: New Evidence from High-Resolution Analyses of Paternal and Maternal Lineages. *Molecular Biology and Evolution*, *26*(9), 2109–2124. https://doi.org/10.1093/molbev/msp120

Vérin, P. (1967). *Les Arabes dans l'Océan Indien et à Madagascar. Arabes et Islamisés à Madagascar et dans L'Océan Indien*. Le Centre d'Archéologie de la Faculté des

Lettres et des Sciences Humaines de l'Université de Madagascar, Antananarivo.

Vérin, P. (1976). The African Element in Madagascar, *Azania: Archaeological Research in Africa*, 11:1, 135-151, DOI: 10.1080/00672707609511234

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, *4*(3), e72. https://doi.org/10.1371/journal.pbio.0040072

Wangkumhang, P., & Hellenthal, G. (2018). Statistical methods for detecting admixture. *Current Opinion in Genetics & Development*, *53*, 121–127. https://doi.org/10.1016/j.gde.2018.08.002

Conclusions and discussion

How to model the human settlement of Madagascar?

How to model the human settlement of Madagascar?

Thanks to this extensive bibliographic work, we were able to better understand the broad spectrum of plausible settlement scenarios that could explain the observed genetic diversity in Madagascar. However, although this bibliographic work provided the necessary context for studying the settlement of the island, numerous questions remain unsolved. We identified and listed these questions as follows:

- Who are the closest genetic parental groups of the Malagasy and when did Malagasy start to diverge from them?
- Did the ancestors of the Malagasy population(s) come from small or large populations?
- How many people originally founded the Malagasy population(s)?
- Was Madagascar settled via large-scale population movements, or through a smaller translocation of individuals?
- What is the possible chronology of admixture and settlement?
- How much time passed between the arrivals to the island and the admixture?

These questions concerning the evolutionary history of Madagascar can be addressed by using computational simulations, combined with the genetic data available. Despite the growing research focused on the settlement of Madagascar, there is still little information about the ancestral populations before the admixture (Douglass et al., 2019; Heiske et al., 2021). The how and when the first Bantu speakers arrived are not known; and the date of arrival of the Austronesian population is even more debated, as this event led this population to move 7000 km from their closest genetic ancestors in South Borneo. Thus, forming a comprehensive view of the migration events that led to the settlement of Madagascar is key to a better understanding of the evolutionary history of the population.

As mentioned before, the patterns of genetic variation across individual genomes are the result of accumulated effects of different past demographic processes (migrations, spatial expansions, etc.) and evolutionary forces (mutation, natural selection, etc.), combined with the stochasticity of inheritance (Loog, 2021). These complex processes and their interactions cannot be easily predicted or interpreted with only analytical genetic

algorithms. In consequence, computer programs have been developed to simulate abstract models of particular systems (Hoban et al., 2012), and can be used for inferring and understanding the functioning of these complex demographic histories. The general principle is to generate *in silico* data sets of genetic polymorphisms under specified scenarios, describing the evolutionary history and genetic diversity of a population (Hoban et al., 2012). In human evolution studies, a scenario might consider population expansions, split times between ancestral and descendant populations, as well as their population sizes and dispersal rates. Similarly, a mutation rate and assumptions about linkage disequilibrium between markers might be considered in these computer simulations. In this way, demographic processes can be inferred by comparing patterns of observed genetic variation to computationally simulated model predictions (Loog, 2021).

This classical approach of system biology can be summarized in **Figure B: Systems biology schema**. The first step is to define alternative models of genetic and demographic history (*i.e.* representing population continuity, replacement or admixture). The second step is to simulate many datasets for each of these alternative models, with each simulation using different values for key parameters, such as within-population or between-population historical events (Cooke & Nakagome, 2018; Hoban, 2014). In the last step, we choose summary statistics to compare the simulated and empirical data. Importantly, the chosen statistics should be informative for distinguishing the genetic variation among competing scenarios that are under consideration, and should be strongly influenced by the parameters of interest. As reviewed in Cooke et al (Cooke & Nakagome, 2018), there are several statistics that are useful to capture different features of patterns of genetic variation, such as statistics that are based on allele-frequency (*i.e.* Tajima's D and Fst) or linkage disequilibrium (r2, decay of extended haplotype homozygosity). Then, we may use *ad hoc* methods to compare the observed data to all the simulated data sets, in order to bound parameter values (Hoban, 2014). Alternatively, the approximate Bayesian computation (ABC) can be implemented for quantifying the fit between observed and simulated data, in order to reconstruct the posterior distribution of parameters and determine the reliability of parameter estimation (Cooke & Nakagome, 2018; Csilléry et al., 2010).

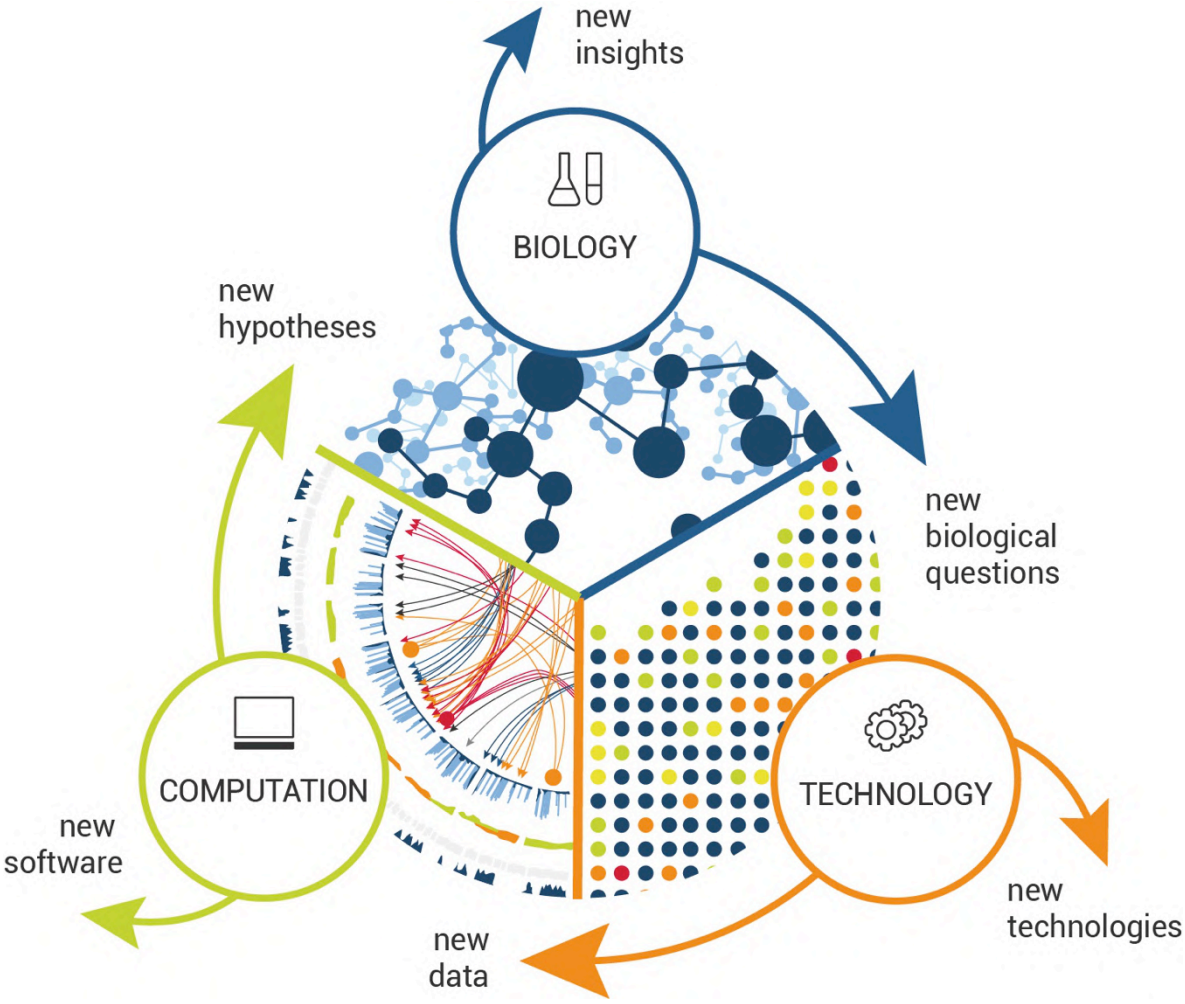


Figure B: Systems biology schema. A common definition of systems biology is the study of a given biological system by the perturbation of a property of that system, the measurement of resulting responses, the integration of these data, and the ultimate modeling of these data to describe the system as well as its response to perturbation (Conesa & Mortazavi, 2014). Image taken from http://www.omicscouts.com/en/disease-and-systems-biology.html.

Current simulation programs compute the frequency of genotypes, haplotypes or alleles that is comparable to the observed frequencies from analysis of samples recovered through field work (Hoban et al., 2012; Kelleher et al., 2016). Simulation programs differ in the software interface, with several simulators coded in high-level programming languages, such as R and Python. Additionally, some of them may include additional data directly from the simulated scenario, such as genetic diversity, demography, migration

events, phylogenetic relationships and mutations, *etc*. These simulation algorithms can be classified into forward and backward categories, they differ in approach, capabilities and computation times, which can lead to particular implementations for addressing different questions (Hoban et al., 2012). For instance, forwards-in-time simulators require defining initial conditions on genetic variation: each individual in the simulated population has to have a genotype, then the individual will follow a life cycle (composed of birth, selection, mating, reproduction, mutation, migration and death) (Hoban et al., 2012), such that the demographic and genetic makeup of subsequent generations is determined by the current generation. On the other hand, backwards-in-time simulators take a lineage approach, where the lineages coalesce progressively with a probability that is influenced by features of the sample and the evolutionary history. This approach considers the genealogy of the sampled DNA fragments (not each single individual in the populations) and follows them back in time to the most recent common ancestor (Hoban et al., 2012). Backwards-in-time simulators are generally faster, as they do not follow each simulated individual.

It has been shown that computer simulations are useful for comparing different models that could explain the genetic variation patterns observed in anthropology studies, allowing to test the plausibility of particular scenarios. Using summary statistics on different models, DeGiorgio and colleagues (DeGiorgio et al., 2009) were able to identify the most probable model explaining the decline of heterozygosity in populations with increasing distance from Africa. The selected model implied serial founder events beginning from an African origin, where a population is formed from a subset of the individuals in the founding population, then, this new population experiences a founder event (founded by a small group); it grows to a larger size, after which a subset of the population becomes the founding group for a third population. The founding process is then iterated. We can interpret the simulation of the serial founder model as follows: when a new colony is founded, it carries only a subset of the diversity from the previous colony, and therefore, a heterozygosity decrease occurs. Thus, if the source in the simulations is placed in Africa, then the prediction of the serial founder model matches the observed pattern of heterozygosity in human populations. This example demonstrates how computer models using heterozygosity as summary statistics can successfully discriminate among the

alternative anthropological scenarios (i.e. independent regions or isolation by distance scenarios).

Based on this summary statistics methodology, and in order to study the evolutionary history of Malagasy and Madagascar settlement, we looked for identifying the plausible models of Madagascar settlement that, at the same time, represented scenarios that can be interrogated using a genetic based modelling approach. We categorized the models according to the event of interest in the evolutionary history of Malagasy, implementing previously proposed settlement scenarios, evaluating the pertinence of precedent genetic results, and testing other potential scenarios based on genetic signals discovered in the laboratory. We analysed the scenarios where the settlement of Madagascar represented an admixture event:

a) through a single founder event to both Asian and African ancestral populations of Malagasy. As proposed by Tofanelli et al, (Tofanelli et al., 2009).

b) involving a strong founder event for the Asian ancestral population of Malagasy. As proposed by Cox and colleagues using mtDNA (Cox et al., 2012)

c) involving a founder event for the African ancestral population and continuous gene flow from the Asian ancestral population inhabiting Indonesia or elsewhere. As mentioned in Cox et al., 2012, another hypothesis is that Indonesian maritime traders initially settled Madagascar, either as a single colonization event or via repeated settlement waves from the same source population.

d) involving a founder event for the Asian ancestral populations (under different demographic trajectories) and a single founder event (or continuous gene flow) for the African ancestral population.

In this manner, we interrogated the possible settlement model of Madagascar by studying and simulating the admixture event and the evolutionary history of African and Asian ancestral populations of Malagasy. With this, we expect to bring together ecological, population genetic, and evolutionary timescales in order to study the effect of multiple events over time on genetic variation, as well as their relationship with major ecological events in Madagascar.

During this 3-year doctoral thesis, we decided to explore further the settlement of Madagascar and implemented computational simulations. Thanks to this strategy, we succeeded in better understanding the functioning of different bioinformatics algorithms. Regarding this aspect, the bibliographic work was essential, as we were able to detect plausible demographic scenarios for explaining the observed genetic diversity in Madagascar. Moreover, it appeared that the construction of plausible models for the settlement of Madagascar necessitates as a prerequisite a realistic model of the evolutionary history of the human species and more particularly information about: (1) how humans settled the world, (2) how humans interacted with their environment during the last 10,000 years, and (3) how these interactions impacted the demographic history of populations. Different disciplines and researchers have addressed these questions, providing diverse scopes and hypotheses. Therefore, considering different layers of information (from a local to a global level) was a useful step for reconstructing the events involved in the settlement of Madagascar.

CHAPTER II

TESTING PLAUSIBLE SCENARIOS FOR THE SETTLEMENT OF MADAGASCAR

CHAPTER II

TESTING PLAUSIBLE SCENARIOS FOR THE SETTLEMENT OF MADAGASCAR

Introduction

On one hand, all of the bibliographic and genetic data available (Douglass et al., 2019; Godfrey et al., 2019; Heiske et al., 2021; Pierron et al., 2017) enabled us to identify and compare several models regarding the settlement of Madagascar. On the other hand, thanks to the innovations in DNA sequencing technologies, and the international consortium MAGE (Madagascar, Anthropologie, Génétique et Ethno-linguistique), the genetic data from Madagascar have been expanded (Pierron et al., 2017). Between 2008 and 2018 the MAGE consortium conducted an extensive sampling of Malagasy genetic diversity across the island: 2,691 mitochondrial samples were fully sequenced; chrY haplogroups were determined for 1,554 male individuals; and using microarray technology, 700 individuals were genotyped across the 22 autosomes, representing 2.2 millions of Single Nucleotide Polymorphisms (SNPs) (Pierron et al., 2017). By generating the bibliographic work and by studying the genetic data available, we investigated the evolutionary history of Malagasy and their ancestors, focusing on the evolutionary forces acting on the populations and their demographic fluctuations. As presented before, we identified and listed the questions that could be addressed with computational simulations and the genetic data available:

- Who are the closest genetic parental groups of the Malagasy and when did Malagasy start to diverge from them?
- Did the ancestors of the Malagasy population(s) come from small or large populations?
- How many people originally founded the Malagasy population(s)?
- Was Madagascar settled via large-scale population movements, or through a smaller translocation of individuals?

- What is the possible chronology of admixture and settlement?
- How much time passed between the arrivals to the island and the admixture?

At the beginning of my thesis, all these questions about the settlement of Madagascar received a new interest. It appeared that modelling human population migration and demography could not only bring information about the history of Madagascar, and the travels across the Indian Ocean, but also might provide important information to understand the impact of human populations when colonizing new environments.

Indeed, the last two millennia have seen on Madagascar a decline in biodiversity, which ultimately resulted in the extinction of all endemic large-bodied vertebrates, such as the giant lemur, the elephant bird, the turtles and the hippopotamus (Douglass et al., 2019; Godfrey et al., 2019; Li et al., 2020). While today Madagascar is still a unique biodiversity hotspot, its landscapes drastically changed since. The reasons of this collapse are heavily discussed, specifically the impact of human presence (Anderson et al., 2018; Hixon, Douglass, Crowley, et al., 2021): these extinctions have been variously attributed to climate change, the arrival of human populations, a cultural transition after the settlement, or a combination of these factors (Crowley, 2010; Godfrey et al., 2019; Voarintsoa et al., 2017). As described in Chapter I, we know that genetic and linguistic analyses have shown that the Malagasy population has emerged from a recent admixture that happened during the last millennium. However, the past of the human Malagasy population is poorly known and there is still no consensus about the past of the ancestral populations before the admixture event. We hypothesized that a better understanding of the demographics of early Malagasy human populations may provide important clues to better understand their possible impact on the island's ecosystem.

Based on a genomic analysis of a large representative sample of Madagascar's inhabitants (MAGE dataset) we intended to infer the demographic trajectories of Malagasy and their ancestors from Africa and Asia. As a first analysis, we inferred changes in effective population size (Ne) over recent times in Madagascar using the haplotype-based IBDNe method (Browning & Browning, 2015) and Ancestry-specific IBDNe-AS (Browning et al., 2018). Our results suggested that the African population that settled Madagascar came from an event of demographic expansion, in concordance with the Bantu

expansion on the African east coast; while the Asian population that arrived in the island derived from a population that suffered a drastic Ne decrease, which trajectory does not correspond to the demographic dynamics observed in the analysed Indonesian populations. Importantly, the IBDNe method has some caveats that need to be taken into account: the Ne estimation procedure cannot fully capture sharp changes; the method does not recover complex demographic histories (as closed and homogenous populations are recommended). We then looked for the basis of the method and discovered that IBDNe-AS relies on the number and length of shared genetic fragments, called Identity-By-Descent haplotypes (IBD) (**Figure C: Visualizing IBD fragments**), which are dependent on the coalescence probabilities and are therefore markers of Ne fluctuations over time. In this manner and complemented with local ancestry inference methods, it was possible to reliably assess the characteristics and demographic parameters of each ancestral population prior to an admixture event (Browning et al., 2018). Therefore, the demographic history depicted by IBDNe-AS for the Asian ancestral population is based on the fact that the average pair of individuals in Madagascar shares more haplotypes of Asian origin in comparison to those from African origin. This overrepresentation of IBD segments of Asian origin in Madagascar was not observed in any other Indonesian populations analysed in the study.

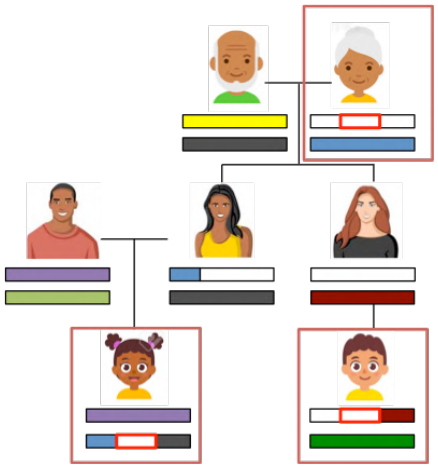


Figure C: Visualizing IBD fragments. Chromosome segments shared between two individuals (red regions shared between two cousins) because of a common ancestor (inherited from their grandmother) are called Identity-by-Descent segments. These haplotypes are useful for measuring genetic similarity between individuals. The number and length of shared IBD fragments in a population are dependent on the coalescence

probabilities and are therefore markers of fluctuations in effective population size (Ne) over time.

We hypothesize that this overrepresentation of IBD could be an important genetic marker as it provided plausible answers to our stated questions. Thus, we conducted a systems biology approach (see **Figure B: Systems Biology schema**) as follows:

1. Based on the bibliography, we modelled settlement scenarios with different demographic parameters to characterize the demographic dynamics of Malagasy's ancestral populations.
2. We characterized the effect of these scenarios and parameters on quantity, length and ancestry of shared chromosome fragments between Malagasy individuals (using the IBD-sharing distribution associated with local ancestry inference as a summary statistic).
3. We compared the results obtained across the diverse models and scenarios with the observed IBD-sharing distribution in our empirical dataset.

This approach allowed us to infer that a genetic admixture event with an African population marked the end of the intriguing thousand-year period of isolation for the Asian ancestral population. This event was followed by a rapid demographic expansion, where the Malagasy Ne increased by a factor of 100 in a period of approx. 700 years. Moreover, this major demographic transition is contemporaneous with several archaeological evidence of a cultural shift through the introduction of cattle (Hixon, Douglass, Crowley, et al., 2021), rice (Crowther et al., 2016), and urbanization on the island (Radimilahy Chantal, 1985). We consider these major events in the history of the Malagasy human population should be analysed together in order to understand the major landscape transformations and the megafauna disappearance, which happened on Madagascar from 500 to 1300 years before present.

The reasoning, methodology and results of this approach are synthesized in the article "The loss of biodiversity in Madagascar is contemporaneous with major historical events", which focuses on the demographic history of Malagasy populations and their relationship with the environment.

Article 3

The loss of biodiversity in Madagascar is contemporaneous with major historical events

Alva Omar, Leroy Anaïs, Heiske Margit, Pereda-Loth Veronica, Tisseyre Lenka, Boland Anne, Deleuze Jean-François, Rocha Jorge, Schlebusch Carina, Fortes-Lima Cesar, Stoneking Mark, Radimilahy Chantal, Rakotoarisoa Jean-Aimé, Letellier Thierry, Pierron Denis

This work was published in Current Biology journal on November 5th of 2022. We send the article on April 7th, 2022 and we received favorable opinions for publication after "major revisions" on May 4th. The article was accepted for publication on september 28th.

Article

Current Biology

The loss of biodiversity in Madagascar is contemporaneous with major demographic events

Highlights

- An Asian population of few hundred individuals was isolated for more than 1,000 years
- This isolation ended around 1,000 years ago by admixture with a small African population
- The newly admixed Malagasy population underwent a rapid demographic expansion
- The population growth coincides with extensive changes in Madagascar's landscape

Authors

Omar Alva, Anaïs Leroy, Margit Heiske, ..., Jean-Aimé Rakotoarisoa, Thierry Letellier, Denis Pierron

Correspondence

denis.pierron@univ-tlse3.fr

In brief

Based on genomic studies, Alva et al. show that the Malagasy ancestral Asian population was isolated for more than 1,000 years. This isolation ended around 1,000 years BP by admixture with a small African population. Around this time, there was a rapid demographic expansion growth, which coincides with extensive changes in Madagascar's landscape.

Alva et al., 2022, Current Biology *32*, 1–11
December 5, 2022 © 2022 Elsevier Inc.
https://doi.org/10.1016/j.cub.2022.09.060

CellPress

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

Current Biology

CellPress

Article

The loss of biodiversity in Madagascar is contemporaneous with major demographic events

Omar Alva,[1] Anaïs Leroy,[1] Margit Heiske,[1] Veronica Pereda-Loth,[1] Lenka Tisseyre,[1] Anne Boland,[2] Jean-François Deleuze,[2] Jorge Rocha,[3,4,5] Carina Schlebusch,[6] Cesar Fortes-Lima,[6] Mark Stoneking,[7,8] Chantal Radimilahy,[9] Jean-Aimé Rakotoarisoa,[9] Thierry Letellier,[1] and Denis Pierron[1,10,11,*]

[1]Équipe de Médecine Evolutive, EVOLSAN faculté de chirurgie dentaire, Université Toulouse III, Toulouse, France
[2]Commissariat à l'Energie Atomique, Institut Génomique, Centre National de Génotypage, 91000 Evry, France
[3]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Portugal
[4]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, 4099-002 Porto, Portugal
[5]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Portugal
[6]Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18C, 75236 Uppsala, Sweden
[7]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany
[8]Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, Villeurbanne, France
[9]Musée d'Art et d'Archéologie, University of Antananarivo, Antananarivo, Madagascar
[10]Twitter: @Evol_San
[11]Lead contact
*Correspondence: denis.pierron@univ-tlse3.fr
https://doi.org/10.1016/j.cub.2022.09.060

SUMMARY

Only 400 km off the coast of East Africa, the island of Madagascar is one of the last large land masses to have been colonized by humans. While many questions surround the human occupation of Madagascar, recent studies raise the question of human impact on endemic biodiversity and landscape transformation. Previous genetic and linguistic analyses have shown that the Malagasy population has emerged from an admixture that happened during the last millennium, between Bantu-speaking African populations and Austronesian-speaking Asian populations. By studying the sharing of chromosome segments between individuals (IBD determination), local ancestry information, and simulated genetic data, we inferred that the Malagasy ancestral Asian population was isolated for more than 1,000 years with an effective size of just a few hundred individuals. This isolation ended around 1,000 years before present (BP) by admixture with a small African population. Around the admixture time, there was a rapid demographic expansion due to intrinsic population growth of the newly admixed population, which coincides with extensive changes in Madagascar's landscape and the extinction of all endemic large-bodied vertebrates. Therefore, our approach can provide new insights into past human demography and associated impacts on ecosystems.

INTRODUCTION

A place of high biodiversity, Madagascar is regarded as one of the last large land masses to have been colonized by humans.[1] Although this island is only 400 km off the coast of East Africa, the question of how and when human populations arrived there is still unresolved, and recent studies raise the question of the impact of human colonization on endemic biodiversity. The last 2 millennia on the island of Madagascar have seen a decline in biodiversity that ultimately resulted in the extinction of all endemic large-bodied vertebrates, such as giant lemurs, elephant birds, turtles, and hippopotami.[1–3] These extinctions have been variously attributed to climate change, the arrival of human populations, a cultural transition after human settlement, or a combination of these factors.[4,5] Resolving the history of Malagasy human populations over the last 2 millennia is crucial to understand the reasons for the collapse of biodiversity. While the earliest human artefacts or features may date back to 4,000 years before present (BP),[6] the human presence since more than 1,000 years BP is debated.[7–9] The settlement of the island has been addressed by multiple disciplines, such as archaeology, anthropology, and linguistics.[10–12] For several centuries, the settlement process of Madagascar has been the subject of much debate. The first European manuscripts mention populations of African and Arab descent,[13] and very early on, an Asian origin was also suspected.[14] At the beginning of the twentieth century, more than 50 authors had already proposed the most diverse origins.[15,16] During the last century, a consensus has emerged in linguistics suggesting that the Malagasy language evolved from an Austronesian background with Bantu and Arab-Swahili contributions.[17–20] This consensus around ancestors from Arab-Swahili-, Bantu-, and

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

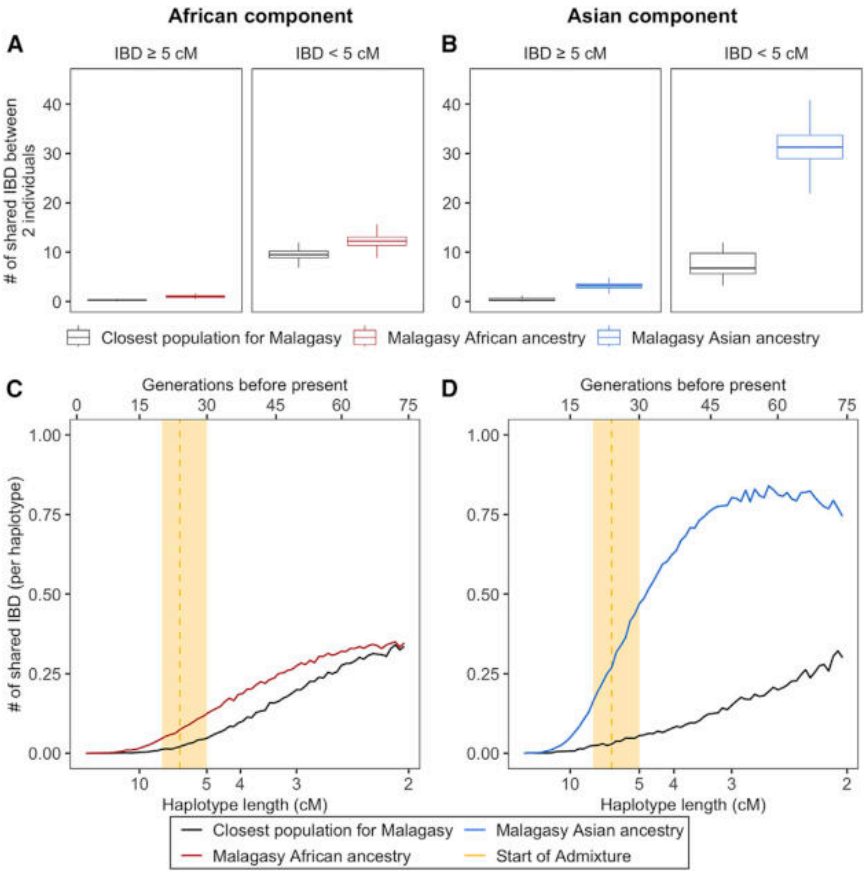


Figure 1. Pairwise IBD sharing in Madagascar, according to inferred local ancestry

(A and B) Boxplot of the distribution of shared IBD segments in Madagascar and their closest genetic population in Africa (A) and Asia (B). The number of segments shared per effective pair (y axis) is shown according to their genetic length (before admixture: <5cM; after admixture: ≥5cM) and their ancestral component.

(C and D) IBD-sharing distribution according to continental origin and length in centiMorgans (cM), along with the distribution detected in the Malagasy's closest genetic populations inhabiting Africa (C) and Asia (D). The x axis represents haplotype length cM, and the age of IBD haplotypes (in generations before present) are labeled on the superior horizontal axis. The number of segments shared per effective pair is shown on the vertical axis (y axis). The yellow dashed line shows the time since admixture in Madagascar estimated by MALDER.[30]

See also Figures S1, S2, and Methods S1.

Austronesian-speaking populations expanded from linguistics and is now widely accepted on the basis of multi- and interdisciplinary approaches.[21–23] Genetic analyses showed that present-day Malagasy populations are the result of an admixture event at the beginning of the last millennium, between Bantu-speaking African and Austronesian-speaking Asian populations.[24,25] But there is still no consensus about the past of the ancestral populations before the admixture event. A better understanding of the demographics of early Malagasy human populations may provide important clues regarding their possible impact on the island's ecosystem.

It has recently been shown that the demographic history of populations over the past 2 millennia can be addressed by studying the sharing of genomic fragments within a population.[26–28] The number and length of shared genetic fragments, called identity-by-descent (IBD) haplotypes, are dependent on the coalescence probabilities and are therefore markers of fluctuations in effective population size (Ne) over time. Importantly, through local ancestry inference methods, it is possible to reliably assess the characteristics and demographic parameters of each ancestral population prior to an admixture event.[26]

In the present study, we show that there are many short IBD segments of Asian ancestry that are shared between Malagasy individuals. Based on extensive demographic simulations, we show that the distribution of Asian IBDs supports only scenarios where the ancestral Austronesian-speaking population genetically isolated itself around 2,000 years BP, remaining as a small population (Ne ≈ 500) for approximately 1,000 years. The smaller number of large IBD fragments that are shared between all Malagasy individuals sampled shows that the population underwent a major demographic expansion around 1,000 years BP, when admixture with the ancestral Bantu-speaking African population is detected. Importantly, this latter time period coincides with the disappearance of the last large-bodied vertebrates in Madagascar.

RESULTS

IBD sharing in Madagascar populations

To better understand the history of Malagasy ancestral populations, we studied the distribution of shared IBD segments from genomic data of 700 individuals sampled all across the island of Madagascar, as well as from genomic data of 3,464 individuals sampled in other populations. (Methods S1A). Using RefineIBD,[29] we first performed an interpopulation analysis to determine the number of IBD segments (>2 cM) shared between Malagasy individuals and individuals from other populations. This analysis confirmed that the populations most closely related to the Malagasy are the Bantu-speaking populations of Eastern Africa and the Austronesian-speaking populations of southern Borneo (Methods S1B; Figure S1). We examined the number of admixture pulses using MALDER[30] and found evidence for a single event of admixture, happening around 25 generations BP. As previously shown using different software,[24,30–32] Malagasy individuals can be grouped into ten clusters based on genetic similarity, and the date of admixture varies (between 20–30 generations BP; Methods S1F) according to the Malagasy genetic cluster. Next, we performed an intrapopulation analysis to determine the number of IBD segments shared between individuals belonging to the same population. By applying local

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

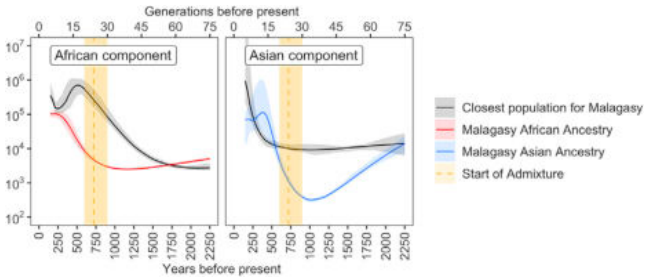


Figure 2. Ancestry-specific effective population size (Ne) estimated for the Malagasy

Left panel: ancestry-specific effective size is shown for African ancestry in the Malagasy populations (red line), with the estimated effective size for Mozambique populations (black line). Right panel: ancestry-specific estimated Ne is shown for the Asian ancestry in the Malagasy population (blue line), as well as the estimated effective size for South Borneo (black line). The y axes show ancestry-specific population size (Ne) plotted on a log scale. The solid lines show estimated effective sizes, and the colored regions show 95% bootstrap confidence intervals. The x axes show years before present, while generations before present are labeled on the superior horizontal axis.
See also Figure S3 and Methods S1.

ancestry inference with RFMix,[33] we characterized the continental origin of each IBD segment (Asian, African, or heterogeneous). We found that on average, two Malagasy individuals share 11.68 ± 2.19 IBD segments, of which 48.67% are of Asian origin, 40.68% of African origin, and 10.65 % of heterogeneous origin (Figure S2). Considering the overall level of ancestry (60.64% African and 39.36% Asian), the number of IBD segments shared between two Malagasy individuals is proportionally higher for those of Asian origin. Specifically, when we focus on IBDs older than the Asian and African admixture (i.e., those shorter than 5 cM and corresponding to 30 generations BP, according to Equation s19 of Al-Asadi et al. (2019)[34]; see STAR Methods) and normalize for the reduced fraction of the ancestry-specific genome considered for each individual's genome (similar to previous studies[26,35]), two Malagasy individuals share about 31.35 ± 4.47 Asian IBD segments, almost three times more than those of African origin (Figures 1A and 1B). Interestingly, the high number of Asian IBD segments is present in all Malagasy genetic clusters previously identified[24] (Methods S1C). In contrast, the overall number of IBD segments observed in the most closely related populations is much lower, i.e., populations in southern Borneo or Mozambique, where individuals share 7.43 ± 2.35 and 9.21 ± 1.80 IBD segments, respectively (Figures 1A and 1B). These results were replicated using two independent IBD and local ancestry detection methods (Methods S1D), confirming that the detection of a high number of Asian IBDs smaller than 5 cM is not an artefact of a particular method.

This result suggests that some demographic or selection event strongly impacted the Asian ancestral population after their separation from the Bornean Austronesian-speaking populations, which are the closest match to the Asian ancestral population in our dataset. We further investigated possible causes of this observation by first interrogating if the signal was present across the genome. As we detected the same signal on all 22 chromosomes (Methods S1E), we ruled out the possibility that it was due to a specific selection pressure on particular genes or pathways, suggesting a global demographic event that impacted the whole genome. To characterize this demographic phenomenon, we studied the IBD-sharing distribution based on the age of haplotypes (Figures 1C and 1D). The excess of Asian IBDs is mainly due to old haplotypes, dating from 45 to 75 generations BP (Figures 1C and 1D). The distribution of recent IBDs (less than 10 generations) is similar to other populations. Given that the numbers and lengths of IBD segments contain information about coalescence probabilities and thus effective population sizes,[26,36,37] this result suggests that the small Asian ancestral population might have drastically increased in size around the time of admixture (30 generations BP).

Malagasy ancestry-specific demographic history

We sought to better characterize the recent demographic history of African and Asian ancestral populations using the ancestry-specific IBDNe pipeline,[26] which allows us to track demographic trajectories from the last 150 to 2,250 years BP. We observed that the admixture event involved two small ancestral populations, followed by a large increase in Ne around the time of admixture (30 generations or 900 years BP) (Figure 2). Notably, the observed increase in effective size is much greater than what would be produced by the simple addition of effective sizes from both populations, showing an intrinsic demographic expansion in addition to the admixture (Methods S1F). Prior to admixture, the size of the ancestral African population appears relatively stable (Figure 2), with only a small reduction (Ne = 2,740–3,170) around 1,200 years BP; regarding the Asian component (Figure 2), the estimated Asian-specific effective size presents a drastic decline for several generations before the time of admixture, reaching the lowest Ne at some hundreds of individuals (Ne = 361–444) around 1,020 years BP. Consistent with the results observed from the Asian IBD-sharing distribution, the pattern of a small population increasing strongly during admixture is common to all Malagasy populations (Methods S1F), and it does not correspond to Ne fluctuations observed in the closest Austronesian-speaking populations from Borneo (Figure S3). Results from IBDNe inferences seem to be robust, as we analyzed a large sample size (700 individuals), and moreover, they were constant through the different genetic clusters described elsewhere.[24]

As pointed out previously, IBDNe estimation tends to smooth over sudden changes in effective size[26,38]; thus, the gradual decline observed between 2,250 and 1,000 years BP for the Asian ancestral population might represent an artefact of the method. Therefore, we used a coalescent-based simulator to model alternative scenarios for the demographic history of the Asian ancestral population,[39] in order to explore their consequences for the expected Asian IBD-sharing distribution. Taking into account the ascertainment bias present in our empirical dataset (see STAR Methods), we compared the impact of multiple parameters on the distribution of IBDs, such as the strength of founder events, the duration of bottleneck events, the amount of gene flow, and the influence of demographic changes during the bottleneck (Ne stable, in expansion, or declining) (Methods S1G–S1I). As expected, all of the 784 tested scenarios produced similar results to those observed in Madagascar in terms of allele frequency spectrum, global ancestry, and the distribution of shared African IBD segments (Methods S1N–S1P). In contrast,

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

the studied parameters heavily influenced the Asian IBD-sharing distribution (Methods S1I and S1J). All scenarios presenting the smallest distance metric compared to the observed Asian IBD-sharing distribution were those in which an Asian ancestral population remained small for approximately 1,000 years following divergence from other Asian populations, receiving minimal or no gene flow (Figure S4; Data S1).

Furthermore, we demonstrated that several previously plausible demographic scenarios are in fact incompatible with the observed distribution of Asian IBD segments. Indeed, the observed distribution of Asian IBD segments is outside five standard deviations of the expected distribution of Asian IBD segments for such scenarios. Thus, these scenarios have a probability of less than 10^{-6} of producing the observed curves (Figure 3; Methods S1H). For example, the high level of shared Asian IBD segments excludes that the population expansion in Madagascar was due to gene flow occurring over several generations (10 generations, between 900 to 600 years BP) from an ancestral southern Borneo population (Figure 3; multiple Asian waves scenario). Scenarios with strong founder events without a bottleneck period before admixture have virtually no probability of yielding the large number of shared Asian IBD segments observed in Madagascar data (Figure 3; Asian founder event and strong Asian founder event scenarios). Similarly, this distribution excludes a gradual decrease in the Asian population size before admixture, as suggested by IBDNe results (Figure 3; slow decrease of a large population scenario). Interestingly, we show that a scenario presenting a 300-year-long bottleneck (duration = 10 generations) with a stringent founder event (Ne = 50), followed by a rapid expansion (reaching Ne = 900 at the admixture time), could produce a similar level of Asian IBDs; nevertheless, IBD segments are larger (hence younger) than observed (Figure 3; bottleneck scenario). Finally, the observed Asian IBD-sharing distribution is only consistent with simulated scenarios where a small Asian population (Ne = 500) has been isolated for at least several centuries (Figure 3; long-term bottleneck scenario). Importantly, identical results are obtained with or without considering ascertainment bias (Methods S1K).

Notably, when including in the model a moderate migration rate from Asian populations (5 or 10 individuals per generation), the number of IBD segments decreased drastically, rendering the scenario of a highly connected ancestral population implausible (Methods S1G; long-term bottleneck scenarios). Furthermore, we show that gene flow from African populations starting before 30 generations BP would produce a date of admixture incompatible with the timing estimated by MALDER (Methods S1G). We also confirmed the strong probability of a bottleneck event for the Asian ancestral population using an approximate Bayesian computation (ABC) approach (approximate posterior probabilities: founder effect = 0.033; multiple Asian waves = 0.00; bottleneck = 0.967; Methods S1G). ABC parameter estimations also support a long bottleneck scenario with limited migration flow based on neural network, local linear regression, and rejection methods (Table S1). For instance, based on local linear regression, a tolerance rate of 5%, we estimated a duration of bottleneck 65 generations (95% confidence interval [CI]: 46–74) and the migration rate between 0–0.02. Interestingly, ABC parameter estimations suggest a Ne at the start of the bottleneck around 930 (95% CI: 109–3,560) and Ne at the end of 615 (95% CI: 505–790). This shows that the population size of the Asian ancestors of the current Malagasy populations just before admixture is of the same order of magnitude as at the beginning of the bottleneck (Ne_startBOT versus Ne endBOT; Table S1). Notably, these sizes are also very small compared to earlier (before bottleneck) and later periods (after admixture). This confirms the IBDNe estimate of a contemporary population expansion of the admixture.

Therefore, our modeling suggests that the current Malagasy population originated from an Austronesian-speaking population who was genetically isolated between 2,000 and 1,000 years. This isolation ended for the Malagasy Asian ancestors around 1,000 years BP, when they admixed with the Malagasy African ancestors. At about the same time, a rapid, large, and intrinsic demographic expansion happened, a date which coincides with extensive changes in the landscape, flora, and fauna (Figure 4).

DISCUSSION

By the analysis of genetic fragments shared across a large representative sample of Madagascar's inhabitants, we show that today's Malagasy populations originate from an admixture between two small populations and that the size of this population grew rapidly during admixture. The large number of short Asian IBDs shows that the admixture event ended a long period of bottleneck and isolation for the Asian ancestral population.

This conclusion is based on the use of deconvolution techniques of the genomes of admixed individuals that allow to identify IBDs and, thanks to proxy populations, identify the portions of genomes that come from each ancestral population before admixture. While these techniques represent a major advance, several major points must be evaluated. A first point to consider is the assumption that if the proxies used are not equally close to the populations that were mixed, this could result in different power to detect African and Asian ancestry in the Malagasy genomes. To limit any possibility of bias in proxy selection, we did not use specific populations as proxies but rather a group of East Asian and African populations (see STAR Methods). Moreover, we have shown previously that for the Malagasy population, the estimation of local ancestry is very stable whatever the combination of proxies and software used.[40] This stability in the attribution of local ancestry is due to the facts that the Malagasy population comes from a admixture of populations that had almost no gene exchange before or after the admixture (African and Asian) and that the separation between the two populations is old (>60,000 years BP) compared to the admixture (1,000 years BP). Therefore, very similar local ancestry results are obtained using multiple pairs: (1) Esan in Nigeria and Bejing Chinese; (2) Esan and Mandar from Sulawesi; (3) Luhya in Webuye, Kenya, and Mandar; (4) Yoruba in Ibadan, Nigeria, and Mandar; (5) a pool of all African individuals in the 1,000 genome dataset as the African source; and (6) a pool of all Asian individuals in the 1,000 genome dataset as the Asian source.[40] The results are also similar when switching local ancestry software; as a result, it is unlikely that the overrepresentation of small Asian IBDs in our dataset is due to misattribution by local ancestry software.

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

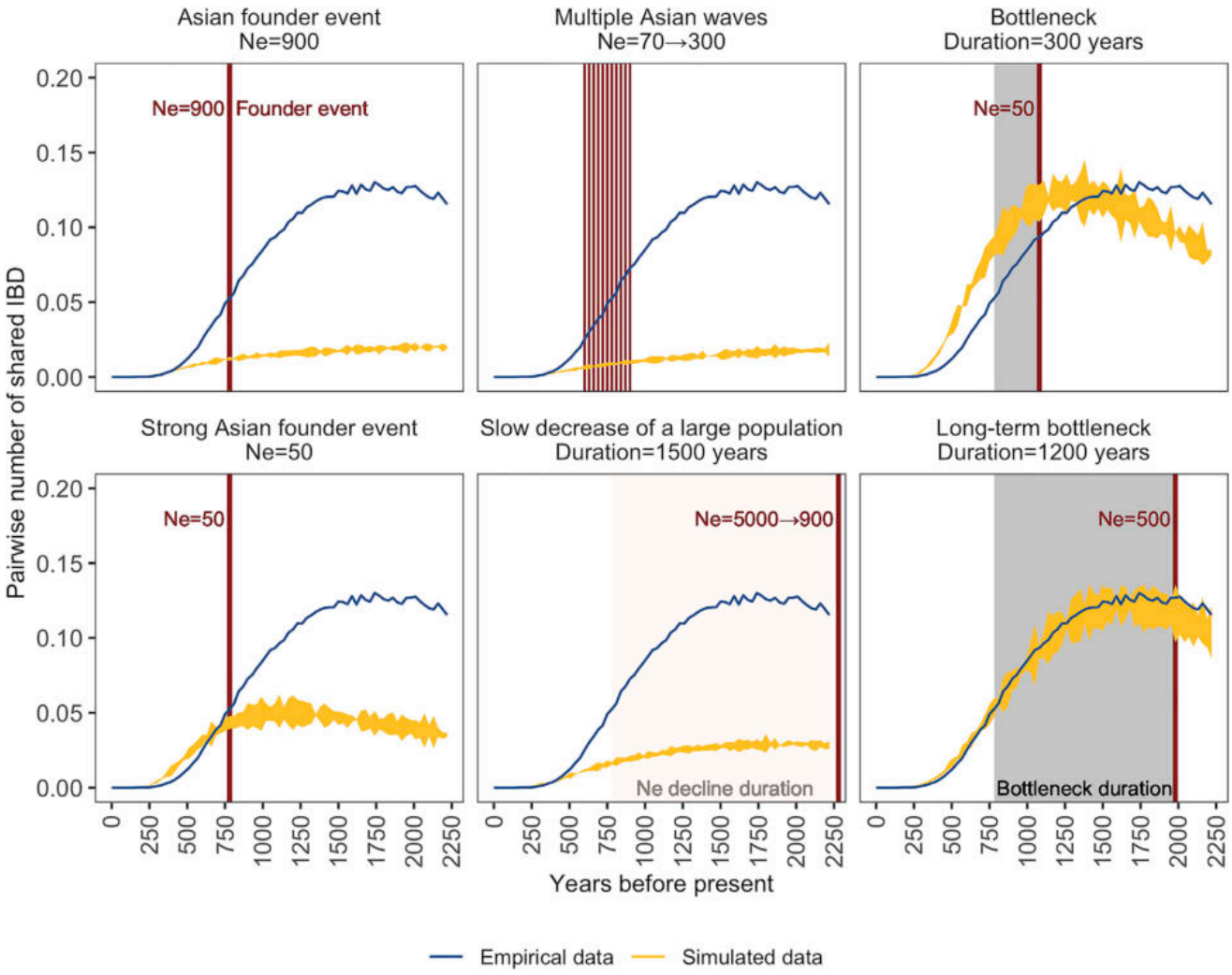


Figure 3. Asian ancestral demographic history simulated scenarios

Distribution of IBD segments of Asian origin shared per generation in Madagascar and in simulated data. Each panel shows the result for simulated scenarios (see Methods S1H), according to the strength of founder events (red line) and the duration of bottleneck events (shaded area). Admixture was simulated between Austronesian-speaking and Bantu-speaking populations (see Methods S1H). The x axis represents the age of IBD haplotypes (in years before present). The number of segments shared per pair is shown on the y axis. Yellow curves show the mean pairwise IBD sharing +/− 5 standard deviations at each generation, based on 3 simulated replicas. Blue lines in each panel represent the observed pairwise IBD sharing of Asian segments from the Malagasy population.
See also Figure S4, Table S1, Data S1, and Methods S1.

A second point to consider is the distortion of the relationship between age of IBDs and their size during population fluctuations. We have shown that the high number of short Asian IBD cannot be explained by any other phenomenon than a long period of bottleneck and isolation. This excess of retained IBD is indeed due to the fact that, in a small isolated population, IBDs tend to accumulate and disappear less frequently than in a large population. We have observed that the time and length relationship is almost absent for IBD accumulated more than 30 generations ago (Methods S1Q). While the ABC estimations does not rely on this relationship, this is nevertheless a limitation to fine estimation of parameters such as bottleneck duration or strength (Methods S1I). Thus, it is possible to consider that the bottleneck lasted between 30 and 70 generations and that the effective population size may have been several hundred individuals (Table S1). Further work is needed to clarify the duration of this long bottleneck.

From an anthropological point of view, an important question is whether the start of the bottleneck corresponds to the arrival in Madagascar of the Asian ancestral population around or before 2,000 years BP or if, alternatively, the migration across the Indian Ocean was a later event during the long-term isolation (two-step scenario), with the Asian ancestral population possibly even arriving just before the admixture. The long-term bottleneck might be more compatible with the settlement and isolation on Madagascar by a small Asian ancestral population, as genetic bottlenecks have been regularly associated with transcontinental migrations.[26,45,46] But bottlenecks have also been reported in Island South East Asia (ISEA), including in other Austronesian populations.[35,47–50] Therefore, a period of genetic

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

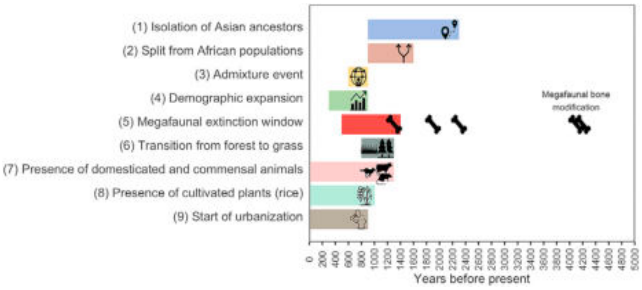


Figure 4. Major demographic events for the Malagasy and extensive changes in Madagascar biodiversity during the last 5,000 years

Settlement evidence and megafaunal extinction windows are illustrated according to the following numbers (y axis): (1) isolation of Asian ancestors (this study) with an effective size of just a few hundred individuals; (2) split from African populations (southern African Bantus) around 1,500 years BP[24]; (3) admixture event happening heterogeneously across the island, between African and Asian populations[24]; (4) demographic expansion (this study) due to intrinsic population growth of the newly admixed population; (5) megafaunal extinction window,[3,5] with bone icons representing the evidence of anthropogenic modifications on large-bodied vertebrates dating from before 1,300 years BP; (6) transition from forest to grass[40] as result of a dramatic increase in the use of fire by humans; (7) presence of domesticated and commensal animals,[4,41,42] where radiocarbon dates confirm that introduced animals were established in Madagascar between 1,200 and 700 years ago; (8) presence of cultivated plants (rice),[9] brought by migrating people on Madagascar; and (9) the start of urbanization, marked by iron smelting practices.[43,44]

isolation before the transcontinental migration event is also possible. Notably, the onset of the bottleneck could mark a geographic migration, but not necessarily to Madagascar, with the possibility of isolation in an unknown location (Indonesia archipelago, Maldivian archipelago, or Eastern African shore). Noteworthy, in the two-step scenario, present and previous results exclude any significant genetic exchange with any other neighboring populations (Asian or African).[24]

The journey of the Austronesians to reach Madagascar is much discussed, and there is a considerable body of work on the process of colonization of Madagascar, although there is no historical record of this phenomenon. Interestingly, an early presence of an Austronesian population in east coast of Africa before or around 2,000 BP is increasingly accepted[51,52,53]. However, many authors doubt that these contacts led to the settlement of Madagascar.[54] The dates most often cited for this settlement vary between 1,600 and 1,200 BP.[23] These dates are supported in particular (but not only) by linguistic works, which are based on the presence in the Austronesian language of Sanskrit words.[55] Moreover, the date of the 7th or 8th century (1,200 years BP) is often retained because it would correspond to the emergence of the Srivijaya empire in Southeast Asia. The expansion of this maritime empire could have caused the migration of people to Madagascar.[55] But the reasons for the migration of Austronesians to Madagascar remain unknown as well as the route taken through the Indian Ocean.[56] Several authors propose a scenario where the Austronesian population would have arrived in East Africa (or the Comoros) and where a first contact with the Africans would have taken place.[9,51] From the 8th century AD (1,200 years BP), an admixed population speaking Malagasy would have entered Madagascar from the west or northwest. It is interesting to note that we find no evidence of ancient genetic admixture (before the 8th century) between African and Asian populations. This emphasizes once again that genetic admixture and linguistic mixing appears to be disconnected in Madagascar. For instance, while linguistics show an origin predominantly Austronesian,[57] genetics suggests that Asian ancestry is only about 40%. Another example of such disconnection is the process of the arrival of Austronesians on the island as waves or as a continuum over several centuries, which is discussed.[58] Our work emphasizes that if these contacts have existed, they have been mainly cultural and have not left any substantial trace on the genome of the current populations. In particular, our work excludes arrivals from different Austronesian populations, because this would have greatly reduced the number of IBD observed. Like admixture phenomena, the history of cultural and genetic connections across the Indian Ocean might be only loosely related. Future ancient DNA evidence and a more extensive sampling in ISEA might shed more light on the Asian ancestral population and the timing of their arrival in Madagascar.

Our estimations suggest that the effective population size of the Asian ancestral population was around 500 individuals. It should be noted that estimates of effective population size based on the coalescence probabilities provide a smaller number than the census population size.[36] Therefore, our work implies that the ancestral Asian population may have reached a few thousand individuals that were genetically isolated over a thousand years, as they should have experienced minimal or no gene flow (migration rate < 0.02; Figure S4). The migration rate in the isolation event is comparable with other previously simulated populations that were separated about 5,000 km (Methods S1G). This isolation, associated with a long-term bottleneck, explains the large genetic distance between the Malagasy population and the closest Austronesian-speaking populations, observed in multiple analyses (IBD sharing, Fst, treemix, and mitochondrial haplotype diversity).[24] Interestingly, this long-term bottleneck might solve one inconsistency in the search for the most related populations of Malagasy in East Asia. Indeed, while Y chromosome and whole genome analyses point to Borneo populations, mitochondrial DNA diversity seems to point to the islands of eastern Indonesia. In particular, there is a high frequency of the so-called "Polynesian motif" in mitochondrial DNA observed among Malagasy maternal lines,[59] whereas this motif is virtually absent in Borneo populations. Our current work opens the possibility that this difference in frequency for the "Polynesian pattern" observed today between Borneo and Madagascar populations is caused by an ancient split between these two populations and a genetic drift associated with the long-term bottleneck experienced by Malagasy ancestors.

The genetic isolation of the ancestral Austronesian population terminated with the gene flow from Bantu-speaking populations around 1,000 years BP. Based on the absence of evidence of similarly admixed populations on African mainland or in Southeast Asia and the heterogeneity of admixture signals across Malagasy population,[24] it is likely that this event happened in Madagascar. The distribution of African IBD fragments suggests that at around this time, the African ancestral population underwent a moderate demographic reduction before a rapid increase. This fluctuation is consistent with the estimated age of

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

separation between the African population ancestral to the Malagasy and other Bantu-speaking populations, as it coincides with the Bantu-speaking populations expansion across Africa.[24,60–62] It should be noted that this does not exclude the presence of older African populations in Madagascar that would not have contributed significantly to the current genetic pool, as has been postulated.[24]

One contribution of the present study is the evidence that both ancestral populations were of limited size a few generations before the start of admixture. In contrast, during the admixture, the Malagasy effective population size increased by a factor of 100 between 1,000 to 300 years BP. In parallel to this human expansion, from 500 to 1,300 years BP[3,5] mega faunal extinctions associated with landscape transformations happened on Madagascar (Figure 4). While various explanations have been proposed to explain the extinctions, recent studies linked these changes to an increase of the human population on the island. But the origin of this increase of human presence is unknown. The present analysis shows that this expansion is not due to the massive arrival of external populations but instead a demographic increase of the newly admixed populations. The high number of 5 centimorgan IBD (African and Asian) suggest that the genetic contribution of any new population arriving on Madagascar island after 1,000 years BP is, at most, limited. In other words, all the current Malagasy descend from the admixture of a small number of individuals of Asian and African origin present on Madagascar around 1,000 years BP.

One may therefore ask why the admixture of these populations is linked to such demographic growth around 1,000 years BP. The question is particularly acute considering that the Asian ancestral population's effective size was limited to a few hundred individuals for ~1,000 years. A first possibility could be that genetic admixture helped the population to overcome new environmental constraints. As previously shown,[40] the African alleles on chromosome 1 were strongly selected for, probably because of the protection against vivax malaria conferred by the Duffy null allele. The selection signal observed is one of the strongest detected in any human population, suggesting that malaria was a heavy burden for the colonizing Asian ancestral population. But beyond genes, the demographic expansion might also be due to the interchanges of cultural practices between populations of Asian and African origins. Genetic admixture started at different times (between 600–1,000 years BP) in different regions of Madagascar. In this period of time, diverse regions presented a transition from forest to grassland,[63,64] and also evidence of a population subsistence strategy shift,[3] coupled with the introduction of domesticated animals, such as cattle,[4] dogs,[41] and rats,[42] as well as cultivated plants, like rice.[9] Importantly, iron smelting practices in the northeastern part shows human occupation of some importance since 800 years BP.[43,44] In consequence, the end of the ancestral Austronesian genetic isolation happened in a period where Malagasy societies seem to have undergone a demographic expansion, marked by the development of agro-pastoralism.

Finally, these changes probably deeply impacted the human demography but also the island. Major landscape transformations were induced due to the usage of fire by humans; it have been proposed that the introduction and expansion of cattle indirectly facilitated the hunting of endemic animals in a context of challenging climatic conditions[5] and also that the introduced herbivores competed with the endemic fauna and ultimately caused a disruption of natural vegetation.[2,3,65] It is now clear that environmental, cultura,l and genetic evidence should be considered together to understand this crucial shifting period around 1,000 years BP in Madagascar.

In conclusion, we show that genetic modeling is a powerful tool for testing the plausibility of a historical scenario. Specifically, the demographic model based on IBD distribution and local ancestry can accurately detect periods of isolation or population expansion due to external migrations. The comparison of these models with environmental and cultural data seems to be particularly insightful. A limitation of this strategy is that the chronology obtained from genetic analyses is different from that obtained in other disciplines. Genetic analysis depends on estimates of recombination and mutation rates and generation times, which have not remained constant through time and may differ between populations.[66] For instance, early extinctions appear to precede genetic mixing and population expansion (Figure 4). This slight lag may be due to imprecise dating estimates, but it may be also an important signal to explore.

Another limitation of this strategy is that Ne estimation does not inform about the presence of individuals and populations that could have potentially been there also, but not contributed to the current studied population. Keeping these limitations in mind, we propose that for the study of the recent past of populations, the strategy used here, based on the IBD-sharing distribution and simulated genetic data, could be useful for other investigations of human demography and their associated ecosystem impact.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection
- METHOD DETAILS
  - Genotyping
- QUANTIFICATION AND STATISTICAL ANALYSES
  - Integrating Malagasy data with SNP array datasets
  - Haplotype inference - Phasing process
  - Relatedness analysis
  - Admixture dates estimation
  - Local ancestry inference
  - Identical-by-descent segment determination
  - Global and ancestry-specific effective size estimation
  - Demographic model
  - Simulated genetic data
  - Simulation of empirical ascertainment bias in simulated genetic data
  - Demographic conditions for the Malagasy's Asian ancestral population

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

- Analysis of demographic simulations
- Effect of the coalescent simulators
- Approximate Bayesian Computation (ABC)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.09.060.

ACKNOWLEDGMENTS

We thank all participants in the study and the students and staff from the Institut de Civilisations/Musée d'Art et d'Archéologie who have contributed to the sampling across Madagascar. This research was funded by Région Aquitaine "Projet MAGE" (Madagascar Genétique et Ethnolinguistique) and French National Research Agency (ANR) grant "MADEOGEN".

AUTHOR CONTRIBUTIONS

Conceptualization, O.A., C.R., T.L., and D.P.; methodology, O.A. and D.P.; software, O.A.; formal analyses, O.A.; investigation, A.L., M.H., V.P.L., T.L., O.A., and D.P.; writing – original draft, O.A. and D.P.; writing – review and editing, T.L., O.A., D.P., and M.S.; resources, A.B., J.F.D., J.R., C.S., C.F.L., C.R., J.A.R., T.L., and D.P.; supervision, M.S., C.R., J.A.R., T.L., and D.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: April 7, 2022
Revised: July 13, 2022
Accepted: September 28, 2022
Published: November 4, 2022

REFERENCES

1. Douglass, K., Hixon, S., Wright, H.T., Godfrey, L.R., Crowley, B.E., Manjakahery, B., Rasolondrainy, T., Crossland, Z., and Radimilahy, C. (2019). A critical review of radiocarbon dates clarifies the human settlement of Madagascar. Quat. Sci. Rev. *221*, 105878. https://doi.org/10.1016/j.quascirev.2019.105878.
2. Burney, D. (2004). A chronology for late prehistoric Madagascar. J. Hum. Evol. *47*, 25–63. https://doi.org/10.1016/j.jhevol.2004.05.005.
3. Godfrey, L.R., Scroxton, N., Crowley, B.E., Burns, S.J., Sutherland, M.R., Pérez, V.R., Faina, P., McGee, D., and Ranivoharimanana, L. (2019). A new interpretation of Madagascar's megafaunal decline: The "Subsistence Shift Hypothesis". J. Hum. Evol. *130*, 126–140. https://doi.org/10.1016/j.jhevol.2019.03.002.
4. Hixon, S.W., Douglass, K.G., Crowley, B.E., Rakotozafy, L.M.A., Clark, G., Anderson, A., Haberle, S., Ranaivoarisoa, J.F., Buckley, M., Fidiarisoa, S., et al. (2021). Late Holocene spread of pastoralism coincides with endemic megafaunal extinction on Madagascar. Proc. Biol. Sci. *288*. 20211204. https://doi.org/10.1098/rspb.2021.1204.
5. Li, H., Sinha, A., Anquetil André, A., Spötl, C., Vonhof, H.B., Meunier, A., Kathayat, G., Duan, P., Voarintsoa, N.R.G., Ning, Y., et al. (2020). A multi-millennial climatic context for the megafaunal extinctions in Madagascar and Mascarene Islands. Sci. Adv. *6*, eabb2459.
6. Dewar, R.E., Radimilahy, C., Wright, H.T., Jacobs, Z., Kelly, G.O., and Berna, F. (2013). Stone tools and foraging in northern Madagascar challenge Holocene extinction models. Proc. Natl. Acad. Sci. USA *110*, 12583–12588. https://doi.org/10.1073/pnas.1306100110.
7. Anderson, A., Clark, G., Haberle, S., Higham, T., Nowak-Kemp, M., Prendergast, A., Radimilahy, C., Rakotozafy, L.M., Ramilisonina, Schwenninger, J.-L., et al. (2018). New evidence of megafaunal bone damage indicates late colonization of Madagascar. PLoS One *13*, e0204368.
8. Mitchell, P. (2020). Settling Madagascar: When Did People First Colonize the World's Largest Island? J. I. Coast Archaeol. *15*, 576–595. https://doi.org/10.1080/15564894.2019.1582567.
9. Crowther, A., Lucas, L., Helm, R., Horton, M., Shipton, C., Wright, H.T., Walshaw, S., Pawlowicz, M., Radimilahy, C., Douka, K., et al. (2016). Ancient crops provide first archaeological signature of the westward Austronesian expansion. Proc. Natl. Acad. Sci. USA *113*, 6635–6640. https://doi.org/10.1073/pnas.1522714113.
10. Beaujard, P. (2011). The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. Archaeological Research in Africa *46*, 169–189. https://doi.org/10.1080/0067270x.2011.580142.
11. Hurles, M.E., Sykes, B.C., Jobling, M.A., and Forster, P. (2005). The Dual Origin of the Malagasy in Island Southeast Asia and East Africa: Evidence from Maternal and Paternal Lineages. Am. J. Hum. Genet. *76*, 894–901. https://doi.org/10.1086/430051.
12. Serva, M. (2012). The Settlement of Madagascar: What Dialects and Languages Can Tell Us. PLoS One *7*, e30666. https://doi.org/10.1371/journal.pone.0030666.
13. Grandidier, A., Charles-Roux, J., Delhorbe, C., Froidevaux, H., and Grandidier, G. (1903). Collection des ouvrages anciens concernant Madagascar: Ouvrages ou extraits d'ouvrages portugais, hollandais, anglais, français, allemands, italiens, espagnols et latins relatifs à Madagascar (1500 à 1613) (Comité de Madagascar).
14. Pyrard, F. (1603). Discours du voyage des François aux Indes Orientales. In (David Le Clerc), p. 371.
15. Grandidier, A., and Grandidier, G. (1885). Histoire physique, naturelle et politique de Madagascar (Imprimerie Nationale).
16. Ferrand, G. (1909). L'Origine africaine des Malgaches. Bul. Mém. Soc. Anthropol. Paris *10*, 22–35. https://doi.org/10.3406/bmsap.1909.8033.
17. Dahl, O.C. (1988). Bantu substratum in Malagasy. Etudes ocean indien, 91–132.
18. Dahl, O.C. (1977). La subdivision de la famille Barito et la place du Malgache. Acta Orient. 77–134.
19. Adelaar, K.A. (1989). Les langues austronésiennes et la place du malagasy dans leur ensemble. Arch. Androl. *38*, 25–52. https://doi.org/10.3406/arch.1989.2588.
20. Simon, P. La langue des ancêtres. (Ny Fitenin-dRazana L'Harmattan).
21. Blench, R.M. (2015). Interdisciplinary approaches to stratifying the peopling of madagascar. Proceedings of the Indian Ocean Conference (British Archaeological Reports).
22. Adelaar, A. (2016). Austronesians in Madagascar: A Critical Assessment of the Works of Paul Ottino and Philippe Beaujard. In Early Exchange between Africa and the Wider Indian Ocean World, G. Campbell, ed. (Springer International Publishing), pp. 77–112.
23. Beaujard, P. (2011). The first migrants to Madagascar and their introduction of plants: linguistic and ethnological evidence. Archaeological Research in Africa *46*, 169–189. https://doi.org/10.1080/0067270x.2011.580142.
24. Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L.M.-A., Rakotomalala, M.M., Razafiarivony, M., Rasoarifetra, B., et al. (2017). Genomic landscape of human diversity across Madagascar. Proc. Natl. Acad. Sci. USA *114*, E6498–E6506. https://doi.org/10.1073/pnas.1704906114.
25. Serva, M., and Pasquini, M. (2020). Dialects of Madagascar. PLoS One *15*, e0240170.
26. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLoS Genet. *14*, e1007385. https://doi.org/10.1371/journal.pgen.1007385.

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

27. Castro e Silva, M.A., Nunes, K., Lemes, R.B., Mas-Sandoval, A., Guerra Amorim, C.E., Krieger, J.E., Mill, J.G., Salzano, F.M., Bortolini, M.C., Pereira, A.C., et al. (2020). Genomic insight into the origins and dispersal of the Brazilian coastal natives. Proc. Natl. Acad. Sci. USA *117*, 2372–2377. https://doi.org/10.1073/pnas.1909075117.

28. Yunusbaev, U., Ionusbaev, A., Han, G., and Kwon, H.W. (2020). Recent effective population size in Eastern European plain Russians correlates with the key historical events. Sci. Rep. *10*, 9729. https://doi.org/10.1038/s41598-020-66734-y.

29. Browning, B.L., and Browning, S.R. (2013). Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics *194*, 459–471. https://doi.org/10.1534/genetics.113.150029.

30. Pickrell, J.K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637. https://doi.org/10.1073/pnas.1313787111.

31. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. PLoS Genet. *8*, e1002453. https://doi.org/10.1371/journal.pgen.1002453.

32. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A Genetic Atlas of Human Admixture History. Science *343*, 747–751. https://doi.org/10.1126/science.1243518.

33. Maples, B., Gravel, S., Kenny, E., and Bustamante, C. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. *93*, 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020.

34. Al-Asadi, H., Petkova, D., Stephens, M., and Novembre, J. (2019). Estimating recent migration and population-size surfaces. PLoS Genet. *15*, e1007908. https://doi.org/10.1371/journal.pgen.1007908.

35. Ioannidis, A.G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Barberena-Jonas, C., Hill, A.V.S., Rodríguez-Rodríguez, J.E., Fox, K., Robson, K., Haoa-Cardinali, S., et al. (2021). Paths and timings of the peopling of Polynesia inferred from genomic networks. Nature *597*, 522–526. https://doi.org/10.1038/s41586-021-03902-8.

36. Browning, S., and Browning, B. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418. https://doi.org/10.1016/j.ajhg.2015.07.012.

37. Palamara, P., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. Am. J. Hum. Genet. *91*, 1150. 822. https://doi.org/10.1016/j.ajhg.2012.11.006.

38. Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Soares-Souza, G.B., Leal, T.P., Furlan, V., Scliar, M.O., Zamudio, R., Zolini, C., et al. (2020). The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proc. Natl. Acad. Sci. USA *117*, 32557–32565. https://doi.org/10.1073/pnas.2013773117.

39. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput. Biol. *12*, e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

40. Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olaso, R., Deleuze, J.-F., et al. (2018). Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat. Commun. *9*, 932. https://doi.org/10.1038/s41467-018-03342-5.

41. Hixon, S.W., Douglass, K.G., Godfrey, L.R., Eccles, L., Crowley, B.E., Rakotozafy, L.M.A., Clark, G., Haberle, S., Anderson, A., Wright, H.T., and Kennett, D.J. (2021). Ecological Consequences of a Millennium of Introduced Dogs on Madagascar. Front. Ecol. Evol. *9*, 689559. https://doi.org/10.3389/fevo.2021.689559.

42. Crowley, B.E. (2010). A refined chronology of prehistoric Madagascar and the demise of the megafauna. Quat. Sci. Rev. *29*, 2591–2603. https://doi.org/10.1016/j.quascirev.2010.06.030.

43. Radimilahy, C. (1985). Contribution à l'étude de l'ancienne métallurgie du fer à Madagascar. Mussé D'Art et d'Archéologie de l'Université de Madagascar. Travaux et Documents XXV. Tananarive.

44. Clist, B.-O. (1995). New Field Data on the Ancient Iron Metallurgy of Madagascar. Nyame Akuma *43*, 23–27.

45. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., The 1000 Genomes Project, Bustamante, C.D., et al. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.

46. Pedersen, C.-E.T., Lohmueller, K.E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H.R., Moltke, I., and Albrechtsen, A. (2017). The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. Genetics *205*, 787–801. https://doi.org/10.1534/genetics.116.193821.

47. Cox, M.P., Hudjashov, G., Sim, A., Savina, O., Karafet, T.M., Sudoyo, H., and Lansing, J.S. (2016). Small Traditional Human Communities Sustain Genomic Diversity over Microgeographic Scales despite Linguistic Isolation. Mol Biol E *33*, 2273–2284. https://doi.org/10.1093/molbev/msw099.

48. Hudjashov, G., Karafet, T.M., Lawson, D.J., Downey, S., Savina, O., Sudoyo, H., Lansing, J.S., Hammer, M.F., and Cox, M.P. (2017). Complex Patterns of Admixture across the Indonesian Archipelago. Mol. Biol. Evol. *34*, 2439–2452. https://doi.org/10.1093/molbev/msx196.

49. Larena, M., Sanchez-Quinto, F., Sjödin, P., McKenna, J., Ebeo, C., Reyes, R., Casel, O., Huang, J.-Y., Hagada, K.P., Guilay, D., et al. (2021). Multiple migrations to the Philippines during the last 50, 000 years. Proc. Natl. Acad. Sci. USA *118*. e2026132118. https://doi.org/10.1073/pnas.2026132118.

50. Tumonggor, M.K., Karafet, T.M., Hallmark, B., Lansing, J.S., Sudoyo, H., Hammer, M.F., and Cox, M.P. (2013). The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. J. Hum. Genet. *58*, 165–173. https://doi.org/10.1038/jhg.2012.154.

51. Allibert, C. (2012). Les réseaux de navigation du début de l'ère chrétienne au XVIe siècle. Rencontre de populations, échanges commerciaux et matrimoniaux, concurrence à l'ouest et à l'est de Madagascar. Topoi. Orient-Occident *11*, 341–357. Autour du Périple de la mer Érythrée. Topoi. Orient-Occident.

52. Beaujard, P. (2013). Madagascar and Africa, Austronesian migration. In The Encyclopedia of Global Human Migration, I. Ness, ed. (Wiley).

53. Blench, R.M. (2008). The Austronesians in Madagascar and their interaction with the Bantu of the East African coast: Surveying the linguistic evidence for domestic and translocated animals. Philippines Journal of Linguistics *18*, 18–43.

54. Blench, R., and Dendo, M. (2006). The Austronesians in Madagascar and on the East African coast: surveying the linguistic evidence for domestic and translocated animals. International Conference on Austronesian Languages Palawan, 17–20.

55. Adelaar, A. (2006). The Indonesian migrations to Madagascar: making sense of the multidisciplinary evidence. In Austronesian diaspora and the ethnogenesis of people in Indonesian Archipelago (Proceedings of the international symposium), pp. 205–232.

56. Fitzpatrick, S.M., and Callaghan, R. (2008). Seafaring simulations and the origin of prehistoric settlers to Madagascar. In Islands of inquiry: Colonisation, seafaring and the archaeology of maritime landscapes (ANU E Press), pp. 47–58.

57. Verin, P., Kottak, C.P., and Gorlin, P. (1969). The Glottochronology of Malagasy Speech Communities. Ocean Ling. *8*, 26. https://doi.org/10.2307/3622902.

58. Beaujard, P. (2003). Les arrivées austronésiennes à Madagascar: vagues ou continuum? Études Océan Indien, 59–147.

59. Razafindrazaka, H., Ricaut, F.-X., Cox, M.P., Mormina, M., Dugoujon, J.-M., Randriamarolaza, L.P., Guitard, E., Tonasso, L., Ludes, B., and Crubézy, E. (2010). Complete mitochondrial DNA sequences provide

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

new insights into the Polynesian motif and the peopling of Madagascar. Eur. J. Hum. Genet. *18*, 575–581. https://doi.org/10.1038/ejhg.2009.222.

60. Boivin, N., Crowther, A., Helm, R., and Fuller, D.Q. (2013). East Africa and Madagascar in the Indian Ocean world. J World Prehist *26*, 213–281. https://doi.org/10.1007/s10963-013-9067-4.
61. Choudhury, A., Sengupta, D., Ramsay, M., and Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. Hum. Mol. Genet. *30*, R56–R63. https://doi.org/10.1093/hmg/ddaa274.
62. Semo, A., Gayà-Vidal, M., Fortes-Lima, C., Alard, B., Oliveira, S., Almeida, J., Prista, A., Damasceno, A., Fehn, A.-M., Schlebusch, C., and Rocha, J. (2020). Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. Mol. Biol. Evol. *37*, 406–416. https://doi.org/10.1093/molbev/msz224.
63. Burns, S.J., Godfrey, L.R., Faina, P., McGee, D., Hardt, B., Ranivoharimanana, L., and Randrianasy, J. (2016). Rapid human-induced landscape transformation in Madagascar at the end of the first millennium of the Common Era. Quat. Sci. Rev. *134*, 92–99. https://doi.org/10.1016/j.quascirev.2016.01.007.
64. Voarintsoa, N.R.G., Railsback, L.B., Brook, G.A., Wang, L., Kathayat, G., Cheng, H., Li, X., Edwards, R.L., Rakotondrazafy, A.F.M., and Madison Razanatseheno, M.O. (2017). Three distinct Holocene intervals of stalagmite deposition and nondeposition revealed in NW Madagascar, and their paleoclimate implications. Clim. Past *13*, 1771–1790. https://doi.org/10.5194/cp-13-1771-2017.
65. Hansford, J.P., Lister, A.M., Weston, E.M., and Turvey, S.T. (2021). Simultaneous extinction of Madagascar's megaherbivores correlates with late Holocene human-caused landscape transformation. Quat. Sci. Rev. *263*, 106996. https://doi.org/10.1016/j.quascirev.2021.106996.
66. Coop, G., and Przeworski, M. (2007). An evolutionary view of human recombination. Nat. Rev. Genet. *8*, 23–34. https://doi.org/10.1038/nrg1947.
67. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.
68. 1000 Genomes Project Consortium, Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.A. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
69. Brucato, N., Kusuma, P., Cox, M.P., Pierron, D., Purnomo, G.A., Adelaar, A., Kivisild, T., Letellier, T., Sudoyo, H., and Ricaut, F.-X. (2016). Malagasy Genetic Ancestry Comes from an Historical Malay Trading Post in Southeast Borneo. Mol Biol *33*, 2396–2400. https://doi.org/10.1093/molbev/msw117.
70. Mörseburg, A., Pagani, L., Ricaut, F.-X., Yngvadottir, B., Harney, E., Castillo, C., Hoogervorst, T., Antao, T., Kusuma, P., Brucato, N., et al. (2016). Multi-layered population structure in Island Southeast Asians. Eur. J. Hum. Genet. *24*, 1605–1611. https://doi.org/10.1038/ejhg.2016.60.
71. Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.-X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.-A., Blench, R.M., et al. (2014). Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc. Natl. Acad. Sci. USA *111*, 936–941. https://doi.org/10.1073/pnas.1321860111.
72. Teo, Y.-Y., Sim, X., Ong, R.T., Tan, A.K., Chen, J., Tantoso, E., Small, K.S., Ku, C.-S., Lee, E.J., Seielstad, M., and Chia, K.S. (2009). Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162. https://doi.org/10.1101/gr.095000.109.
73. Busby, G.B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V.D., Amenga-Etego, L.N., Enimil, A., Apinjoh, T., Ndila, C.M., et al.; Malaria Genomic Epidemiology Network (2016). Admixture into and within sub-Saharan Africa. Elife *5*, e15266. https://doi.org/10.7554/elife.15266.
74. May, A., Hazelhurst, S., Li, Y., Norris, S.A., Govind, N., Tikly, M., Hon, C., Johnson, K.J., Hartmann, N., Staedtler, F., and Ramsay, M. (2013). Genetic diversity in black South Africans from Soweto. BMC Genom. *14*, 644. https://doi.org/10.1186/1471-2164-14-644.
75. Montinaro, F., Busby, G.B.J., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., Destro-Bisol, G., Pascali, V.L., and Capelli, C. (2017). Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. Genetics *205*, 303–316. https://doi.org/10.1534/genetics.116.189209.
76. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S., Thomas, M., Luiselli, D., et al. (2012). Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. Am. J. Hum. Genet. *91*, 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015.
77. Patin, E., Siddle, K.J., Laval, G., Quach, H., Harmant, C., Becker, N., Froment, A., Régnault, B., Lemée, L., Gravel, S., et al. (2014). The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. Nat. Commun. *5*, 3163. https://doi.org/10.1038/ncomms4163.
78. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B., et al. (2012). Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science *338*, 374–379. https://doi.org/10.1126/science.1227721.
79. Schlebusch, C.M., Prins, F., Lombard, M., Jakobsson, M., and Soodyall, H. (2016). The disappearing San of southeastern Africa and their genetic affinities. Hum. Genet. *135*, 1365–1373. https://doi.org/10.1007/s00439-016-1729-8.
80. Brucato, N., Fernandes, V., Mazières, S., Kusuma, P., Cox, M.P., Ng'ang'a, J.W., Omar, M., Simeone-Senelle, M.-C., Frassati, C., Alshamali, F., et al. (2018). The Comoros Show the Earliest Austronesian Gene Flow into the Swahili Corridor. Am. J. Hum. Genet. *102*, 58–68. https://doi.org/10.1016/j.ajhg.2017.11.011.
81. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559.
82. Browning, S.R., and Browning, B.L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. *81*, 1084–1097. https://doi.org/10.1086/521987.
83. Dias-Alves, T., Mairal, J., and Blum, M.G.B. (2018). Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species. Mol. Biol. Evol. *35*, 2318–2326. https://doi.org/10.1093/molbev/msy126.
84. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2008). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326. https://doi.org/10.1101/gr.081398.108.
85. R Core Team (2021). R: A language and environment for statistical (R Foundation for Statistical Computing).
86. Li, S., Schlebusch, C., and Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc. R. Soc. B *281*, 20141448. https://doi.org/10.1098/rspb.2014.1448.
87. Ralph, P., and Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. PLoS Biol. *11*, e1001555. https://doi.org/10.1371/journal.pbio.1001555.
88. Lipson, M., Loh, P.-R., Patterson, N., Moorjani, P., Ko, Y.-C., Stoneking, M., Berger, B., and Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. Nat. Commun. *5*, 4689. https://doi.org/10.1038/ncomms5689.
89. Soares, P.A., Trejaut, J.A., Rito, T., Cavadas, B., Hill, C., Eng, K.K., Mormina, M., Brandão, A., Fraser, R.M., Wang, T.-Y., et al. (2016). Resolving the ancestry of Austronesian-speaking populations. Hum. Genet. *135*, 309–326. https://doi.org/10.1007/s00439-015-1620-z.

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

90. Ko, A.S., Chen, C.-Y., Fu, Q., Delfin, F., Li, M., Chiu, H.-L., Stoneking, M., and Ko, Y.-C. (2014). Early Austronesians: Into and Out Of Taiwan. Am. J. Hum. Genet. *94*, 426–436. https://doi.org/10.1016/j.ajhg.2014.02.003.

91. Choin, J., Mendoza-Revilla, J., Arauna, L.R., Cuadros-Espinoza, S., Cassar, O., Larena, M., Ko, A.M.-S., Harmant, C., Laurent, R., Verdu, P., et al. (2021). Genomic insights into population history and biological adaptation in Oceania. Nature *592*, 583–589. https://doi.org/10.1038/s41586-021-03236-5.

92. Xu, S., Pugach, I., Stoneking, M., Kayser, M., and Jin, L.; The HUGO Pan-Asian SNP Consortium (2012). Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. Proc. Natl. Acad. Sci. USA *109*, 4574–4579. https://doi.org/10.1073/pnas.1118892109.

93. Zhou, Y., Qiu, H., and Xu, S. (2017). Modeling Continuous Admixture Using Admixture-Induced Linkage Disequilibrium. Sci. Rep. *7*, 43054. https://doi.org/10.1038/srep43054.

94. Gignoux, C.R., Henn, B.M., and Mountain, J.L. (2011). Rapid, global demographic expansions after the origins of agriculture. Proc. Natl. Acad. Sci. USA *108*, 6044–6049. https://doi.org/10.1073/pnas.0914274108.

95. Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Mol. Biol. Evol. *27*, 2534–2547. https://doi.org/10.1093/molbev/msq148.

96. Gopalan, S., Berl, R.E., Myrick, J.W., Garfield, Z.H., Reynolds, A.W., Bafens, B.K., Belbin, G., Mastoras, M., Williams, C., Daya, M., et al. (2022). Hunter-gatherer genomes reveal diverse demographic trajectories during the rise of farming in Eastern Africa. Curr. Biol. *32*, 1852–1860.e5. https://doi.org/10.1016/j.cub.2022.02.050.

97. Hudjashov, G., Karafet, T.M., Lawson, D.J., Downey, S., Savina, O., Sudoyo, H., Lansing, J.S., Hammer, M.F., and Cox, M.P. (2017). Complex Patterns of Admixture across the Indonesian Archipelago. Mol. Biol. Evol. *34*, 2439–2452. https://doi.org/10.1093/molbev/msx196.

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The Malagasy genetic data	[24]	EGA : EGAS00001002549
Software and algorithms		
MSPRIME	[39]	https://tskit.dev/msprime/docs/stable/installation.html
IBDne	[26]	https://faculty.washington.edu/browning/ibdne.html
Refined IBD	[29]	https://faculty.washington.edu/browning/refined-ibd.html
RFMIX	[33]	https://github.com/slowkoni/rfmix

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Denis Pierron (denis.pierron@univ-tlse3.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate new unique dataset. The Malagasy genetic data described in this article are available from EGA (EGAS00001002549). Codes and scripts for Msprime simulations can be found in Methods S1.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection

The genome wide SNP data from 700 individuals analyzed here were produced by the project MAGE.[24] Samples were collected from 2007–2014 utilizing an extensive grid-based approach, in which individuals were sampled from 257 villages (2.8 ± 0.7 individuals per village) from all over Madagascar. All individuals were unrelated and all of them received detailed information on the study and gave written consent prior to the donation of their sample. This study was approved by the Human Subjects'Ethics Committee of the Health Ministry of Madagascar as well as by French Ethics Committees (Committees of Protection of Persons, and French National Commission on Informatics and Liberty).

METHOD DETAILS

Genotyping

Genome-wide SNP data (>2.3 million markers) was generated using the Illumina Human Omni 2.5-8 (Omni 2.5) BeadChip. Filtering and quality controls were performed using VCFTools,[67] with SNPs excluded that failed the Hardy-Weinberg exact (HWE) test (P-value < 1x10-5) for the 700 Malagasy individuals or had missing data rates >0.10 across all samples.

QUANTIFICATION AND STATISTICAL ANALYSES

Integrating Malagasy data with SNP array datasets

We merged the Malagasy genome wide SNP data with the following datasets:

- 1000 Genomes Project.[68]
- Asian individuals previously genotyped by different studies.[24,40,48,69–72]
- African individuals from different studies.[62,73–80]

From the 1000 Genomes Project dataset, we selected individuals of African (ESN, YRI, GWD and LWK) and Asian (CHB, CDX, CHS and KHV) ancestries. Before merging population genetic datasets, we performed the following steps in order to reduce possible

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

batch effects produced by different genotyping platforms. First, we used bcftools for checking the correspondence between the reference allele (from human genome assembly GRCh37) and the presentation of alleles in genotype calls. Then, we performed a quality control procedure on the original datasets, removing all SNPs with missing call rates and sites failing Hardy-Weinberg equilibrium (HWE) analyses (p-value < 1e-5). Following this step, we performed a missing genotype filter of 90% for the merged dataset. The datasets were then merged with high quality genotyping data. Finally, we applied a relatedness filter using KING[81] for removal of possible duplicated samples between original datasets. We removed 15 individuals presenting kinship coefficients above 0.354 (putative twin/duplicated samples).

We generated two datasets (Methods S1A) considering the density of SNPs and samples sizes:

→ Density Dataset (311,411 SNPs and 3168 individuals).
→ Diversity Dataset (78,906 SNPs and 4212 individuals).

Haplotype inference - Phasing process

For both datasets, imputation of missing genotypes and estimation of phased haplotypes were obtained with BEAGLE v5.1,[82] using the HapMap GRCh37 genetic map.

Relatedness analysis

We identified the maximal set of unrelated individuals using kinship analyses implemented in KING[81] for removal of possible related individuals. We considered that two individuals were related (third degree or closer) if the kinship coefficient was greater than 0.0442. In order to apply the IBDNe Ancestry-specific method, we followed the same steps as in the original publication. More precisely, we discarded all IBD segments belonging to 114 and 268 couples from the Density and Diversity datasets, respectively. From these couples, 27 and 26 belonged to Malagasy in the Density and Diversity data set, respectively. In total, we removed 4,234 IBD from the Density dataset and 13,330 IBD from the Diversity dataset where respectively 25.88% and 6.82% were from Malagasy couples. The resulting IBD segments were used for IBD-sharing analyses (using both datasets), IBDNe and IBDNe-AS estimation (using only the Density dataset).

Admixture dates estimation

We inferred the time of admixture in the Malagasy population using MALDER[30] with default parameters. The African reference panel is composed of 419 individuals (ESN, YRI, GWD and LWK) and the east-Asian reference panel of 400 individuals (CHB, CDX, CHS and KHV). These 819 samples from the 1000 Genomes Consortium Data were used as a proxy for the Malagasy source populations in all following analyses.

Local ancestry inference

Local ancestry analyses were performed on admixed individuals using the software package RFMix[33] and LOTER.[83] RFMix results were produced using one expectation-maximization (EM) step and a window size of 0.2 cM. For both methods, the time since admixture was set at 30 generations before present. We used as a reference panel the African (ESN, YRI, GWD and LWK) and Asian (CHB, CDX, CHS and KHV) population from the 1000 Genomes Project.

Identical-by-descent segment determination

We used RefinedIBD[29] for Identity-By-Descent (IBD) detection. We filled a gap between two detected IBD segments if the length of the gap was less than 0.6 cM, allowing a number of discordant sites =1 [26]. For confirmation purposes, we also inferred IBD haplotypes using Germline[84] with the following parameters: bits = 32, err_hom = 1, err_het = 1 as in Browning, et al.[29] Only IBD segments greater than 2 cM were used in the experiments and IBD segments from related individuals were removed from the analyses. The R programming language[85] was used for performing the following analyses on detected IBD segments:

1. We approximated the age of IBD segments using equation s19 from,[34] under the assumption the population is sufficiently large:
   E = 75 x (2/L), where
   E: is the time in generations to the most recent ancestor
   L: is the length of the IBD segment in centiMorgans.
2. We estimated the ancestry of each IBD segment by mapping its coordinates to local ancestry assignments from the sample haplotypes. As previous studies[24,40,69,71,80] identified two main ancestries contributing to the admixture (Bantu from southeast Africa and Austronesian from Indonesia), ancestry proportions were calculated using the number of genotyped locus (SNP) assigned as Asian (ASI) or African (AFR) divided by the total number of genotyped locus spanning the IBD region in both haplotypes. IBD segments were classified as African (AFR locus $\geq$ 90%), Heterogeneous (10% < ASI locus < 90%) or Asian (ASI locus $\geq$ 90%).
3. After IBD ancestry assignment, we filtered out admixed segments showing discordance between the origins of haplotypes. We kept segments with >50% concordance, based on:

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

Concordance = 1 - [|(n - m)| / s], where
n: number of sites of ANC origin on haplotype 1
m: number of sites of ANC origin on haplotype 2
s: number of sites covering the IBD segment.

4. We compared the level of IBD sharing within populations using the mean number of segments shared between individuals. To calculate this mean for each population of size N, the denominator was the number of pairs in that group, specified by the equation $(N^2 - N)/2$.
5. We compared the level of ancestry-specific IBD sharing within populations using the mean number of segments of a particular ancestral origin shared between individuals. We normalized this mean using as the denominator the ancestry-adjusted number of effective pairs, as described previously.[26,35] If there are N individuals, then the ancestry adjusted number of pairs of haplotypes is defined by, where
pi = the global ancestry of individual i
pj = the global ancestry of individual j

Global and ancestry-specific effective size estimation

To estimate whether the Malagasy ancestral populations came from small or large populations, we ran IBDNe Ancestry-Specific (AS) for the population following the pipeline described in Browning et al.[26] This method allows inferring ancestry-specific recent effective population size (Ne) by integrating inferred IBD segments and local ancestry calls. Based on this information, IBD segments can be classified based on ancestry and then used to infer the ancestry-specific Ne history, based on the relation between the number and length of IBD segments and the probability of coalescence. Following Browning et al.,[26] we implemented the IBDNe-AS method using IBD segments greater than 2cM.

To better place the demographic history of Madagascar in a global context, we calculated the overall recent demographic history with the software IBDNe[36] for populations across the Indian Ocean, using only the inferred IBD segments greater than 4cM. IBDNe authors recommend a cutoff between 3-6 cM for SNP array data, and 4 cM has been used for other populations in order to avoid the underestimation of effective population size.[27] As the density of SNPs is an important factor for these inferences, we only applied these methods to the Density dataset. All effective size analyses were limited to populations with sample sizes greater than 39 individuals and we inferred Ne only between 4 and 100 generations before the present, as in Borda et al.[38]

Demographic model

We used Msprime[39] to simulate three continental populations (Africa, Europe and Asia), two Bantu populations (Bantu, Mozambique), two Austronesian populations (Austronesian and South Borneo) and an admixed population (Madagascar). Our pre-admixture model for the three continental populations is based on a published model inferred from the 1000 Genomes project data[45] and also described in Browing et al.[26] This model simulates the origin of modern humans, the Out of Africa (OoA) event, the Asian/European split from the OoA population, and finally, the split between Asian and European populations.

The dispersal of Bantu lineages is the result of a population expansion that started in the Nigeria-Cameroon borderlands around 4,000 to 5,000 YBP.[62] Regarding the pre-admixture model for the Bantu expansion, we simulated an initial Bantu population (BAN) that diverged from the African continental population at 167 generations before present, starting with 2,200 individuals.[86] There are diverse hypotheses about the migration route(s) of the Bantu expansion; nevertheless, our aim in this study is to evaluate the demographic dynamics of the African ancestral population of Madagascar. Then, we split a second Bantu population (MOZ) representing the Mozambique population; which was sampled and is accessible through the publication.[62] Using the method euroIBD,[87] this date was estimated to be 60 generations BP (~ 1,800 years BP) (Methods S1L); IBDNe analyses showed an initial population size at this time of 3,000 individuals. Migration rates between African and Bantu populations were set to 0.0037 as in[86]. All African simulated populations were programmed to reach an effective size of 14,000 individuals at present.

For the Austronesian expansion event, evidence from linguistics and archaeology indicates that the expansion began around 4,000 to 5,000 years BP[88,89] with an initial effective size of 5,000 individuals.[89] Thus, we simulated an Austronesian population (AUS) that diverged from Mainland Southeast Asia (MSEA) based on these parameters: divergence from MSEA populations at 166 generations BP (~ 5,000 years BP, assuming a generation time of 30 years), with an initial Ne of 5,000 individuals. It should be noted that recent evidence suggest that Austronesians arrived in Taiwan in the north ~6,000 years BP,[90,91] nevertheless, this date would not affect our more recent inferences. Then, we approximated a South Borneo population (SBO) by simulating a population that diverged from the Austronesian population (AUS) 100 generations before present, as the euroIBD results over our Density dataset show a split between South Borneo and Philippines at this time (Methods S1L). South Borneo population correspond to an admixture of the Austronesian population and Mainland Southeast Asia. Based on IBDNe analysis, we set the Ne initial for South Borneo at 10,000 individuals. Migration rates between Asian and Austronesian populations were set to 0.0016.[92] All Asian simulated populations must reach an effective size of 45,000 individuals at present.

Finally, multiple scenarios were created and evaluated to simulate the history of the Malagasy population. We created this admixed population under two different admixture models – single pulse of admixture vs. multiple pulses of admixture (representing admixture occurring over several generations).[93] Regarding the single pulse model, we simulated an admixture event occurring at 20, 25 and 30 generations before present, between populations from MOZ and SBO at a proportion of 62.5%and 37.5%, respectively. The initial Ne

Please cite this article in press as: Alva et al., The loss of biodiversity in Madagascar is contemporaneous with major demographic events, Current Biology (2022), https://doi.org/10.1016/j.cub.2022.09.060

was set at 2500 individuals. We also simulated multiples pulses of admixture happening between the periods 20-25, 20-30 and 20-40 generations. Under this scenario, the admixed population begins with 900 individuals of AUS origin who then receive MOZ gene flow at a constant rate during 5, 10 or 20 generations, in order to reach the global ancestry proportions observed in Madagascar of ~3/8 Asian ancestry and ~5/8 African ancestry. Based on IBDne results, we simulated for the admixed populations a current Ne of 150,000 individuals. We consider this Ne realistic taking into account the whole Malagasy population over the past 100 years; in contrast to long-term Ne sizes that are much smaller, this Ne is in the range previously estimated for human populations in the Holocene period.[94]

Simulated genetic data

We simulated sequence data with a mutation rate of 1.25×10^{-8} per base pair per meiosis, using the HapMap GRCh37 genetic map for each autosomal chromosome. To replicate the real data available, we sampled 700 individuals from the admixed population, 161 from the Mozambique population, 91 from the South Borneo population, 100 from the Bantu population, 100 from the Austronesian population, 400 from East-Asia, 400 from Europe and 419 from Africa.

Given that Msprime reports the ancestral recombination graph for each individual's haplotypes, the output genetic data represent the ancestral and derived states with certainty of the haplotypes involved. Thus, as the output VCF report phased data, we did not directly perform a second pashing step with Beagle.

Simulation of empirical ascertainment bias in simulated genetic data

As Msprime produce a VCF file covering all the positions declared in the HapMap genetic map, we extracted all sites that lie in the genomic regions covered by our Density dataset. Particularly, we defined the genomic region covered by the Density dataset as the coordinates produced by the first SNP in the dataset and the last one for each chromosome. In this manner, we subset all the simulated positions declared in the HapMap genetic map for reducing the computation time of following steps. It has been observed that ascertainment bias can influence analyses based on individual SNPs, such as PCA and Fst analyses,[95] and thus can skew comparisons between empirical and simulated data. However, we have not found studies reporting the possible effect of ascertainment bias on IBD analyses. In principle, IBD analyses are based on genomic segments whose identification should not be skewed by SNP ascertainment bias. Nevertheless, we reproduced the observed ascertainment bias in our empirical datasets and confirmed that SNP ascertainment bias does not influence the results (Methods S1K). More specifically, we applied diverse filters based on minor allele frequency (MAF) in order to match the allele frequency spectrum between the empirical and simulated data (as done previously[95]). First, we applied different MAF filters (0.0, 0.01, 0.03 and 0.05) based on control or reference populations (African, European and Asian continental populations, from the 1000 Genomes panel) that were external to the history of Malagasy populations. Next, we obtained the union or the intersection between the set of filtered SNPs according to each reference population and MAF filter. By analysing 108 possible MAF filter combinations, we found that applying a MAF $\geq$ 0.03 and MAF $\geq$ 0.01 in Africa and Europe simulated populations, respectively, could reproduce the allele frequency spectrum observed in the African and Asian populations from 1000 Genomes in our empirical dataset (Methods S1M). Finally, we randomly selected a certain number of SNPs in a 10kb window, in order to match the SNP density of our Density Dataset. We note that even when ascertainment bias is neglected, we still found the same results for IBD detection using the scenarios from Figure 3 (Methods S1K). This result is in accordance with a similar demonstration performed by Gopalan et al.[96]

Demographic conditions for the Malagasy's Asian ancestral population

Given that we found an excess of IBD of Asian origin dated before the admixture, we hypothesize that this overrepresentation is possibly the result of the high coalescence rate between haplotypes due to founder or bottleneck events happening before the admixture. First, we compared the impact of multiple parameters on the distribution of IBDs, such as the strength of founder events (Ne at the beginning of bottleneck: Ne_startBOT), the duration of bottleneck events (Tdur_BOT), the amount of gene flow (Migration rate during the bottleneck: Mig_BOT), and the influence of demographic changes during the bottleneck i.e stable, in expansion or declining (Ne at the beginning of admixture : Ne_endBOT).

Due to similar results obtained using the whole genome or a subset of chromosomes (Methods S1E), we explored the Asian IBD distribution using only chromosomes 22 and 20, allowing us to run each of the 784 different scenarios three times, in order to compute a standard deviation and probability distribution. More specifically, we interrogated this signal by simulating a scenario of single-pulse admixture (happening at 25 generation BP) and other scenarios of multiple-waves of admixture (happening during 10 generations, between 20 to 30 generations BP). For both scenarios, Ne trajectories for the Asian ancestral population of Malagasy were simulated, varying the initial effective population size (Ne_startBOT = 50, 100, 300, 500, 1000, 5000 or 10000 individuals), the time of isolation (Tdur_BOT = 1, 5, 10, 20, 30, 40, 50, 70 generations before the admixture), the influence of gene flow (Mig_BOT = 1, 5 or 10 migrants per generation) and the possibility of demographic expansion/decline (reaching an Ne_endBOT = 900 individuals).

The maximal time of isolation and bottleneck tested is 70 generations (4,000 y. BP). This limit has been selected because it corresponds to the arrival of Austronesians in the Southeast Asian island and the admixture between AUS and MSEA.[97] Present Malagasy and present Borneo populations share this common admixture signal but not the bottleneck signal. This allows the maximal time of isolation/bottleneck as the split time between AUS and SBO (common ancestral population of simulated Malagasy and Borneo populations). Also, in all scenarios, bottleneck and isolation periods stop simultaneously at the beginning of the admixture with MOZ. By definition the isolation stops at the admixture. IBDne and IDBNe-AS showed that the admixture happened between two small populations. Since our method would not detect a limited expansion a few generations before the admixture, we use the start

of admixture as the final point for bottleneck which allows us to keep constant the african demographic parameters (estimated through IBDne-AS).

We evaluated different demographic scenarios by calculating an indicator based on the difference between the number of observed and simulated IBD segments. Specifically, for each of the last 75 generations, we calculated the absolute difference between the number of observed and simulated IBD, and then computed the mean to obtain the final indicator.

Based on this exploratory analysis, we simulated all 22 chromosomes to implement IBD-sharing analyses over a set of specific demographic scenarios, which were chosen based on anthropological hypotheses (Methods S1H). The objective of this step is to assess the likelihood of each scenario according to the Asian IBD-sharing distribution. More precisely, we considered the following scenarios for the Asian population ancestral to the Malagasy:

- Split from Southern Borneo at 25 generations BP through a founder event followed immediately by the admixture (Ne_endBOT=900, Ne_startBOT=900, Tdur_BOT =0, Mig_bot=0).
- Split from Southern Borneo at 26 generation BP through a strong founder event followed by admixture one generation after (Ne_endBOT=900, Ne_startBOT =50, Tdur_BOT =1, Mig_BOT =0).
- Admixture through multiple pulses of gene flow (Ne=70 to 300) coming from an ancestral Southern Borneo population, between 20-30 generations before present.
- Split from Southern Borneo through a founder event of Ne_startBOT =5,000 individuals, followed by a slow decrease of Ne during 50 generations (Tdur_BOT =50, Mig_BOT =0) until reaching Ne_endBOT=900 (before the admixture at 25 generations BP).
- Split from Southern Borneo through a strong founder event of Ne_startBOT =50, with effective population size expansion during 10 generations (Tdur_BOT =10, Mig_BOT =0), reaching Ne_endBOT=900 (before the admixture at 25 generations BP).
- Split from Southern Borneo through a founder event of Ne_startBOT =500, with effective population size constant (Ne_endBOT=500) during 40 generations (Tdur_BOT =40, Mig_BOT =0), receiving no gene flow before the admixture. It is important to notice that this was the best candidate based on the exploratory simulation analysis (Methods S1H, S1I, and Data S1).

Analysis of demographic simulations

We analyzed all demographic scenarios with an automated bioinformatics pipeline, implementing the same methods and parameters used for the observed data. More precisely, we performed local ancestry inference using the software package RFMix. The time since admixture was set to 30 generations before present and a window size of 0.2 cM. We used as a reference panel the AFR and MSEA populations from the simulated data. Regarding the IBD segment determination, we used RefinedIBD and filled the gaps between segments with the merge-ibd script, specifying a gap's length of 0.6 cM and a number of discordant sites =1. In order to assess the relationship between age of IBDs and their size, we also recovered the IBD segments shared between simulated Malagasy individuals from Msprime simulated tree sequences (chromosome 22 and 20, scenario 6, Methods S1Q).

Effect of the coalescent simulators

To control the effect of the coalescent simulators, we simulated the same demographic scenario (Ne_startBOT=500; Ne_endBOT=500; Tdur_BOT=40g, Mig_BOT=0.0) under the Standard Coalescent model, the discrete time Wright-Fisher model and a combination of these two (DFTW during the first 500 generations and the Standard Coalescent until the end of simulation) on chromosome 22 and 20. All coalescence models produce similar results, that is a similar accumulation of short IDB in a long bottleneck scenario (Methods S1R).

Approximate Bayesian Computation (ABC)

We implemented the Approximate Bayesian Computation (ABC) for estimating confidence intervals. We defined three alternative models of genetic and demographic history: demographic model A: a founder effect for the Asian ancestral population, demographic model B : multiple founder effects through multiple Asian waves of gene flow, demographic model C: bottleneck event for the Asian ancestral population (Methods S1S). For each model, 500 simulations were generated, and for each simulation different parameter values were randomly chosen, as shown in Table S1. We choose as summary statistics the pairwise IBD-sharing, as it should be informative for distinguishing among competing scenarios that are under consideration, and should be strongly influenced by the parameters of interest. Model choice was performed using a rejection method (package abc from R) based on 500 simulations for each studied demographic scenario. First, we performed a cross-validation step based on 100 pseudo-observed simulations with the rejection method, considering a tolerance rate of 5% (Methods S1S). Next, we estimated the posterior probability of each demographic scenario, using the rejection method and retaining 10% of the simulations with the closest summary statistic values. Based on the best supported demographic model, we performed parameter estimation with ABC from 5668 simulations. In particular, we estimated the Ne_startBOT, Ne_endBOT, Tdur_BOT and Mig_BOT parameters, using the neural network, local linear regression and rejection algorithms, implemented in the abc package from R, under a tolerance rate of 0.05.

Conclusions and perspectives

Conclusions and perspectives

The timing of Madagascar's initial settlement and the involved consequences are topics of current concern among multiple disciplines, as there is particular interest on the rate and manner in which human settlers altered the island's landscapes and ecologies. We have demonstrated that haplotype-based inference, coupled with simulated genetic data, can provide crucial clues to reconstruct the demographic history of human populations that have participated in recent settlement events. By interrogating the impact of different demographic scenarios on the Asian IBD-sharing of Malagasy populations, we detected a long-term bottleneck for the Asian ancestral population of Malagasy, and estimated the genetic separation timeline from the source population (timing of the founder event) and the reduction in population size (strength of founder event). We propose that this complex demographic process occurred before the admixture with Bantu-speaking populations, describing at least 1,200 years of evolutionary history shared by the Asian ancestors of Malagasy. By integrating these results into our model, we found that the bottleneck event for the Asian ancestral population before the admixture was crucial to describe the system as well as its response to perturbations.

We complemented our demographic conclusions bridging ecological, population genetics and evolutionary timescales to understand the possible relationship between human activities, species extinctions and climate change. Beyond genes, contact between populations also brings interchanges of cultural practices, which ultimately can lead to a different exploitation of the land. Our study suggests that the admixture event is contemporaneous to drastic changes for both ancestral populations as well as for Madagascar's environment. Firstly, the Asian ancestral population ended their bottleneck when they entered in contact with Bantu-speaking populations. During this time, different regions of Madagascar presented a transition from forest to grassland (Burns et al., 2016; Voarintsoa et al., 2017), but also a subsistence strategy shift (Godfrey et al., 2019), coupled with the introduction of vertebrates, such as cattle (Hixon, Douglass, Crowley, et al., 2021), dogs(Hixon, Douglass, Crowley, et al., 2021; Hixon, Douglass, Godfrey, et al., 2021) and rats (Crowley, 2010), as well as cultivable plants of Asian origin, like rice (Crowther et al., 2016). In this manner, major ecological transformations in Madagascar were encompassed

by the use of land during the contact of its ancestral populations, favouring agropastoralism, trade and urbanism.

In addition, we were able to inspect other interrogations that are key elements for the study of Madagascar settlement. According to the list of questions previously presented, we inspected in the first place who are the closest genetic parental groups of the Malagasy and when did Malagasy ancestors start to diverge from them. Thus, we assessed genetic ancestral contributions from Africa and Asia by computing the number of IBD segments shared between Malagasy and other populations. We found that populations that inhabit Mozambique and South Borneo share the largest quantity of IBD segments with Malagasy individuals, pointing to these populations as the closest genetic parental groups of Malagasy. Regarding the divergence time between these populations, previous analyses have proposed a divergence time of 1,500 and 2,000-3,000 years ago between Malagasy and their African and Asian ancestors, respectively.

In a second time, we investigated if the ancestors of the Malagasy population(s) came from small or large populations. Thus, we inferred changes in effective population size (Ne) over recent time in Madagascar using the haplotype-based IBDNe method (Browning & Browning, 2015) and Ancestry-specific IBDNe application (Browning et al., 2018). Considering each ancestry in turn, we observed similarities and differences in the estimated pre-admixture effective population sizes: in the African component, we observed a reduction in effective size during the period of 1200-3000 years ago, going from 8,290 to 2,490 individuals, and in the Asian component, we saw a severe reduction in Ne for the period 1,020-3000 years ago, going from 89,400 individuals to 319 individuals. Interestingly, results based on computational simulations suggest a bottleneck of ~500 individuals during at least 1,200 years for the Asian ancestral population. To summarize, we showed that current-day populations in Madagascar come from an admixture of individuals coming from small African and Asian ancestral populations, although the evolutionary history of each one is drastically different.

In a third time, we inspected how many people originally founded the Malagasy population(s). Using the results from IBDNe and our simulations, we found that Malagasy populations present an effective size at the beginning of admixture of 2,500 individuals, where approximately 37.5% of individuals were of Asian ascendance and 62.5% of African

ascendance. We coded these results on our simulations and produced the same admixture date and global ancestry profile observed in empirical data. It is important to mention that the effective size at the admixture varies according to the Malagasy genetic group analysed, which implies that these observations must be treated and interpreted carefully, as they can be impacted by evolutionary events in the ancestral source populations, but also by demographic and migration events during and after the settlement of Madagascar.

In a fourth analysis, we wanted to inspect if Madagascar was settled via large-scale population movements, or through a smaller translocation of individuals. Regarding the Asian component, our results demonstrate that prior hypothesized scenarios are insufficient to explain the overrepresentation of Asian IBD segments, particularly where Madagascar is settled by large-scale population movements (modeled through a continuous gene flow from a large Asian ancestral population inhabiting Indonesia). Thus, our analyses suggest that the effective size of the Asian ancestral source population was limited to a few hundreds, implying that they may have reached a few thousand of individuals in census size. We do not know if the Asian ancestors arrived to the island driven by extrinsic factors (such as wind, ocean currents, etc) or by specialized one-way movements (Cayuela et al., 2018). Previous genetic analyses have discarded recurrent, two-way out and back genetic exchanges (Pierron et al., 2017). Importantly, knowing if this small translocation of individuals occurred during the scale of one or multiple generations could give us insight about the distance and context of the complex dispersal of Malagasy Asian ancestors that ultimately reached the island. Regarding the African component, we explored if the model of admixture affected the African IBD-sharing distribution in Madagascar. We found that simulated scenarios where admixture occurred through the translocation of African individuals ($N_e \approx 1{,}500$) or where continuous gene flow lasted 150-600 years can reproduce the admixture dates and the global ancestry profile observed in Malagasy population (with small variations in IBD-sharing). Importantly, the effective size of these continuous population movements goes from ~75 to ~500 individuals. In this case, both scenarios seem possible for the African ancestors of Malagasy, indicating a complex scenario of connections between populations from Madagascar and the eastern Africa around the date of admixture.

Finally, we tried to reconstruct the possible chronology of Madagascar settlement and admixture. Firstly, our modeling suggested that the Asian ancestral population ended their long-term bottleneck (~1,000 years) when they entered in contact with ancestral Bantu-speaking populations. Based on genetic analyses, this contact most likely occurred between 800-1100 years BP, after which the Malagasy population underwent a large demographic expansion. Our genetic analyses found that ancestry-specific demographic trajectories correspond with megafaunal extinction timing, the introduction of foreign species (cattle, rice, etc.) (Crowley, 2010; Crowther et al., 2016; Hixon, Douglass, Crowley, et al., 2021; Hixon, Douglass, Godfrey, et al., 2021); and important transitions in land-use (Burns et al., 2016; Voarintsoa et al., 2017), supported by a change of subsistence strategies (Godfrey et al., 2019). We wanted to know how much time passed between the arrivals of Malagasy ancestors to the island and the admixture, nevertheless, this question requires the complementarity of archaeological analyses and genetic studies based on contemporary and ancient DNA samples.

We are looking forward to expanding our computational model to explore other axes that could give us a wider insight regarding the settlement of Madagascar. First, it would be necessary to better characterize the African contribution to Malagasy genetic pool, inspecting if more than one ancestral source in the African eastern coast participated in the admixture, as well as when these movements could have happened and how many people arrived. For achieving this purpose, it would be necessary to simulate more than one bantu population inhabiting the African eastern coast, where the divergence time and gene flow between them are essential; then, different scenarios should be designed for testing if the African contribution to Malagasy genetic pool comes from one or multiple pulses of admixture, and if these pulses represent individuals from one same homogenous population. We can evaluate the pertinence of these scenarios by comparing simulated and empirical data concerning the dates of admixture, ADMIXTURE profiles, IBD-sharing between Malagasy individuals and eastern Africa individuals and global ancestry profiles.

Secondly, we would like to investigate the genetic contributions coming from Europe, Middle East and South Asia, in order to characterize the possible dates of gene flow. For this objective, we could add these genetic contributions to our computation model and test if they occur around or after the admixture of Austronesian-speaking and Bantu-

speaking populations. Thus, it will be necessary to add these populations to the empirical dataset; simulate their evolutionary history in our demographic model; generate different scenarios where gene flow to Malagasy populations occurs at different generations and rates; finally, the plausibility of these genetic contributions can be evaluated comparing simulated and empirical data, using as summary statistics the global ancestry profile of individuals and the IBD segments shared (in terms of quantity and length) between Malagasy and abroad populations.

A third point essential to understand the settlement involves the modelling of Malagasy population structure. As previously done in other studies (Ioannidis et al., 2021; Ralph & Coop, 2013), we can use the IBD-sharing between individuals in order to infer the divergence time between populations. This part would be complex, as it requires modelling different dates and proportions of admixture, which also seem to be correlated with geography. Once we have detailed the demographic parameters of ancestral populations, it will be necessary to interrogate how we can explain the genetic structure of Malagasy populations. Thus, different admixture scenarios should be considered for each genetic cluster previously identified in Madagascar, varying the initial Ne of Malagasy population and the global proportions of admixture. At the end, we can compare the admixture dates, global ancestry profile and the IBD-sharing between Malagasy individuals for detecting the set of scenarios that could reproduce the observed genetic diversity in Madagascar. At this point, it would be useful to implement an analysis for inspecting the male/female demographic configuration of African and Asian ancestors that participated in the admixture, through the analysis of mitochondrial DNA and the analyses of haplotypes from chromosomes X and Y. Finally, the computational model will complement the selection analyses on Malagasy genetic variation, as it provides an estimate of the expected diversity due to only demographic factors. Importantly, the same model can be coded under different simulation algorithms, which can help us to implement particular scenarios of selective pressure in order to evaluate potential selection signals coming from Malagasy genomes.

In most anthropological research studies, an essential factor for interrogating hypotheses is how the data was collected. In fact, the sampling design determines which people or populations are to be sampled and how large the samples should be, which can directly affect the general use of this data in research on human genome variation and the

different sets of hypotheses that can be tested (National Research Council, 1997). Thanks to the efforts done by the consortium MAGE, the grid-based sampling of Malagasy populations allowed us to study the patterns of variation in the genome, as well as the patterns of gene flow and the structure of the population. However, the interpretation of sequence variation is still a challenge, since the production of data has increased but no much progress towards mapping and understanding complex genetic relationships has been done. Therefore, the scientific work done during this doctoral thesis delves into the systems biology approaches for interpreting genetic variation from a functional point of view, studying the interactions between parts of the system using experimental and computational methods (Conesa & Mortazavi, 2014; Sunyaev & Roth, 2013). This strategy relies on genetic modelling, which is a powerful tool for testing the likelihood of a historical scenario. Specifically, the demographic model based on IBD-sharing distribution and local ancestry can accurately differentiate between founder effects and bottleneck events, while can also detect periods of isolation or population expansion due to external migrations.

Nevertheless, we acknowledge that this strategy presents limitations that must be explored in the future. First, this systems biology approach was applied and refined focusing on the Malagasy population, using the genetic data available and the previous research done on the settlement of Madagascar. As such, the model did not include other parameters that may interfere with mating behavior, social structuring, sex-related events, generation overlap, *etc*. These possible confounding factors (which represent genuine questions) should be addressed by expanding the assumptions and evolutionary events of the computational model created during this doctoral thesis. In the meantime, the anthropological interpretations based on the proposed evolutionary parameters and their confidence intervals must be taken carefully. Secondly, it is important to keep in mind that the estimated dates of evolutionary events are based on the duration of one generation in humans, which can make it difficult to compare with other dating methods like those used in archaeology. For example, genetic analysis depends on estimates of recombination and mutation rates and generation times, which have not remained constant through time and may differ between populations. In any case, the described evolutionary events in this chapter happened in a time scale of centuries, which offer a confident overlapping signal with previous archeological evidence. Thirdly, the effective size estimation does not inform

about the presence of individuals and populations that could have potentially been there also, but didn't contribute to the current genetic pool of the studied population. Given the complexity of Madagascar settlement, it will be necessary a more extensive sampling of contemporaneous populations around the Indian Ocean, but it will be crucial to access ancient DNA in order to elucidate the past evolutionary history of Malagasy. Taking into account these limitations, we have limited our conclusion to the demonstration of the long-term bottleneck and isolation of the Malagasy Asian ancestors.

CHAPTER III

THE DEMOGRAPHIC HISTORY AND MUTATIONAL LOAD OF MALAGASY POPULATIONS

CHAPTER III

THE DEMOGRAPHIC HISTORY AND MUTATIONAL LOAD OF MALAGASY POPULATIONS

Introduction

Current genetic diversity observed in human populations is the result of past evolutionary processes, geographic expansions and demographic histories. In particular, recent studies suggest that past demographic histories can differentially impact the load and the distribution of deleterious variants across populations (Henn et al., 2015). The burden of deleterious mutations of a population is defined as mutational load, and depicts the reduction in the average fitness of a population due to deleterious mutations compared with the theoretical optimal fitness (Grossen et al., 2020; Lopez et al., 2018). It has been shown that the mutational load of deleterious mutations may vary according to the effective size (Ne) trajectory of the population, including founder and bottleneck events (Grossen et al., 2020). A mentioned example is that during a prolonged bottleneck, a fraction of deleterious mutations may shift between being weakly selected to being effectively neutral and drift to fixation, thus increasing the load (Glémin et al., 2003; Lopez et al., 2018). Indeed, depending on the effect of deleterious mutations, the genetic drift and the effective size of the population, these alleles can reach different frequencies and prevalence scenarios. For instance, the functionality of mutations (deleterious effect and expression) might be differently impacted by evolutionary forces, *i.e.,* the probability of a highly deleterious pathogenic mutation to be fixed in a population is *null*, as they tend to be eliminated and rarely rise to high frequencies (Henn et al., 2015). In contrast, a neutral or weaker effect mutation can fluctuate to higher frequencies due to random drift. Interestingly, the probability to rapidly disappear or to be fixed for such mutations is higher in small populations. Thus, the mutation load of deleterious mutations in a given population may reflect its demographic history.

Several recent papers have tested if human populations carry differential burdens of deleterious mutations due to differences in demographic histories, reaching a variety of contradictory conclusions (Henn et al., 2015, 2016; Simons et al., 2014). For instance, the Out of Africa (OOA) dispersal ~60,000 years ago is characterized by a series of founder events as modern humans expanded into multiple continents (DeGiorgio et al., 2009). When these events are modelled through computational simulations, the observed heterozygosity in deleterious sites is greater in OOA populations, particularly for moderate and large effect mutations, suggesting an accumulation of deleterious alleles with the OOA dispersal (particularly under a model of where deleterious mutations are recessive). Importantly, the dynamics of the mutational load depends on the assumed fitness effect of an allele, where an additive model of dominance yields small differences in load between populations; while under a recessive model the mutational load varies according to specific demographic histories (Lopez et al., 2018). These two models assume different distributions of dominance across the genome: the additive model measures the effect of dominance on fitness by assuming that deleterious mutations exhibit some penetrance; on the contrary, recessive models assume that deleterious mutations show no penetrance (in heterozygotes).

Many studies have used summary statistics to approximate the mutational load under different allele dominance models: Pedersen and collaborators (Pedersen et al., 2017) interrogated the genetic consequences of population bottleneck of ~20,000 years and found that assuming an additive effect of deleterious alleles, the Inuit show a slight increase in load (< 4%) compared to European, East Asian and African populations. In contrast, the genetic load under a recessive model suggested a higher load (~20%) compared to other less bottlenecked human populations (Pedersen et al., 2017). Importantly, through forward simulations they found that there is a small but significant increase of 4% in genetic load, suggesting that small population sizes over a long period of time can lead to increased genetic load even if alleles have an additive effect. In another study from 2018, Lopez and colleagues interrogated if different past demographic histories (representing different subsistence strategies) have affected the efficacy of selection on the distribution of deleterious alleles across hunter-gatherer and farmers populations (Lopez et al., 2018). They showed that the distribution of deleterious alleles was compatible with a similar

efficacy of selection to remove deleterious variants with additive effects, predicting with simulations that present-day mutational loads are almost identical between hunter-gatherer and farmer populations. Under the recessive model of dominance, the trajectory of mutational loads indicated that the population decline experienced by Europeans (proxy of OOA population) and African hunter-gatherers led to a surge in the load; where its long duration led to an increase in load contributed by weakly deleterious mutations. Nevertheless, they propose that this increase has been partially counteracted by strong gene flow from expanding farmers.

As presented in Chapter II, the demographic history of African and Asian ancestors of Malagasy has been addressed using genome-wide data, the sharing of chromosome segments between individuals (Identity-by-Descent), local ancestry information and simulated genetic data. By modelling alternative scenarios for the demographic history of the Asian ancestral population, exhaustive simulation analyses suggested that the current Malagasy population originated from an Austronesian-speaking population who was genetically isolated for around >1,000 years. We aim to explore if this isolation, associated with a long-term bottleneck, could have affected the efficacy of selection, yielding to an increase of deleterious alleles. In order to better understand the evolutionary forces acting on Madagascar, we report whole-genome sequencing (WGS) data from 67 Malagasy individuals and analyse patterns of deleterious mutations in the Malagasy populations, inspecting possible differences in the ancestry-specific distribution of allele frequencies.

We assessed the deleteriousness of variants using a method based on sequence conservation (genomic evolutionary rate profiling-rejected substitution (GERP-RS)) (Cooper et al., 2005; Goode et al., 2010). As used in previous studies, these conservation score reflects various levels of constraint within a mammalian phylogeny, and is used to categorize mutations by their predicted deleterious effect: the more phylogenetically conserved a site is, the more likely it is that a new allele is deleterious and has a high GERP-RS score (Henn et al., 2015). In this manner, we compared different proxies for genetic load between Malagasy and other populations. For additive models, we compared the number of derived mutations (assigned to different GERP-RS scores) per individual across Malagasy and 1000 Genomes Project populations (The 1000 Genomes Project Consortium et al., 2015). For the recessive model, we compared the number of

homozygous-derived genotypes per individual, which has been equivalently used to quantify genetic load under a recessive model (Henn et al., 2016; Lopez et al., 2018; Pedersen et al., 2017). In addition, we looked at the mutational load over a set of putatively loss of function (LoF) variants proposed by functional annotation algorithms (Balasubramanian et al., 2017; Cingolani et al., 2012). We explored possible signals of past demographic events on present-day genetic variation in Madagascar at the whole genome level, producing a catalogue of derived mutations with potentially deleterious effects, which can be interrogated in future work.

Methodology

Samples

In this study, we analysed 67 genomes that were collected by the MAGE (Madagascar Genetic and Ethnolinguistic) consortium during 2007 – 2014, with ethical approval by the Human Subjects' Ethics Committees of the Health Ministry of Madagascar and by French committees (Ministry of Research, National Commission for Data Protection and Liberties and Persons Protection Committee). Sampling was done using an extensive grid based approach, in which individuals were sampled from 257 villages (2.8 ± 0.7 individuals per village) from all over Madagascar. All individuals were unrelated; all of them received detailed information on the study and gave a written consent prior to the donation of their sample. Additional samples were used in order to understand Madagascar's diversity in an ocean Indian perspective. High-coverage data from 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015) were recovered as follows: 99 ESN, 99 LWK and 108 YRI individuals, representing the Esan, Luhya and Yoruba populations in Africa, respectively. While for the east-Asian populations we worked with 99 CDX, 103 CHB and 99 KHV samples, corresponding to the Chinese Dai, Han Chinese and Kinh populations. As reference populations for SNP ancestry determination we built a set of African (n=306) and East-Asian (n=295) individuals from merging the samples mentioned above.

Read Processing

Whole-genome sequencing was performed in the Centre National de Recherche en Génomique Humaine (CNRGH) using Illumina technology. Raw sequencing reads were mapped to the hg19 reference genome using BWA (mem algorithm and default parameters) (Li et al., 2009). Duplicates were removed with Picard tools. Best-practice recommendations from GATK 4.1 (McKenna et al., 2010) were followed for local realignment around indels, recalibration and base calling. Per-individuals gVCF files were generated using GATK HaplotypeCaller. Each individual's gVCFs was combined into multisample variants database using GenomicsDBImport, and joint genotyping was performed using GATK GenotypeGVCFs in order to output all sites to a multi-sample VCF. Then we filtered out genotypes using the GATK Variant Quality Score Recalibration tool and used the call set from the 99.0 quality tranche. We recovered 20,631,821 SNP and indel variants.

Phasing Analysis

For haplotype determination we kept 15,437,487 SNPs bi-allelic fully genotyped in our data set and in 1000 Genomes data, removing all indels. Phasing analysis was performed with BEAGLE version 5.0 (Browning & Browning, 2007), parameters par default, using the HapMap recombination map for GRCh37 assembly (The International HapMap Consortium, 2007).

Local Ancestry Estimation

Local Ancestry Inference was performed on the 67 Malagasy Individuals, using as proxy of the ancestral populations the reference data set of African and East Asian individuals. Local Ancestry analyses were performed using the software RFMix v1.5.4 (Maples et al., 2013). Results from RFMix were obtained using the TrioPhased algorithm with a window size of 0.2 and 27 generations since admixture and n=5. Global ancestry proportions for each Malagasy individual were calculated by dividing the number of SNPs belonging to a particular ancestry by the total number of SNPs.

Variant Annotation

Ancestral state was inferred using 1000 Genomes annotated VCFs, which is based on orthologous regions in a great ape and rhesus macaque phylogeny, as described in Henn *et*

al (2015). As in Lucas-Sanchez *et al*, sites where the ancestral state was unknown were removed from the dataset for all analyses (Lucas-Sánchez et al., 2021), yielding a total of 14,987,217 annotated sites. Deleteriousness of each variant was assessed using GERP RS scores, which is a method to predict the effect of allele substitutions based on sequence conservation across different taxa of mammals (Cooper et al., 2005; Davydov et al., 2010; Goode et al., 2010). GERP RS scores were collected and annotated with respect to the derived allele, using ANNOVAR database (Wang et al., 2010) and resulting in a final dataset of 1,111,408 sites. Then, we classified each individual's variants according to local ancestry assignments inferred at these regions, mapping for each allele the ancestral assignment (African or Asian) done by RFMix software. On a second step, we approximated deleterious mutations by annotating the variants associated with a loss-of-function (LoF) effect, such as stop codon gain, loss of start codon, mutations in splice donor and acceptor sites, in all consensus coding sequences (CCDS). snpEFF and aLoFT software were used for this purpose (Balasubramanian et al., 2017; Cingolani et al., 2012). Entrez database was used for recovering the genes associated to a particular locus (Maglott et al., 2011). KEGG database was used for retrieving annotated information for that gene (Kanehisa et al., 2016).

Statistical and Frequency Analysis

Genomic data management, allelic frequency and individual's count of derived alleles were performed with VCFtools v0.1.13 (Danecek et al., 2011). Importantly, based on allele frequencies across populations, we kept variants that presented a derived allele in all subsets of individuals from Malagasy, African and Asian populations, given that we ignored: (1) if this site presented an alternative allele in the rest of sub-sampled individuals; (2) if this allele was truly non-variant or a sequencing error, as we did not verified coverage levels in the 1000 genomes panel for assuring the genotype state. Statistical analysis (mean ancestries, sd, quantile determination) and data visualization were done in R and plotted with the ggplot package (R Core Team, 2021).

Mutation load

We approximated genetic load using the GERP score load, described in Henn, et al (2015). This statistic transforms each GERP-RS category into a selection coefficient: the moderate

category ($2 \leq$ GERP-RS < 4) is assigned to $4.5x10^{-4}$, the large category ($4 \leq$ GERP-RS < 6) is assigned to $4.5x10^{-3}$, and extreme effects category (GERP-RS ≥ 6) is assigned to $1x10^{-2}$. Then, the genetic load per site is estimated according to the formula (Kimura et al., 1963):

$$\text{Load} = 1 - w = 1 - (1 - 2q(1 - q)sh - sq^2)$$

In this formula, q is the allele frequency, h corresponds to the dominance model (h=0 if recessive and h=0.5 if additive), while s denotes the selection coefficient previously assigned based on GERP-RS categories. The total GERP mutation load corresponds to the sum of GERP score at each site.

Results

Population history and global patterns of genetic diversity

In order to better understand the evolutionary history of the Malagasy, we have analysed whole genome data from 67 non-related Malagasy individuals sampled all over Madagascar (Fig. 1a). After applying quality filtering, we recovered 20,631,821 variants; about 3 Millions variants were not present in the 1000 Genomes Consortium data (Figure 1). For quality and time purposes, we focalized only on biallelic SNPs (15,437,487), as indels and multi-allelic SNP increase the error rate step between diverse databases. Multiallelic sites can provide unique insights into human migration and disease; however, more consistent and accurate annotation of such sites is required to fully exploit this potential. After phasing, we determined the ancestry proportions for each of the 67 Malagasy individuals using 15.43 millions SNPs across the 22 autosomal chromosomes. We applied RFMix software (Maples et al., 2013) to calculate local ancestry inference. Reference panels were constructed for African (n=306) and East-Asian (n=295) samples taken from 1,000 Genomes Project populations (The 1000 Genomes Project Consortium et al., 2015) to approximate the ancestral populations for admixed Malagasy population similarly to a previous study (Pierron et al., 2017). Among the individuals, African (AFR) ancestry percentage varied from 27.12% to 86.75% and Asian percentage from 13.25% to 72.88%,

with average ancestry proportions of 66.09% for African and 33.91% for Asian ancestry (sd=12.17%). We observed a heterogeneous ancestry distribution across the island, where higher proportions of Asian ancestry are located in the center, while African ancestry is predominant in the rest of Madagascar, with the highest values present in the north coast of the island (Figure 1). This whole genome analysis confirms the ancestry diversity at the island level found by previous analyses performed on microarray data (Pierron et al., 2014, 2017).

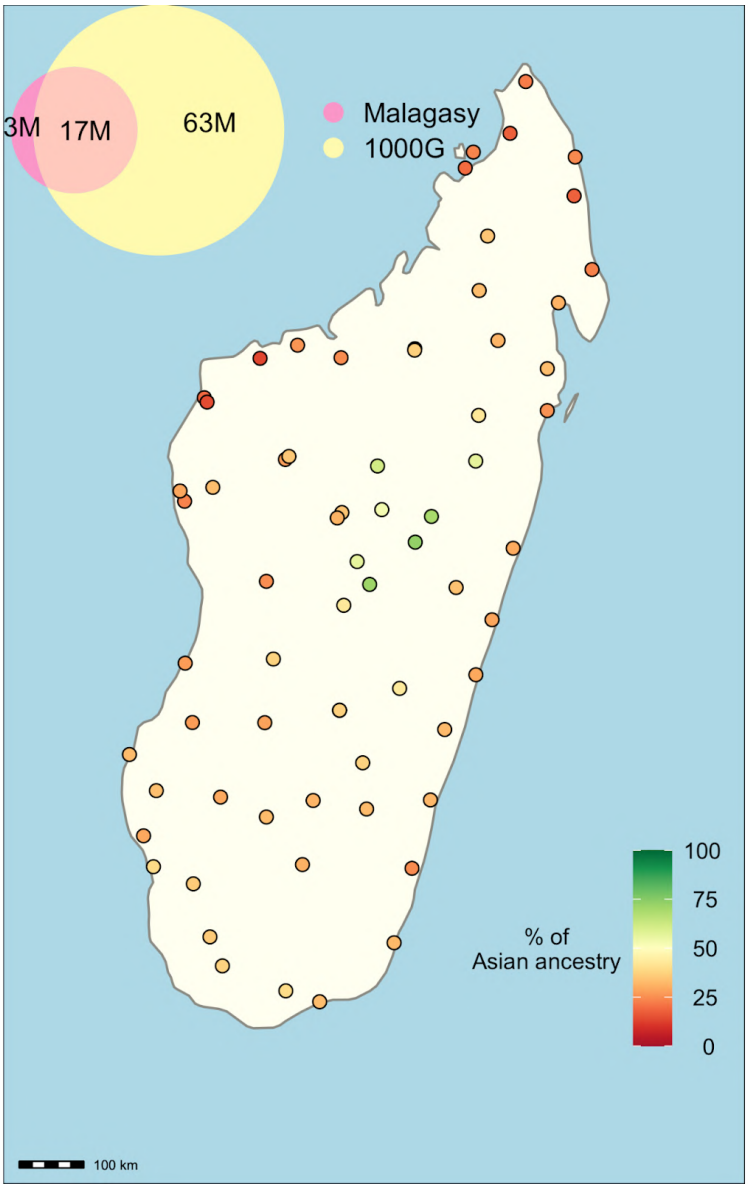


Figure 1. Sampling locations of whole-genome sequenced individuals and their estimated global ancestry profile. Sampling locations of sequenced Malagasy individuals

and their global ancestry information. Venn diagram depicts the variants recovered used in this study after whole genome sequencing.

Site Frequency spectrum across populations for deleterious derived mutations

We effectively annotated 14,987,217 sites with the ancestral allele information (we discarded sites where the ancestral allele was missing), nevertheless, our dataset was reduced by the database of associated GERP-RS scores, which is limited to regions of the genome that can be reliably aligned across typically as much as 35 mammal species and, for computational optimization, only reports sites with scores > 2. The underlying assumption is that high conservation of sequences (with little variation across mammal) is due to selective constraint, and that any changes would be potentially deleterious (i.e. strong negative selection). In this manner, the GERP score was available for 1,111,408 SNVs, allowing an explorative view of the load and distribution of deleterious variation in the studied populations. As we kept sites where the derived allele was present in all populations, we recovered a total of 370,766 variants for our exploratory dataset. We then classified these SNVs into 3 categories based on GERP-RS scores, reflecting the likely severity of mutational effects: moderate ($2 \leq$ GERP-RS < 4), large ($4 \leq$ GERP-RS < 6) and extreme (GERP-RS ≥ 6). We found that the majority of SNVs are of moderate effect (SNVs=324,241), followed by 45,728 variants of large effect and only a small fraction (SNVs=797) of variants representing extreme effects.

In order to quantify and summarize the possible effect of population-specific history on deleterious mutations, we explored the site frequency spectrum (SFS) for summarizing the distribution and enrichment of deleterious mutations across diverse populations (Fig. 2). Using 1000 Genomes Project data, we sampled 67 individuals from the Yoruba, Luhya, Esan, Han, Dai, and Kinh populations, and compared the SFS with Malagasy genetic variation. It is important to notice that we report the distribution of alleles according to the derived allele frequency (DAF), which produces a different SFS pattern regarding the minor allele frequency (MAF). As recapitulated elsewhere (Henn et al., 2015; Laval et al., 2010), demography results in different SFS for each population: we observe that Asian populations have many more derived fixed variants than any other population (average 8.54% of total sites; sd=0.23%), which can be in concordance with the OOA dispersal

through serial founder events resulting in the fixation by genetic drift. On the other hand, African populations present less sites with fixed derived variants (average 1.36% of total sites; sd=0.19%), implying that a vast quantity of novel variants have appeared without reaching fixation, which is expected for populations with large long-term effective sizes. Regarding Madagascar, we detected a similar SFS distribution to African and Asian proxies; nevertheless, the quantity of fixed derived alleles is the lowest between all populations (0.03% of total sites). Importantly, this distribution of fixed derived alleles seems to be conserved across moderate, large and extreme deleteriousness categories.

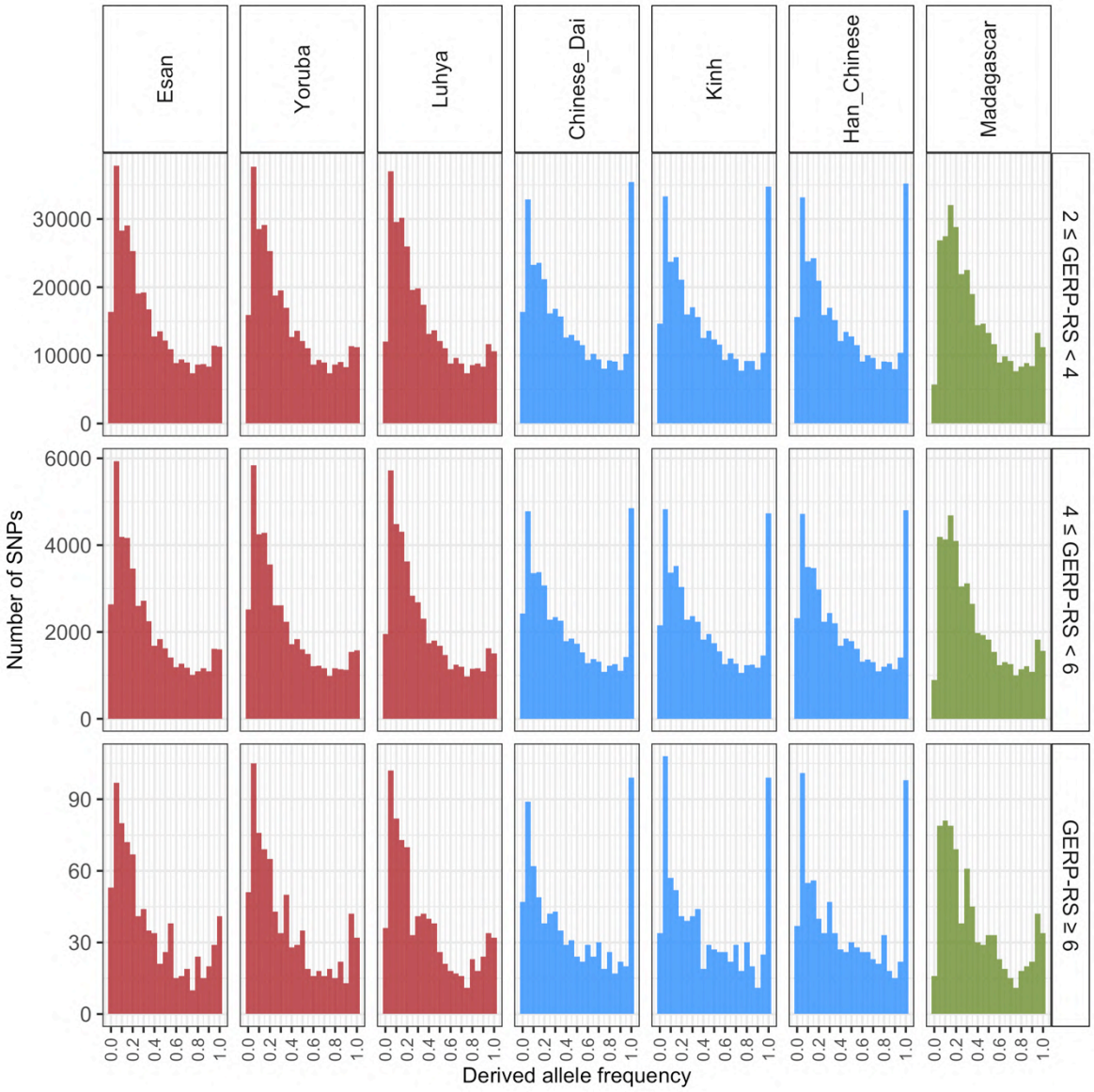


Figure 2. Differences of the site frequency spectrum (SFS) for derived alleles across populations. Derived variants were annotated with GERP-RS score and classified

according to the predicted deleterious effect: moderate ($2 \leq$ GERP-RS < 4), large ($4 \leq$ GERP-RS < 6) and extreme (GERP-RS ≥ 6).

We were interested in genomic sites with predicted extreme deleterious effects, as new mutations will be held at low frequencies if their effect is deleterious, and eventually be eliminated (Henn et al., 2015). Thus, we investigated the distribution of these sites on the genome based on the number of derived alleles, which is summarized in figure 3. We found mutations associated with extreme deleterious effects (GERP-RS $\geq$ 6) in all chromosomes, where chromosome 2 and 22 present the highest (n=117) and lowest (n=3) number of annotated sites, respectively. Since this could be related to chromosome specific gene density and chromosome size, further analyses are needed to detect selective processes.

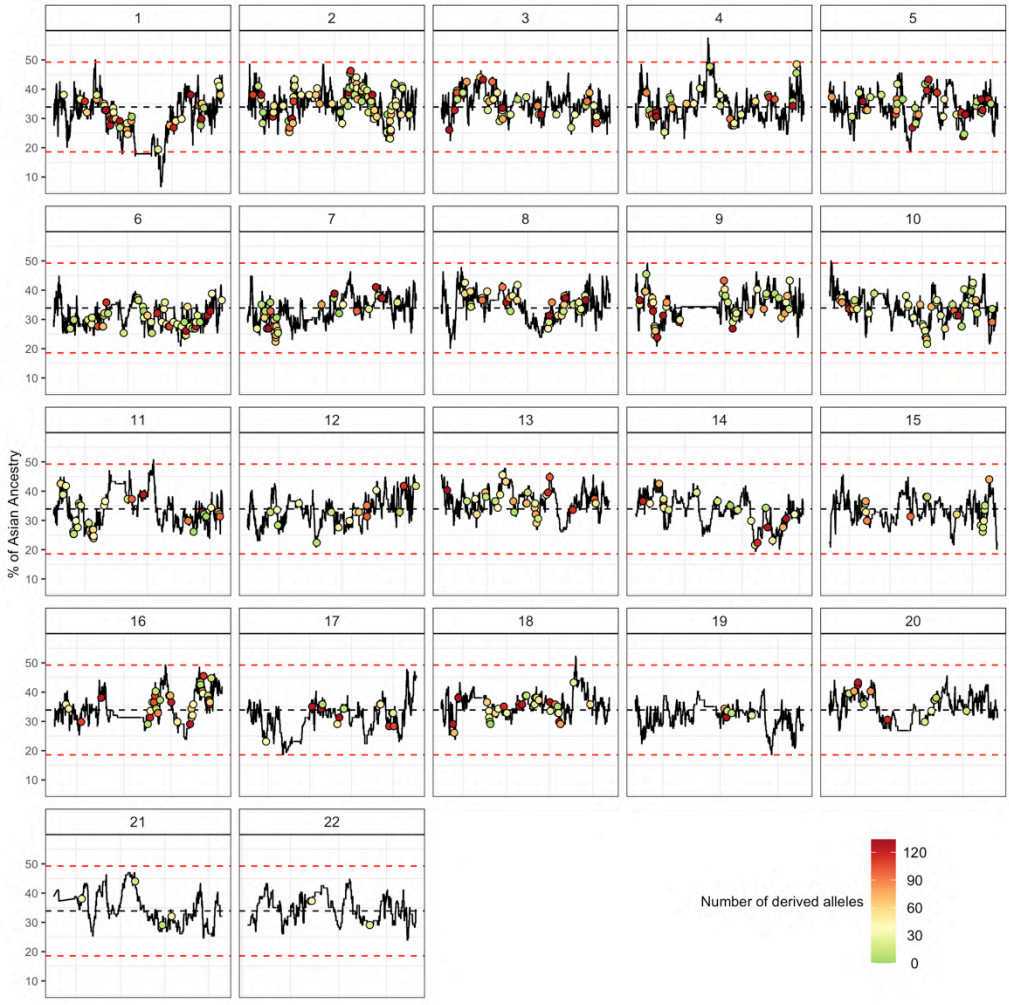


Figure 3. Genomic localization of variants with predicted extreme deleterious effects (GERP-RS ≥ 6). Each predicted deleterious mutation is represented as a circle, where the colours correspond to the number of derived alleles found in the Malagasy population (N=67). Black line corresponds to local Ancestry across 22 autosomal chromosomes, depicting for each chromosomal position (x-axis) the average Asian ancestry proportion across Malagasy individuals (y-axis). Black dotted horizontal line shows average Asian ancestry across the genome (mean=33.91%), and red dotted lines represent 3 standard deviations (sd=5.11%) from average Asian ancestry across the genome.

Consequences for disease mapping

From figure 3, we found that many mutations supposed to be deleterious presented a high frequency in Madagascar, which is suspicious given that these variants showed conservation status across different taxa. Therefore, we identified and characterized the 10 most frequent SNVs in Madagascar predicted to be of high impact (Table 1). We found that the majority of these sites were almost fixed in Madagascar and abroad, suggesting that these mutations, presumably common in all proxies, are not deleterious; they could cause a dramatic functional change in a protein, but not affect the fitness of the carrier; or, alternatively, they could be even beneficial in a particular context for Homo sapiens, since the high conservation status is computed across mammals and the derived allele almost reach fixation.

Indeed, variants that are shared across populations are also typically older (given the split between African and Out-Of-Africa populations ~60,000 years ago): if these mutations are benign or neutral, they can be maintained over long periods of time and in multiple populations (Henn et al., 2015). We therefore included an additional parameter, i.e. mutations relatively rare in the world (DAF <5%), but still present in Madagascar (Table 2). Rare variants tend to be geographically restricted, more recent and are more likely to occur at predicted functional sites (Henn et al., 2015). We found mutations localized in regions that map to genes involved in transcription factors and DNA reparation pathways. It is important to mention that we detected 346 SNVs with low frequencies in our dataset, which in addition are localized in conserved genomic regions across different taxa. One

important signal found in the 10 most frequent candidates was the genotype counts in the population, where some of these sites had no derived homozygous.

Table 1. Highly frequent extreme deleterious mutations in Malagasy genomes. GRCh37 genomic information depicts chromosome, position, rsID, GERP-RS, ancestral allele (AA), and derived/annotated allele (DA) of locus. Malagasy genetic diversity columns show results based on 67 genomes, showing the average Asian ancestry, number of derived homozygotes, number of derived alleles, and ancestry-specific count of derived alleles for each locus. Derived allele frequency is shown for 67 individuals sampled from 1,000 Genomes Panel and Malagasy populations. Functional annotation corresponds to database retrieved annotations for the gene associated to the SNV, showing the NCBI and KEGG identifier. The nomenclature, orthology, pathway, additional information and disease were annotated related to the gene from KEGG database.

GRCh37 genomic information						Malagasy genetic diversity					Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
8	93655802	rs115627342	6.06	C	G	0.3134	1	66	133	91	42	0.993	1	1	1	0.963	0.985	0.978							
9	3969142	rs141364611	6.17	A	G	0.3657	1	66	133	84	49	0.993	1	1	1	0.978	1	1	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
2	63220969	rs141864934	6.16	T	C	0.2985	1	66	133	93	40	0.993	1	1	1	0.963	0.978	0.993	ncbi-geneid:23301	hsa:23301	EHBP1, HPC12	EH domain-binding protein 1	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Hereditary prostate cancer
7	82059070	rs143388418	6.03	G	A	0.3881	1	66	133	81	52	0.993	1	1	1	0.985	0.985	0.985	ncbi-geneid:781	hsa:781	CACNA2D1, CACNA2	voltage-dependent calcium channel alpha-2/delta-1	Ion channels [BR:hsa04040]	Signal transduction, Endocrine system, Circulatory system, Cardiovascular disease, Protein families: signaling and cellular processes	
4	23830299	rs146691710	6.06	A	G	0.3209	1	66	133	90	43	0.993	1	1	1	0.993	0.978	0.993	ncbi-geneid:10891	hsa:10891	PPARGC1A, LEM6	alpha peroxisome proliferator-activated receptor gamma coactivator 1-alpha	Mitochondrial biogenesis [BR:hsa03029]	Signal transduction, Endocrine system, Aging, Environmental adaptation, Neurodegenerative disease, Endocrine and metabolic disease, Protein families: genetic information processing	
7	19761304	rs556495664	6.02	A	G	0.2687	1	66	133	97	36	0.993	1	1	1	1	1	1	ncbi-geneid:256130	hsa:256130	TMEM196	transmembrane protein 196		Poorly characterized	
6	98298715	rs606090715	6.08	C	A	0.291	1	66	133	94	39	0.993	1	1	1	0.948	0.918	0.948							
9	14096784	rs78364791	6.08	T	C	0.3657	1	66	133	85	48	0.993	0.955	0.985	0.985	0.993	1	0.985	ncbi-geneid:4781	hsa:4781	NFIB, CTF	nuclear factor I/B	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	
1	77088902	rs78888615	6.07	A	G	0.3284	1	66	133	90	43	0.993	1	1	1	0.993	0.985	0.985	ncbi-geneid:256435	hsa:256435	ST6GALNAC3, PRO7177	N-acetylgalactosaminide alpha-2,6-sialyltransferase (sialyltransferase 7C) [EC:2.4.3.7]	Enzymes [BR:hsa01000], Glycosyltransferases [BR:hsa01003]	Glycan biosynthesis and metabolism, Protein families: metabolism, Glycosyltransferases	
8	124988166	rs80140850	6.17	T	C	0.3657	1	66	133	85	48	0.993	1	1	1	0.993	0.97	0.97	ncbi-geneid:654463	hsa:654463	FER1L6, C8ORFK23	fer-1-like protein 6	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	

Table 2. Variants present in Malagasy genomes and abroad (DAF < 0.05) predicted to have a large and extreme deleterious effect. Legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
5	78964729	rs10514164	4.32	T	G	0.3731	11	0	11	6	5	0.082	0.03	0.037	0.045	0.022	0.03	0.045	ncbi-geneid:167153	hsa:167153	TENT2, APD4	poly(A) RNA polymerase GLD2 [EC:2.7.7.19]	Enzymes [BR:hsa01000], Messenger RNA biogenesis [BR:hsa03019]	Protein families: genetic information processing, Transferring phosphorus-containing groups	
1	2444470	rs114254488	5.59	C	T	0.2985	10	0	10	10	0	0.075	0.015	0.03	0.037	0.007	0.022	0.022	ncbi-geneid:55229	hsa:55229	PANK4, CTRCT49	bifunctional damage-control phosphatase, subfamily II, fusion protein		Unclassified: metabolism	Cataract
5	106016334	rs12514358	6.06	C	T	0.4254	9	0	9	3	6	0.067	0.037	0.045	0.037	0.015	0.007	0.007							
16	7223641	rs17142766	5.73	A	G	0.3955	12	0	12	7	5	0.09	0.03	0.037	0.007	0.022	0.007	0.015	ncbi-geneid:54715	hsa:54715	RBFOX1, 2BP1	RNA binding protein fox-1	Spliceosome [BR:hsa03041]	Protein families: genetic information processing	
3	114379193	rs17681207	6.16	C	T	0.3134	5	0	5	4	1	0.037	0.03	0.045	0.007	0.037	0.037	0.022	ncbi-geneid:26137	hsa:26137	ZBTB20, DPZF	zinc finger and BTB domain-containing protein 20	Transcription factors [BR:hsa03000], Ubiquitin system [BR:hsa04121]	Protein families: genetic information processing	Primrose syndrome
5	137031775	rs2905612	4.16	C	T	0.2612	14	0	14	7	7	0.104	0.007	0.007	0.015	0.045	0.007	0.022	ncbi-geneid:26249	hsa:26249	KLHL3, PHA2D	kelch-like protein 2/3	Ubiquitin system [BR:hsa04121]	Protein families: genetic information processing	Hyperkalemic distal renal tubular acidosis (RTA type 4),Renal tubular acidosis
2	146219068	rs56673819	6.02	T	C	0.403	5	0	5	1	4	0.037	0.045	0.015	0.015	0.007	0.037	0.037							
6	38729511	rs61748600	6	T	C	0.306	5	0	5	2	3	0.037	0.03	0.015	0.015	0.045	0.037	0.045	ncbi-geneid:1769	hsa:1769	DNAH8, ATPase	dynein axonemal heavy chain	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Neurodegenerative disease, Protein families: signaling and cellular processes	Spermatogenic failure
1	107289830	rs61788201	4.76	A	T	0.2537	8	1	10	10	0	0.075	0.007	0.022	0.007	0.022	0.015	0.015							
22	33199362	rs62232903	4.62	G	A	0.3507	16	2	20	10	10	0.149	0.03	0.015	0.03	0.037	0.022	0.007	ncbi-geneid:8224	hsa:8224	SYN3	synapsin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
22	33199362	rs62232903	4.62	G	A	0.3507	16	2	20	10	10	0.149	0.03	0.015	0.03	0.037	0.022	0.007	ncbi-geneid:7078	hsa:7078	TIMP3, HSMRK222	metalloproteinase inhibitor 3	Peptidases and inhibitors [BR:hsa01002]	Cancer: overview, Protein families: metabolism	Sorsby fundus dystrophy
11	103624943	rs71482405	5.06	A	G	0.3284	12	0	12	7	5	0.09	0.022	0.045	0.015	0.022	0.045	0.022							
16	50263762	rs7203951	6.05	A	G	0.291	8	0	8	4	4	0.06	0.015	0.045	0.015	0.03	0.015	0.007	ncbi-geneid:64282	hsa:64282	TENT4B, PAPD5	non-canonical poly(A) RNA polymerase PAPD5/7 [EC:2.7.7.19]	Enzymes [BR:hsa01000], Messenger RNA biogenesis [BR:hsa03019], Transfer RNA biogenesis [BR:hsa03016]	Folding, sorting and degradation, Protein families: genetic information processing, Transferring phosphorus-containing groups	
8	21984765	rs73549523	5.17	C	T	0.3284	13	0	13	6	7	0.097	0.007	0.007	0.037	0.03	0.022	0.007	ncbi-geneid:55806	hsa:55806	HR, ALUNC	[histone H3]-dimethyl-L-lysine9 demethylase [EC:1.14.11.65]	Enzymes [BR:hsa01000], Chromosome and associated proteins [BR:hsa03036]	Protein families: genetic information processing, Acting on paired donors, with incorporation or reduction of molecular oxygen	Atrichia with papular lesions,Hypotrichosis,Marie-Unna hereditary hypotrichosis,Alopecia universalis
15	93536173	rs76621355	5.25	C	T	0.3582	12	0	12	1	11	0.09	0.015	0.022	0.015	0.022	0.045	0.022	ncbi-geneid:1106	hsa:1106	CHD2, DEE94	chromodomain-helicase-DNA-binding protein 2 [EC:5.6.2.-]	Enzymes [BR:hsa01000], Chromosome and associated proteins [BR:hsa03036]	Protein families: genetic information processing, Isomerases altering macromolecular conformation	Early infantile epileptic encephalopathy,Epileptic encephalopathy, childhood-onset
2	99483990	rs77944740	4.05	A	C	0.3881	9	1	11	1	10	0.082	0.03	0.045	0.045	0.03	0.007	0.007	ncbi-geneid:343990	hsa:343990	CRACDL, C2orf55				

Therefore, we looked for departures of Hardy Weinberg equilibrium, as SNVs showing extreme values implies a difference between expected and observed genotypes frequencies (Figure 4). From these analyses, we show the 15 SNVs with the lowest p-value (Table 3), which indicates sites that do not respect Hardy-Weinberg equilibrium. Interestingly, we found two mutations in chromosome 6 that are only present in heterozygous state in the Malagasy population, with the allele frequencies around 0.5 for Malagasy and proxy population. As in Malagasy, the 1000 genomes reference population presented almost only heterozygous genotypes for these alleles. These could indicate a genotyping error or a complex region in the genome where machine sequencing can be outperformed. Previous studies have signalled that PRIM2 are among the most difficult genes to reconstruct using de novo assembly from short sequence reads due to the presence of duplicated regions along the genome (three partial copies of the PRIM2 gene in chromosome 6 and 3) (Alkan et al., 2011; Genovese et al., 2013).

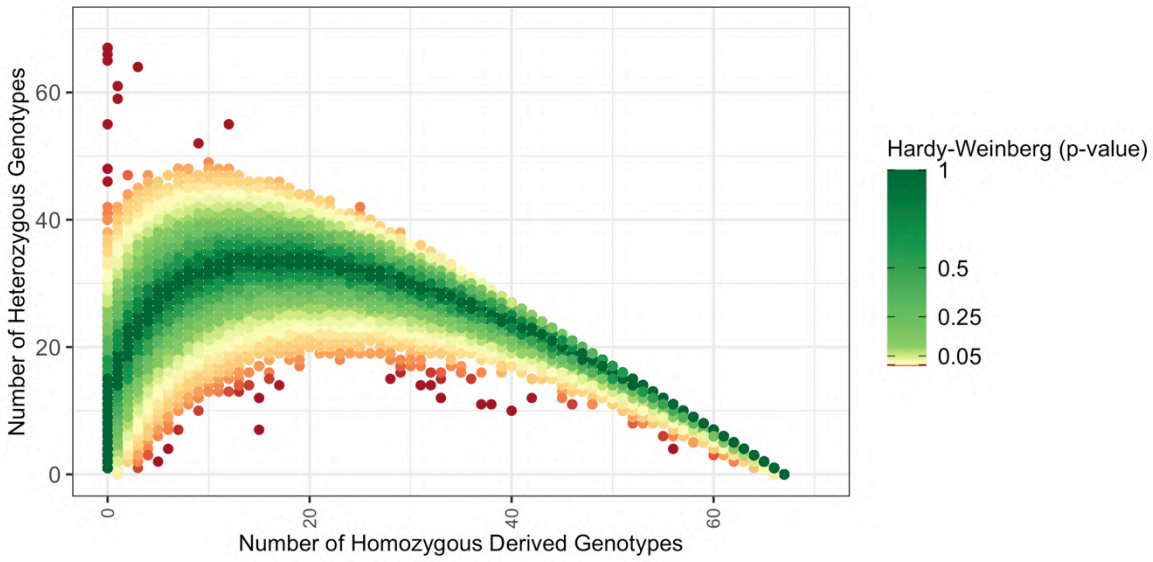


Figure 4. Hardy-Weinberg Equilibrium and deleterious variants We compared the observed number of heterozygous (y-axis) and derived homozygous (x-axis) genotypes in Madagascar. Each variant is plotted as a point, with colour reflecting HWE p-values estimated with VCFtools (Danecek et al., 2011). Specifically, variants in the upper left zone resent an excess of heterozygotes compared to expected values.

Furthermore, we detected a signal on chromosome 13 showing a p-value of $7.24x10^{-8}$, where the observed heterozygotes show 19 more genotypes than expected. These SNVs mapped to the region 39917029 - 40177330, where the ENTREZ database recovered the annotated gene *Lhfpl6*. The *Lhfp* gene is localized on the long arm of chromosome 13, a region recurrently targeted by chromosomal aberrations in lipomas (Petit et al., 1999). More recently, a co-expression network analysis revealed that *Lhfp* was strongly connected to genes involved in osteoblast differentiation (Mesner et al., 2019), identifying *Lhfp* as a regulator of osteoblast activity and bone mass in mice. As these genomic regions are conserved across mammals, we can imagine that *Lhfp* locus in humans might be important, as in mice, for osteoblast function and bone mineral density. Alternatively, the variant described could have fluctuated in allele frequency for genetic drift influence.

Table 3. Variants with predicted deleterious effects that significantly deviate from Hardy-Weinberg equilibrium. Legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)						Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	HWE	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
6	257186	rs10111327	2.11	T	C	0.2985	1.28E-12	61	1	63	46	17	0.47	0.5	0.59	0.5	0.418	0.455	0.448							
6	32487462	rs115001626	2.61	G	T	0.2985	3.99E-09	7	15	37	29	8	0.276	0.448	0.373	0.187	0.194	0.231	0.284	ncbi-geneid:3127	hsa:3127	HLA-DRB5, HLA-DRB5*	MHC class II antigen	Exosome [BR:hsa04147]	Signaling molecules and interaction, Transport and catabolism, Immune system, Infectious disease: viral, Infectious disease: bacterial, Infectious disease: parasitic, Immune disease, Cardiovascular disease, Endocrine and metabolic disease, Protein families: signaling and cellular processes	Primary central nervous system lymphoma
9	29095016	rs1342251	2.66	T	A	0.3433	1.59E-07	12	33	78	58	20	0.582	0.784	0.836	0.776	0.642	0.537	0.687							
6	294268	rs15111159	2.78	T	C	0.2985	4.58E-11	59	1	61	44	17	0.455	0.53	0.575	0.545	0.507	0.463	0.433	ncbi-geneid:56940	hsa:56940	DUSP22, JKAP	atypical dual specificity phosphatase [EC:3.1.3.16 3.1.3.48]	Enzymes [BR:hsa01000], Protein phosphatases and associated proteins [BR:hsa01009]	Protein families: metabolism, Acting on ester bonds	Anaplastic large-cell lymphoma,Peripheral T cell lymphoma
1	61288285	rs1996168	2.92	C	T	0.4776	2.35E-09	55	0	55	32	23	0.41	0.209	0.246	0.254	0.455	0.44	0.47							
8	11247075	rs2572442	2.26	A	T	0.291	7.44E-07	4	56	116	77	39	0.866	1	1	1	0.552	0.731	0.627							
6	57290485	rs35444204	3.66	T	C	0.3358	9.85E-20	67	0	67	45	22	0.5	0.515	0.493	0.493	0.5	0.493	0.522	ncbi-geneid:5558	hsa:5558	PRIM2, PRIM2A	DNA primase large subunit	DNA replication proteins [BR:hsa03032]	Replication and repair, Protein families: genetic information processing	
6	349343	rs3778605	2.27	T	A	0.2985	6.97E-16	64	3	70	49	21	0.522	0.515	0.478	0.507	0.507	0.507	0.522	ncbi-geneid:56940	hsa:56940	DUSP22, JKAP	atypical dual specificity phosphatase [EC:3.1.3.16 3.1.3.48]	Enzymes [BR:hsa01000], Protein phosphatases and associated proteins [BR:hsa01009]	Protein families: metabolism, Acting on ester bonds	Anaplastic large-cell lymphoma,Peripheral T cell lymphoma
1	61288266	rs3936928	2.12	A	G	0.4776	2.35E-09	55	12	79	37	42	0.59	0.791	0.769	0.746	0.552	0.56	0.53							
13	39934649	rs55958075	2.01	C	T	0.3433	7.25E-08	10	40	90	57	33	0.672	0.903	0.948	0.881	0.828	0.866	0.91	ncbi-geneid:10186	hsa:10186	LHFPL6, LHFP	LHFPL tetraspan subfamily member protein	Transporters [BR:hsa02000], Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	
8	25068604	rs60821775	2.37	G	C	0.3731	9.97E-08	11	37	85	50	35	0.634	0.903	0.94	0.963	0.373	0.336	0.336	ncbi-geneid:80005	hsa:80005	DOCK5	dedicator of cytokinesis protein 5	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
6	22053672	rs6927366	2.1	A	G	0.2836	1.59E-07	11	38	87	62	25	0.649	0.172	0.209	0.201	0.679	0.493	0.604							
20	52480410	rs73912566	2.61	T	C	0.3955	3.36E-18	66	0	66	41	25	0.493	0.328	0.358	0.358	0.478	0.5	0.485							
6	57295212	rs75428242	2.16	A	T	0.3358	9.85E-20	67	0	67	43	24	0.5	0.522	0.5	0.5	0.53	0.5	0.493	ncbi-geneid:5558	hsa:5558	PRIM2, PRIM2A	DNA primase large subunit	DNA replication proteins [BR:hsa03032]	Replication and repair, Protein families: genetic information processing	
10	47107666	rs9423080	2.93	G	T	0.3433	5.84E-17	65	0	65	40	25	0.485	0.47	0.455	0.44	0.448	0.47	0.47							

We also looked if it was possible to detect mutations that could have been acting on the same biological function (i.e. epistasis). Thus, we sought to identify highly differentiated variants (HDVs) showing substantially different allele frequencies (> 40%) between African and Asian proxies populations (Choudhury et al., 2021). We considered sites with scores GERP-RS $\geq$ 2) and recovered 20,247 variants. We added a last filter, looking for sites presenting 0 homozygous derived genotypes in Madagascar (252 SNVs detected) and for which KEGG pathway data was available (Kanehisa et al., 2016). We decided to look at the pathways presenting the highest number of genes touched by these mutations (SNVs > 5). We summarized this information in table 4, where we found 37

mutations in different chromosomes accounting for genes involved in enzymes, transcription factors, membrane trafficking and domain-containing proteins not elsewhere classified.

Table 4. Variants showing highly differentiated allele frequencies between African and Asian populations. Legend is the same described in Table 1.

GRCh37 genomic information							Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene					
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
15	71654056	rs10775203	2.3	T	C	0.3209	23	0	23	7	16	0.172	0.522	0.507	0.537	0.03	0.052	0.022	ncbi-geneid:79875	hsa:79875	THSD4, AAT12	thrombospondin type-1 domain-containing protein 4	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Protein families: signaling and cellular processes	Familial thoracic aortic aneurysm and dissection
21	32659284	rs11702036	2.38	T	C	0.291	20	0	20	8	12	0.149	0.448	0.515	0.478	0.03	0.037	0.022	ncbi-geneid:7074	hsa:7074	TIAM1, NEDLDS	T-lymphoma invasion and metastasis-inducing protein 1	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Cellular community - eukaryotes, Cell motility, Immune system, Cancer: overview, Protein families: signaling and cellular processes	
7	147677544	rs12532921	2.07	C	A	0.2985	17	0	17	4	13	0.127	0.597	0.552	0.56	0.045	0.007	0.037	ncbi-geneid:26047	hsa:26047	CNTNAP2, AUTS15	contactin associated protein-like 2	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signaling molecules and interaction, Protein families: signaling and cellular processes	Pitt-Hopkins syndrome,Autism
11	21178116	rs16907565	2.03	A	T	0.3731	22	0	22	5	17	0.164	0.515	0.478	0.478	0.075	0.097	0.112	ncbi-geneid:4745	hsa:4745	NELL1, IDIUG8,	protein kinase C-binding protein NELL	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
19	4312103	rs3745989	2.84	C	A	0.2985	18	0	18	6	12	0.134	0.53	0.552	0.612	0.09	0.037	0.075	ncbi-geneid:79187	hsa:79187	FSD1, GLFND	fibronectin type III and SPRY domain-containing protein 1	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
15	71642403	rs3784384	3.86	C	A	0.3209	23	0	23	7	16	0.172	0.537	0.485	0.537	0.022	0.045	0.022	ncbi-geneid:79875	hsa:79875	THSD4, AAT12	thrombospondin type-1 domain-containing protein 4	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Protein families: signaling and cellular processes	Familial thoracic aortic aneurysm and dissection
3	185484257	rs4686388	3.77	A	G	0.3284	27	0	27	10	17	0.201	0.612	0.612	0.59	0.067	0.045	0.112	ncbi-geneid:10644	hsa:10644	IGF2BP2, IMP-2	insulin-like growth factor 2 mRNA-binding protein 2	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	Type 2 diabetes mellitus
4	21315062	rs6812694	2.87	C	T	0.2836	16	0	16	8	8	0.119	0.448	0.463	0.448	0.03	0.022	0.045	ncbi-geneid:80333	hsa:80333	KCNIP4, CALP	Kv channel-interacting protein	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
18	59763727	rs10432259	2.97	T	G	0.291	23	0	23	11	12	0.172	0.582	0.575	0.619	0.112	0.082	0.067	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
18	59791520	rs10503065	2.16	A	G	0.291	23	0	23	9	14	0.172	0.537	0.552	0.582	0.119	0.082	0.06	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
18	59779231	rs12607624	3.17	G	A	0.291	23	0	23	10	13	0.172	0.59	0.575	0.619	0.112	0.075	0.06	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
4	47589203	rs1316874	3.13	G	A	0.2985	11	0	11	2	9	0.082	0.56	0.478	0.478	0.03	0.007	0.037	ncbi-geneid:57205	hsa:57205	ATP10D, ATPVD	phospholipid-translocating ATPase [EC:7.6.2.1]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Catalysing the translocation of other compounds	
10	90216485	rs2312610	3.63	C	T	0.3582	25	0	25	7	18	0.187	0.515	0.5	0.507	0.03	0.052	0.06	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
10	90257225	rs2576171	2.87	T	C	0.3582	22	0	22	6	16	0.164	0.448	0.448	0.455	0.007	0.022	0.03	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
14	74538706	rs4646861	5.28	A	G	0.3806	16	0	16	3	13	0.119	0.582	0.567	0.5	0.037	0.007	0.007	ncbi-geneid:4329	hsa:4329	ALDH6A1, MMSADHA	malonate-semialdehyde dehydrogenase (acetylating) / methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.18 1.2.1.27]	Enzymes [BR:hsa01000]	Carbohydrate metabolism, Amino acid metabolism, Metabolism of other amino acids, Acting on the aldehyde or oxo group of donors	Methylmalonate semialdehyde dehydrogenase deficiency
10	90216054	rs4934409	2.63	T	A	0.3582	25	0	25	7	18	0.187	0.515	0.5	0.507	0.03	0.052	0.06	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
2	64804495	rs10180097	2.89	A	G	0.3358	13	0	13	3	10	0.097	0.493	0.53	0.478	0.045	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
3	32024251	rs12629793	2.18	C	G	0.2985	23	0	23	7	16	0.172	0.44	0.522	0.463	0.007	0.015	0.037	ncbi-geneid:114884	hsa:114884	OSBPL10, ORP10	oxysterol-binding protein-related protein 9/10/11	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64803774	rs12713520	2.38	A	G	0.3358	13	0	13	3	10	0.097	0.493	0.53	0.478	0.06	0.052	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64773779	rs1978404	2.12	A	G	0.3358	12	0	12	4	8	0.09	0.478	0.53	0.478	0.045	0.06	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
21	35934476	rs2251817	3.31	C	T	0.3134	22	0	22	6	16	0.164	0.627	0.597	0.612	0.067	0.104	0.104	ncbi-geneid:1827	hsa:1827	RCAN1, ADAPT78	calcipressin-1	Membrane trafficking [BR:hsa04131]	Endocrine system, Infectious disease: viral, Protein families: genetic information processing	Down syndrome
2	64783870	rs7562829	2.14	T	G	0.3358	13	0	13	4	9	0.097	0.493	0.522	0.478	0.052	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64752872	rs7585497	2.03	G	A	0.3358	12	0	12	3	9	0.09	0.47	0.522	0.463	0.045	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	144400860	rs78141710	2.42	C	A	0.3657	18	0	18	4	14	0.134	0.448	0.381	0.418	0.015	0.082	0.022	ncbi-geneid:55843	hsa:55843	ARHGAP15, BM046	Rho GTPase-activating protein 15	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
21	35938024	rs8130808	2.29	G	T	0.3134	22	0	22	6	16	0.164	0.604	0.582	0.612	0.067	0.104	0.104	ncbi-geneid:1827	hsa:1827	RCAN1, ADAPT78	calcipressin-1	Membrane trafficking [BR:hsa04131]	Endocrine system, Infectious disease: viral, Protein families: genetic information processing	Down syndrome
13	24701318	rs9553191	2.59	A	G	0.3731	22	0	22	11	11	0.164	0.463	0.396	0.433	0.03	0.045	0.037	ncbi-geneid:221178	hsa:221178	SPATA13, ARHGEF29	Rho guanine nucleotide exchange factor 4/29	Membrane trafficking [BR:hsa04131]	Cell motility, Protein families: genetic information processing	
14	65555471	rs10143198	2.55	T	C	0.3284	22	0	22	6	16	0.164	0.649	0.604	0.575	0.097	0.06	0.097	ncbi-geneid:4149	hsa:4149	MAX, bHLHd4	Max protein	Transcription factors [BR:hsa03000]	Signal transduction, Cancer: overview, Cancer: specific types, Protein families: genetic information processing	Malignant paraganglioma
18	53370201	rs1037430	2.1	T	C	0.3358	16	0	16	7	9	0.119	0.425	0.41	0.493	0.03	0.015	0.015	ncbi-geneid:6925	hsa:6925	TCF4, CDG2T	transcription factor 4/12	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Pitt-Hopkins syndrome,Fuchs corneal dystrophy
3	32024251	rs12629795	2.18	C	G	0.2985	23	0	23	7	16	0.172	0.44	0.522	0.463	0.007	0.015	0.037	ncbi-geneid:344787	hsa:344787	ZNF860,NA	KRAB domain-containing zinc finger protein	Transcription factors [BR:hsa03000]	Infectious disease: viral, Protein families: genetic information processing	
10	3821561	rs17731	3.18	G	A	0.403	17	0	17	3	14	0.127	0.493	0.44	0.47	0.03	0.022	0.015	ncbi-geneid:1316	hsa:1316	KLF6, BCD1	krueppel-like factor 6/7	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Prostate cancer
7	42067356	rs2237425	5.15	C	G	0.2463	13	0	13	7	6	0.097	0.41	0.328	0.418	0.007	0.007	0.007	ncbi-geneid:2737	hsa:2737	GLI3, ACLS	zinc finger protein GLI3	Transcription factors [BR:hsa03000]	Signal transduction, Cancer: overview, Cancer: specific types, Protein families: genetic information processing	Pallister-Hall syndrome,Polysyndactyly,Postaxial polydactyly,Greig cephalopolysyndactyly syndrome,Preaxial polydactyly
6	45545074	rs3899143	2.72	C	T	0.306	15	0	15	4	11	0.112	0.463	0.418	0.425	0.015	0.015	0.007	ncbi-geneid:860	hsa:860	RUNX2, AML3	runt-related transcription factor 2	Transcription factors [BR:hsa03000]	Endocrine system, Cancer: overview, Protein families: genetic information processing	Cleidocranial dysplasia
9	4169816	rs4740753	2.16	T	C	0.3209	26	0	26	10	16	0.194	0.664	0.694	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4170020	rs4740755	2.55	A	C	0.3209	25	0	25	9	16	0.187	0.664	0.687	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4169621	rs7045178	3.48	T	A	0.3209	26	0	26	10	16	0.194	0.664	0.687	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4170457	rs7853070	2.04	A	G	0.3209	26	0	26	9	17	0.194	0.664	0.687	0.627	0.067	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
19	51829903	rs924150	3.3	A	C	0.3358	20	0	20	7	13	0.149	0.485	0.463	0.53	0.06	0.022	0.045	ncbi-geneid:57616	hsa:57616	TSHZ3, TSH3	teashirt	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	

In order to identify the possible demographic impact on function/disease mapping, we searched SNVs annotated as deleterious by two other approaches in complementarity with GERP-RS score: snpEFF (Cingolani et al., 2012) and aLoFT (Balasubramanian et al., 2017) software annotation. These programs predict the result in the loss of function (LoF) of human genes, providing annotations for putative protein-damaging variants. We detected 1,673 and 1,310 SNVs with snpEFF and aLoFT, respectively. We searched for SNVs predicted to be deleterious by evolutionary conservation state and LOF prediction methods, detecting a total of 27 SNVs. We present this information in Table 5. Importantly, we detected 7 sites for which the allele annotated for the predicted deleterious effect does not correspond between these annotation methods. This could mean an artefact of methods or that the ancestral allele could be involved in an effect on function. The rest of sites

(SNVs=20) could be considered as potential signals for evaluating functional roles of genes and variants.

Table 5. Potentially deleterious variants based on three different annotation software. GRCh37 genomic information depicts chromosome, position, rsID, GERP-RS, annotated effect allele (EA) according to snpEFF, ancestral allele (AA), and derived allele (DA) of locus. The rest of the legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene					
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGV	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
8	124154697	rs10101626	5.36	G	T	0.3806	28	2	32	24	8	0.239	0.209	0.142	0.112	0.343	0.321	0.261	ncbi-geneid:93594	hsa:93594	TBC1D31, Gm85				
19	1534042	rs10853954	2.29	T	C	0.291	37	10	57	27	30	0.425	0.575	0.575	0.694	0.269	0.373	0.313	ncbi-geneid:126520	hsa:126520	PLK5, PLK-5	inactive serine/threonine-protein kinase PLK5	DNA repair and recombination proteins [BR:hsa03400]	Protein families: genetic information processing	
8	143922642	rs114099899	2.55	A	G	0.4328	2	65	132	74	58	0.985	1	1	1	0.97	0.978	0.97	ncbi-geneid:2765	hsa:2765	GML, LY6DL				
14	57948380	rs1152522	2.75	T	C	0.2612	29	12	53	23	30	0.396	0.91	0.993	0.978	0.187	0.172	0.209	ncbi-geneid:55195	hsa:55195	CCDC198, C14orf105				
1	40773150	rs12077871	3.92	G	A	0.3358	14	0	14	5	9	0.104	0.119	0.09	0.082	0.127	0.104	0.097	ncbi-geneid:1298	hsa:1298	COL9A2, DJ396222.4	collagen type IX alpha	Proteoglycans [BR:hsa00535]	Signal transduction, Signaling molecules and interaction, Cellular community - eukaryotes, Digestive system, Infectious disease: viral, Protein families: signaling and cellular processes	Multiple epiphyseal dysplasia,Stickler syndrome
1	5935162	rs1287637	4.93	T	A	0.2985	11	2	15	13	2	0.112	0.224	0.231	0.276	0.127	0.097	0.067	ncbi-geneid:261734	hsa:261734	NPHP4, POC10	nephrocystin-4	Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	Nephronophthisis,Senior-Loken syndrome
14	24470138	rs1811890	2.84	C	T	0.3881	19	3	25	10	15	0.187	0.269	0.209	0.134	0.142	0.022	0.03	ncbi-geneid:10901	hsa:10901	DHRS4, CR	dehydrogenase/reductase SDR family member 4 [EC:1.1.-.-]	Enzymes [BR:hsa01000]	Metabolism of cofactors and vitamins, Transport and catabolism, Acting on the CH-OH group of donors	
14	24470138	rs1811890	2.84	C	T	0.3881	19	3	25	10	15	0.187	0.269	0.209	0.134	0.142	0.022	0.03	ncbi-geneid:317749	hsa:317749	DHRS4L2, SDR25C3	dehydrogenase/reductase SDR family member 4-like protein 2 [EC:1.1.-.-]	Enzymes [BR:hsa01000]	Metabolism of cofactors and vitamins, Acting on the CH-OH group of donors	
18	25616451	rs1944294	3.2	A	T	0.3134	22	5	32	13	19	0.239	0.254	0.351	0.321	0.037	0.067	0.082	ncbi-geneid:1000	hsa:1000	CDH2, ACOGS	cadherin 2, type 1, N-cadherin	Chromosome and associated proteins [BR:hsa03036], Cell adhesion molecules [BR:hsa04515], CD molecules [BR:hsa04090]	Signaling molecules and interaction, Cardiovascular disease, Protein families: genetic information processing, Protein families: signaling and cellular processes, (CDH)	Arrhythmogenic right ventricular cardiomyopathy,Agenesis of corpus callosum, cardiac, ocular, and genital syndrome
1	179858444	rs2245425	5.26	G	A	0.291	30	23	76	46	30	0.567	0.776	0.731	0.761	0.328	0.425	0.507	ncbi-geneid:26092	hsa:26092	TOR1AIP1, LAP1	torsin-1A-interacting protein	Chaperones and folding catalysts [BR:hsa03110]	Protein families: genetic information processing	Limb-girdle muscular dystrophy
1	236706300	rs2273865	2.59	T	A	0.3507	16	3	22	17	5	0.164	0.149	0.201	0.216	0.172	0.216	0.261	ncbi-geneid:3964	hsa:3964	LGALS8, Gal-8	galectin-8	Membrane trafficking [BR:hsa04131], Lectins [BR:hsa04091]	Protein families: genetic information processing, Protein families: signaling and cellular processes, (LGALS)	
4	111542154	rs2278782	2.67	G	A	0.2761	15	2	19	14	5	0.142	0.172	0.104	0.201	0.112	0.179	0.127	ncbi-geneid:5308	hsa:5308	PITX2, ARP1	paired-like homeodomain transcription factor 2	Transcription factors [BR:hsa03000]	Signal transduction, Protein families: genetic information processing	Axenfeld-Rieger syndrome,Ring dermoid of cornea,Anterior segment dysgenesis
14	88862529	rs3179969	2.45	A	G	0.2761	36	20	76	58	18	0.567	0.664	0.537	0.59	0.343	0.485	0.425	ncbi-geneid:55812	hsa:55812	SPATA7, HEL-S-296	spermatogenesis-associated protein 7	Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	Leber congenital amaurosis
5	74965122	rs34358	4.88	G	A	0.4104	34	11	56	37	19	0.418	0.396	0.455	0.418	0.597	0.515	0.493	ncbi-geneid:728780	hsa:728780	ANKDD1B,NA				
8	99205612	rs3735887	3.65	T	C	0.291	27	19	65	48	17	0.485	0.485	0.328	0.433	0.522	0.545	0.493	ncbi-geneid:79815	hsa:79815	NIPAL2, NPAL2	magnesium transporter	Transporters [BR:hsa02000]	Protein families: signaling and cellular processes	
19	51535130	rs3745540	4.27	A	G	0.3209	34	9	52	26	26	0.388	0.567	0.657	0.687	0.358	0.306	0.321	ncbi-geneid:43849	hsa:43849	KLK12, KLK-L5	kallikrein 12 [EC:3.4.21.-]	Enzymes [BR:hsa01000], Peptidases and inhibitors [BR:hsa01002]	Protein families: metabolism, Acting on peptide bonds (peptidases)	
15	31294714	rs3784589	5.05	C	A	0.2985	4	0	4	3	1	0.03	0.007	0.007	0.007	0.112	0.045	0.037	ncbi-geneid:4308	hsa:4308	TRPM1, CSNB1C	transient receptor potential cation channel subfamily M member 1	Ion channels [BR:hsa04040]	Protein families: signaling and cellular processes	Congenital stationary night blindness,Chromosome 15q13.3 microdeletion syndrome
1	228469903	rs3795786	4.21	A	T	0.3134	12	1	14	3	11	0.104	0.269	0.313	0.254	0.022	0.007	0.037	ncbi-geneid:84033	hsa:84033	OBSCN, ARHGEF30	obscurin-RhoGEF [EC:2.7.11.1]	Enzymes [BR:hsa01000], Protein kinases [BR:hsa01001], Membrane trafficking [BR:hsa04131]	Protein families: metabolism, Protein families: genetic information processing, Transferring phosphorus-containing groups	
1	67242087	rs3816989	5.92	G	A	0.3731	14	2	18	12	6	0.134	0.112	0.157	0.134	0.097	0.052	0.037	ncbi-geneid:200132	hsa:200132	DYNLT5, TCTEX1D1	dynein light chain Tctex-type 5	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Protein families: signaling and cellular processes	
7	156468559	rs3823617	5.22	T	C	0.4179	10	0	10	9	1	0.075	0.127	0.112	0.097	0.097	0.149	0.127	ncbi-geneid:140545	hsa:140545	RNF32, FKSG33				
7	156468559	rs3823617	5.22	T	C	0.4179	10	0	10	9	1	0.075	0.127	0.112	0.097	0.097	0.149	0.127	ncbi-geneid:64327	hsa:64327	LMBR1, ACHP	limb region 1 protein	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Acheiropodia,Triphalangeal thumb-polysyndactyly syndrome,Syndactyly,Preaxial polydactyly,Laurin-Sandrow syndrome
1	247719769	rs56043070	3	G	A	0.4254	2	0	2	2	0	0.015	0.022	0.015	0.022	0.022	0.045	0.022	ncbi-geneid:148823	hsa:148823	GCSAML, C1orf150				
1	225285200	rs5744168	4.67	G	A	0.3507	2	0	2	2	0	0.015	0.045	0.045	0.06	0.022	0.015	0.03	ncbi-geneid:7100	hsa:7100	TLR5, MELIOS	toll-like receptor 5	Pattern recognition receptors [BR:hsa04054]	Immune system, Infectious disease: bacterial, Immune disease, Protein families: signaling and cellular processes	Systemic lupus erythematosus
15	55722882	rs57809907	4.68	C	A	0.403	25	6	37	34	3	0.276	0.007	0.022	0.007	0.463	0.396	0.619	ncbi-geneid:161582	hsa:161582	DNAAF4, CILD25	dynein axonemal assembly factor 4	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Protein families: signaling and cellular processes	Primary ciliary dyskinesia,Dyslexia
1	47080679	rs6671527	2.2	A	G	0.3582	30	3	36	31	5	0.269	0.06	0.075	0.075	0.44	0.5	0.448	ncbi-geneid:8569	hsa:8569	MKNK1, MNK1	MAP kinase interacting serine/threonine kinase [EC:2.7.11.1]	Enzymes [BR:hsa01000], Protein kinases [BR:hsa01001]	Signal transduction, Endocrine system, Protein families: metabolism, Transferring phosphorus-containing groups	
1	47080679	rs6671527	2.2	A	G	0.3582	30	3	36	31	5	0.269	0.06	0.075	0.075	0.44	0.5	0.448	ncbi-geneid:148932	hsa:148932	MOB3C, MOB1E				
8	100133706	rs7460625	3.32	G	T	0.291	27	8	43	30	13	0.321	0.5	0.463	0.455	0.396	0.187	0.313	ncbi-geneid:157680	hsa:157680	VPS13B, BLTP3B	vacuolar protein sorting-associated protein 13B	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Cohen syndrome
10	90354423	rs74777687	3.87	G	C	0.3582	8	0	8	6	2	0.06	0.007	0.022	0.037	0.037	0.067	0.03	ncbi-geneid:142910	hsa:142910	LIP3, LIPL1	lipase member J [EC:3.1.1.-]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on ester bonds	
1	158549492	rs863362	4.18	C	T	0.0821	33	24	81	73	8	0.604	0.634	0.515	0.604	0.552	0.627	0.604	ncbi-geneid:128367	hsa:128367	OR10X1, OR1-13	olfactory receptor	G protein-coupled receptors [BR:hsa04030]	Sensory system, Protein families: signaling and cellular processes	
21	31744127	rs877346	2.31	A	T	0.3209	24	10	44	21	23	0.328	0.246	0.366	0.336	0.09	0.142	0.179	ncbi-geneid:337959	hsa:337959	KRTAP13-2, KAP13-2				

Comparison of different approximations of genetic load in Malagasy and other populations

Finally, we explored if past demographic histories have influenced the genetic load in Malagasy and proxies populations. Importantly for this thesis, we wanted to explore and frame our results (based on whole-genome sequencing) according to previous studies. In consequence, we first interrogated if whole-genome sequencing results are similar to analyses based on exome data for African and Asian populations (such data do not exist on Madagascar population). Thus, we approximated genetic load using the "GERP score load", where each GERP-RS category is translated into a selection coefficient (Pedersen et

al., 2017). In this manner, the total mutation load for each population is the sum of the GERP-RS load score from all loci. Up to date, there is often no concrete evidence of the genetic mode of inheritance of genetic variants for complex disease genes (Zhao et al., 2016). Thus, most studies test multiple genetic models to explore the biological rationale behind the dominant, additive and recessive genetic models, as the appropriate selection of genetic models in functional studies can enhance the detection of the risks related to allelic variants (Setu & Basak, 2021). Thus, we analysed the potential deleterious impact of variants assuming (a) that two copies of allele B are required for increased risk (recessive model of dominance) and (b) that the genotypes 'AA' *versus* 'AB' *versus* 'BB' all present a different risk associated to the allele B (additive model) (Lewis, 2002). We consider that a mutation cannot be deleterious and dominant, since the sequences analysed in this exploratory work were retrieved from supposedly healthy individuals

We then compared the mutational load between populations assuming two different models of dominance (i.e. additive and recessive). As previously shown by Henn and colleagues (Henn et al., 2015, 2016) using whole-exome sequence data from 1000 Genomes Project, we detected that the mutation load is higher under the additive model of dominance than under the recessive model (Madagascar genetic load = 90.09 and 84.52, respectively) (Figure 5). Nevertheless, the differences in load between African and Asian populations under the additive model of dominance are greater than previous estimations. We suppose that this increase can be due to the extending dataset coming from WGS, depicting more derived variants in comparison to exome data. In addition, the quantity of load for Yoruba populations seems to be > ~6 times higher than the reported based on whole-exome data (Henn, 2015). Particularly, we observed differences on mutation load between populations when the dominance effect is set to additive (i.e. Yoruba = 85.14; Han_Chinese = 111.67, Madagascar = 90.09). Regarding the mutational load under the recessive model, we noted the same differences between populations (i.e. Yoruba = 79.97; Han_Chinese = 106.73, Madagascar = 84.45). Previous studies have stated that we must observe higher differences when the effect of deleterious variants is set to recessive dominance (Henn et al., 2015; Pedersen et al., 2017), nevertheless, we did not recover such observations using whole-genome sequence data. These differences made us look to other

statistics to evaluate genetic load, such as the number of deleterious alleles across individuals.

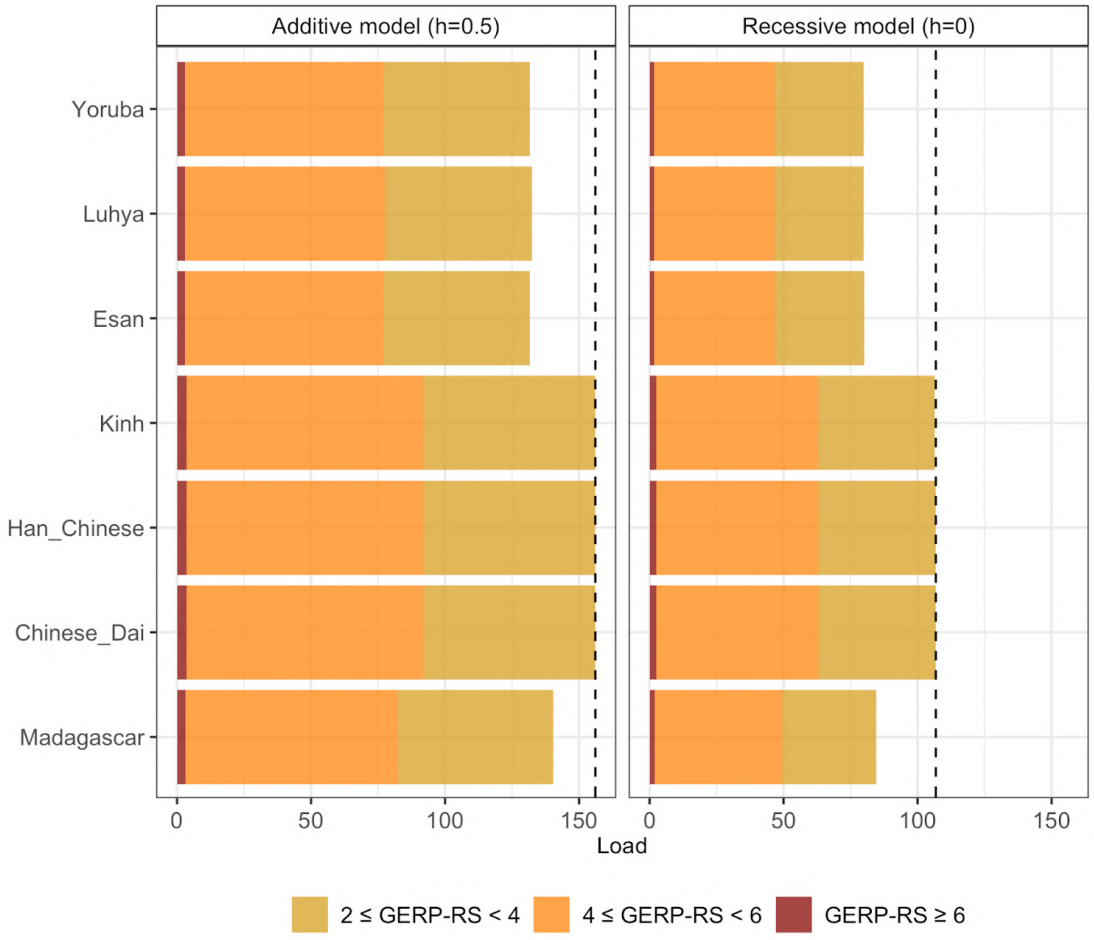


Figure 5. Estimated mutational load under an additive and recessive. The GERP score load is the approximation to the genetic load based on annotated GERP-RS scores from ANNOVAR database, converted to selection coefficients using the approach of Henn and colleagues (Henn et al., 2015) (see Methods). Importantly, we performed the calculation based on allele frequency of the annotated effect variant (derived allele). The proportion of predicted deleteriousness corresponds to moderate ($2 \leq$ GERP-RS < 4), large ($4 \leq$ GERP-RS < 6) and extreme effects (GERP-RS ≥ 6). The mutational load under an additive model of dominance (h=0.5) is shown in the left panel (assuming deleterious variants show some penetrance); the mutational load under a recessive model of dominance (h=0) is shown in the right panel.

Next, we estimated the ratio of derived alleles per individual, taking into account the GERP-RS category of each site (Figure 6). The results from figure 6 show that the load across Malagasy individuals is somewhere between the mutation loads observed in African and Asian proxies populations, which seems logical due to the admixture event. We also took into account local ancestry assignments for each derived allele and compared the ancestry-specific genetic load in Madagascar. We detected a higher proportion of derived alleles on Asian loci, which varies more between individuals in comparison to the African component (Figure 6D-F). Based on these differences, we made a comparison between ancestry-specific load in Madagascar against proxy populations in Africa and Asia. According to an additive effect of dominance, this was done by calculating the ratio (normalized by average ancestry proportion) of derived alleles per individual, for making comparisons between population as follows: a ratio > 1 indicates that the average number of derived alleles in Malagasy ancestral population (African or Asian) is higher than the corresponding proxy population (African or Asian), implying a greater genetic load.

Regarding the ancestry-specific mutation load at Madagascar under the additive dominance effects, we found that the both ancestral components presented a similar load pattern to those from African and Asian populations (Figure 7), with subtle differences between them: while the Malagasy African load is > 1 across all GERP-RS categories and African populations, the Malagasy Asian load is < 1 across all GERP-RS categories. Both ancestral components show a considerable standard deviation for sites predicted as extremely deleterious. Thus, it is possible that Asian ancestral population accumulated deleterious mutations during the long-term bottleneck, where some deleterious alleles could drift to higher frequencies, but not enough to surpass Asian proxies populations (Figure 6D-F; and Figure 4: Asian component). Moreover, if there was a transient increase in the number of deleterious homozygotes during bottleneck, it is possible that positive selection might have had an impact in removing possibly damaging variation (Lopez et al., 2018; Lucas-Sánchez et al., 2021; Pedersen et al., 2017; Simons et al., 2014). We are looking forward to implementing the adequate statistical tests in order to find if these differences are significant (i.e. bootstrapping the dataset). Also, these results must be addressed and extended with the help of computational simulations and statistical validation, in order to

evaluate the possible influence of ancestry-specific demography on current genetic diversity.

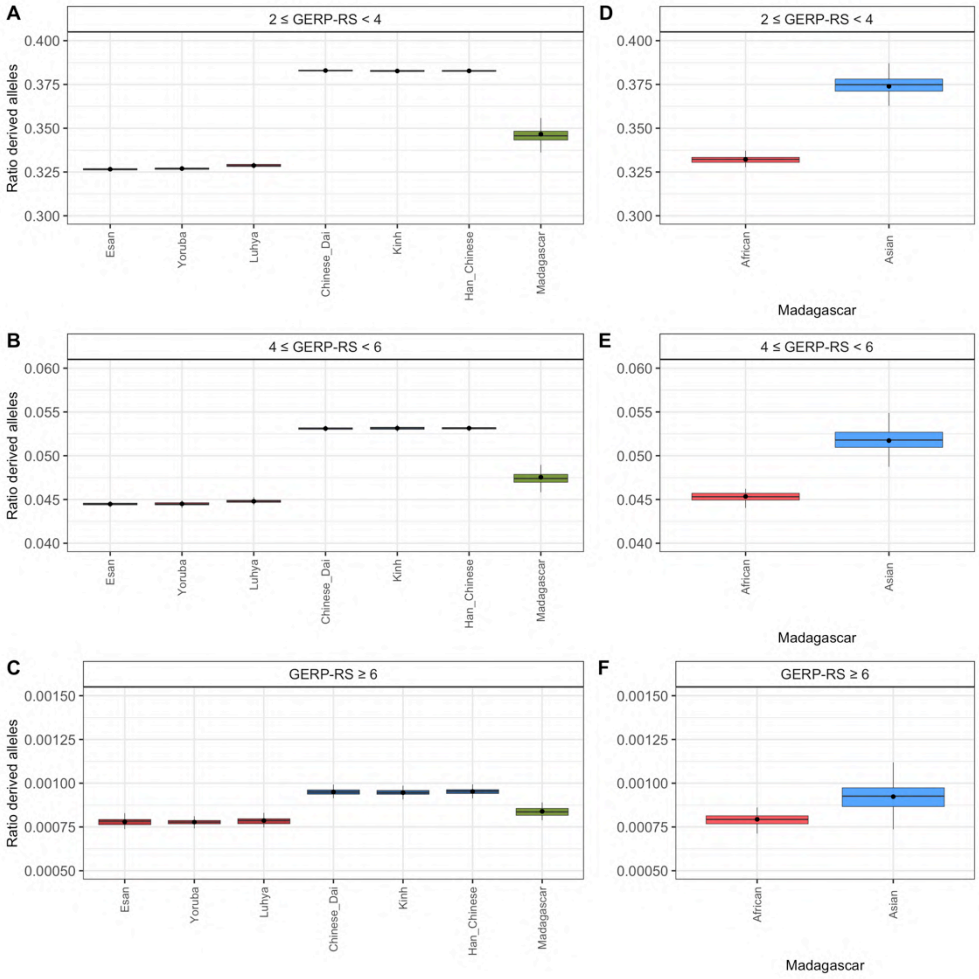


Figure 6. Individual proportion of derived alleles. The proportion of derived alleles is plotted for each population (x-axis) according GERP-RS score categories. (A-C) For each individual's genome, we estimate the proportion of derived alleles. (D-F) Ancestry-specific proportion of derived alleles. For each Malagasy genome, we count the number of derived alleles assigned to African or Asian regions; we then normalize by the global ancestry of each individual.

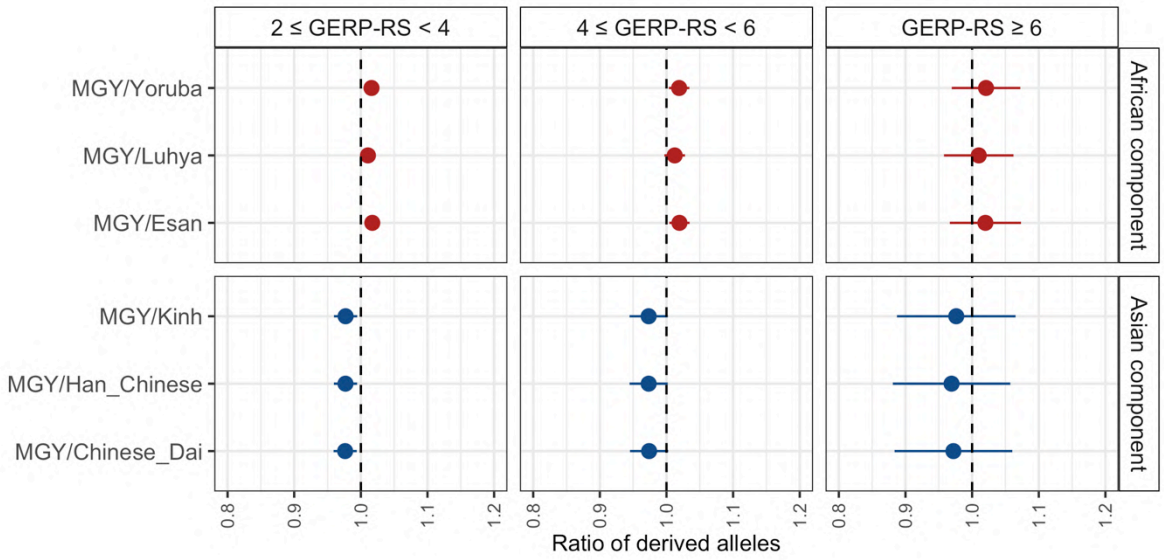


Figure 7. Ancestry-specific comparison of the per-individual number of derived alleles and homozygous derived genotypes across populations and GERP-RS categories. Mean ratio differences of derived alleles between populations are plotted as points, with error bars depicting standard deviation. The African component panel the mean number of derived alleles mapping to African ancestry per-individual (normalized by ancestry proportions), divided by the mean number of derived alleles in African populations. The Asian component panel represents the mean number of derived alleles mapping to Asian ancestry per-individual (normalized by ancestry proportions), divided by the mean number of derived alleles in Asian populations.

Discussion

We explored how the past demographic history of Malagasy population has impacted the distribution of deleterious alleles, particularly whether the long-term bottleneck for the Asian ancestral population had led to an accumulation of derived mutations in evolutionary conserved genomic regions. First, we studied previous experiments done on genetic data and applied the same strategies, inspecting the frequency of derived mutations according to the evolutionary genomic context of each variant using the GERP-RS score. This score depicts the intensity of rejecting new mutations in genomic regions shared in a phylogenetic tree of 35 mammals: higher the value, higher the predicted deleterious impact

of new mutations. Thus, we performed comparisons between mutation load of each ancestral component and proxy populations across these evolutionary categories. Our ancestry-specific approach for genetic load estimation showed subtle differences between ancestral components and proxies populations in Africa and Asia: we observe that the number of derived mutations on African assigned regions is slightly higher than in African proxies (Figure 7); while the number of derived mutations in Asian assigned regions is slightly lower than in the Asian proxies (Figure 7). We cannot conclude if the reduction of Asian derived alleles in Malagasy genomes (in comparison to Asian proxies) is the result of the past bottleneck for the Asian ancestral population, but this point should be explored in the future. Particularly, it will be necessary to generate a wider and more diverse database of individuals, including populations from African eastern coast en Island Southeast Asia (ISEA). In addition, we must be careful when generating the genomic database, as different batch effects can surge from merging different datasets, generating false positive signals. Also, we must properly design quality filters in order to detect rare variants with potential deleterious effects, without neglecting sites with substantial information (i.e. sites that do not vary or vary only in some populations). In addition, according to different software or databases, there could be discrepancies between the choice of annotating the reference, the alternative, the ancestral, the derived or the effect allele. Thus, it will be necessary to take into account these differences for the interpretation of mutation load, as we can rapidly arrive at the challenge of knowing which alleles are deleterious and what should be the effect under the particular dominance assumptions (i.e. if it is necessary an homozygous genotype for affecting the fitness).

This study allowed us to better understand that variation generated by mutations is the result of complex mechanisms, which can affect the evaluation of mutation's deleteriousness and bias the approaches for calculating the mutation load. By inspecting the allele and genotype frequencies of variants predicted to be deleterious, we narrowed the spectrum of mutations for studying potentially functional/deleterious effects. In this manner, we compared Malagasy and proxies populations and recovered mutations that followed particular variation patterns, such as: variants present at Madagascar but almost absent abroad, mutations showing departures from Hardy-Weinberg Equilibrium (even if not statistically significant in the 67 Malagasy samples), mutations with highly

differentiated allele frequencies (> 40%) between Africa and Asia, sites showing less than one homozygous for the deleterious allele in Madagascar, sites annotated as deleterious by more than one software (GERP-RS, snpEFF and aLoFT). The results of these criteria are shown in Tables 2-5. All these patterns helped us to detect variants with potential effects on function, for which we propose that they could have an effect at the molecular level (i.e. truncation of a protein) without affecting the fitness of the individual, or alternatively, these variants interact with others and leverage the expected effect. The link of these mutations on health or disease present in the Malagasy population should be investigated in the future.

In summary, we produced a database of 370,766 SNVs annotated with the ancestral and derived allele, along with the GERP-RS score. We inspected different summary statistics (i.e. number of derived alleles per individual) for studying the relationship between human demographic history and the distribution of mutations. Furthermore, we inspected genomic variation based on allele and genotype frequencies, providing a catalogue of SNVs that could represent potential deleterious variants. This catalogue, as well as the algorithms and databases, can be interrogated and accessed for future experiments in the laboratory. This chapter constitutes an effort for bridging and exploring demographic history and functional variation in human populations. Eventually, this strategy can be complemented and implemented for other populations around the world, procuring as best-practices: (1) the proper merge of whole-genome sequencing data coming from different anthropological/genetic studies, (2) declaring the predicted deleteriousness of variants independently for each population, annotating the variant under an evolutionary scope (i.e. ancestral/derived allele, conservation status across taxa, etc.) and with more than one software; (3) the implementation of computational simulations for discerning between demographic and evolutionary selection effects. With all this in mind, we expect to arrive at proper methodology in order to propose solid conclusions regarding the complex interactions of evolutionary forces under particular demographic scenarios. Making the following steps towards the analysis of whole-genome data will be demanding (i.e. storage memory, processing capacity, etc.), but it will allow us to exploit at maximum the genomic signals coming from discrepancies, even minimal, between empirical and simulated data.

References

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61–65. https://doi.org/10.1038/nmeth.1527

Balasubramanian, S., Fu, Y., Pawashe, M., McGillivray, P., Jin, M., Liu, J., Karczewski, K. J., MacArthur, D. G., & Gerstein, M. (2017). Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nature Communications*, *8*(1), 382. https://doi.org/10.1038/s41467-017-00443-5

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Choudhury, A., Sengupta, D., Ramsay, M., & Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*, *30*(R1), R56–R63. https://doi.org/10.1093/hmg/ddaa274

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118}; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, *15*(7), 901–913. https://doi.org/10.1101/gr.3577405

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

DeGiorgio, M., Jakobsson, M., & Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences*, *106*(38), 16057–16062. https://doi.org/10.1073/pnas.0903341106

Genovese, G., Handsaker, R. E., Li, H., Altemose, N., Lindgren, A. M., Chambert, K., Pasaniuc, B., Price, A. L., Reich, D., Morton, C. C., Pollak, M. R., Wilson, J. G., & McCarroll, S. A. (2013). Using population admixture to help complete maps of the human genome. *Nature Genetics*, *45*(4), 406–414. https://doi.org/10.1038/ng.2565

Glémin, S., Ronfort, J., & Bataillon, T. (2003). Patterns of Inbreeding Depression and Architecture of the Load in Subdivided Populations. *Genetics*, *165*(4), 2193–2212. https://doi.org/10.1093/genetics/165.4.2193

Goode, D. L., Cooper, G. M., Schmutz, J., Dickson, M., Gonzales, E., Tsai, M., Karra, K., Davydov, E., Batzoglou, S., Myers, R. M., & Sidow, A. (2010). Evolutionary

constraint facilitates interpretation of genetic variation in resequenced human genomes. *Genome Research*, *20*(3), 301–310. https://doi.org/10.1101/gr.102210.109
Grossen, C., Guillaume, F., Keller, L. F., & Croll, D. (2020). Purging of highly deleterious mutations through severe bottlenecks in Alpine ibex. *Nature Communications*, *11*(1), 1001. https://doi.org/10.1038/s41467-020-14803-1
Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., & Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature Reviews Genetics*, *16*(6), 333–343. https://doi.org/10.1038/nrg3931
Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R., Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., & Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences*, *113*(4). https://doi.org/10.1073/pnas.1510805112
Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070
Kimura, M., Maruyama, T., & Crow, J. F. (1963). THE MUTATION LOAD IN SMALL POPULATIONS. *Genetics*, *48*(10), 1303–1312. https://doi.org/10.1093/genetics/48.10.1303
Laval, G., Patin, E., Barreiro, L. B., & Quintana-Murci, L. (2010). Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. *PLoS ONE*, *5*(4), e10284. https://doi.org/10.1371/journal.pone.0010284
Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, *3*(2), 146–153. https://doi.org/10.1093/bib/3.2.146
Lopez, M., Kousathanas, A., Quach, H., Harmant, C., Mouguiama-Daouda, P., Hombert, J.-M., Froment, A., Perry, G. H., Barreiro, L. B., Verdu, P., Patin, E., & Quintana-Murci, L. (2018). The demographic history and mutational load of African hunter-gatherers and farmers. *Nature Ecology & Evolution*, *2*(4), 721–730. https://doi.org/10.1038/s41559-018-0496-4
Lucas-Sánchez, M., Font-Porterias, N., Calafell, F., Fadhlaoui-Zid, K., & Comas, D. (2021). Whole-exome analysis in Tunisian Imazighen and Arabs shows the impact of demography in functional variation. *Scientific Reports*, *11*(1), 21125. https://doi.org/10.1038/s41598-021-00576-0
Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, *39*(Database), D52–D57. https://doi.org/10.1093/nar/gkq1237
Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *The American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110
Mesner, L. D., Calabrese, G. M., Al-Barghouthi, B., Gatti, D. M., Sundberg, J. P.,

Churchill, G. A., Godfrey, Dana. A., Ackert-Bicknell, C. L., & Farber, C. R. (2019). Mouse genome-wide association and systems genetics identifies Lhfp as a regulator of bone mass. *PLOS Genetics*, *15*(5), e1008123. https://doi.org/10.1371/journal.pgen.1008123

Pedersen, C.-E. T., Lohmueller, K. E., Grarup, N., Bjerregaard, P., Hansen, T., Siegismund, H. R., Moltke, I., & Albrechtsen, A. (2017). The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics*, *205*(2), 787–801. https://doi.org/10.1534/genetics.116.193821

Petit, M. M. R., Schoenmakers, E. F. P. M., Huysmans, C., Geurts, J. M. W., Mandahl, N., & Van de Ven, W. J. M. (1999). LHFP,a Novel Translocation Partner Gene ofHMGICin a Lipoma, Is a Member of a New Family ofLHFP-like Genes. *Genomics*, *57*(3), 438–441. https://doi.org/10.1006/geno.1999.5778

Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L. M.-A., Rakotomalala, M. M., Razafiarivony, M., Rasoarifetra, B., Raharijesy, M. A., Razafindralambo, L., Ramilisonina, Fanony, F., Lejamble, S., Thomas, O., Mohamed Abdallah, A., Rocher, C., Arachiche, A., … Letellier, T. (2017). Genomic landscape of human diversity across Madagascar. *Proceedings of the National Academy of Sciences*, *114*(32), E6498–E6506. https://doi.org/10.1073/pnas.1704906114

Pierron, D., Razafindrazaka, H., Pagani, L., Ricaut, F.-X., Antao, T., Capredon, M., Sambo, C., Radimilahy, C., Rakotoarisoa, J.-A., Blench, R. M., Letellier, T., & Kivisild, T. (2014). Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. *Proceedings of the National Academy of Sciences*, *111*(3), 936–941. https://doi.org/10.1073/pnas.1321860111

R Core Team. (2021). R: A language and environment for statistical. *R Foundation for Statistical Computing*. https://www.R-project.org/

Setu, T. J., & Basak, T. (2021). An Introduction to Basic Statistical Models in Genetics. *Open Journal of Statistics*, *11*(06), 1017–1025. https://doi.org/10.4236/ojs.2021.116060

Simons, Y. B., Turchin, M. C., Pritchard, J. K., & Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nature Genetics*, *46*(3), 220–224. https://doi.org/10.1038/ng.2896

The 1000 Genomes Project Consortium, Corresponding authors, Auton, A., Abecasis, G. R., Steering committee, Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603

Zhao, F., Song, M., Wang, Y., & Wang, W. (2016). Genetic model. *Journal of Cellular and Molecular Medicine*, *20*(4), 765–765. https://doi.org/10.1111/jcmm.12751

CHAPTER IV

CONCLUSIONS AND PERSPECTIVES

CHAPTER IV

CONCLUSIONS AND PERSPECTIVES

In this thesis, we studied the consequences of human evolutionary history on genetic diversity and its impact on the health of current populations. We were particularly interested in the effects of admixture processes (events involving the formation of a population by the genetic addition from at least two source populations) and the arrival to new environments. We focused on the human population currently living in Madagascar. Previous genetic and linguistic analyses have shown that the Malagasy population is the result of a genetic admixture that has occurred over the last millennium, between Bantu-speaking populations and Austronesian-speaking populations (Heiske et al., 2021; Pierron et al., 2017). The genetic study of Malagasy populations, as well as the settlement of the island, has been addressed during the last decades using genetic data, where the efforts of the MAGE consortium helped to collect genomic information at a high-resolution level based on a sample of 257 villages spread across the entire island. We conducted this project studying the genetic diversity based on 2.5 million SNPs on the nuclear genome and whole-genome sequencing data. We therefore sought to understand how the human history of the settlement of the island has influenced the genetic diversity of Malagasy individuals. Particularly, we studied the effective population size of Malagasy African and Asian ancestors before the admixture. This information influences the genetic variation in populations and can help to infer the expected neutral genetic diversity due to only mutation and genetic drift (Henn et al., 2015). In addition, it can serve to extrapolate data regarding hypotheses about innovations in long-distance navigation, vessel construction and possibly subsistence strategies of African and Asian source populations. In this manner, we investigated the past evolutionary history of Malagasy using genetic data and computational simulations, creating a computational model that helped us to better understand the demographic trajectories of Malagasy genetic ancestors before the admixture event. In particular, by studying the sharing of chromosome segments between individuals (IBD determination), local ancestry information and simulated genetic data, we inferred that the ancestral Asian population of Malagasy was isolated for more than 1000

years with an effective size of just a few hundred individuals. This isolation ended around 1000 years before present (BP) by admixture with Bantu-Speaking African populations in Madagascar. We further investigated the effects of these demographic processes on the effects of natural selection. We inspected different summary statistics (i.e. number of derived alleles per individual) for studying the relationship between human demographic history and the distribution of deleterious mutations. Furthermore, we inspected variation based on allele and genotype frequencies in Madagascar at whole-genome-sequencing level, providing a catalogue of SNVs that could represent potential deleterious variants. In summary, based on genomic data, we generated a computational model for studying the settlement of Madagascar, which can be helpful for the study of micro evolutionary processes —mutation, drift, migration, and selection— operating on human populations.

These results are a small piece of a much larger puzzle in the evolutionary history of human populations. Humans are a young species that has undergone recent dispersal out of Africa, moving into a wide range of very different environments (with different climates, altitudes, food sources, and pathogens) and followed by a marked population growth beginning in the Neolithic period (Mark Jobling et al., 2014). The human population is clearly not in a state of equilibrium, as recent patterns of migration (i.e. extensive intercontinental migrations) and population growth (i.e. Neolithic, industrialization) have emerged (Mark Jobling et al., 2014). Quantifying these past events is challenging, but can be approached by using genetic data and bioinformatics methods. Importantly, the genetic approach based on the genotyping of a large number of individuals has many advantages, such as an increased statistical power, the representativeness of the diversity of the studied population, the possibility of testing a larger number of scenarios and hypotheses, etc. For all these reasons, we observe that large-scale genetic approaches are more commonly used in anthropology and evolutionary medicine. However, these approximations based on big-data analysis should be used with caution and with awareness of the many scientific, sociological and ethical limitations that they can pose.

First, we must consider that in order to reconstruct past demographic and evolutionary processes across genetically diverse regions (i.e. long-range migrations, admixtures, adaptation process, etc.), we need a comprehensive representation of genetic variation across the geographical areas of interest. Thus, it is necessary to foresee scientific

collaborations and obtain the adequate permissions in order to access published genetic data coming from other anthropological studies or projects. Importantly, these proceedings involve ethical considerations that have to be taken into account in order to promote fairer international collaborations. These ethical considerations have been addressed more openly during the last years by diverse academics, highlighting the importance of informed consent and the sovereignty of populations and individuals who acceded to participate. For example, French law considers genetic data as an element of the human body, which "belongs" to a "donor", but is not property of the individual (a consequence of the principle non-patrimoniality of the human body) (Stoeklé et al., 2018). Thus, each french person can have and use their own genetic data but in a very specific space defined by France, such as medical research. Interestingly, the ethical and legal principle for using genetic data is different across countries. As another example, USA abrogates for the principle of auto-determination over the non-patrimoniality. In consequence, genetic data is considered as an economically valuable asset, and the individual is able to manage their genetic information under property rights (Stoeklé et al., 2018). As in the way France or the USA has decided how to regulate the use of genetic data, each population, country or nation can have their legislations regarding the sovereignty of data. In any case, it is the responsibility of the researcher to conduct their investigations with scientific integrity, condemning abuses, mischievous agreements and extractivism practices. Regarding the sharing of data, it must be done in compliance with the rules of the right to the protection of privacy, in particular the protection of personal data, as the handling of a massive amount of personal data can potentially harbor very informative and delicate information about the participant, his family, and the population where the individual comes from.

Second, given that the identification of past evolutionary forces and demographic parameters inferred to lead to known genetic diversity depends totally on the set of individuals analysed, we must be careful when reporting the results of a particular population. Indeed, during the investigation based on genomic data we can detect a large proportion of rare variants that are likely to have occurred during or after population divergence (Henn et al., 2015). These variants can be population-specific and found at very low frequencies. Thus, it is important that the research group properly conduct their investigations assuring that confidentiality is a crucial requirement, ensuring that the

information about an individual's genome will not be accidentally released, sold, lost or stolen. One example or recommendations can be found at the National Centre for Indigenous Genomics in Australia, which is governed by a majority-Indigenous board and has approached communities to ask what they wish to do with their samples. In addition, it can be an interest in keeping disease susceptibilities private. As others authors have mentioned, some results can be meaningless, overwhelming, or unnecessarily worrying for the participants or even the population. Additionally, we must consider that the researchers are not extent of misclassifications regarding the individual's assignments to a particular population and that the history reconstructed from the genome does not necessarily have all the history of a given population. A first approximation for addressing these issues could be carrying out these projects in close collaboration with archeological, historical and social disciplines, as the link with other approaches and methodologies may allow focusing and discerning on the history of settlement and the way of life of individuals, proposing coherent systems that represent human populations.

As mentioned elsewhere, the idea that DNA variants respect the borders of recent constructs such as nation states is implausible (Mark Jobling et al., 2014). In addition, each country has different laws and requirements for the access, management, and ownership of individual's genetic data. Having access to a worldwide genetic database opens the door for studying the evolutionary history of populations in a comprehensive view of genetic variation, allowing the implementation of computational simulations and systems biology approaches. Nevertheless, the definition of populations, the collection of data and the reporting of results are at the origin of many ethical questions. Moreover, we must also anticipate the computation needs (i.e. storage and processing memory) for implementing computational simulations on millions of genetic markers coming from thousands of individuals. It is possible that having access to a curated database in the laboratory can facilitate the design of experiments (i.e. characterize demographic histories, scan for selection signals), as well as the robustness of conclusions, but it will demand hard bioinformatics and computational work. For instance, the sequencing of the first human genome began in the early 1990s (by a large consortium), but in 2020 the aggregated potential sequencing capacity of 33 European institutions (public and private) was approximately ~198, 000 WGS per year (Narayanasamy et al., 2020). The main objective

of analysing this massive quantity of data would be the depth study of diversity from complete genome information, which will help in the imputation on a larger scale of previous anthropological studies based on microarray data, as well as a high-resolution research of genetic variation. We have demonstrated that genomic-based inference, coupled with simulated genetic data, can provide crucial clues to reconstruct the demographic history of human populations that have participated in recent settlement events. Through this doctoral thesis we looked to compare theoretical diversity with the observed genetic diversity, seeking whether the different alleles accumulated on a population can help to elucidate the settlement of a territory, as well as the migration and demographic history experienced through time. The study of this information made it possible to propose a list of the loci that might have undergone selection pressures in the recent past of populations. We acknowledge that the model is far from being completed, with diverse aspects open for amelioration. We discussed some of the ethical, sociological and scientific issues that arise when performing population genetics studies and evolutionary medicine inferences based on genetic data. As mentioned elsewhere, we must have in mind that the populations who have collaborated with their DNA are equal participants in the construction of knowledge. Final points that we could discuss are the political implications of this kind of work (the study of human populations) and the role of the researcher who develops such questions and experiments (past evolutionary events based on genetic data). Particularly, it is important to be assertive and critic, trying to avoid (1) pushing some beliefs in order to justify that a theory is true and (2) chasing questions essentially different to those shared by the people who actually represent the studied population.

In conclusion, despite all these limitations (technical, ethical, social, etc.) the study of large-scale genetic collections is one of the major scientific breakthroughs of the last decade in the study of the history of populations. New technological advances in genomics (single-molecule genomic sequencing, microfluidic platforms, etc.), the possibility of having access to a large number of individuals and populations, as well as the advances in bioinformatics (computational power and software development), will probably allow scientists in the coming decades to refine the evolutionary history of our species, both on a global (i.e. Out-of-Africa peopling of the world) and local scale (i.e. settlement of new territories). The work for inferring evolutionary history in human populations can only be

achieved through a multidisciplinary approach (biology, computer science, mathematics, anthropology, archaeology, medicine, etc.), opening the doors for diverse projects, not only in anthropology, but also in many fields such as medicine, ecology, history, etc. This new approach therefore leaves room for many new scientific projects and collaborations, going from the development of new bioinformatics methods (i.e. to study the relationship of physiology-pathology of certain deleterious variants) to the design and execution of anthropological studies (fieldwork, sample collection, etc.).

BIBLIOGRAPHY

Bibliography consulted for the redaction of Introduction, Chapter I (introduction, discussion and conclusions), Chapter II (introduction, discussion and perspectives) and Chapter IV of this doctoral thesis.

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., Kyriazis, C. C., Ragsdale, A. P., Tsambos, G., Baumdicker, F., Carlson, J., Cartwright, R. A., Durvasula, A., Gronau, I., Kim, B. Y., McKenzie, P., Messer, P. W., Noskova, E., Ortega-Del Vecchyo, D., … Kern, A. D. (2020). A community-maintained standard library of population genetic models. *ELife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Anderson, A., Clark, G., Haberle, S., Higham, T., Nowak-Kemp, M., Prendergast, A., Radimilahy, C., Rakotozafy, L. M., Ramilisonina, Schwenninger, J.-L., Virah-Sawmy, M., & Camens, A. (2018). New evidence of megafaunal bone damage indicates late colonization of Madagascar. *PLOS ONE*, *13*(10), e0204368. https://doi.org/10.1371/journal.pone.0204368

Arenas, M., & Posada, D. (2014). Simulation of Genome-Wide Evolution under Heterogeneous Substitution Models and Complex Multispecies Coalescent Histories. *Molecular Biology and Evolution*, *31*(5), 1295–1301. https://doi.org/10.1093/molbev/msu078

Assael, Y., Sommerschield, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, *603*(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z

Browning, S. R., & Browning, B. L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *The American Journal of Human Genetics*, *97*(3), 404–418. https://doi.org/10.1016/j.ajhg.2015.07.012

Browning, S. R., Browning, B. L., Daviglus, M. L., Durazo-Arvizu, R. A., Schneiderman, N., Kaplan, R. C., & Laurie, C. C. (2018). Ancestry-specific recent effective population size in the Americas. *PLOS Genetics*, *14*(5), e1007385. https://doi.org/10.1371/journal.pgen.1007385

Burns, S. J., Godfrey, L. R., Faina, P., McGee, D., Hardt, B., Ranivoharimanana, L., & Randrianasy, J. (2016). Rapid human-induced landscape transformation in Madagascar at the end of the first millennium of the Common Era. *Quaternary Science Reviews*, *134*, 92–99. https://doi.org/10.1016/j.quascirev.2016.01.007

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). *Mitochondrial DNA and human evolution*. 6.

Cayuela, H., Rougemont, Q., Prunier, J. G., Moore, J.-S., Clobert, J., Besnard, A., & Bernatchez, L. (2018). Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review. *Molecular Ecology*, *27*(20), 3976–4010. https://doi.org/10.1111/mec.14848

Choudhury, A., Sengupta, D., Ramsay, M., & Schlebusch, C. (2021). Bantu-speaker migration and admixture in southern Africa. *Human Molecular Genetics*, *30*(R1), R56–R63. https://doi.org/10.1093/hmg/ddaa274

Conesa, A., & Mortazavi, A. (2014). The common ground of genomics and systems biology. *BMC Systems Biology*, *8*(Suppl 2), S1. https://doi.org/10.1186/1752-0509-8-S2-S1

Cooke, N. P., & Nakagome, S. (2018). Fine-tuning of Approximate Bayesian Computation for human population genomics. *Current Opinion in Genetics & Development*, *53*, 60–69. https://doi.org/10.1016/j.gde.2018.06.016

Cox, M. P., Nelson, M. G., Tumonggor, M. K., Ricaut, F.-X., & Sudoyo, H. (2012). A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1739), 2761–2768. https://doi.org/10.1098/rspb.2012.0012

Crowley, B. E. (2010). A refined chronology of prehistoric Madagascar and the demise of the megafauna. *Quaternary Science Reviews*, *29*(19–20), 2591–2603. https://doi.org/10.1016/j.quascirev.2010.06.030

Crowther, A., Lucas, L., Helm, R., Horton, M., Shipton, C., Wright, H. T., Walshaw, S., Pawlowicz, M., Radimilahy, C., Douka, K., Picornell-Gelabert, L., Fuller, D. Q., & Boivin, N. L. (2016). Ancient crops provide first archaeological signature of the westward Austronesian expansion. *Proceedings of the National Academy of Sciences*, *113*(24), 6635–6640. https://doi.org/10.1073/pnas.1522714113

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

DeGiorgio, M., Jakobsson, M., & Rosenberg, N. A. (2009). Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences*, *106*(38), 16057–16062. https://doi.org/10.1073/pnas.0903341106

Douglass, K., Hixon, S., Wright, H. T., Godfrey, L. R., Crowley, B. E., Manjakahery, B., Rasolondrainy, T., Crossland, Z., & Radimilahy, C. (2019). A critical review of radiocarbon dates clarifies the human settlement of Madagascar. *Quaternary Science Reviews*, *221*, 105878. https://doi.org/10.1016/j.quascirev.2019.105878

Godfrey, L. R., Scroxton, N., Crowley, B. E., Burns, S. J., Sutherland, M. R., Pérez, V. R., Faina, P., McGee, D., & Ranivoharimanana, L. (2019). A new interpretation of Madagascar's megafaunal decline: The "Subsistence Shift Hypothesis". *Journal of Human Evolution*, *130*, 126–140. https://doi.org/10.1016/j.jhevol.2019.03.002

Gross, M. (2012). The evolution of writing. *Current Biology*, *22*(23), R981–R984. https://doi.org/10.1016/j.cub.2012.11.032

Heiske, M., Alva, O., Pereda-Loth, V., Van Schalkwyk, M., Radimilahy, C., Letellier, T., Rakotarisoa, J.-A., & Pierron, D. (2021). Genetic evidence and historical theories of the Asian and African origins of the present Malagasy population. *Human Molecular Genetics*, *30*(R1), R72–R78. https://doi.org/10.1093/hmg/ddab018

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., & Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature Reviews Genetics*, *16*(6), 333–343. https://doi.org/10.1038/nrg3931

Hixon, S. W., Douglass, K. G., Crowley, B. E., Rakotozafy, L. M. A., Clark, G., Anderson, A., Haberle, S., Ranaivoarisoa, J. F., Buckley, M., Fidiarisoa, S., Mbola, B., & Kennett, D. J. (2021). Late Holocene spread of pastoralism coincides with endemic megafaunal extinction on Madagascar. *Proc. R. Soc. B*, *288*, 20211204. https://doi.org/10.1098/rspb.2021.1204

Hixon, S. W., Douglass, K. G., Godfrey, L. R., Eccles, L., Crowley, B. E., Rakotozafy, L. M. A., Clark, G., Haberle, S., Anderson, A., Wright, H. T., & Kennett, D. J. (2021). Ecological Consequences of a Millennium of Introduced Dogs on Madagascar. *Frontiers in Ecology and Evolution*, *9*, 689559. https://doi.org/10.3389/fevo.2021.689559

Hoban, S. (2014). An overview of the utility of population simulation software in

molecular ecology. *Molecular Ecology*, *23*(10), 2383–2401. https://doi.org/10.1111/mec.12741

Hoban, S., Bertorelle, G., & Gaggiotti, O. E. (2012). Computer simulations: Tools for population and evolutionary genetics. *Nature Reviews Genetics*, *13*(2), 110–122. https://doi.org/10.1038/nrg3130

Hudjashov, G., Karafet, T. M., Lawson, D. J., Downey, S., Savina, O., Sudoyo, H., Lansing, J. S., Hammer, M. F., & Cox, M. P. (2017). Complex Patterns of Admixture across the Indonesian Archipelago. *Molecular Biology and Evolution*, *34*(10), 2439–2452. https://doi.org/10.1093/molbev/msx196

Ioannidis, A. G., Blanco-Portillo, J., Sandoval, K., Hagelberg, E., Barberena-Jonas, C., Hill, A. V. S., Rodríguez-Rodríguez, J. E., Fox, K., Robson, K., Haoa-Cardinali, S., Quinto-Cortés, C. D., Miquel-Poblete, J. F., Auckland, K., Parks, T., Sofro, A. S. M., Ávila-Arcos, M. C., Sockell, A., Homburger, J. R., Eng, C., … Moreno-Estrada, A. (2021). Paths and timings of the peopling of Polynesia inferred from genomic networks. *Nature*, *597*(7877), 522–526. https://doi.org/10.1038/s41586-021-03902-8

Jacobs, G. S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C. C., Lawson, D. J., Mondal, M., Pagani, L., Ricaut, F.-X., Stoneking, M., Metspalu, M., Sudoyo, H., Lansing, J. S., & Cox, M. P. (2019). Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*, *177*(4), 1010-1021.e32. https://doi.org/10.1016/j.cell.2019.02.035

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

Larena, M., Sanchez-Quinto, F., Sjödin, P., McKenna, J., Ebeo, C., Reyes, R., Casel, O., Huang, J.-Y., Hagada, K. P., Guilay, D., Reyes, J., Allian, F. P., Mori, V., Azarcon, L. S., Manera, A., Terando, C., Jamero, L., Sireg, G., Manginsay-Tremedal, R., … Jakobsson, M. (2021). Multiple migrations to the Philippines during the last 50,000 years. *Proceedings of the National Academy of Sciences*, *118*(13), e2026132118. https://doi.org/10.1073/pnas.2026132118

Li, H., Sinha, A., Anquetil André, A., Spötl, C., Vonhof, H. B., Meunier, A., Kathayat, G., Duan, P., Voarintsoa, N. R. G., Ning, Y., Biswas, J., Hu, P., Li, X., Sha, L., Zhao, J., Edwards, R. L., & Cheng, H. (2020). A multimillennial climatic context for the megafaunal extinctions in Madagascar and Mascarene Islands. *Science Advances*, *6*(42), eabb2459. https://doi.org/10.1126/sciadv.abb2459

Loog, L. (2021). Sometimes hidden but always there: The assumptions underlying genetic inference of demographic histories. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1816), 20190719. https://doi.org/10.1098/rstb.2019.0719

Mark Jobling, Edward Hollox, & Chris Tyler-Smith. (2014). *Human Evolutionary Genetics* (2nd ed.). Garland Sciences.

Narayanasamy, S., Markina, V., Thorogood, A., Blazkova, A., Shabani, M., Knoppers, B. M., Prainsack, B., & Koesters, R. (2020). Genomic Sequencing Capacity, Data Retention, and Personal Access to Raw Data in Europe. *Frontiers in Genetics*, *11*, 303. https://doi.org/10.3389/fgene.2020.00303

National Research Council. (1997). *Evaluating Human Genetic Diversity* (p. 5955). National Academies Press. https://doi.org/10.17226/5955

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

Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., Rakotozafy, L. M.-A., Rakotomalala, M. M., Razafiarivony, M., Rasoarifetra, B., Raharijesy, M. A., Razafindralambo, L., Ramilisonina, Fanony, F., Lejamble, S., Thomas, O., Mohamed Abdallah, A., Rocher, C., Arachiche, A., … Letellier, T. (2017). Genomic landscape of human diversity across Madagascar. *Proceedings of the National Academy of Sciences*, *114*(32), E6498–E6506. https://doi.org/10.1073/pnas.1704906114

Radimilahy Chantal. (1985). *Contribution à l'étude de l'ancienne métallurgie du fer à Madagascar. Mussé D'Art et d'Archéologie de l'Université de Madagascar, Travaux et Documents XXV. Tananarive*.

Ralph, P., & Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. *PLoS Biology*, *11*(5), e1001555. https://doi.org/10.1371/journal.pbio.1001555

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H., & Jakobsson, M. (2012). Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. *Science*, *338*(6105), 374–379. https://doi.org/10.1126/science.1227721

Stoeklé, H.-C., Forster, N., Turrini, M., Charlier, P., Hervé, C., Deleuze, J.-F., & Vogt, G. (2018). La propriété des données génétiques: De la donnéeà l'information. *médecine/sciences*, *34*(12), 1100–1104. https://doi.org/10.1051/medsci/2018291

Sunyaev, S. R., & Roth, F. P. (2013). Systems biology and the analysis of genetic variation. *Current Opinion in Genetics & Development*, *23*(6), 599–601. https://doi.org/10.1016/j.gde.2013.11.010

Tofanelli, S., Bertoncini, S., Castri, L., Luiselli, D., Calafell, F., Donati, G., & Paoli, G. (2009). On the Origins and Admixture of Malagasy: New Evidence from High-Resolution Analyses of Paternal and Maternal Lineages. *Molecular Biology and Evolution*, *26*(9), 2109–2124. https://doi.org/10.1093/molbev/msp120

Voarintsoa, N. R. G., Railsback, L. B., Brook, G. A., Wang, L., Kathayat, G., Cheng, H., Li, X., Edwards, R. L., Rakotondrazafy, A. F. M., & Madison Razanatseheno, M. O. (2017). Three distinct Holocene intervals of stalagmite deposition and nondeposition revealed in NW Madagascar, and their paleoclimate implications. *Climate of the Past*, *13*(12), 1771–1790. https://doi.org/10.5194/cp-13-1771-2017

Yuan, X., Miller, D. J., Zhang, J., Herrington, D., & Wang, Y. (2012). An Overview of Population Genetic Data Simulation. *Journal of Computational Biology*, *19*(1), 42–54. https://doi.org/10.1089/cmb.2010.0188

ANNEXES

Annex A: Supplementary information from article entitled “The loss of biodiversity in Madagascar is contemporaneous with major historical events”

Alva Omar, Leroy Anaïs, Heiske Margit, Pereda-Loth Veronica, Tisseyre Lenka, Boland Anne, Deleuze Jean-François, Rocha Jorge, Schlebusch Carina, Fortes-Lima Cesar, Stoneking Mark, Radimilahy Chantal, Rakotoarisoa Jean-Aimé, Letellier Thierry, Pierron Denis

This work was published in Current Biology journal on November 5th of 2022. We send the article on April 7th, 2022 and we received favorable opinions for publication after “major revisions” on May 4th. The article was accepted for publication on september 28th.

Annex A. Supplementary information from article entitled "The loss of biodiversity in Madagascar is contemporaneous with major historical events".

Figure S1. **Sampling locations included in the study. (A)** Map showing the locations of samples used in the Density dataset. **(B)** Map showing the locations of samples used in the Diversity dataset. Colors indicate regional affiliation of populations. Detailed sample information is given in supplementary table S1.

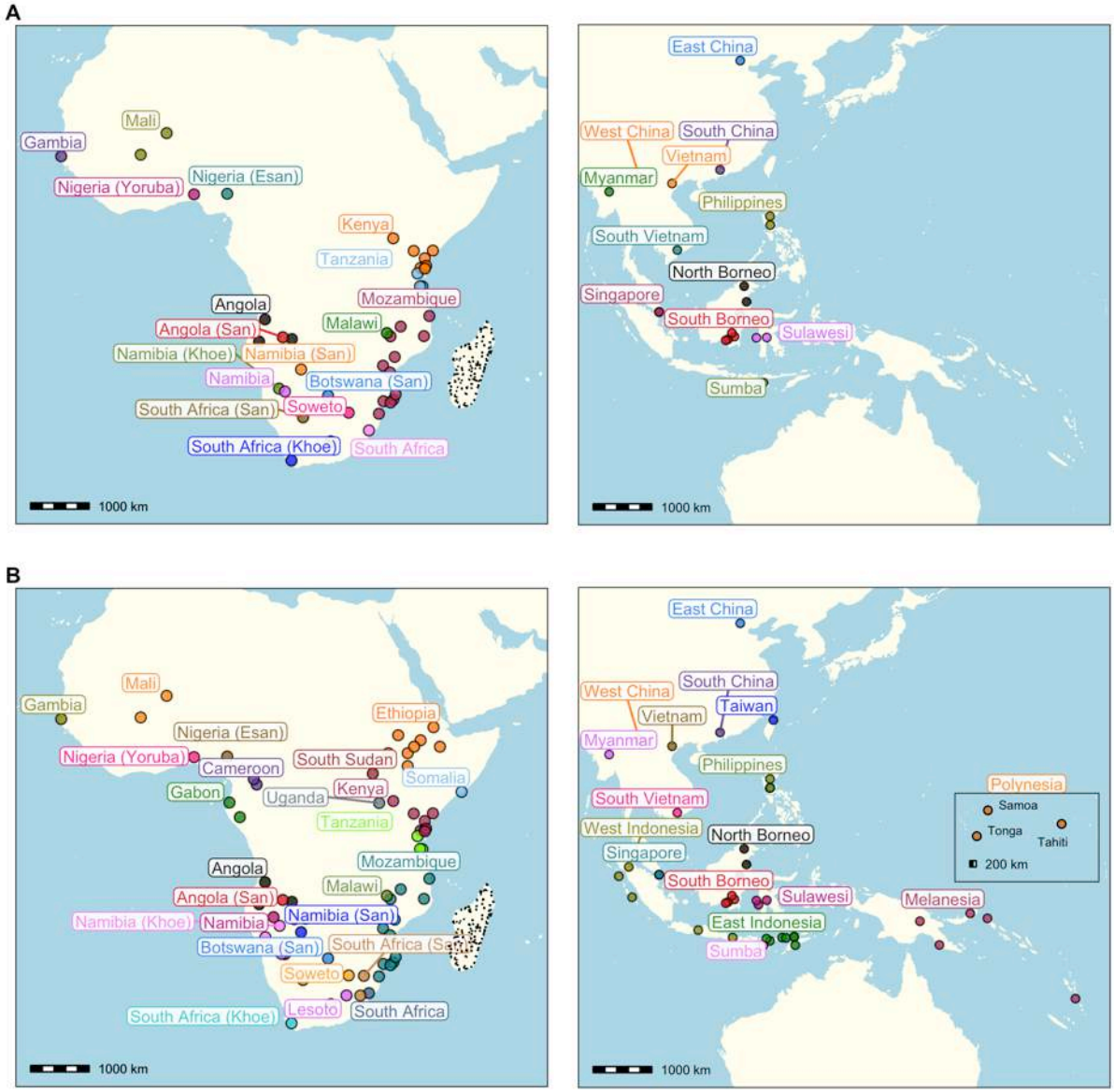


Figure S2. Boxplot of the distribution of the average number of IBD segments shared, in the density dataset, between a Malagasy individual and an individual from Africa in red (*upper panel*) or from Asia in blue (*lower panel*). For each population *IBDs are divided in into 2 length bin <5cm (left) and >5cm (right).*Comparisons were made using the density dataset, comprising 311,411 SNPs and 3168 individuals. In this study, we classified the Khoe-San populations based on the designations previously proposed[61], when referring collectively to the traditional hunter-gatherers of southern Africa as San; and when referring to traditional pastoralist groups as Khoe. Regarding the linguistic component of these populations, we notice that the Khwe population speak a click language from the Khoe subfamily.

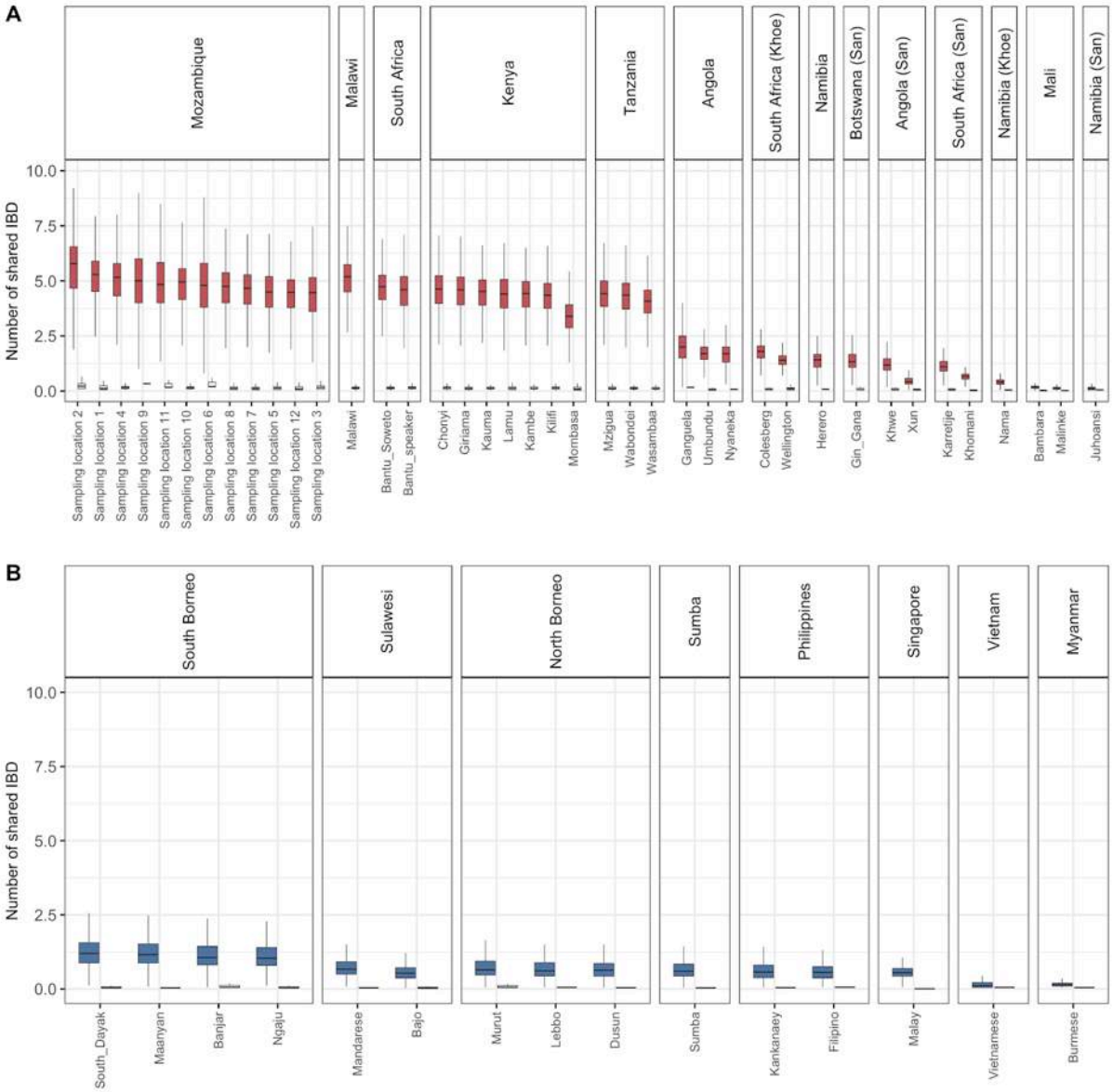


Figure S3. Boxplot of the distribution of the average number of IBD segments shared, in the diversity dataset between a Malagasy individual and an individual from Africa in red (*upper panel*) or from Asia in blue (*lower panel*). For each population *IBDs are divided in into 2 length bin <5cm (left) and >5cm (right).* Comparisons were made using the diversity dataset, comprising 78,906 SNPs and 4,164 individuals. In this study, we classified the Khoe-San populations based on the designations previously proposed[61], when referring collectively to the traditional hunter-gatherers of southern Africa as San; and when referring to traditional pastoralist groups as Khoe. Regarding the linguistic component of these populations, we notice that the Khwe population speak a click language from the Khoe subfamily.

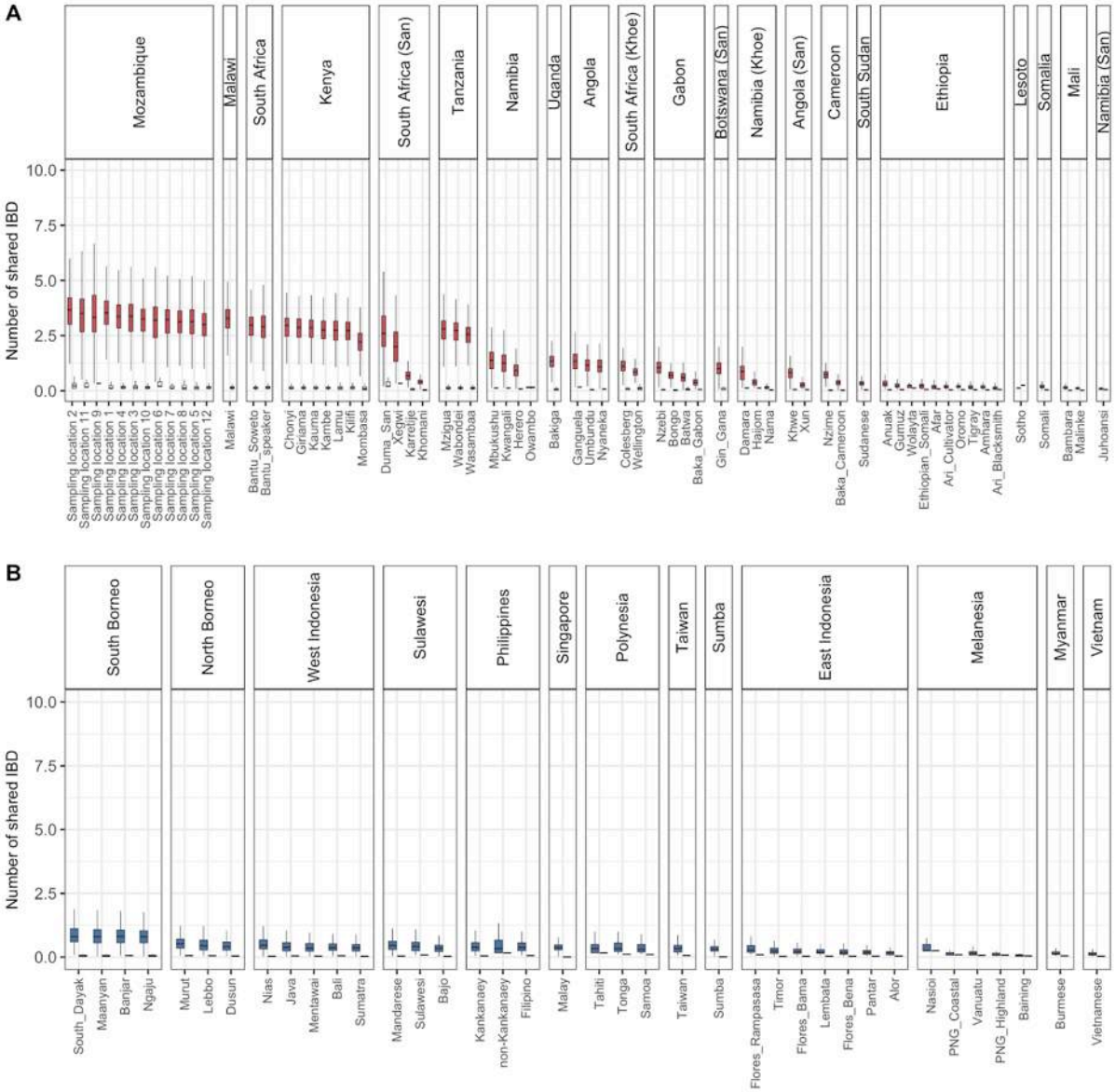


Figure S4. Distribution of the average number of shared IBD segments between two Malagasy individuals, according to the inferred continental ancestry of each IBD segment. A) boxplot of the global number of shared IBD B) Distribution of shared IBD according their size. We estimated the ancestry of each IBD segment by mapping its coordinates to local ancestry assignments from sample haplotypes. The ancestry proportion was calculated using the number of Asian (ASI) sites divided by the total number of SNPs spanning the IBD region in both haplotypes. IBD segments were classified as African (AFR SNPs ≥ 90%), Heterogenous (10% < ASI SNPs < 90%) or Asian (ASI SNPs ≥ 90%).

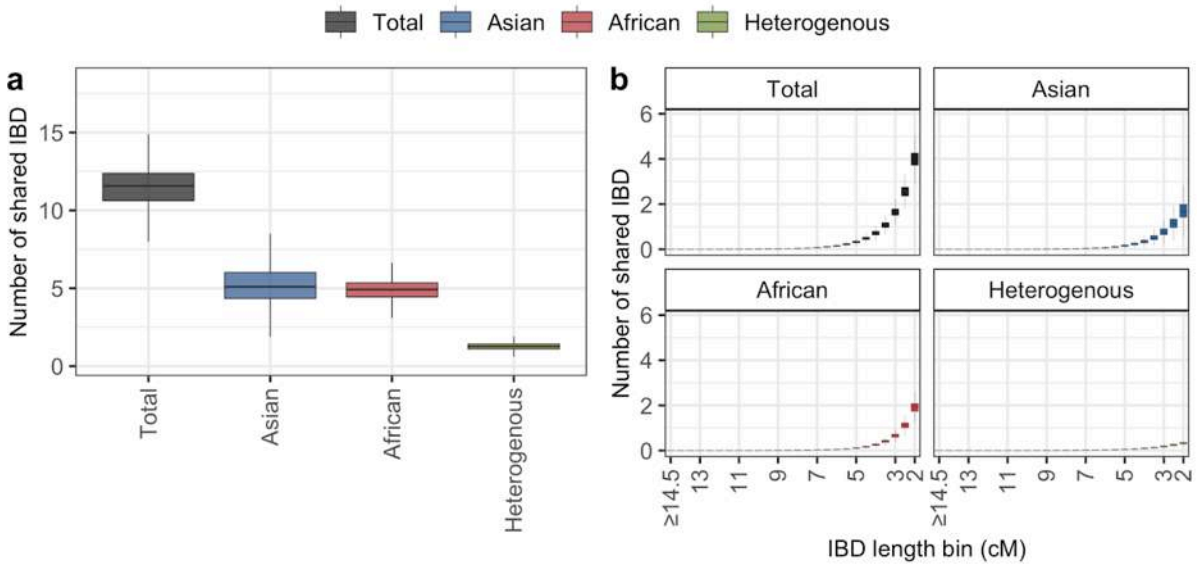


Figure S5. Average number of shared IBD segments between two Malagasy individuals according to ancestral origin and genetic clustering across Madagascar. Each panel depicts the location of a genetic grouping of individuals as identified previously[12], and boxplots of the distribution of shared IBD segments per effective pair of individuals, according to their continental ancestry (African or Asian) and genetic length (IBD < 5cM or IBD ≥ 5cM). We normalized the ancestry-specific numbers of shared IBD fragments by taking into account the reduced fraction of the genome represented by the specific ancestry.

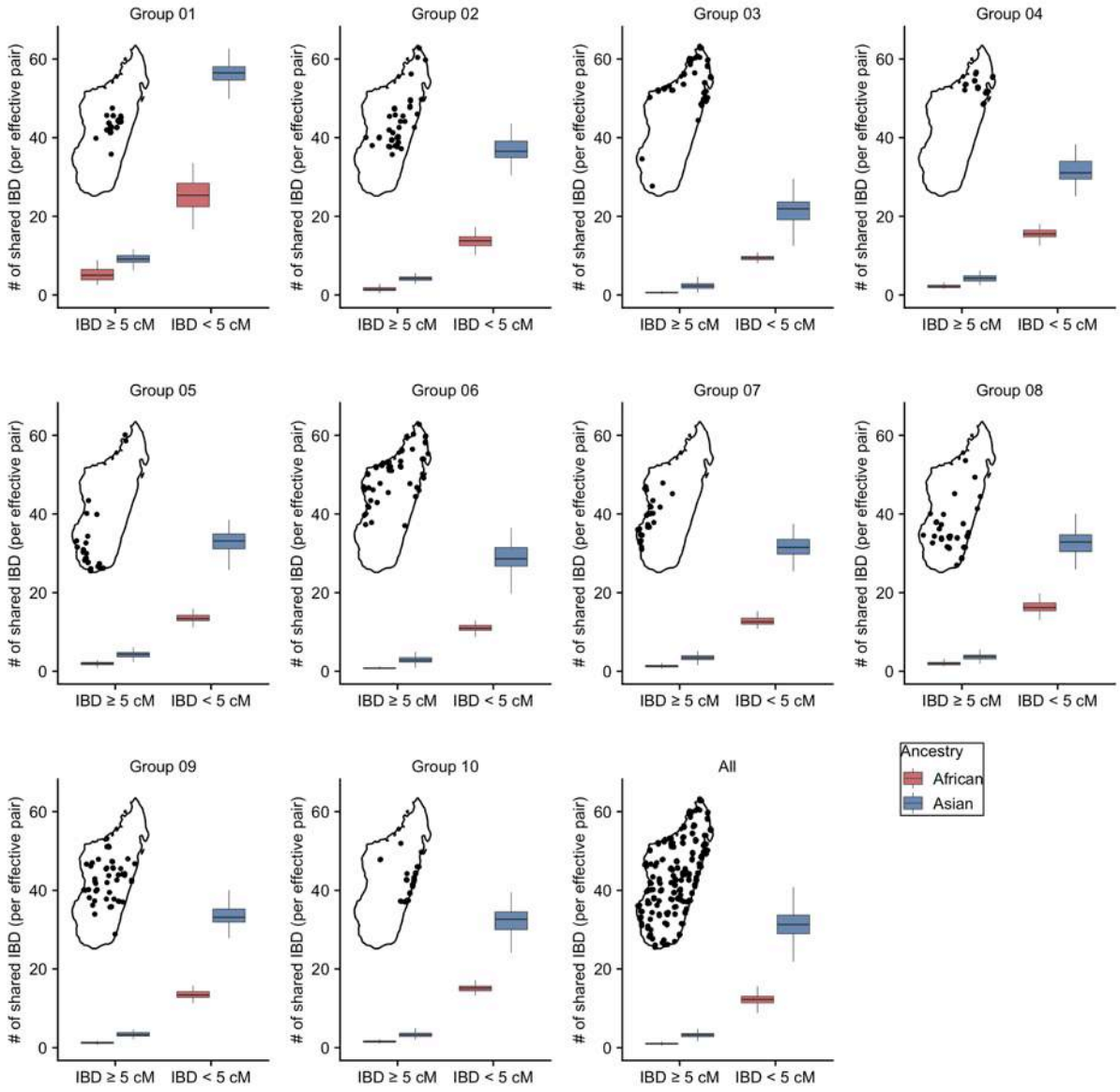


Figure S6. Replication analysis of pairwise IBD sharing in Madagascar, according to inferred local ancestry based on LOTER and Germline methods. (A and B) Boxplot of the distribution of shared IBD segments in Madagascar and the closest genetic population in Africa and Asia (left and right panel, respectively) according to genetic length (< 5cM or ≥ 5cM). **(C and D)** IBD-sharing distribution according to continental origin and length in centiMorgans (cM), along with the distribution detected in the Malagasy's closest genetic populations inhabiting Africa and Asia (right and left panel, respectively). The x-axis represents haplotype length cM, and the age of IBD haplotypes (in generations before present) are labelled on the superior horizontal axis. The number of segments shared per effective pair is shown on the vertical axis (y axis). The yellow dashed line shows the time since admixture in Madagascar estimated by MALDER. Germline and LOTER were used for IBD determination and local ancestry inference, respectively.

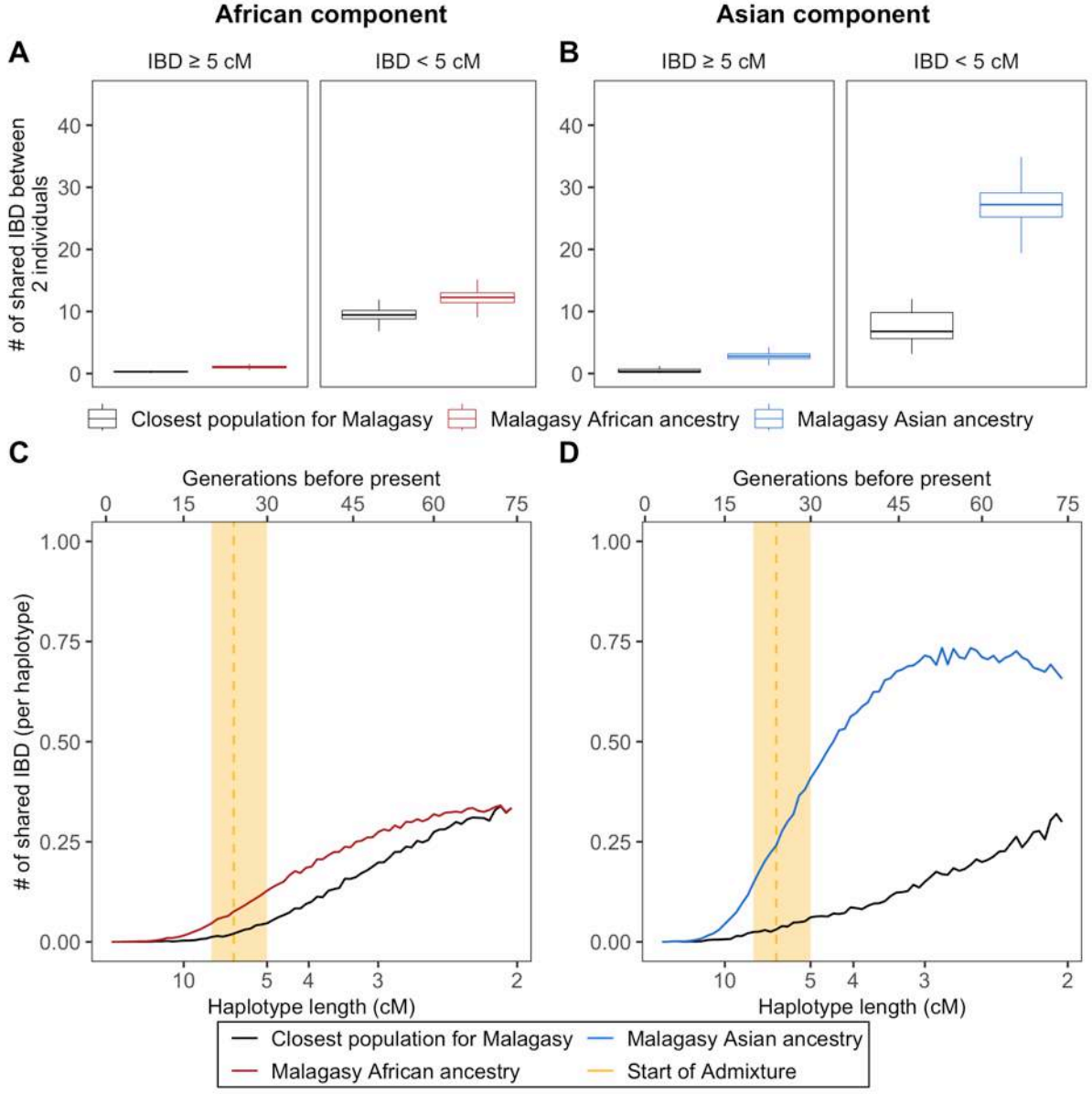
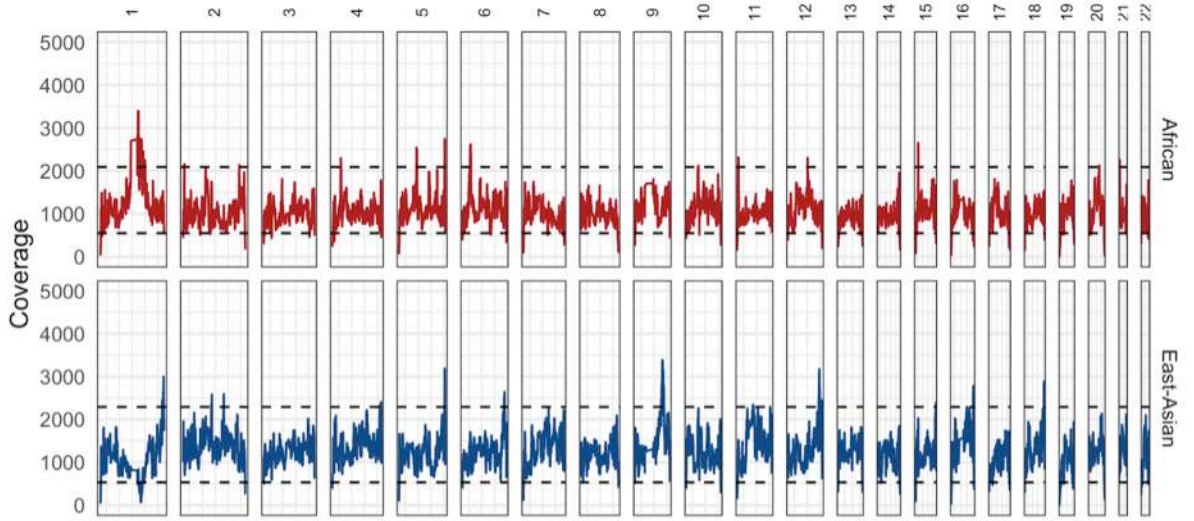


Figure S7. Distribution of IBD segments across the genome of Malagasy individuals. A) global coverage of shared IBD by genomic locus according to continental ancestry (African in red and Asian in blue). Using the R package "Genomic ranges", we estimated for each locus (x-axis) the ancestry-specific coverage from detected IBD segments (y-axis). Upper and lower black dotted lines represent the 0.99 and 0.01 quantiles of data, respectively. B) Average number of shared IBD segments between two Malagasy individuals according to ancestral origin and for each chromosome. Each panel represents a chromosome, showing the boxplot of the distribution of shared IBD segments per effective pair of individuals, according to continental ancestry (African or Asian) and genetic length (IBD < 5cM or IBD ≥ 5cM). We normalized the ancestry-specific number of shared IBD fragment by taking into account the reduced fraction of the genome represented by each specific ancestry.

A)



B)

Figure S8. Estimated ancestry-specific effective population size (Ne) for Madagascar over the last 100 generations. Related Malagasy samples (k > 0.0442) were removed from the data. Using IBDNe, ancestry specific and overall effective population sizes were estimated. The y-axes show Ne values, plotted on a log scale. The x-axes show generations before present. **(A)** Historical effective population size trajectory for Madagascar. **(B)** Historical Ne trajectories for individual genetic groups described previously[12]. For both (A and B), the lines show estimated effective population sizes. The colored regions represent the 95% bootstrap confidence interval.

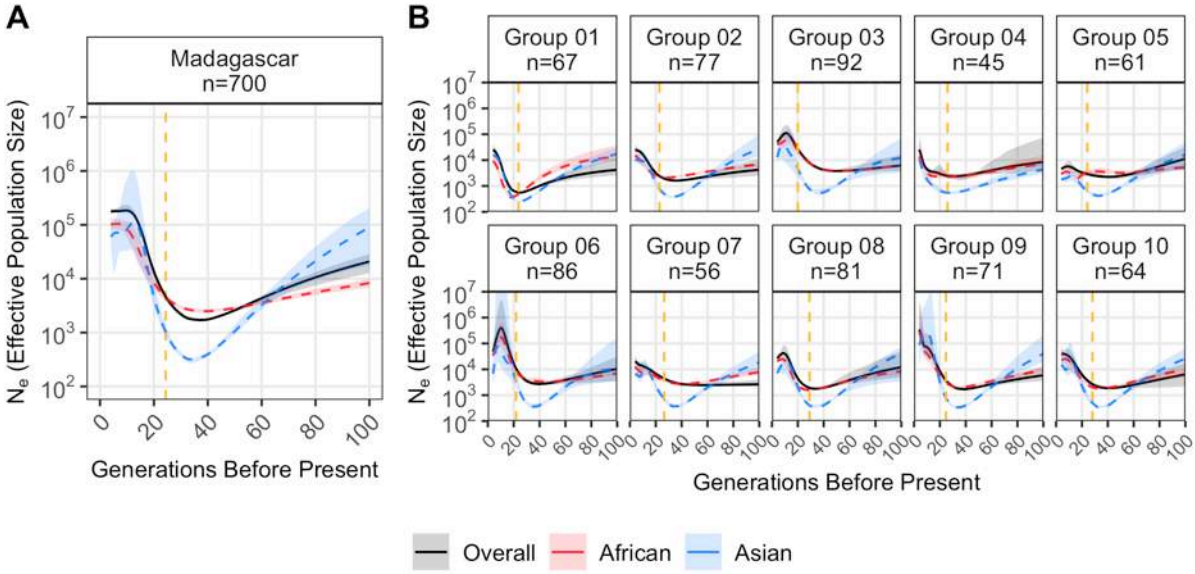


Figure S9. Estimated effective population size (Ne) over the last 100 generations for the African populations analysed in this study. The red line in each plot shows the estimated effective population size, and the coloured regions are bootstrap 95% confidence intervals. The y-axes (effective population size) are plotted on a logarithmic scale.

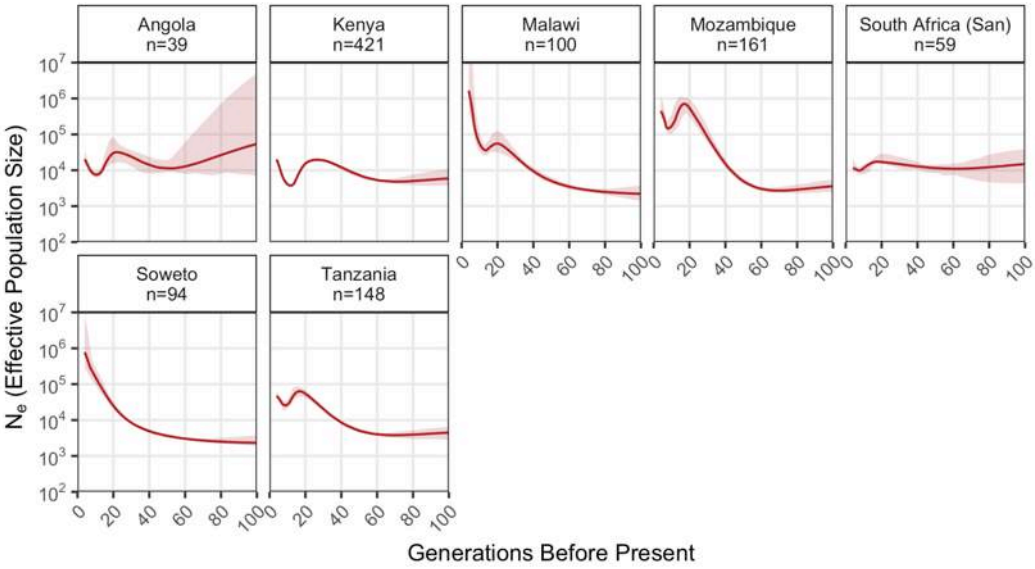


Figure S10. Estimated effective population size (Ne) over the last 100 generations for the Asian populations analyzed in this study. The blue line in each plot shows the estimated effective population size, and the coloured regions are bootstrap 95% confidence intervals. The y-axes (effective population size) are plotted on a logarithmic scale.

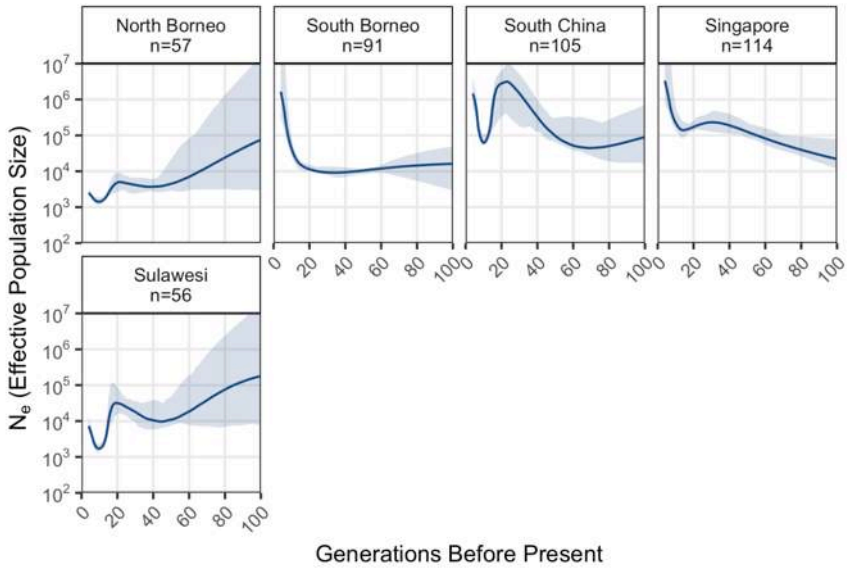
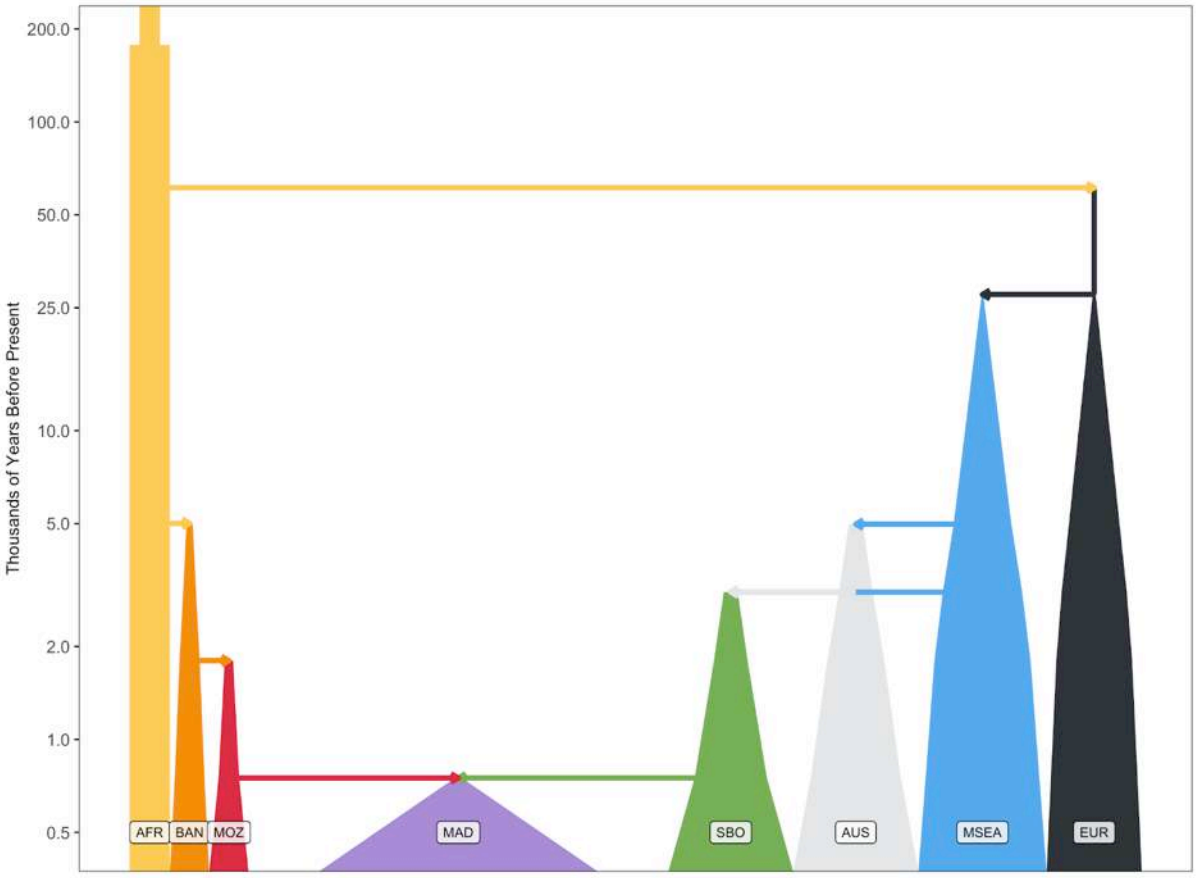


Figure S11. Demographic parameters used for Malagasy demographic history. We used Msprime[27] (Kelleher et al., 2016) to simulate three continental populations (AFR, EUR and ASI), two Bantu populations (BAN, MOZ), two Austronesian populations (AUS and SBO) and an admixed population (MAD). Our pre-admixture model for the African (AFR), European (EUR) and Asian (ASI) continental populations is based on a published model inferred from the 1000 Genomes project data[28] and also described elsewhere[14]. This model simulates the advent of modern humans, the Out of Africa (OoA) event, and the Eurasian bottleneck after the OoA and the split between Asian and European populations. Regarding the Bantu-speaking and Austronesian-speaking expansion, we simulated an initial Bantu population (BAN) that diverged from the African continental population and an Austronesian population (AUS) diverging from the Asian continental population (ASI). We approximated a South Borneo population (SBO) by simulating a population that diverged from AUS at 100 generations BP; we approximated the time of Bantu occupation of eastern Africa by simulating a Mozambique population (MOZ) splitting from BAN at 60 generations BP. Finally, we simulated an admixture event between MOZ and SBO populations at a proportion of 5/8 and 3/8, in order to reach the global ancestry proportions observed in Madagascar: 1/3 Asian ancestry and 2/3 African ancestry. A) illustration of the model in a direct admixture case, Figure S11 B) illustration of the model in a post bottleneck admixture case.

A)



B)

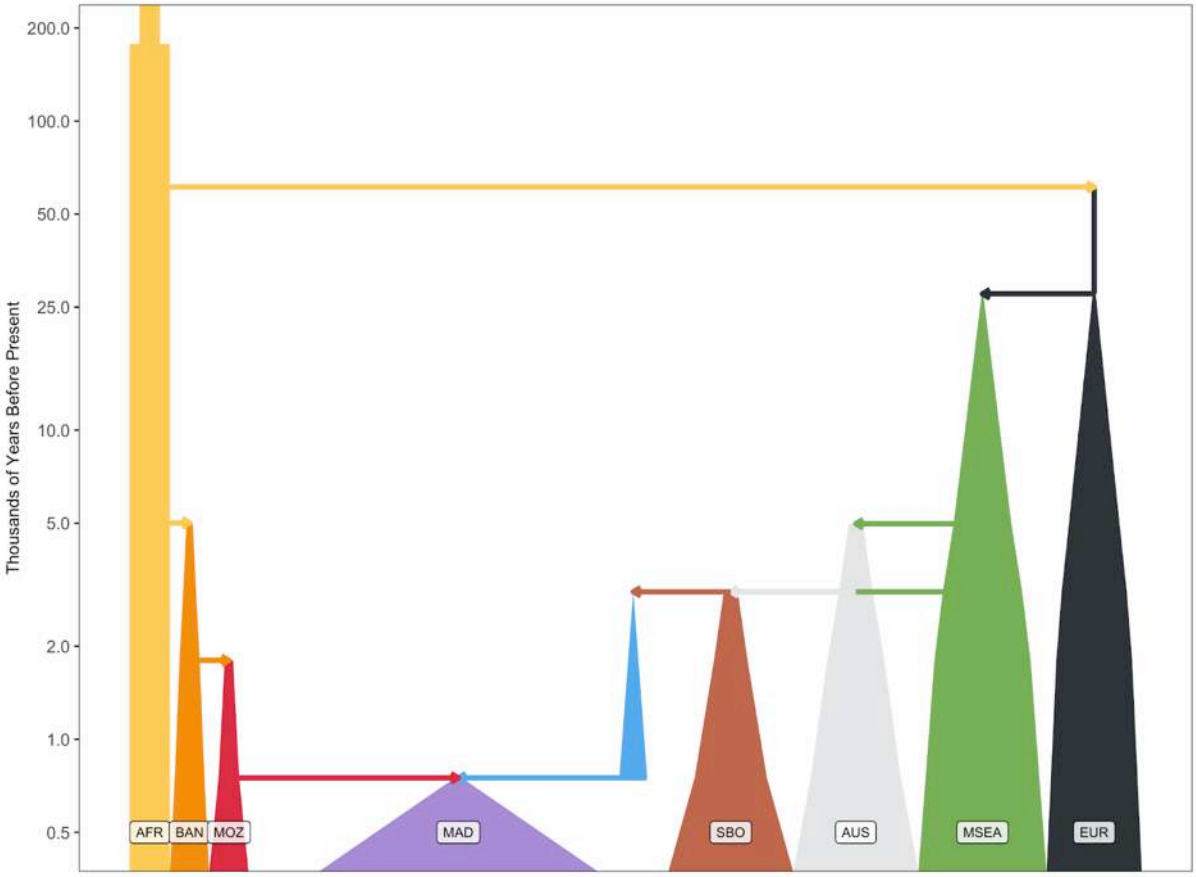
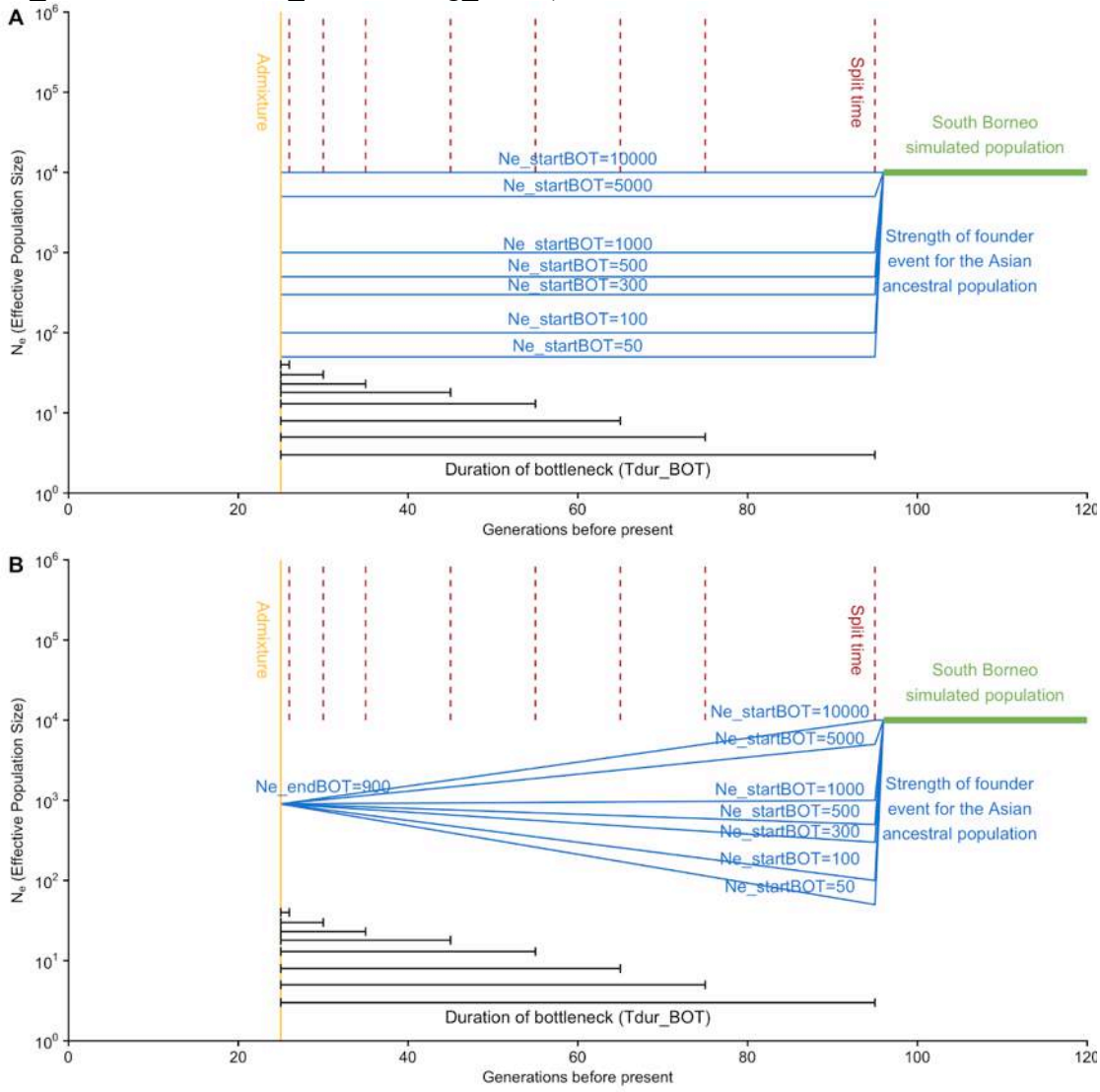


Figure S12. Graphical summary of simulated demographic scenarios for the Malagasy ancestral Asian population. A) Exploratory step with population with constant size B) Exploratory step with population in expansion or decline C) Compatibility step**.** Given that the effective size of the Malagasy's ancestral Asian population underwent a severe decline for several generations, we hypothesize that many of the possible coalescences between haplotypes occurred in the pre-admixture period through founder effect or bottleneck events. In the exploratory step (A and B), we simulated an admixture event (immediate admixture at 25 generations BP or continuous gene flow between 20-30 generations BP) for particular scenarios regarding the Asian ancestral population. We varied the strength of founder events (Ne_startBOT= 50, 100, 300, 500, 1000, 5000 or 10000 individuals) at the different split dates from South Borneo, the duration of the bottleneck (Tdur_bot = 1, 5, 10, 20, 30, 40, 50, or 70 generations before the admixture), and the influence of gene flow (1, 5 or 10 migrants per generation). We allowed for the possibility of changes in demography during the bottleneck by adding a constant Ne *Ne_endBOT*= Ne_startBOT**)** or an expansion/decline in Ne **(***Ne_endBOT= 900***)**. In the compatibility step, we simulated 6 scenarios (Table S4) 1: Funder event (Ne_endBOT=900, Ne_startBOT=900, Tdur_bot=0, Mig_bot=0), 2: Strong funder event (Ne_endBOT=900, Ne_startBOT =50, Tdur_bot=1, Mig_bot =0), 3: Multiple Asian waves, 4: Slow decrease of a large population (Ne_startBOT =5,000 ,Tdur_bot=50, Mig_bot =0, Ne_endBOT=900), 4: Bottleneck (Tdur_bot=10, Mig_bot =0), reaching

Ne_endBOT==900), 5: long-term bottleneck Ne_startBOT =500, Ne_endBOT=500,Tdur_bot=40, Mig_bot=0).



C

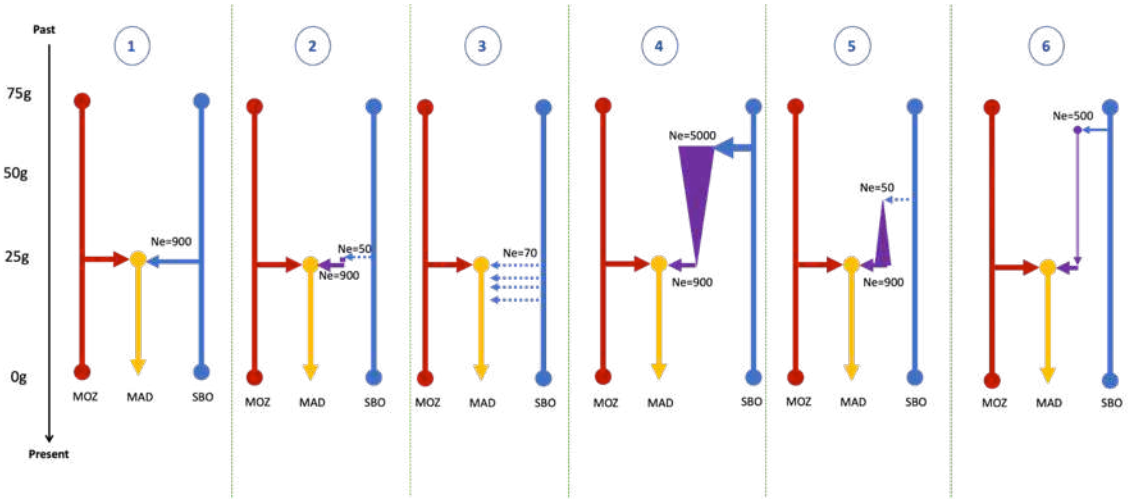
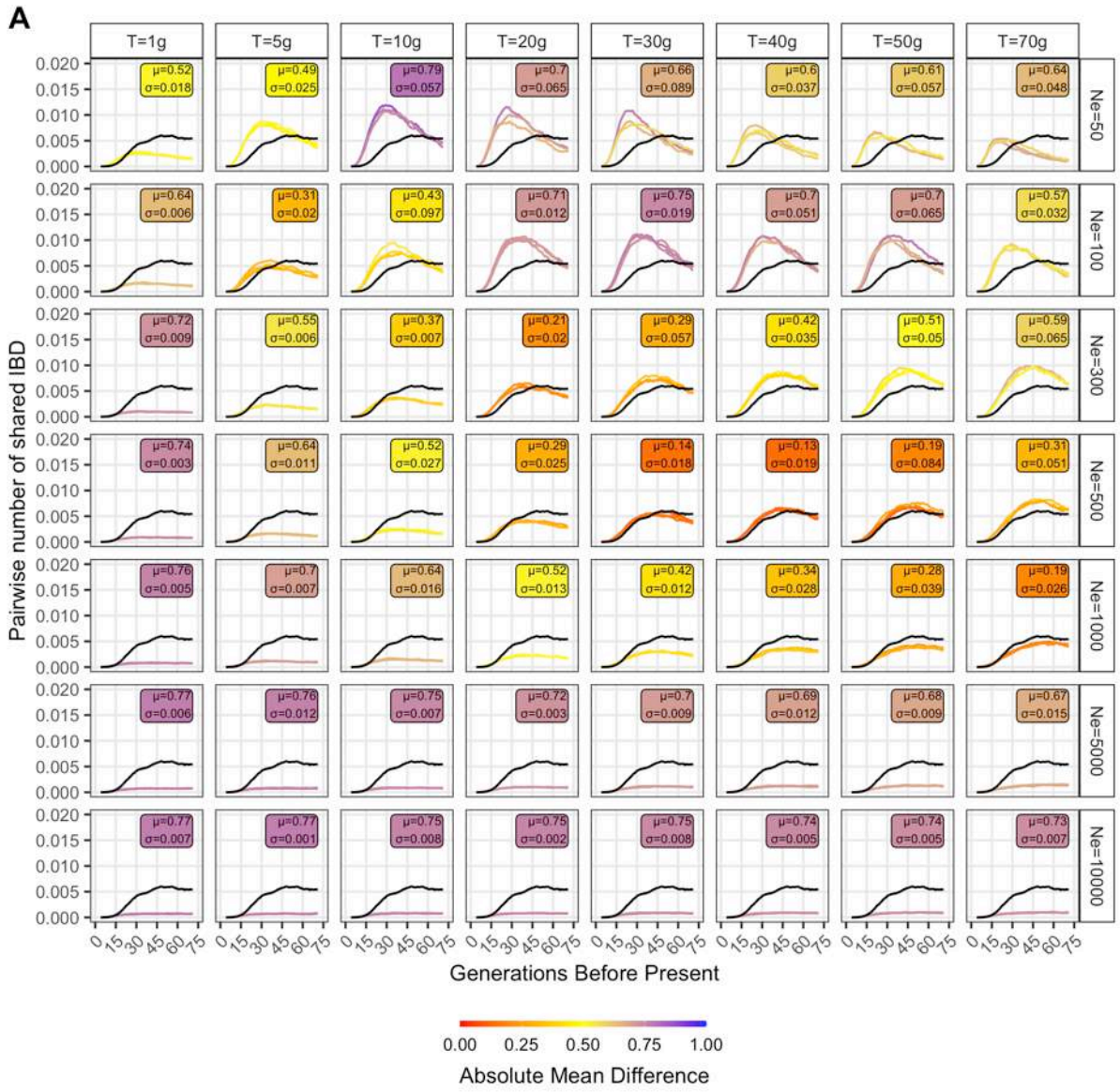
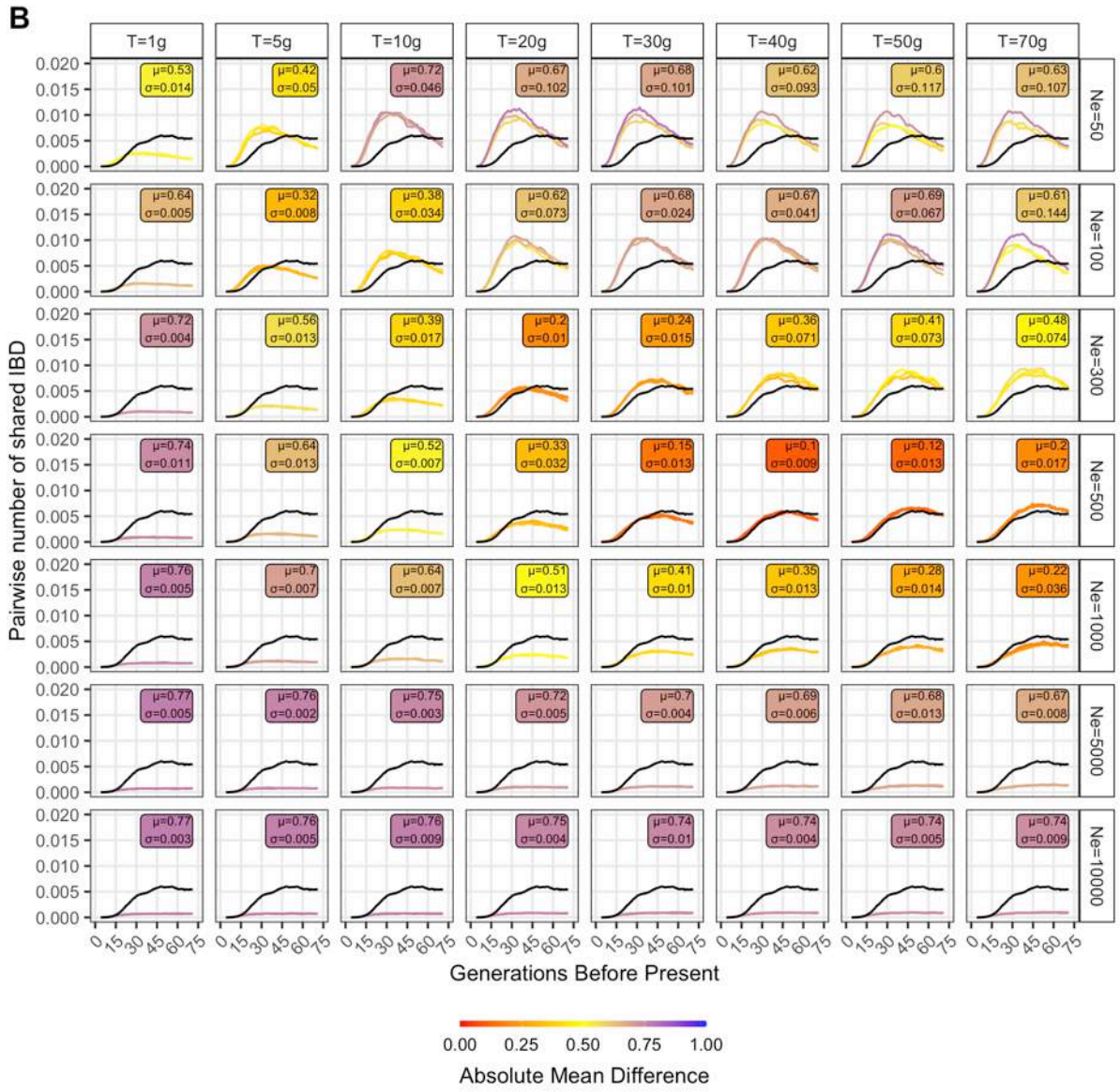
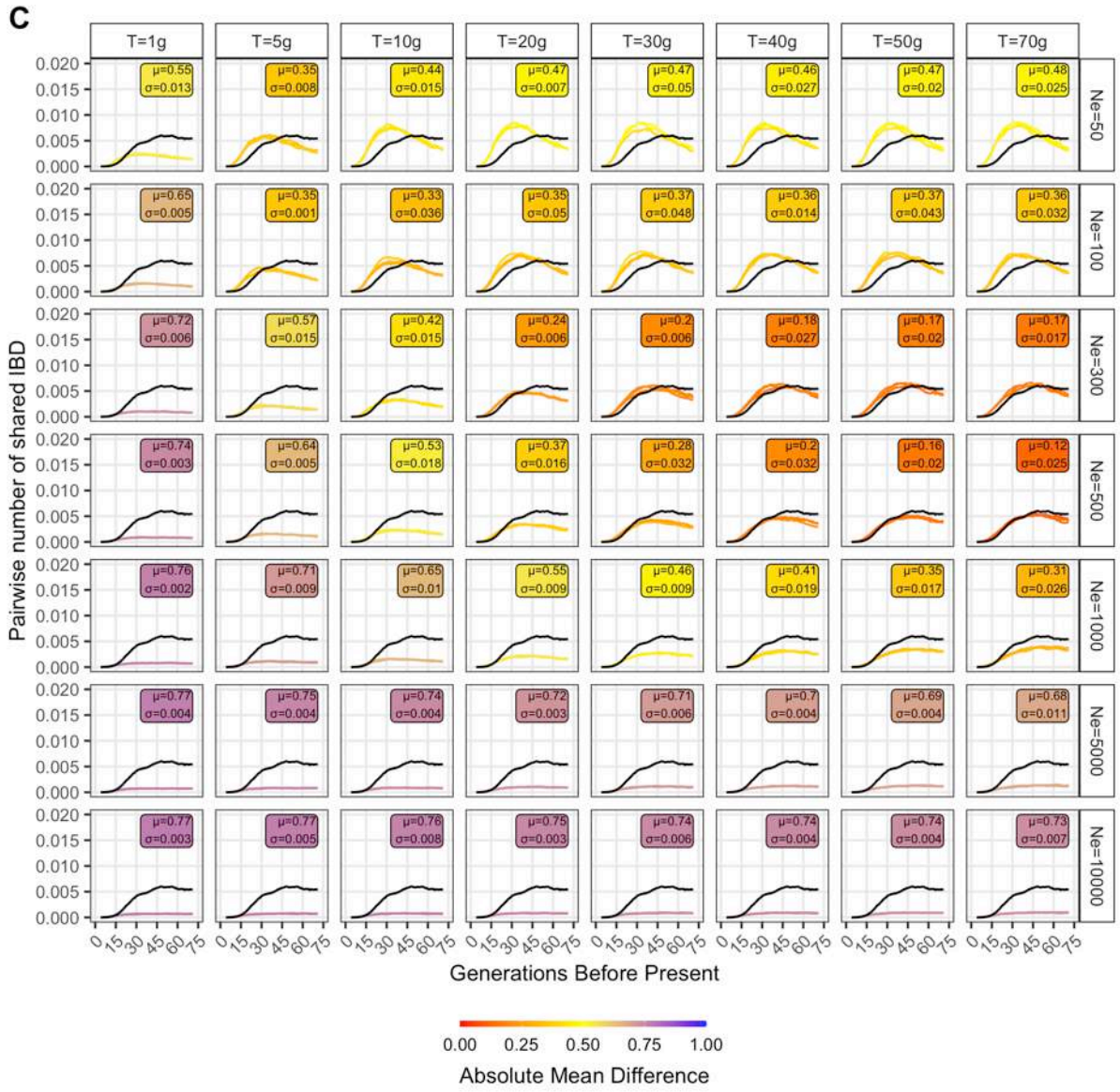
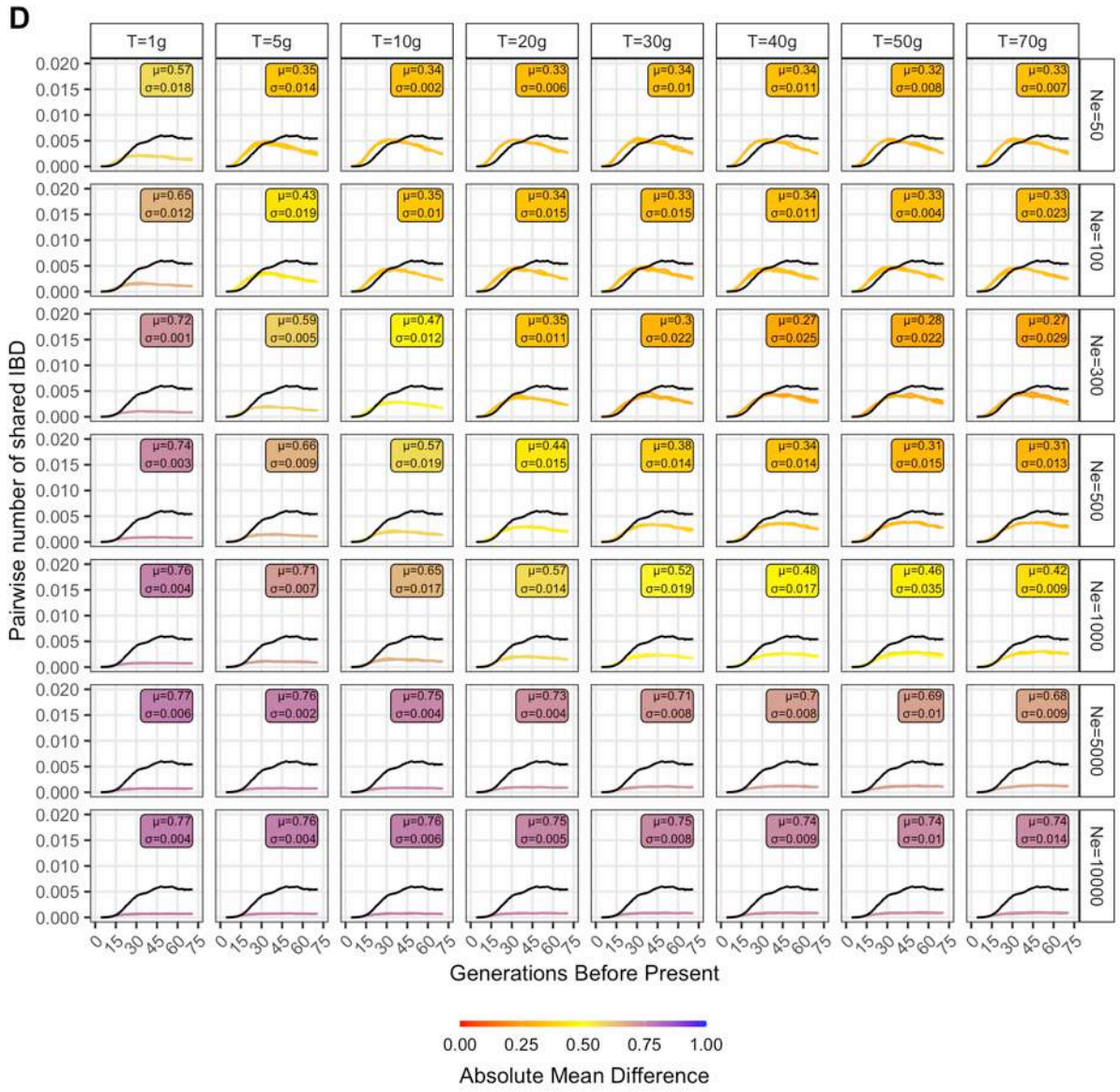


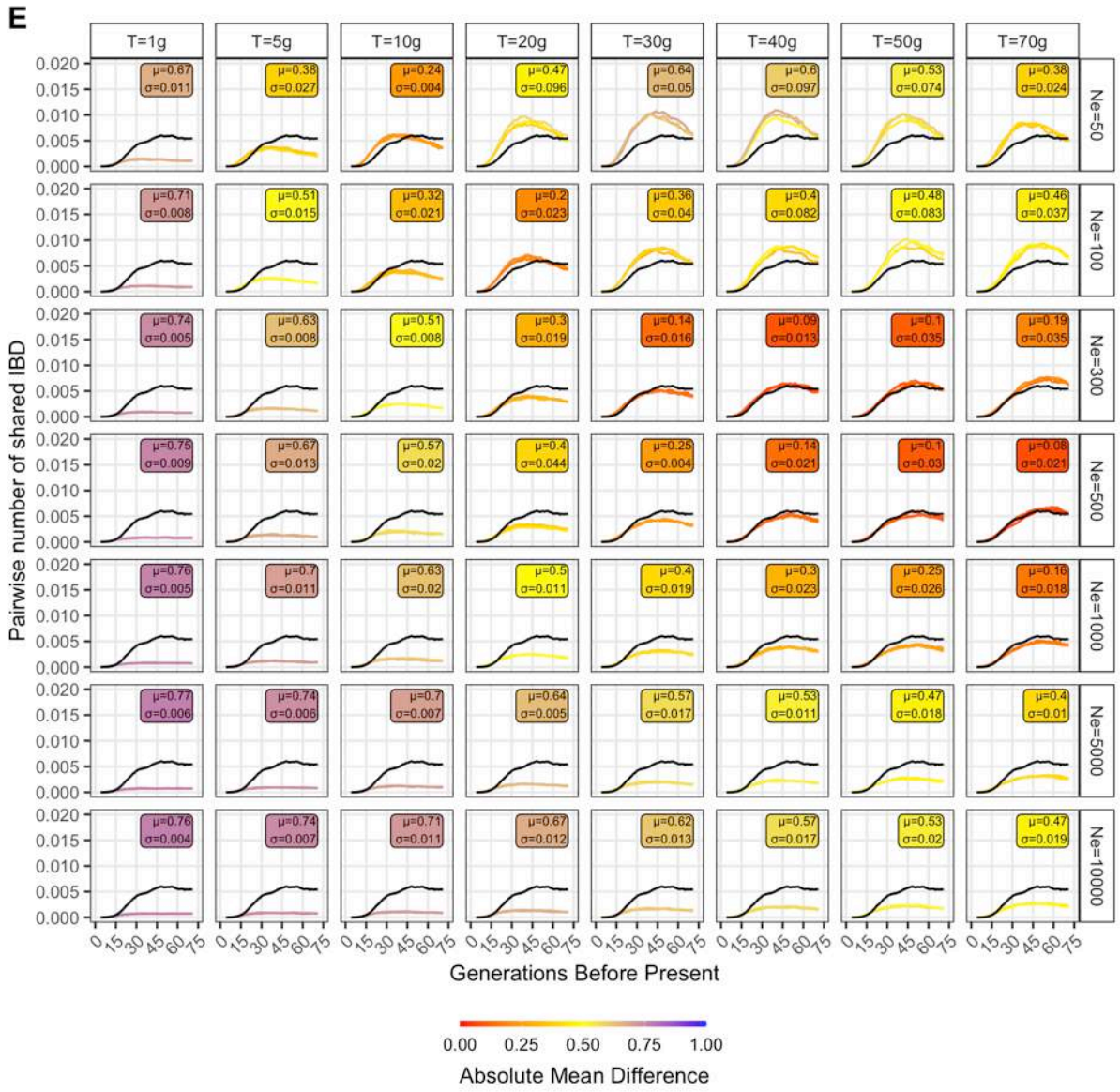
Figure S13. Effect of different demographic scenarios on the simulated IBD segments of Asian origin shared in Madagascar. Each panel shows the result for simulated scenarios, according to the strength of founder events (Ne = 50, 100, 300, 500, 1000, 5000 or 10000 individuals) and the duration of the bottleneck (T = 1, 5, 10, 20, 30, 40, 50, or 70 generations before the admixture) for the Malagasy Asian ancestral population. The x-axis represents the age of IBD haplotypes (in generations before present). The number of segments shared per pair is shown on the y axis. Coloured lines show the pairwise IBD-sharing from simulations at each generation, while black lines represent the observed pairwise IBD-sharing. The average score and standard deviation using 3 replicates are shown in the upper-right legend. **(A to D)**. Before the admixture event, we simulated an isolated Austronesian population with stable Ne during the bottleneck. Migration rates to the isolated population were set to 0 (A), 1 (B), 5 (C) and 10 (D) individuals per generation. **(E to H)** We simulated an isolated Austronesian population with changing Ne during the bottleneck. Migration rates to the isolated population were set to 0 (E), 1 (F), 5 (G) and 10 (H) individuals per generation. **(I and J)** We simulated a South Bornean population with initial Ne=5,000 and growth rate of 0.0219. From this population, an Austronesian population diverged with stable Ne during a bottleneck event. Migration rates to the isolated population were set to 0 or 1 individuals per generation (I and J, respectively). **(K and L)** We simulated a South Bornean population with initial Ne=5,000 and growth rate of 0.0219. From this population, an Austronesian population diverged with changing Ne during a bottleneck event. Migration rates to the isolated population were set to 0 or 1 individuals per generation (K and L, respectively). **(M and N)** Admixture was simulated under the Continuous Gene Flow model, with the Austronesian ancestral population receiving African gene flow during the period 20-30 generations BP. Before the admixture event, we simulated a bottleneck event for the Austronesian population with stable (M) or changing (N) Ne during the bottleneck. Migration rates to the isolated population were set to 0 individuals per generation.

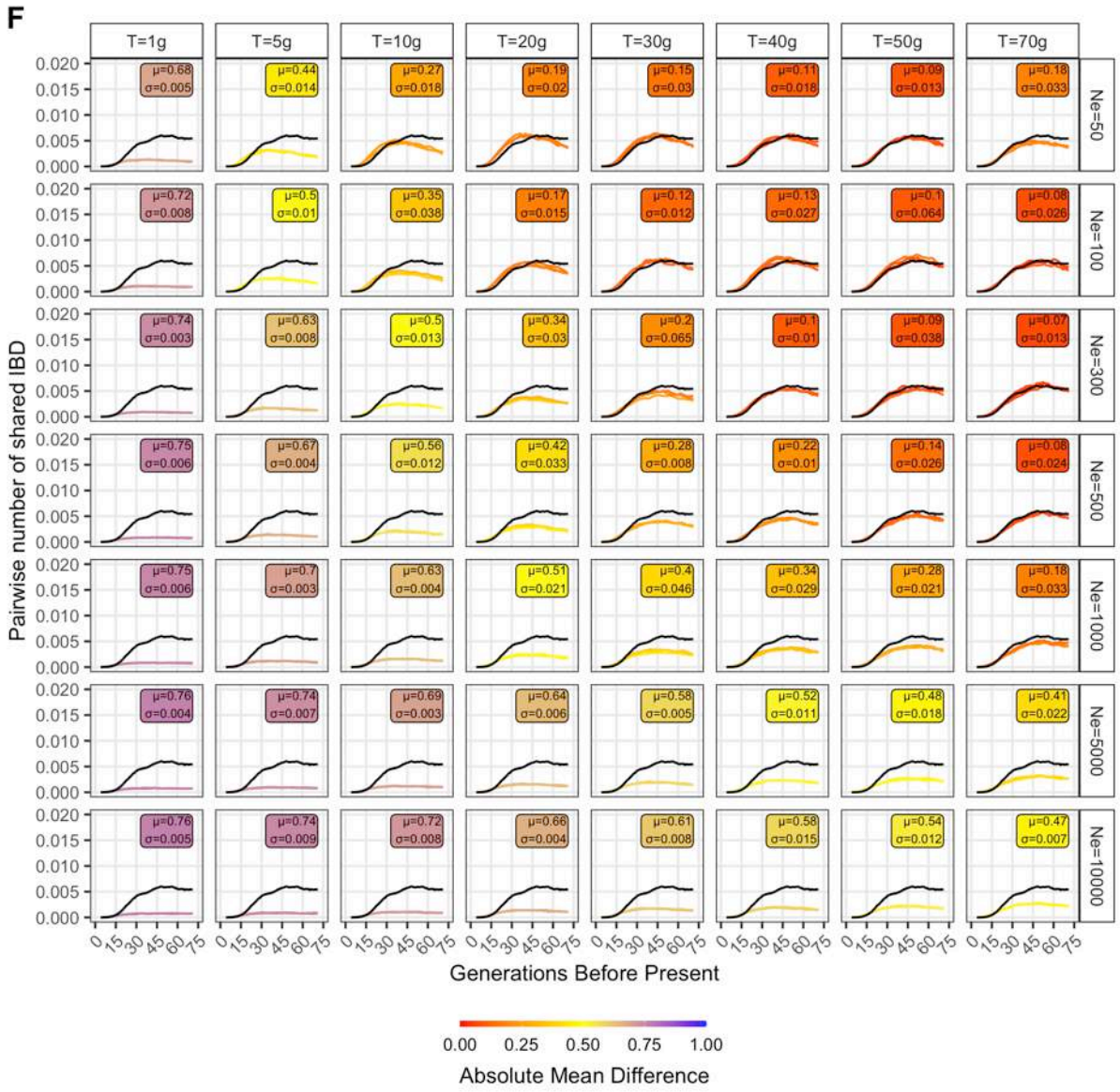

A
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

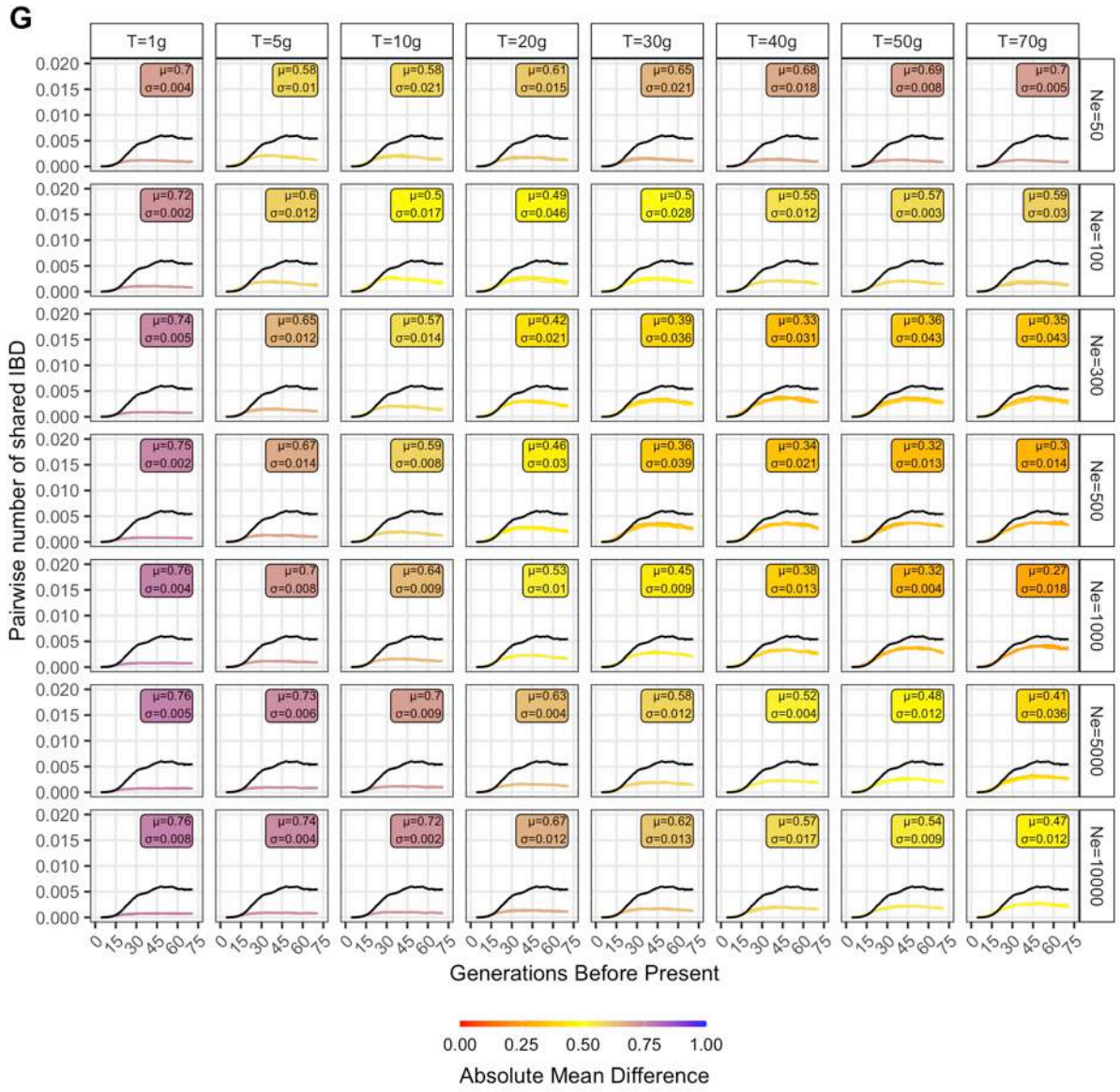

B
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

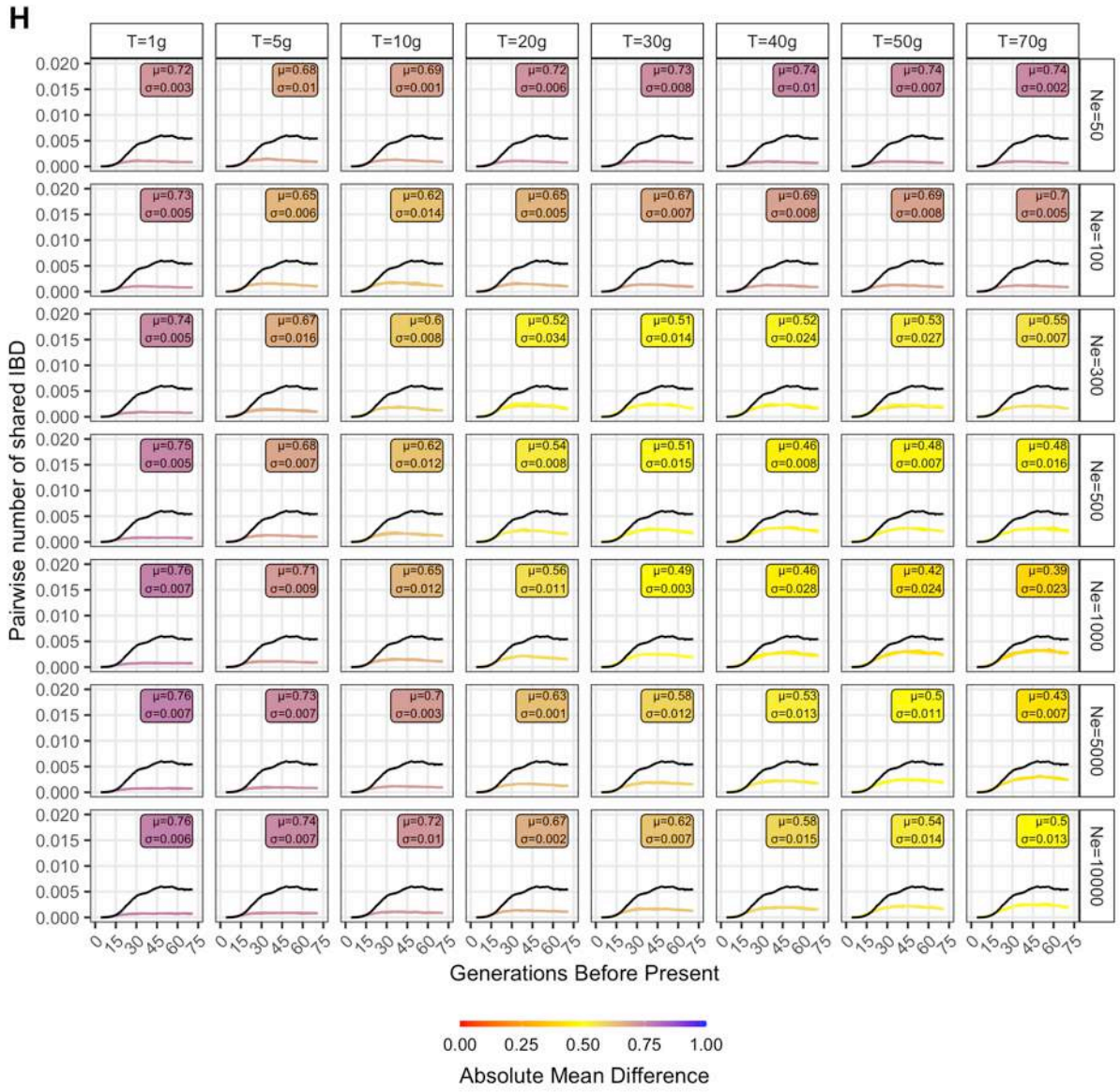

C
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00


D
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

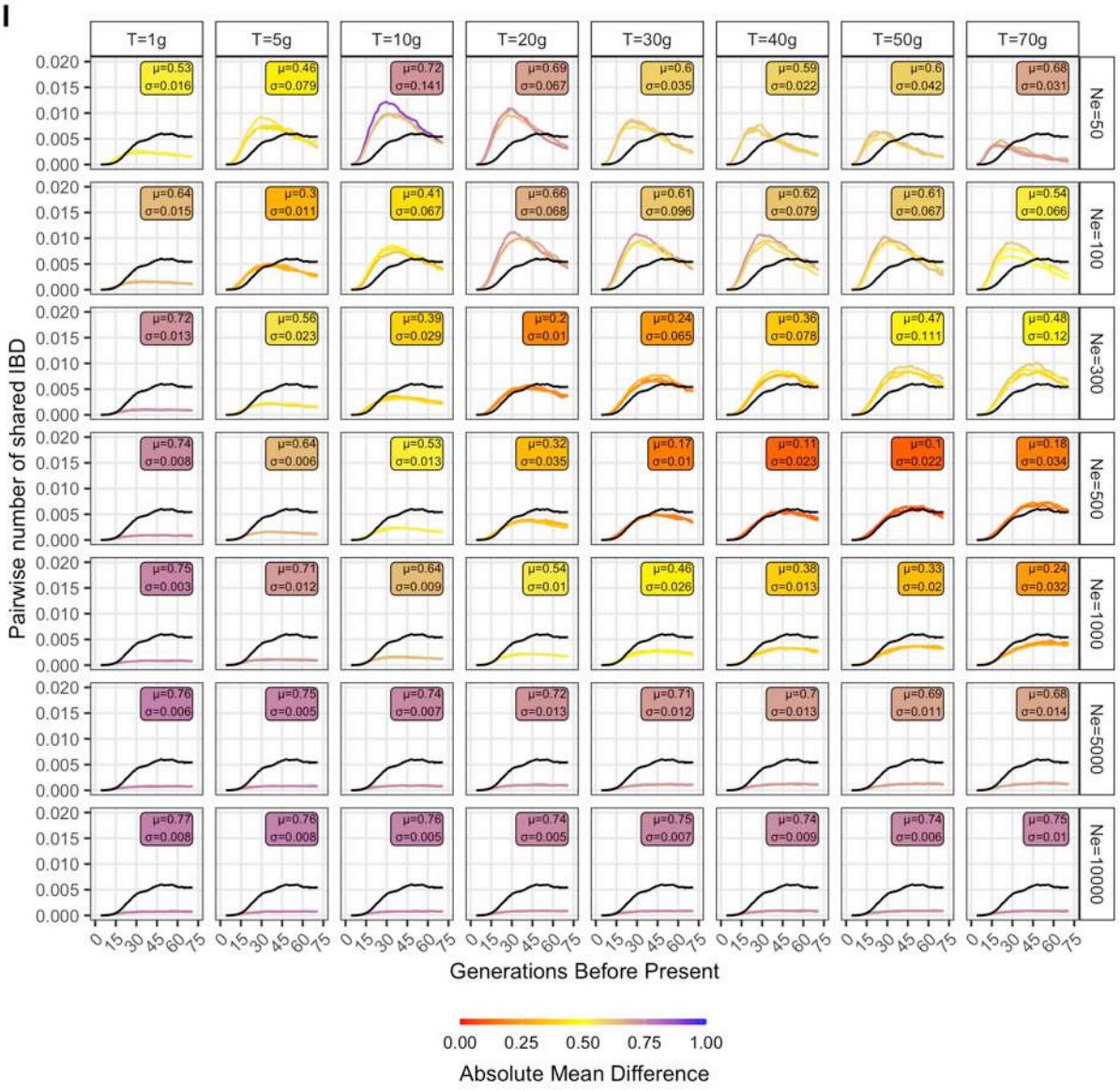

E
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

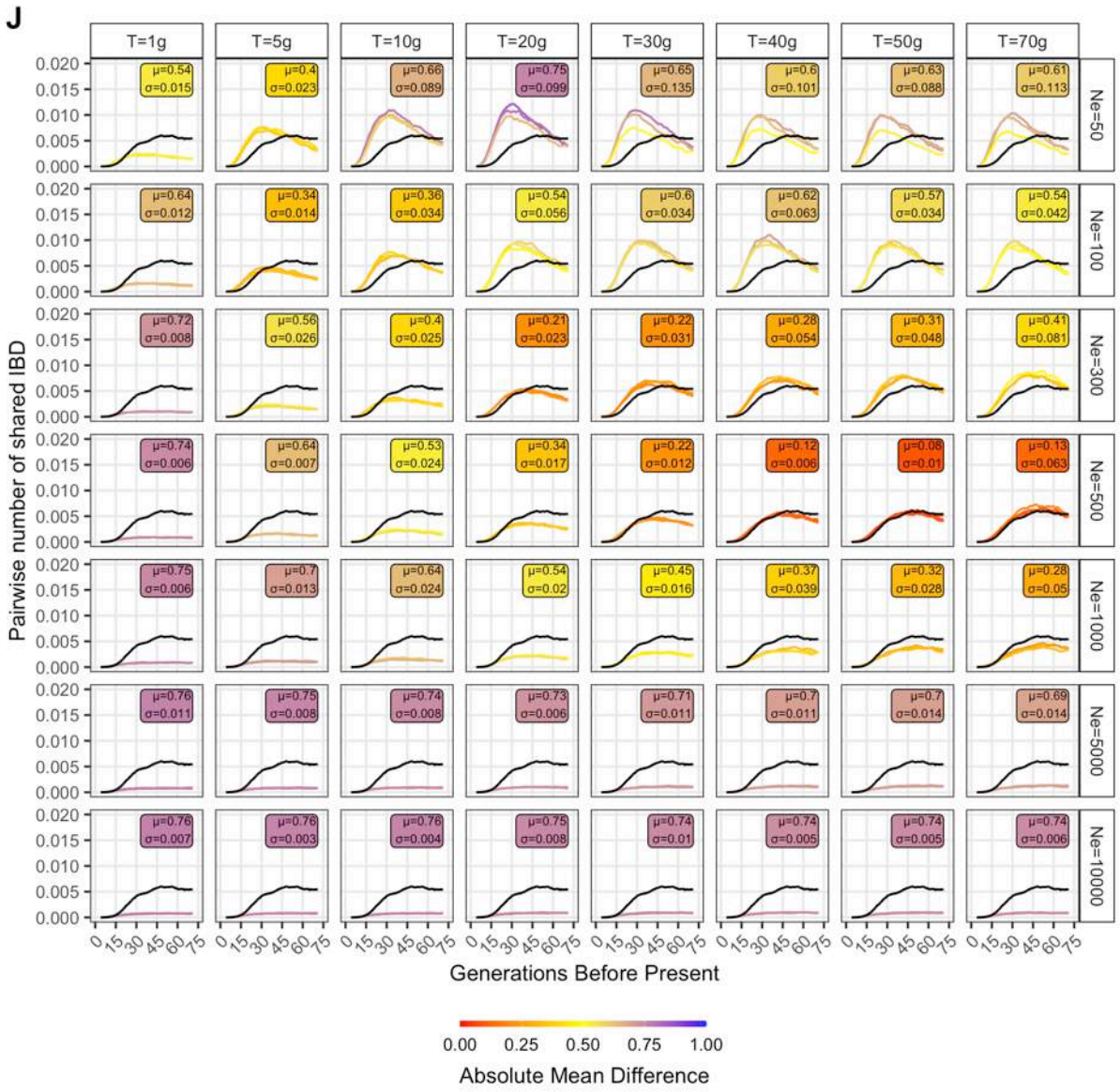

F
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

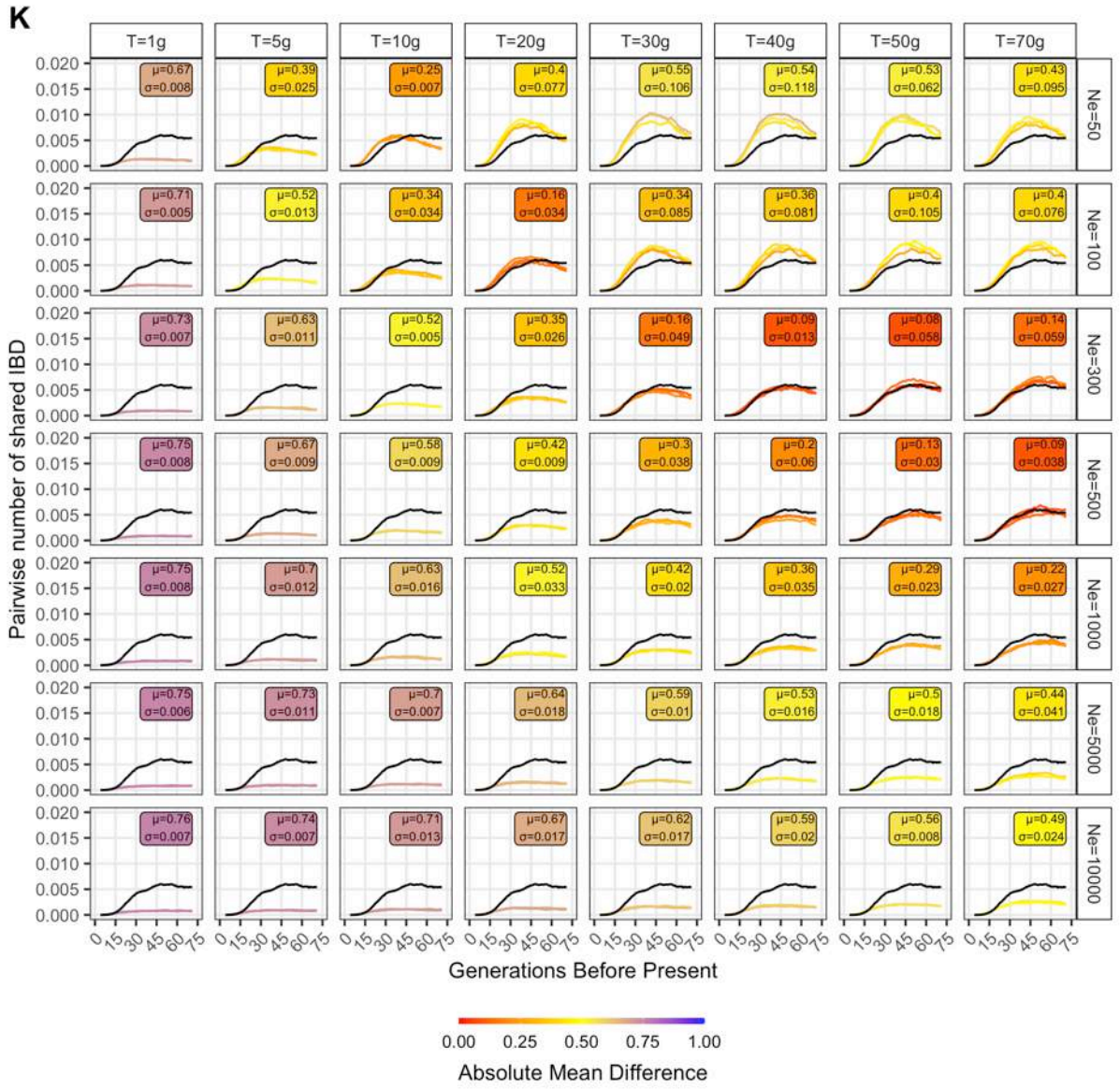

G
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

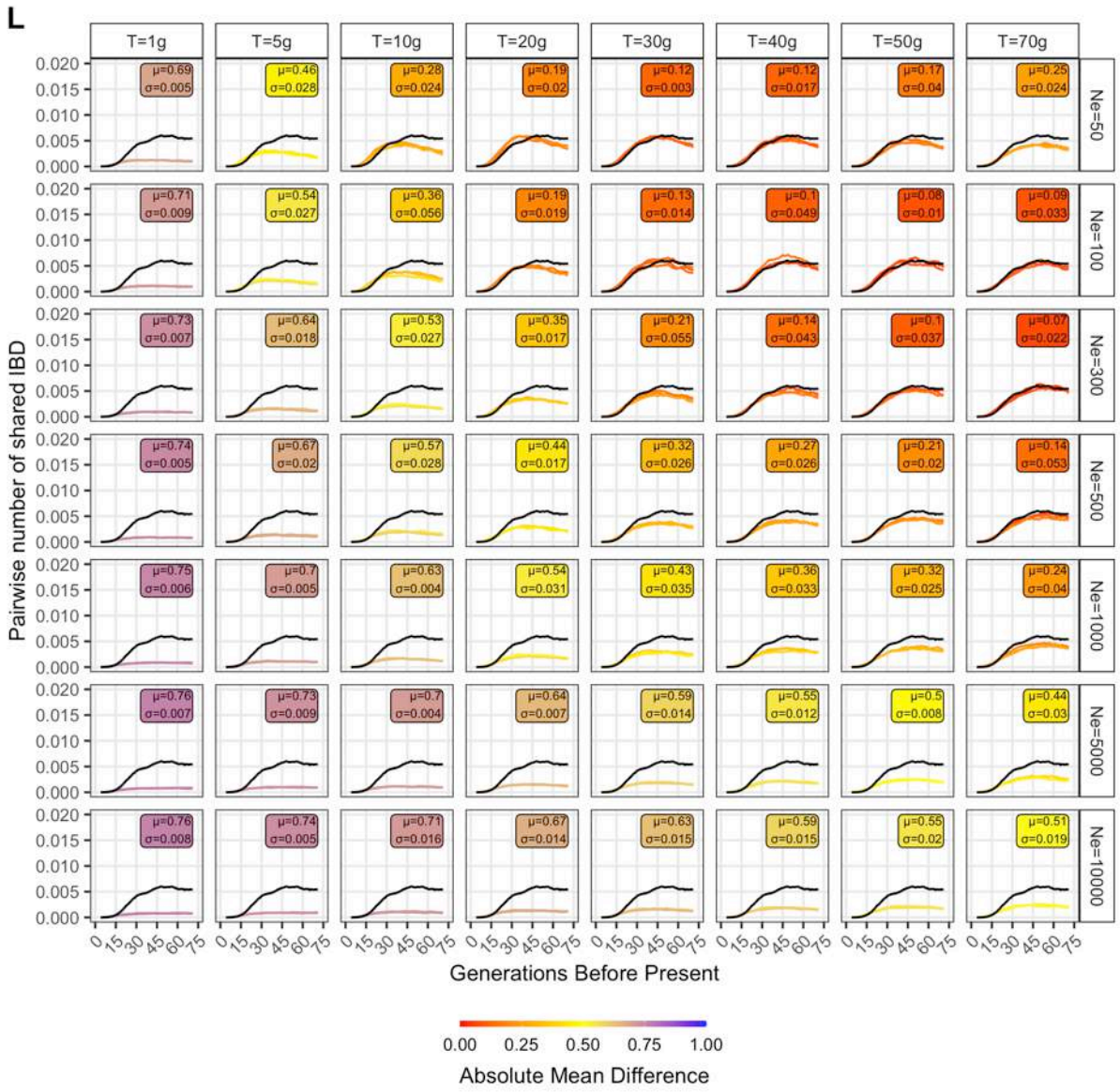

H
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

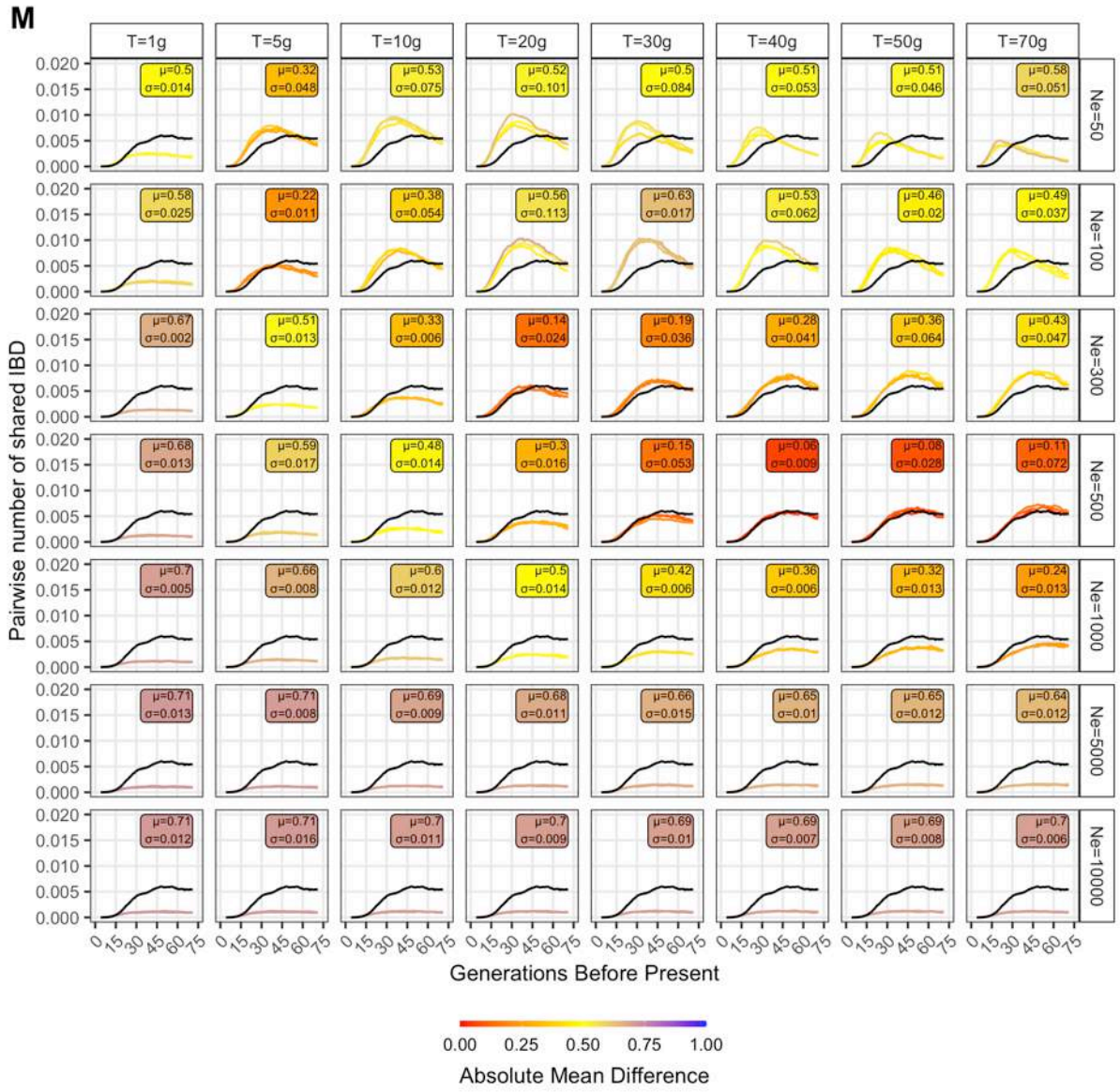
I


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00


J
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00


K
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00


L
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00


M
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
Generations Before Present
Absolute Mean Difference
0.00
0.25
0.50
0.75
1.00

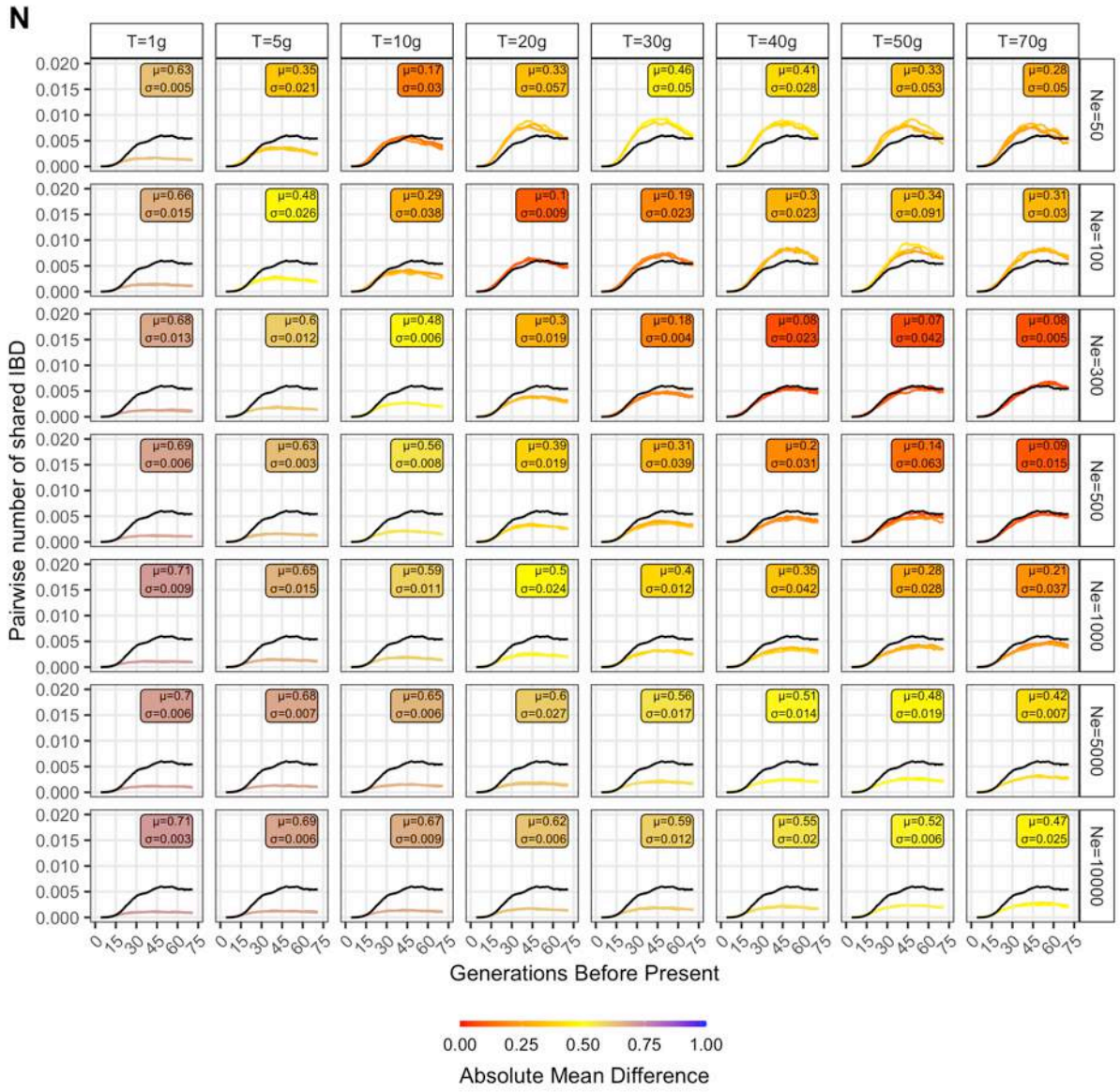


Figure S14. Distribution of the absolute mean differences scores between simulated and observed IBD-sharing of Asian origin in Madagascar, according to different simulated demographic scenarios We evaluated different demographic scenarios using the average

absolute difference between the number of observed and simulated Asian IBD segments during the last 75 generations. **A)** Distribution of scores according to different bottleneck durations. **B)** Distribution of scores according to different strengths in founder events. **C)** Distribution of scores according to different migration rates. The y-axes show the number of scenarios simulated. The x-axes show generations before present.

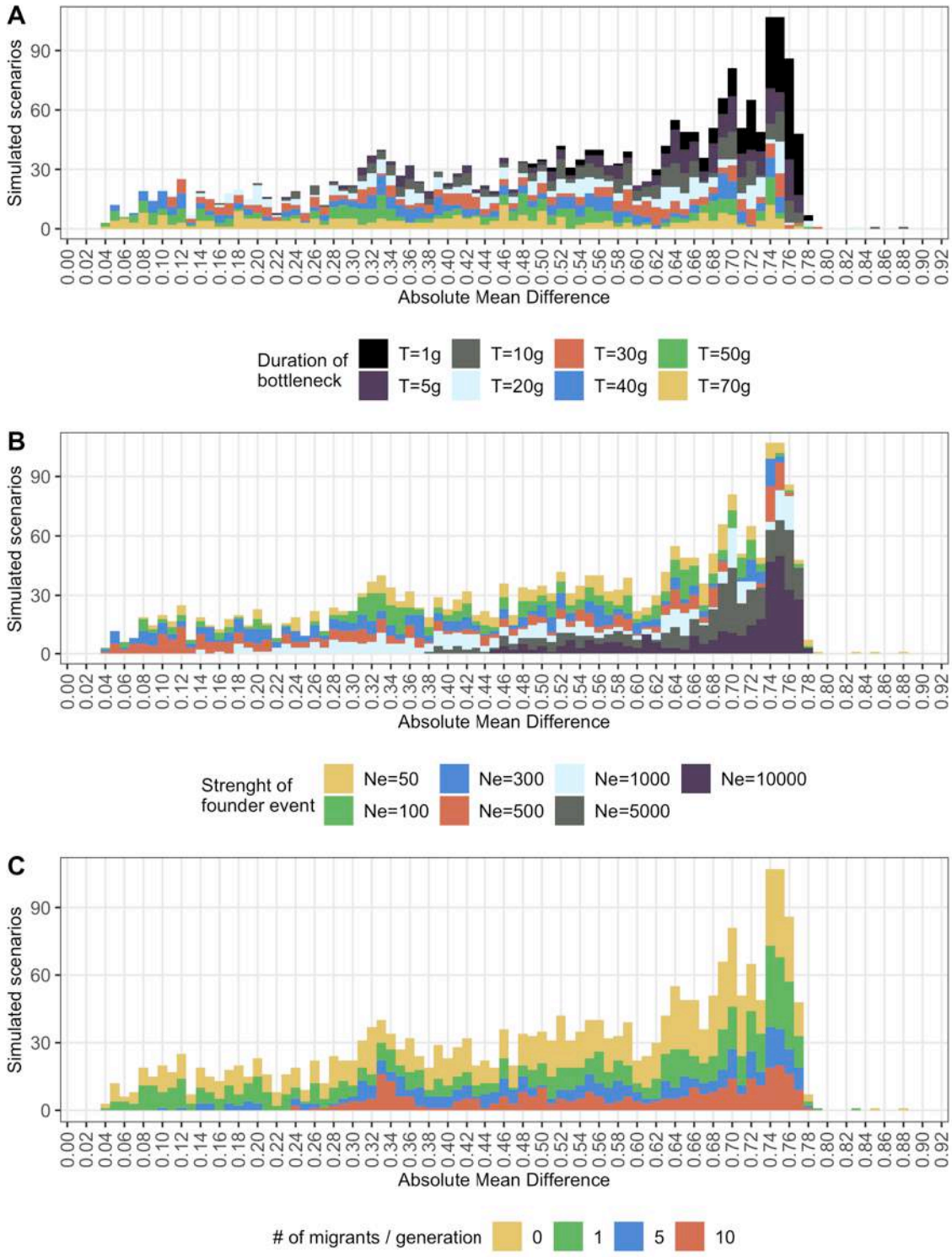


Figure S15. Effect of gene flow during the long-term bottleneck on the IBD segments of Asian origin shared in Madagascar. Each panel shows the result for a Long-term bottleneck scenario according to different simulated migration rates. The x-axis represents the age of IBD haplotypes (in years before present). The number of segments shared per pair is

shown on the y-axis. Yellow curves show the mean pairwise IBD-sharing at each generation +/- 5 standard deviations, based on 3 simulated replicas. Blue lines in each panel represent the observed pairwise IBD-sharing of Asian segments in the Malagasy population.

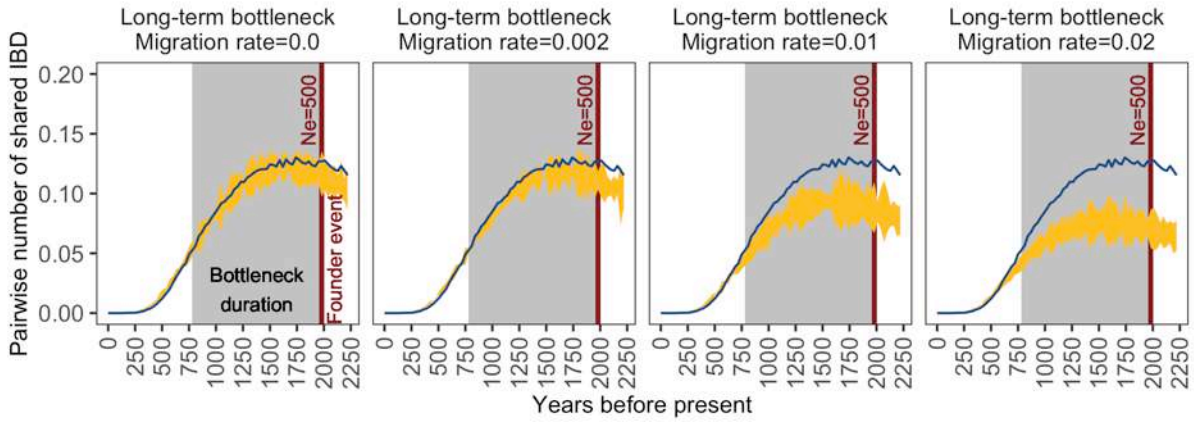


Figure S16. Asian ancestral demographic history simulated scenarios and Malagasy settlement without inclusion of ascertainment bias. Each panel shows the result for simulated scenarios (see Supplementary Table S4), according to the strength of founder events (red line) and the duration of bottleneck events (shaded area). The x-axis represents the age of IBD haplotypes (in years before present). The number of segments shared per pair is shown on the y-axis. Yellow curves show 5 standard deviations from the mean pairwise IBD-sharing at each generation across 3 simulated replicas **with ascertainment bias.** Green curves show 5 standard deviations from the mean pairwise IBD-sharing at each generation across 3 simulated replicas **without ascertainment bias.**

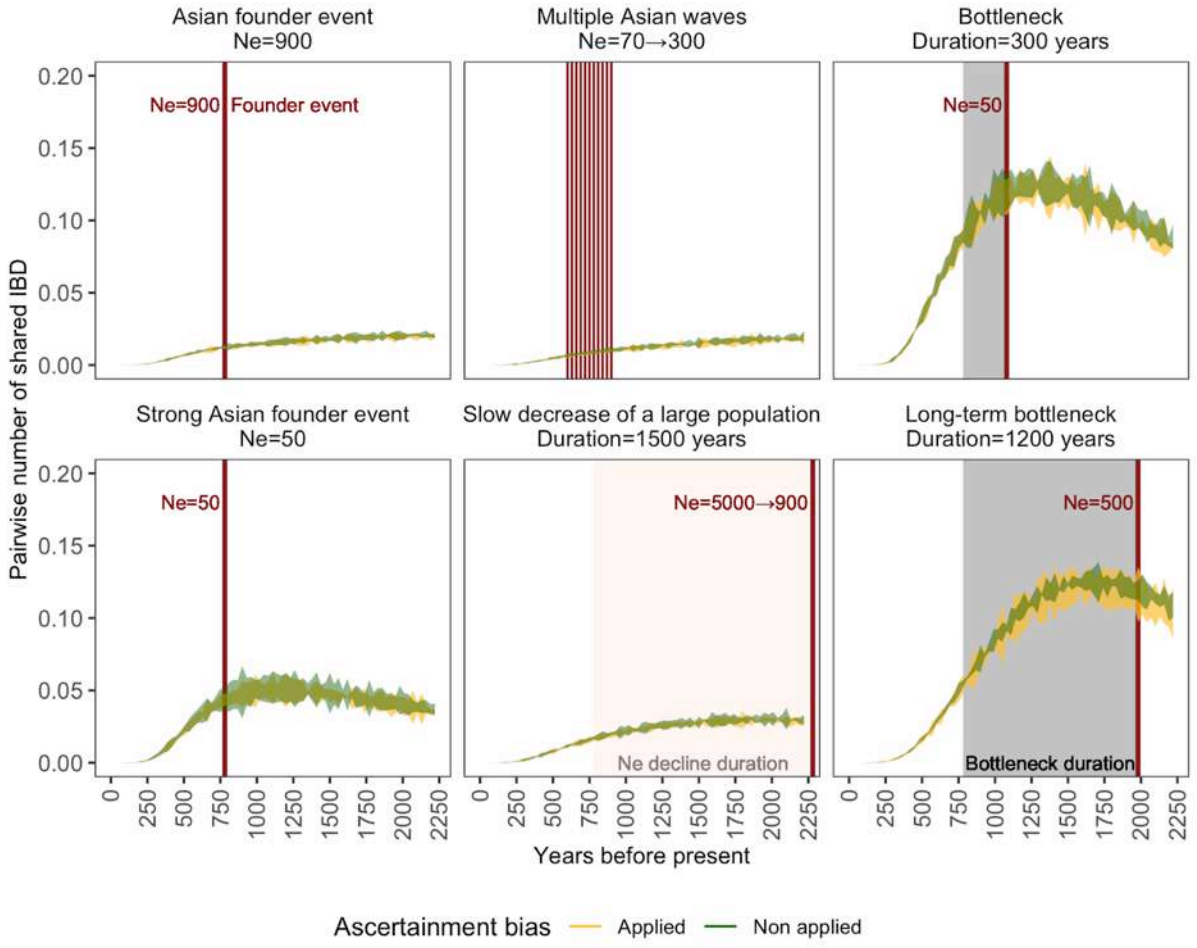


Figure S17. Estimated average number of most recent genetic common ancestors per generation back through time according to different pairs of populations. Timeline of the point (dotted line) and smooth (solid line) estimations of the average number of common ancestors shared between Malagasy and South Borneo, Malagasy and Mozambique, Mozambique and Angola populations, and South Borneo and Philippines, as estimated by shared IBD segments from genome-wide data.

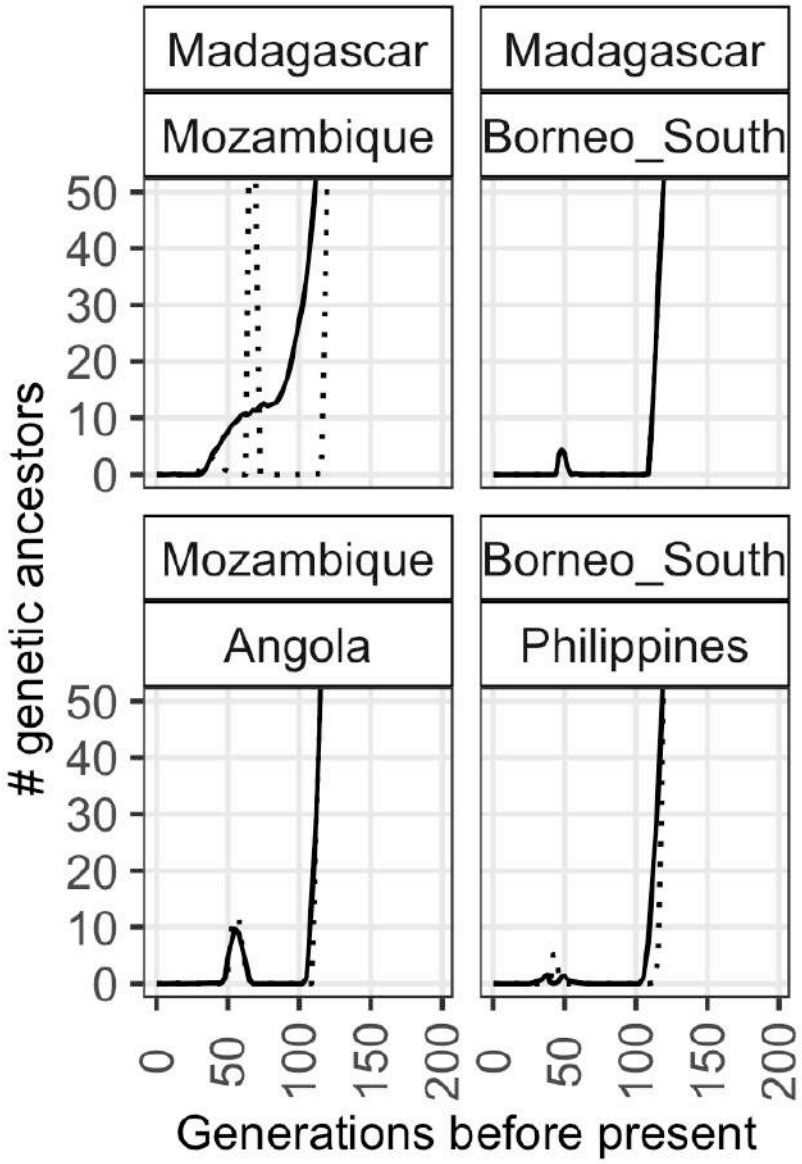


Figure S18. Effect of minor allele frequency (MAF) filters on allele frequency spectrum of simulated reference populations. Coloured bars show the allele frequency distribution from simulations in African (A) and Asian (B) reference populations, while red lines represent the observed allele frequency distribution in our density dataset.

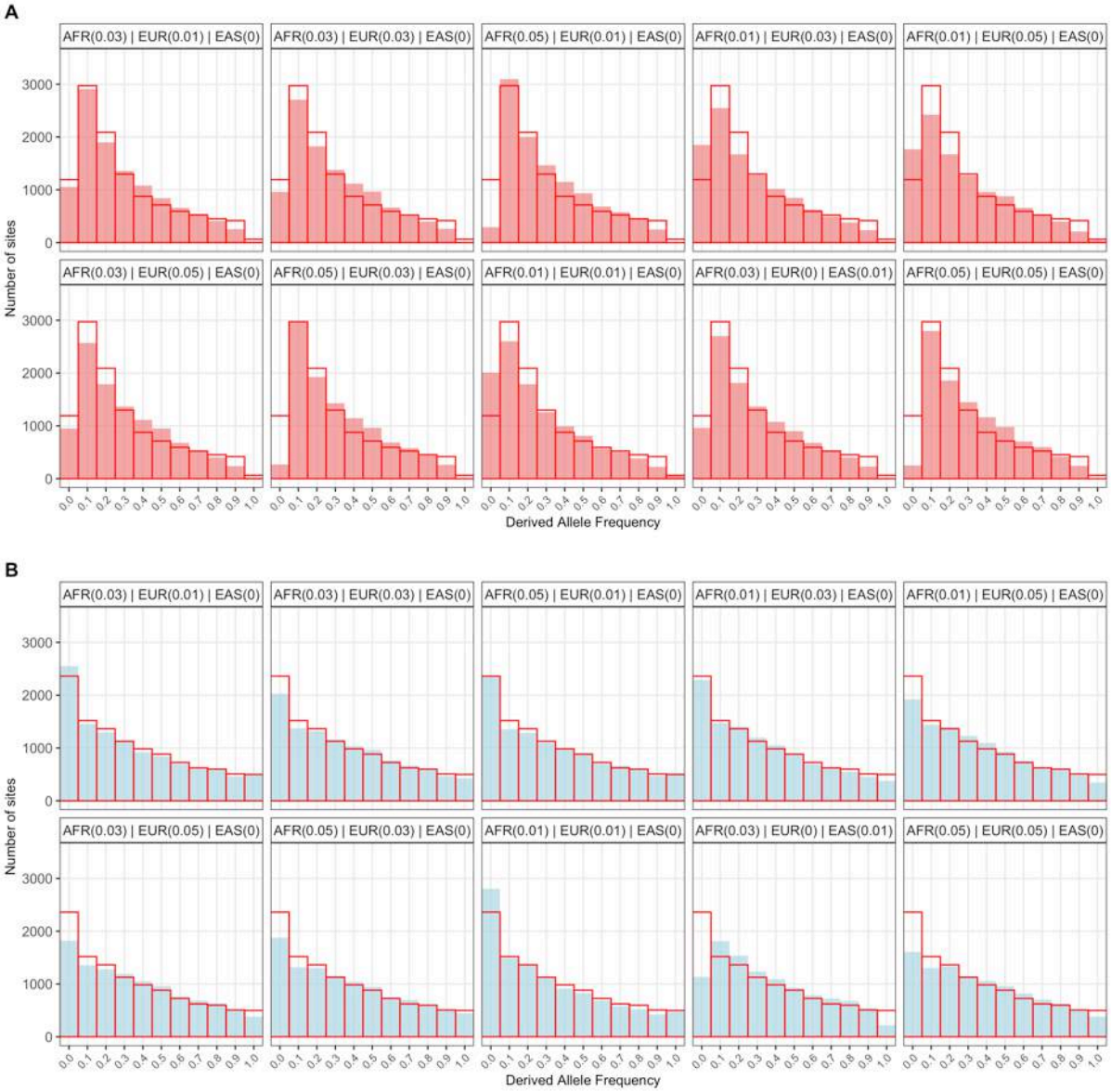
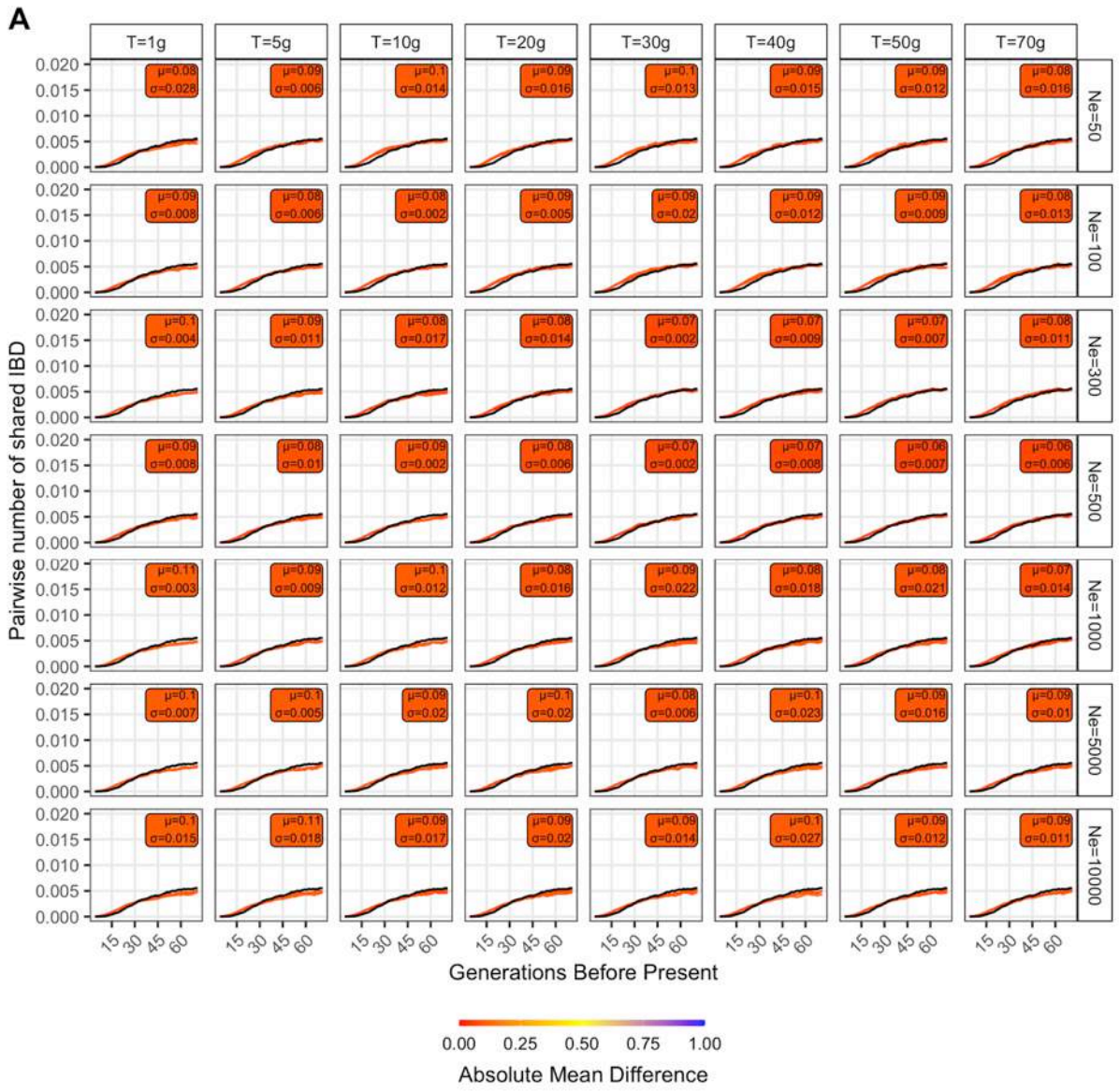
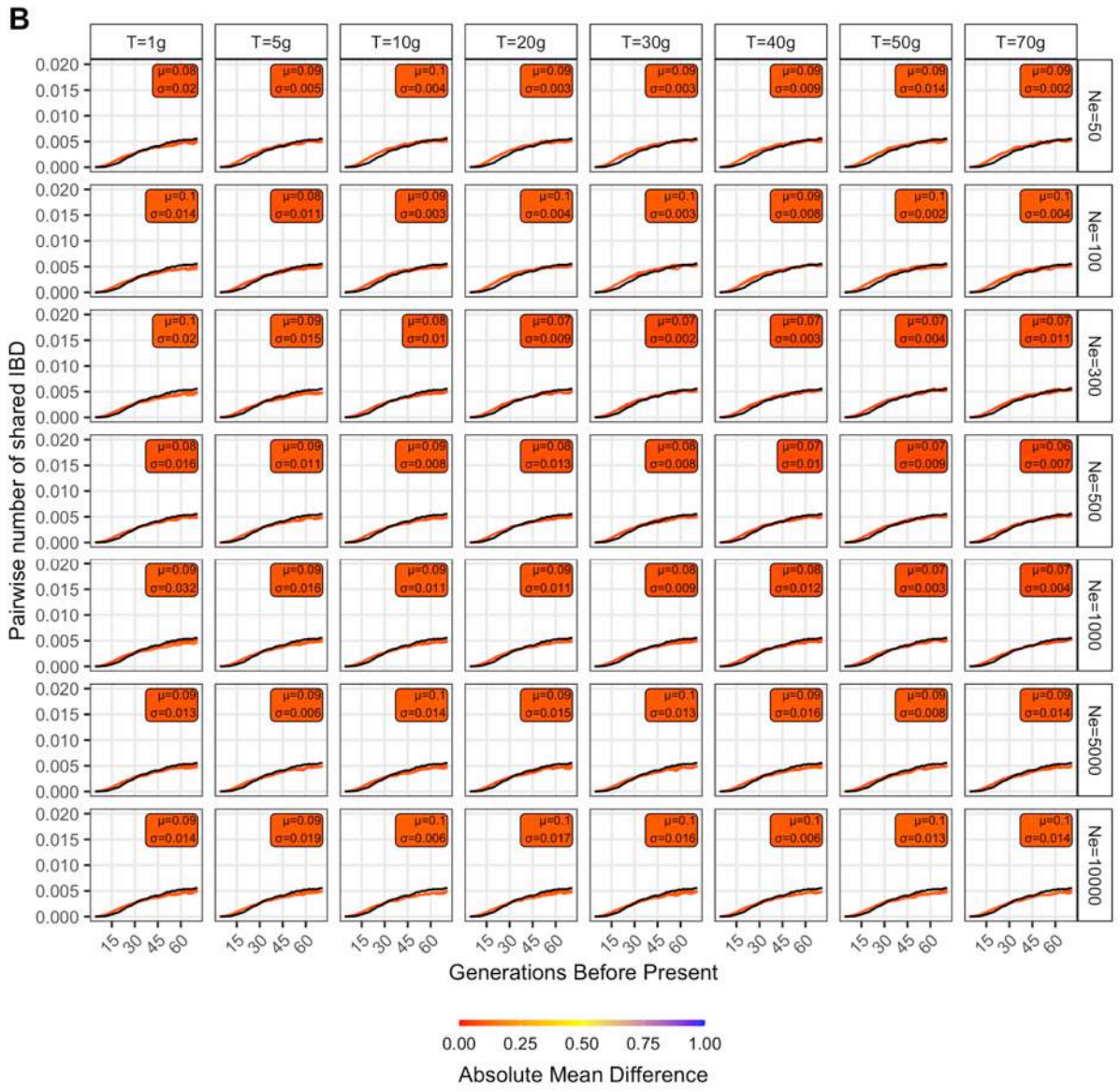
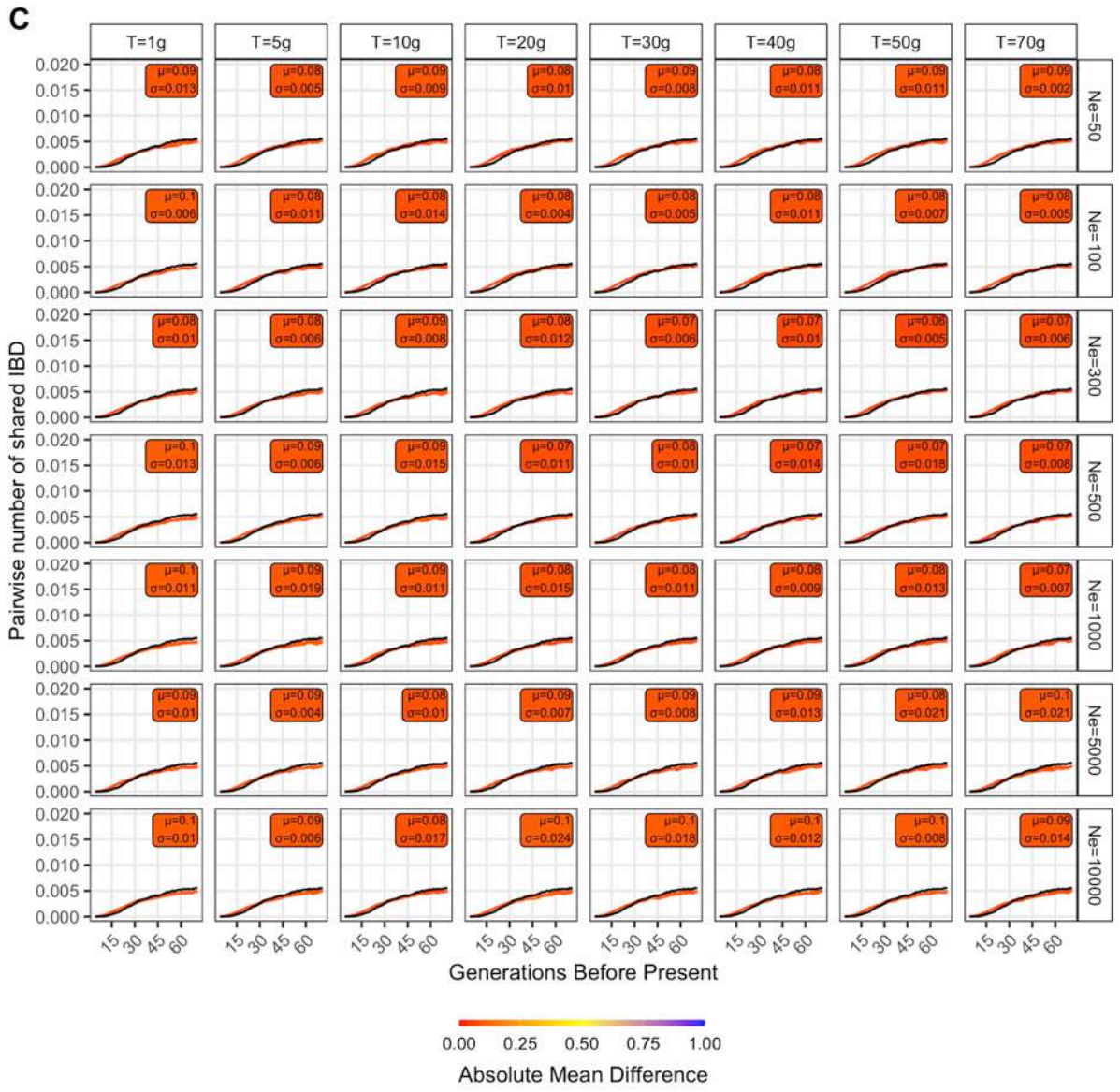
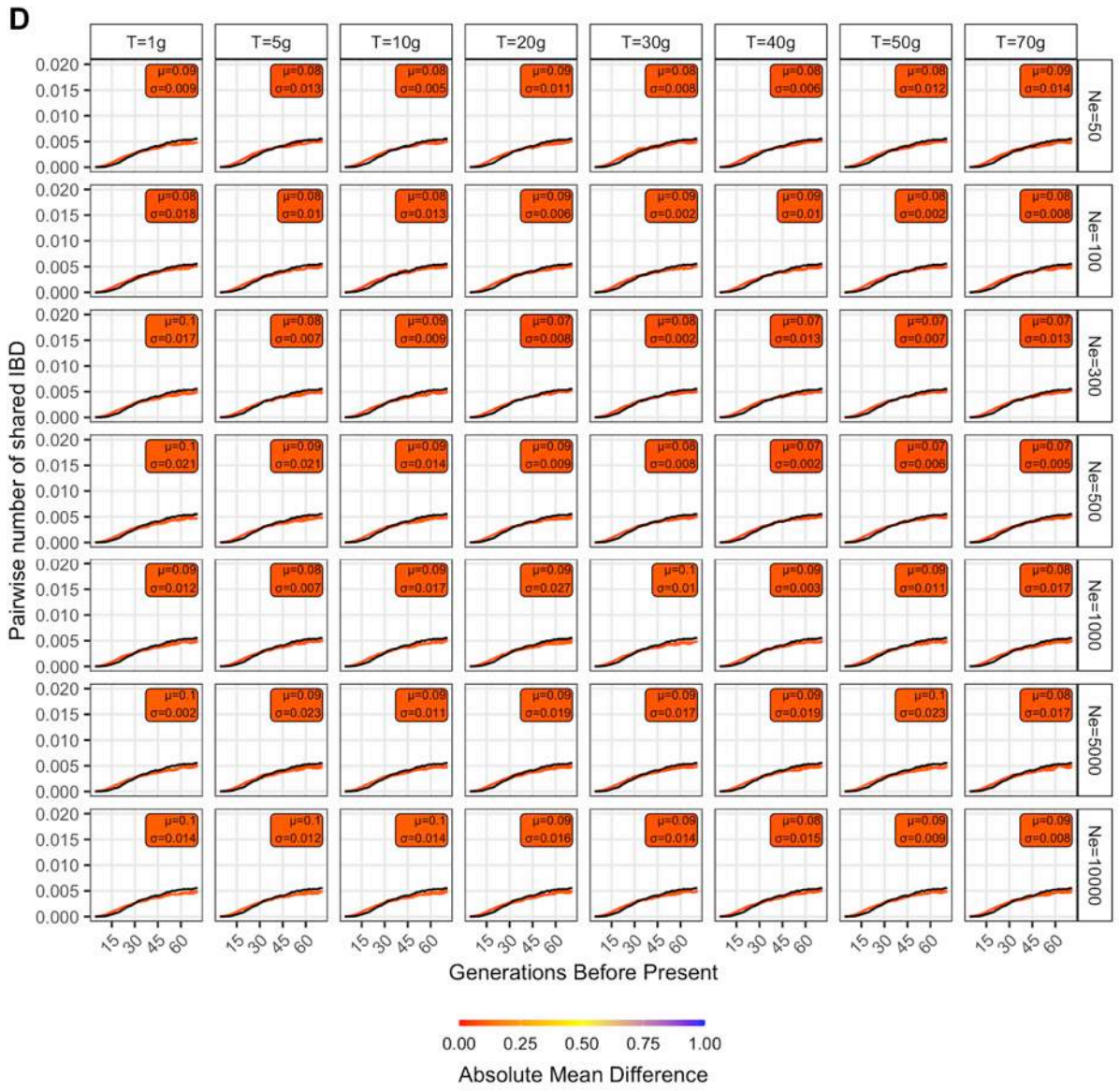


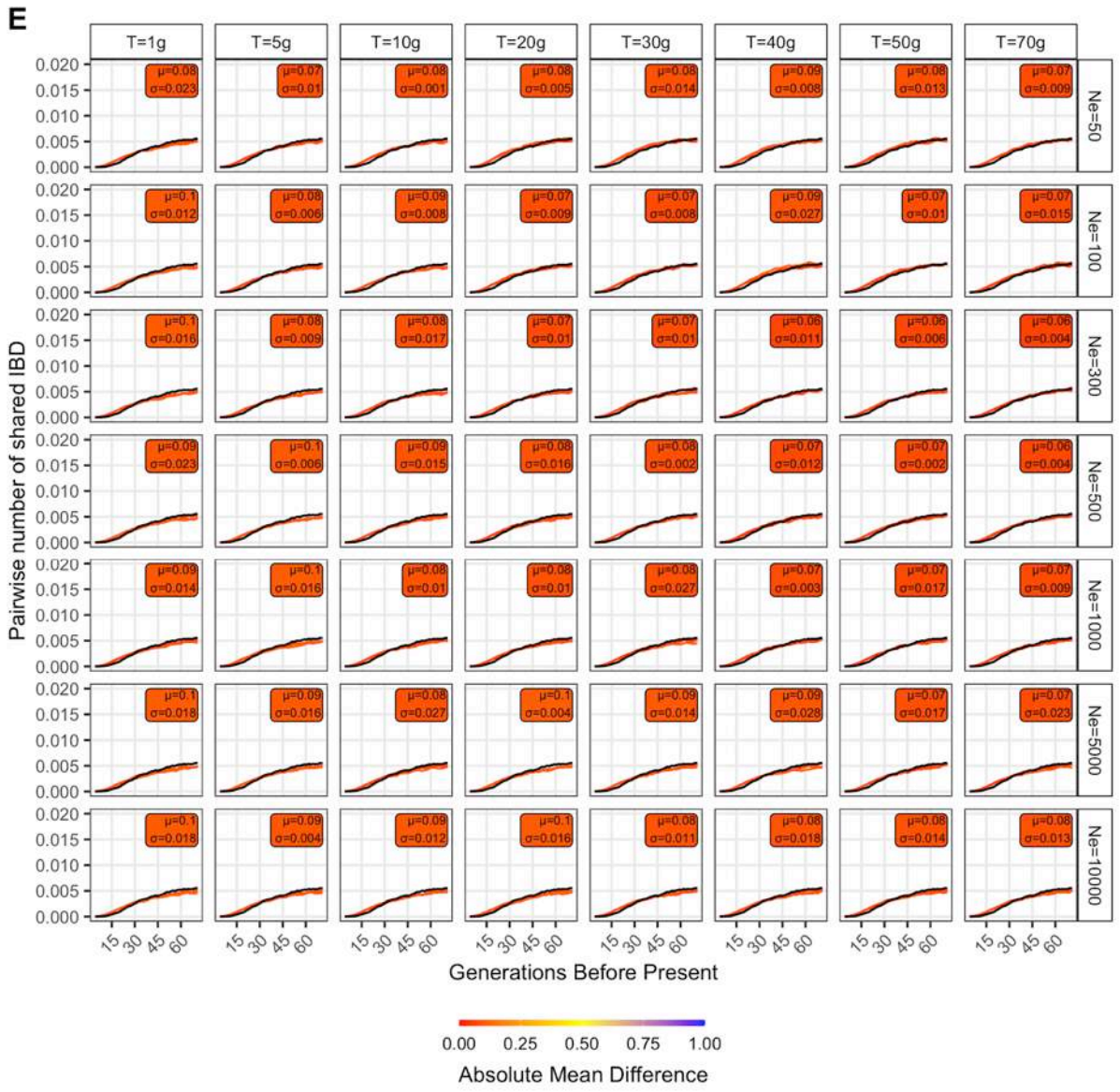
Figure S19. Effect of different demographic scenarios on the simulated IBD segments of African origin shared in Madagascar. Each panel shows the result for simulated scenarios, according to the strength of founder events (Ne = 50, 100, 300, 500, 1000, 5000 or 10000 individuals) and the duration of bottleneck (T = 1, 5, 10, 20, 30, 40, 50, or 70 generations before the admixture) for the Asian ancestral population. The x-axis represents the age of IBD haplotypes (in generations before present). The number of segments shared per pair is shown on the vertical axis (y axis). Coloured lines show the pairwise IBD-sharing from simulations at each generation, while black lines represent the observed pairwise IBD-sharing. The average score and standard deviation using 3 replicates are shown in the upper-right legend. (**A to D**). Before the admixture event, we simulated an isolated Austronesian population with stable Ne during the bottleneck. Migration rates to the isolated population were set to 0 (A), 1 (B), 5 (C) and 10 (D) individuals per generation. (**E to H**) We simulated an isolated Austronesian population with changing Ne during the bottleneck. Migration rates to the isolated population were set to 0 (E), 1 (F), 5 (G) and 10 (H) individuals per generation. (**I and J**) We simulated a South Bornean population with initial Ne=5,000 and growth rate of 0.0219. From this population, we an Austronesian population diverged with stable Ne during a bottleneck event. Migration rates to the isolated population were set to 0 (I) or1 (J) individuals per generation. (**K and L**) We simulated a South Bornean population with initial Ne=5,000 and growth rate of 0.0219. From this population, an Austronesian population diverged with changing Ne during a bottleneck event. Migration rates to the isolated population were set to 0 and 1 individual per generation (K and L, respectively). (**M and N**) Admixture was simulated under the Continuous Gene Flow model, with the Austronesian ancestral population receiving African gene flow during the period 20-30 generations BP. Before the admixture event, we simulated a bottleneck event for the Austronesian population with stable (M) or changing (N) Ne during the bottleneck. Migration rates to the isolated population were set to 0 individuals per generation.

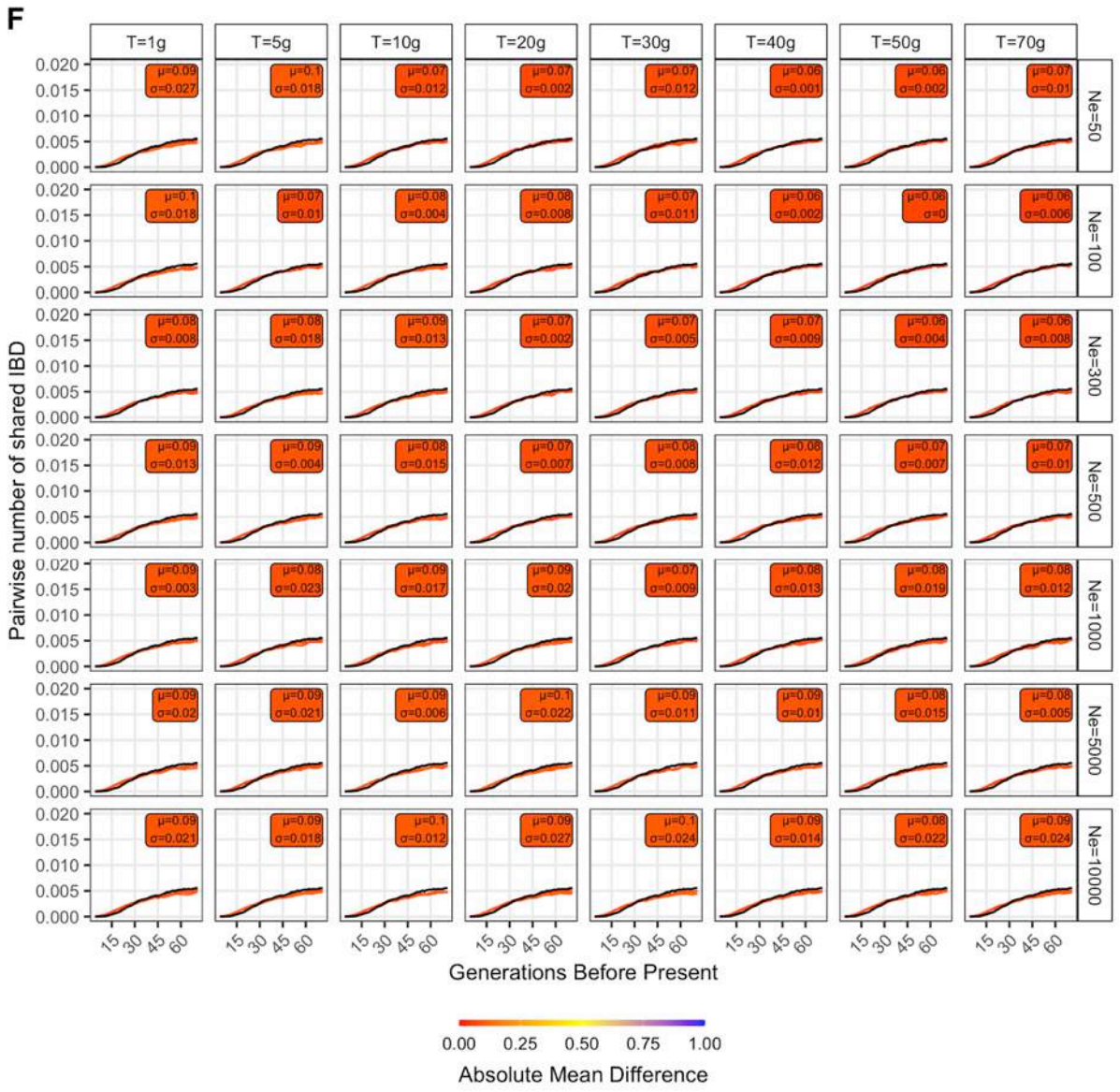

A
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15 30 45 60
Generations Before Present
0.00 0.25 0.50 0.75 1.00
Absolute Mean Difference

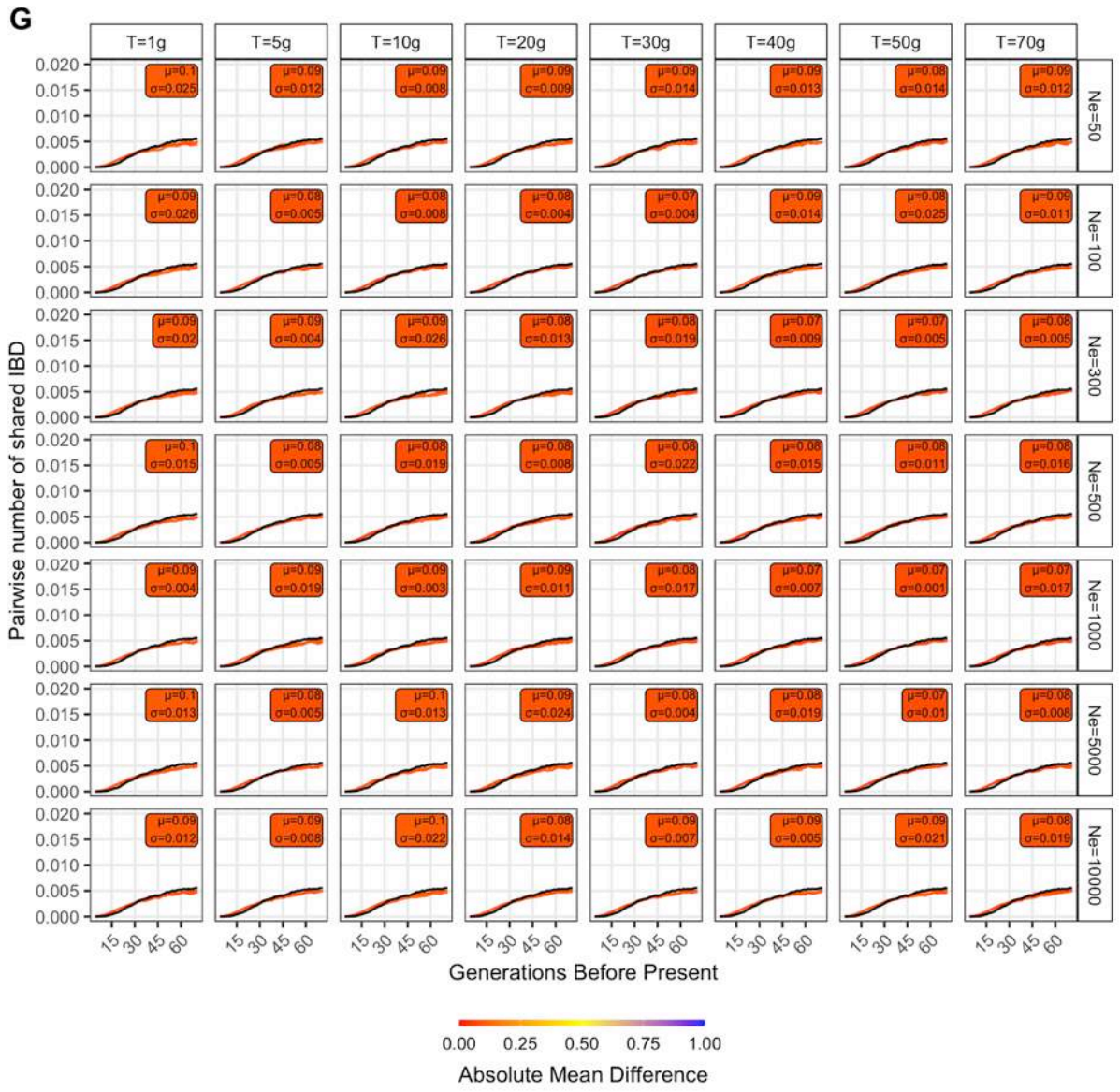

B
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

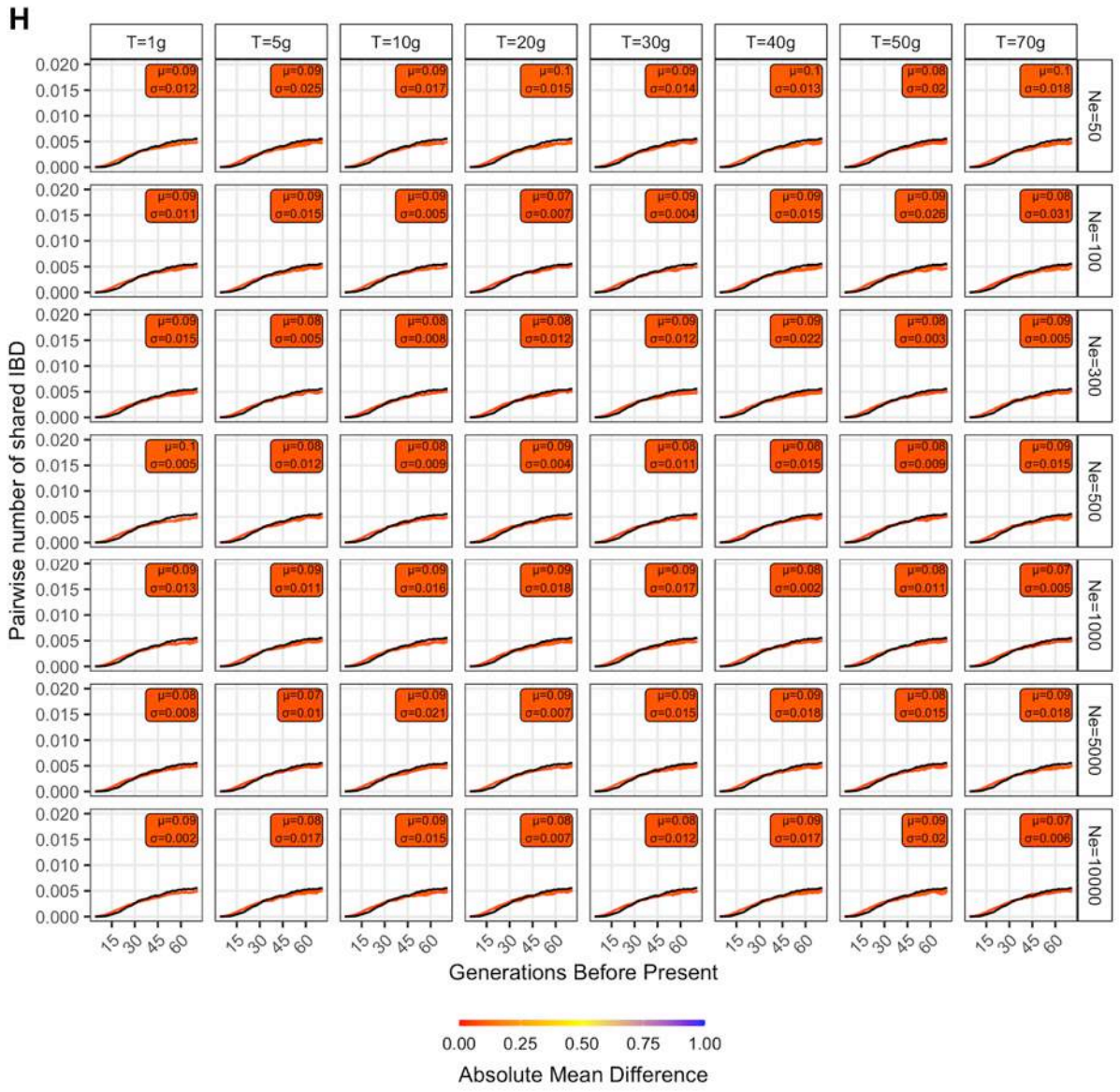

C
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference


D
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

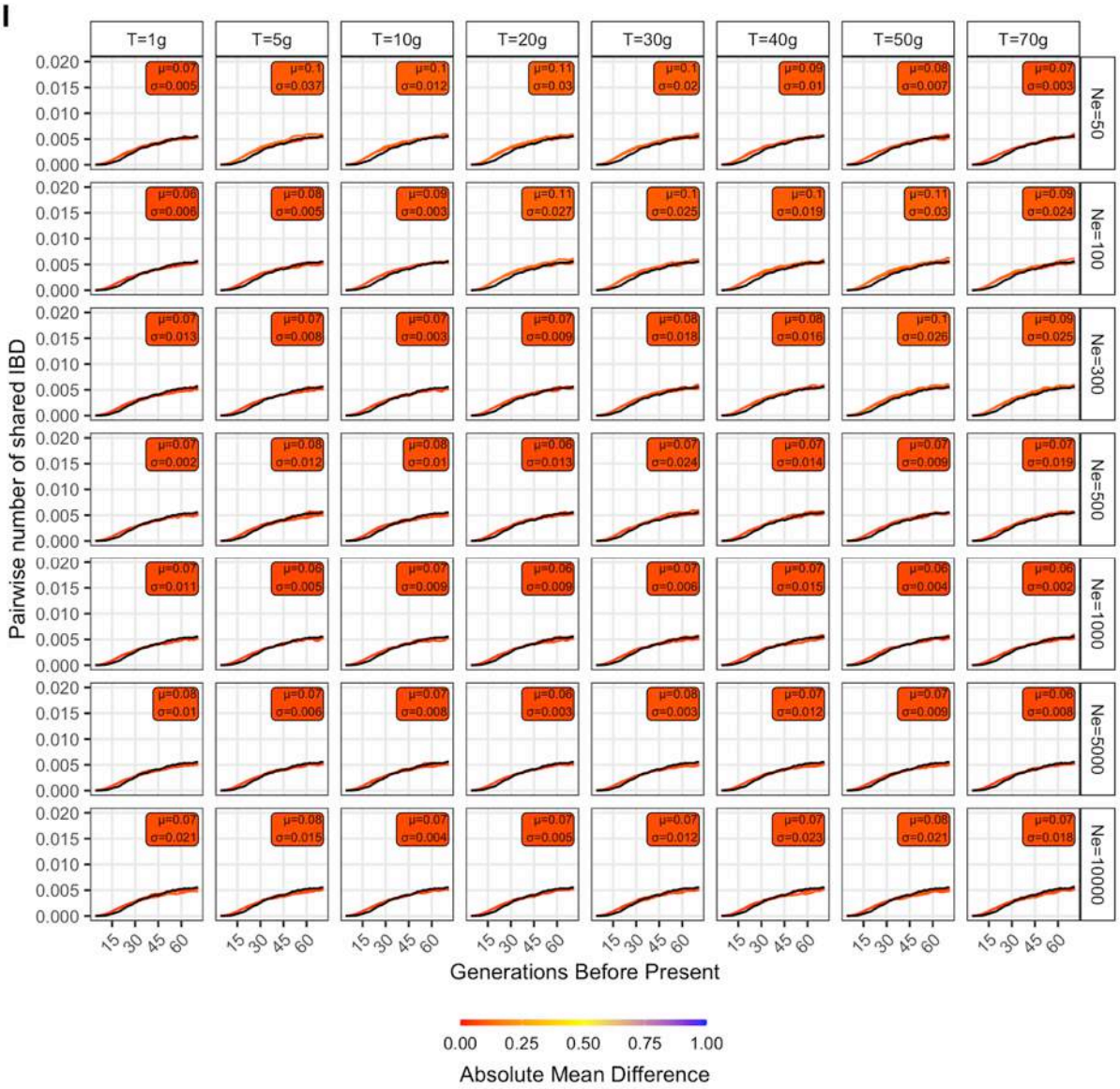

E
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

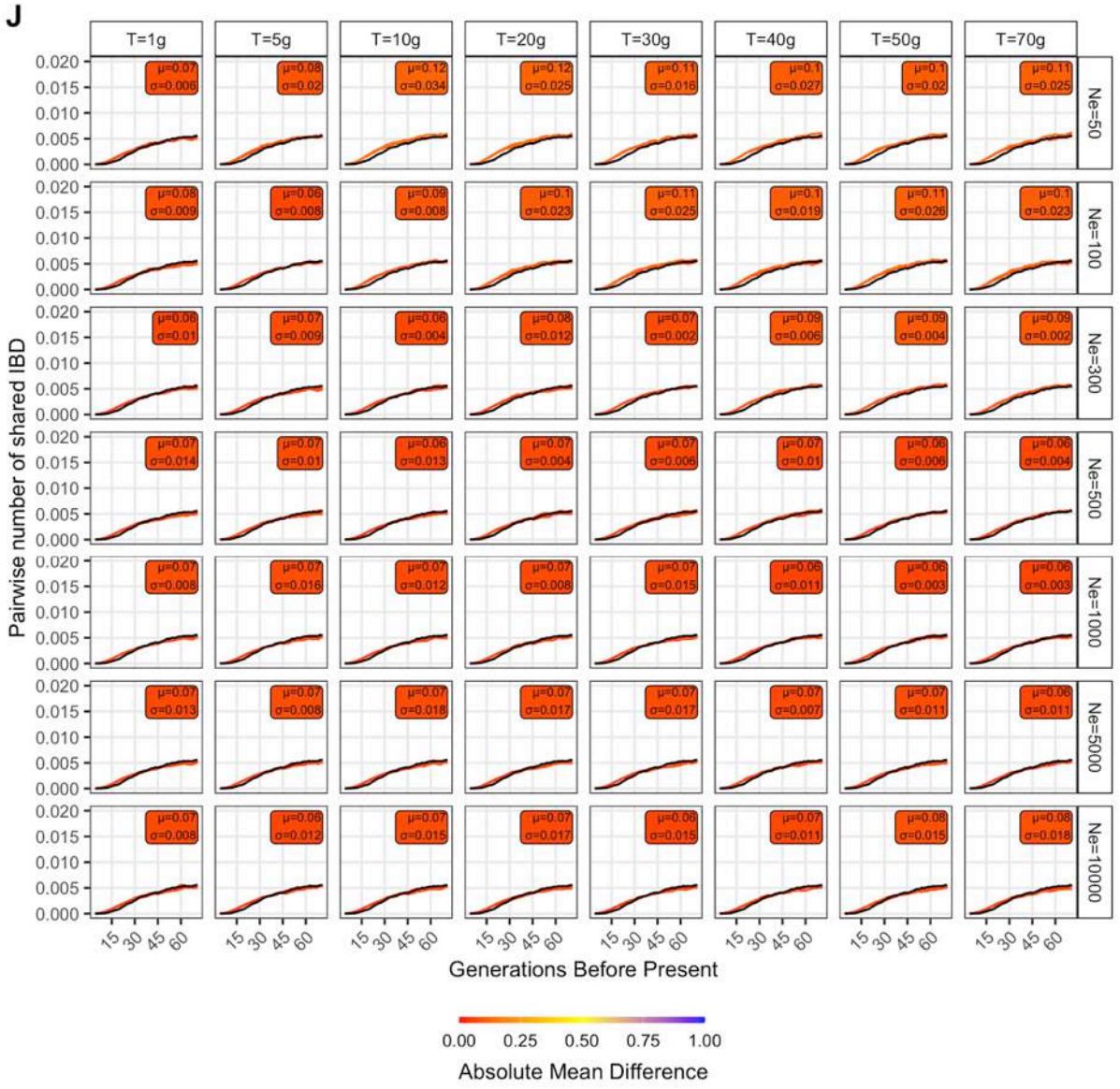

F
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

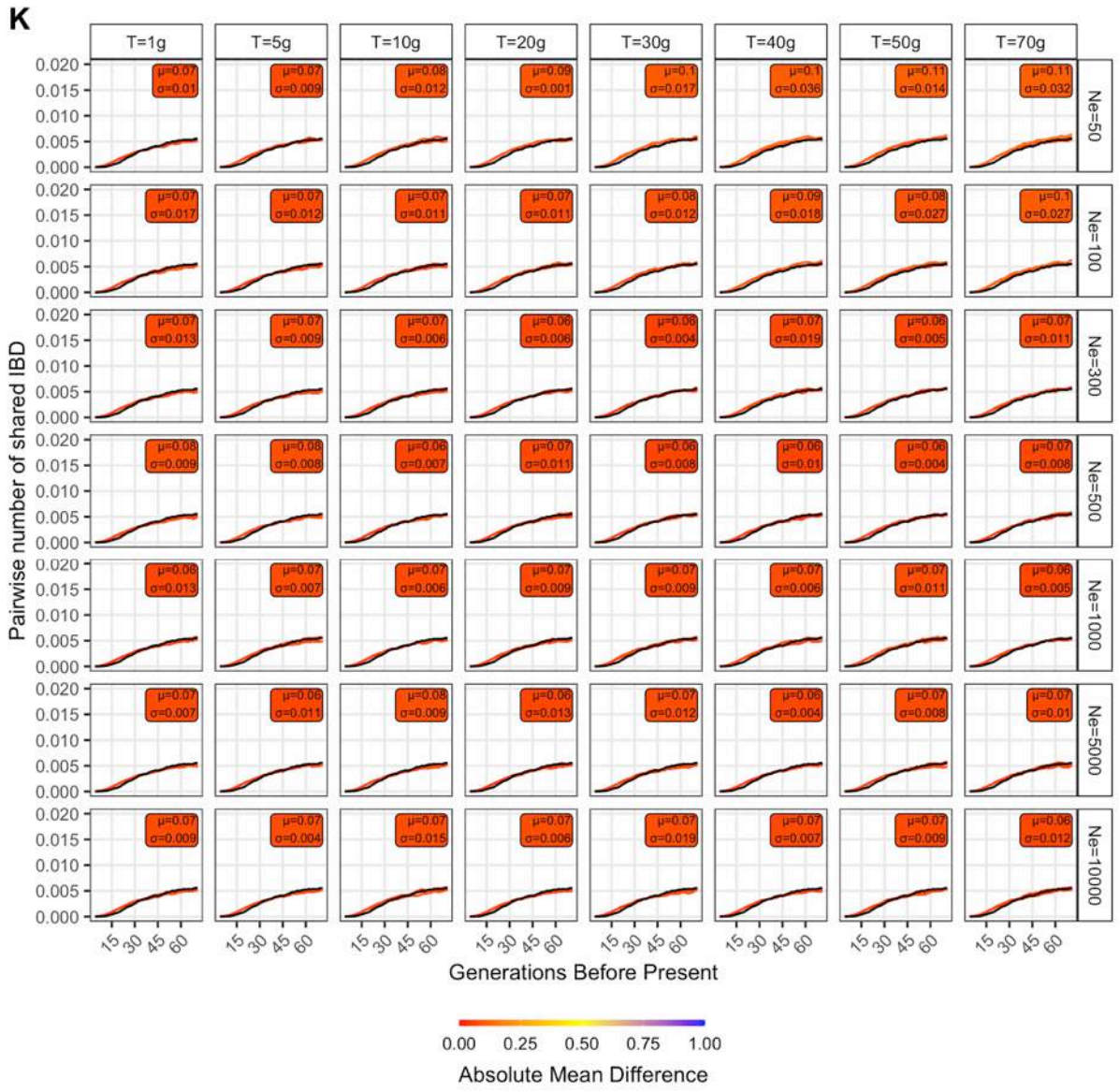

G
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

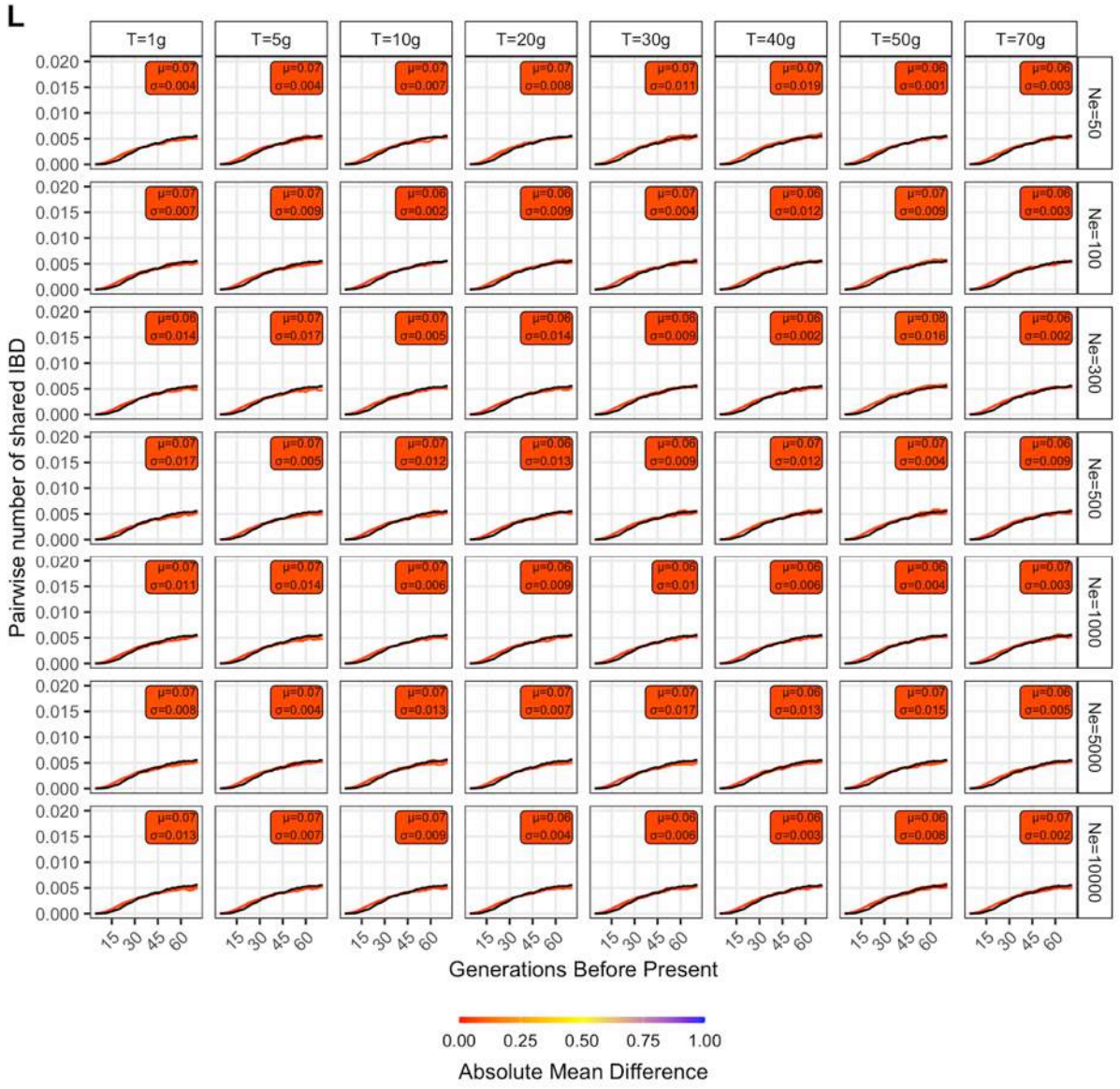

H
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

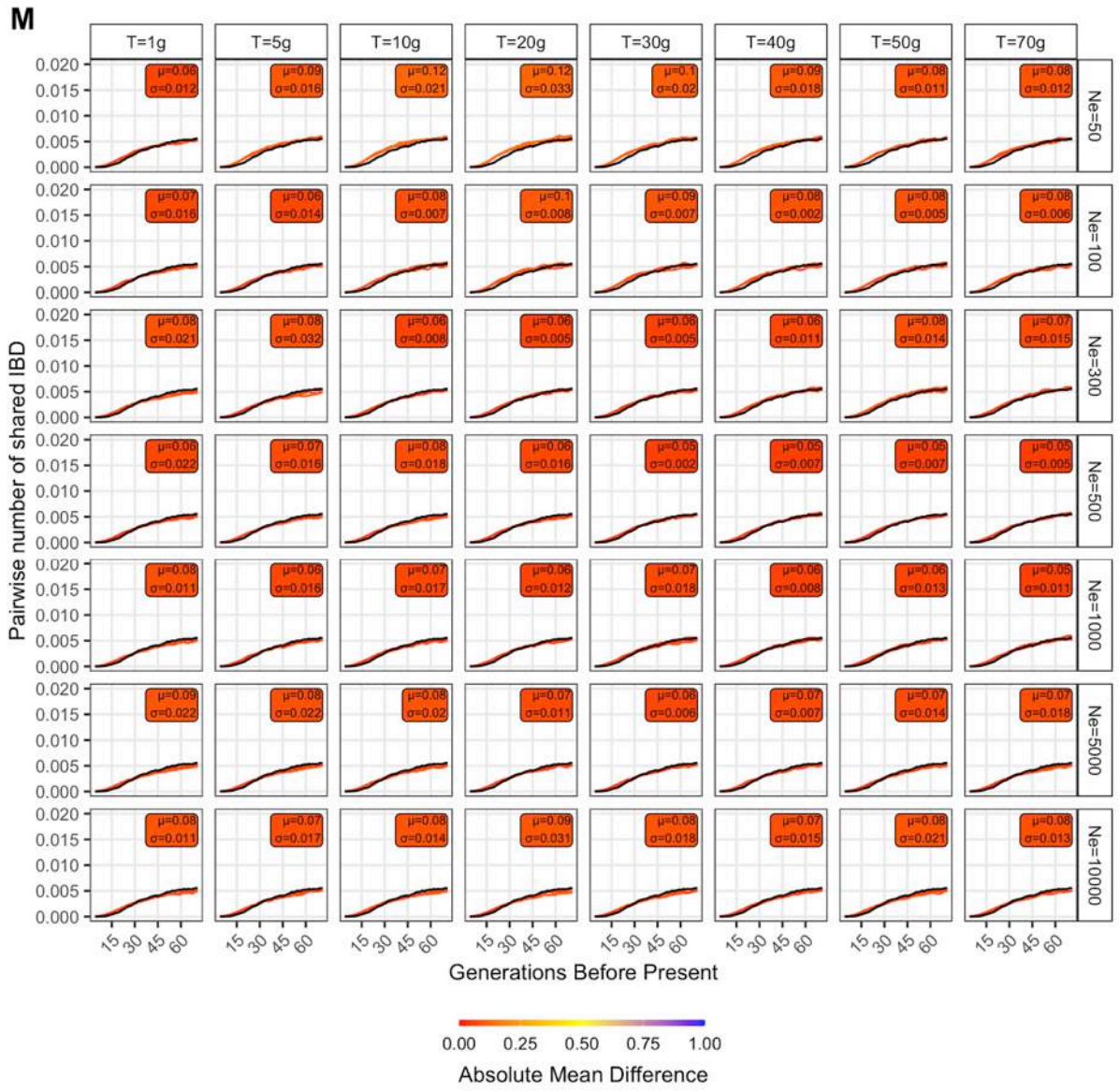
I


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference


J
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference


K
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference


L
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference


M
T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Pairwise number of shared IBD
0.020
0.015
0.010
0.005
0.000
15
30
45
60
Generations Before Present
0.00
0.25
0.50
0.75
1.00
Absolute Mean Difference

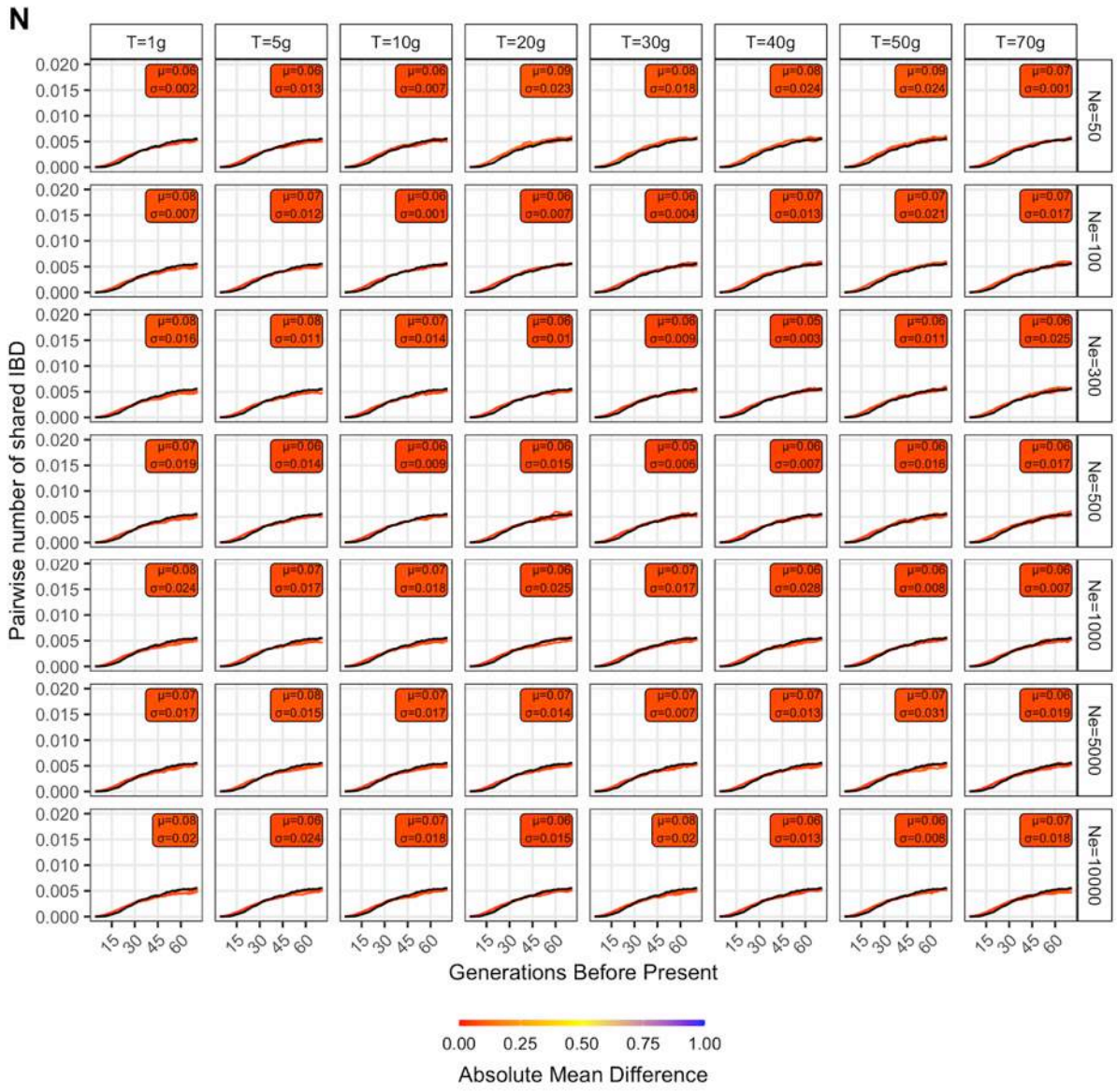
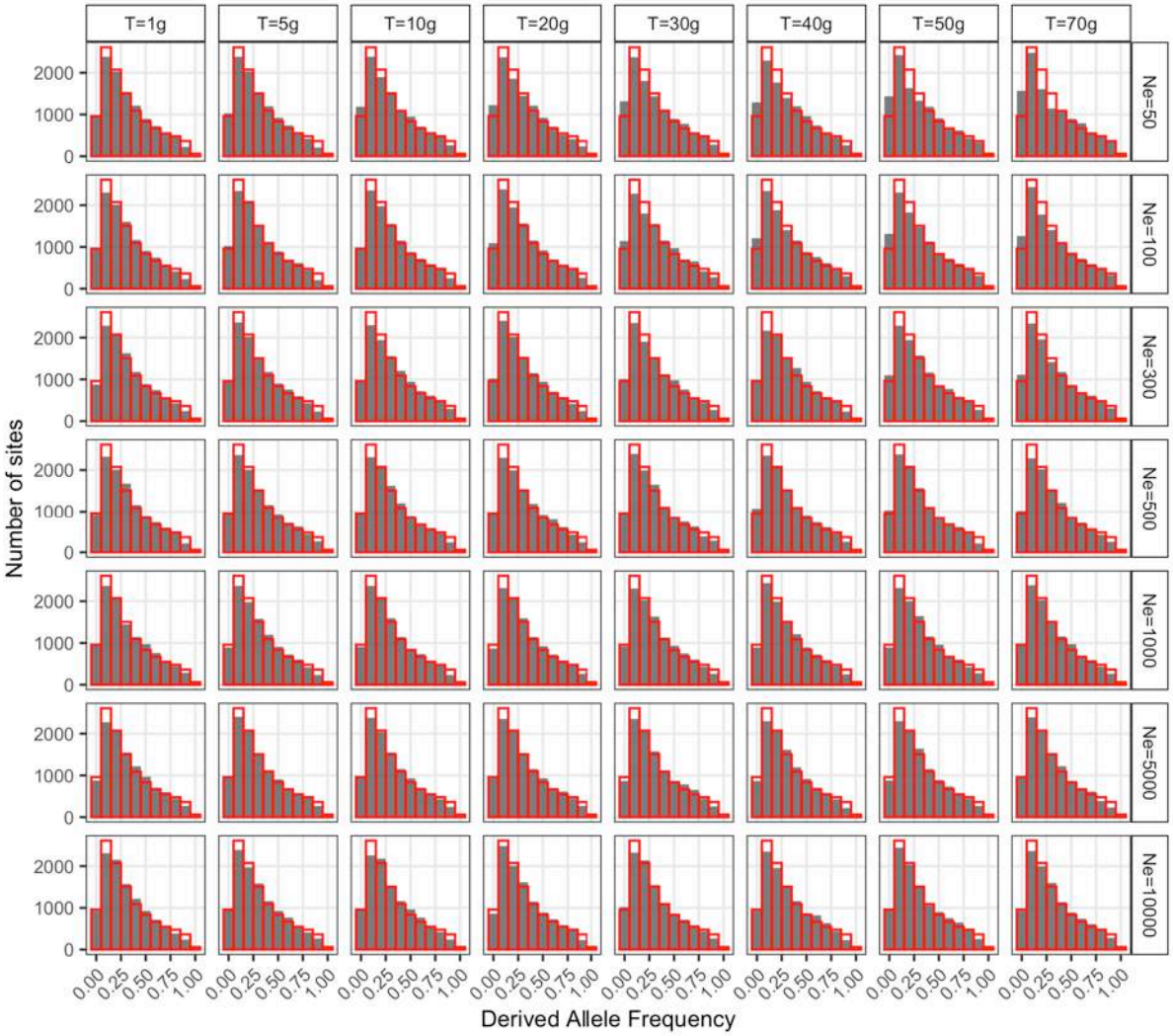
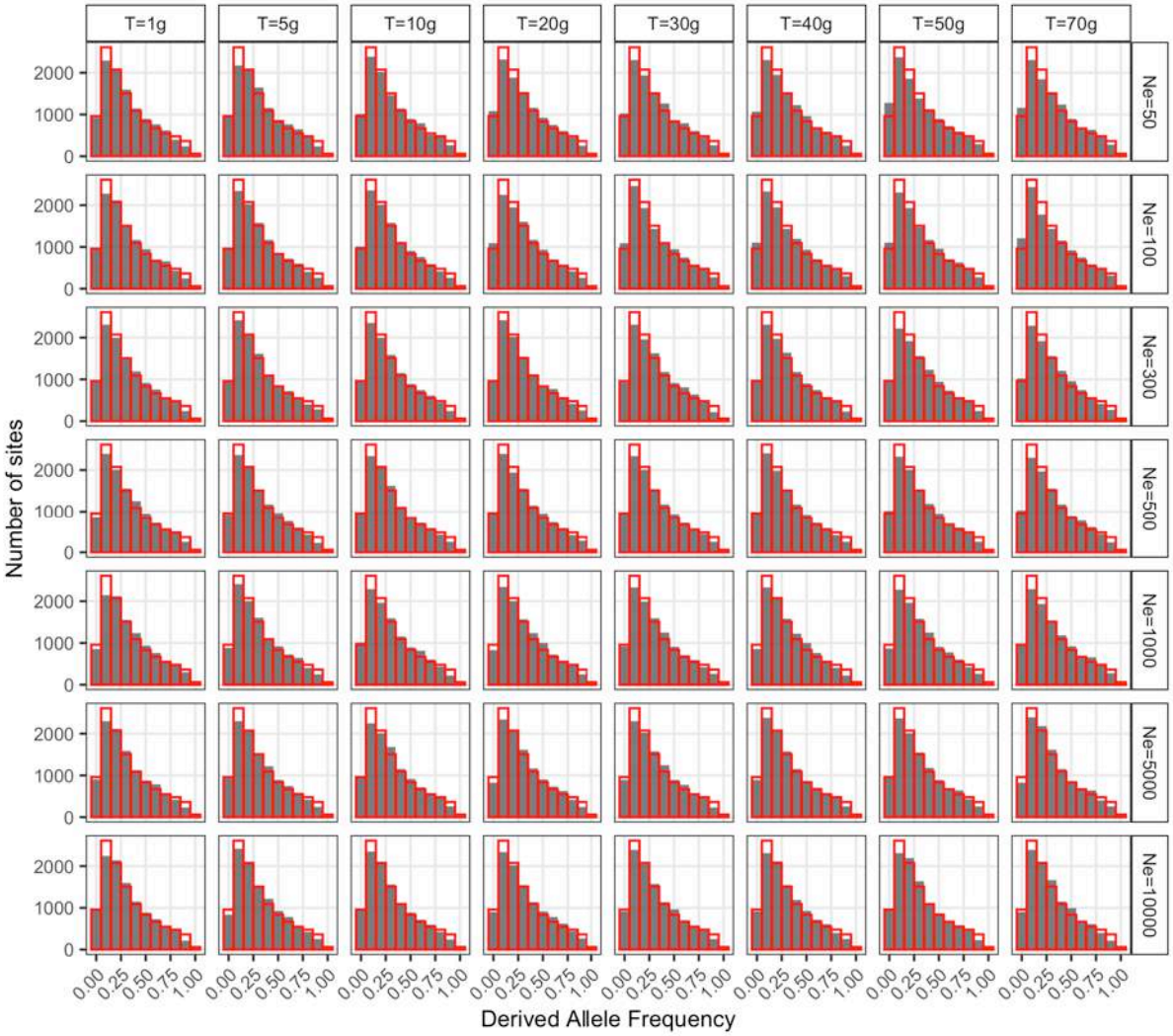


Figure S20. Effect of different demographic scenarios on the simulated allele frequency spectrum in Madagascar. Each panel shows the result for simulated scenarios. The details of these correspond to the legend from Fig. S19.

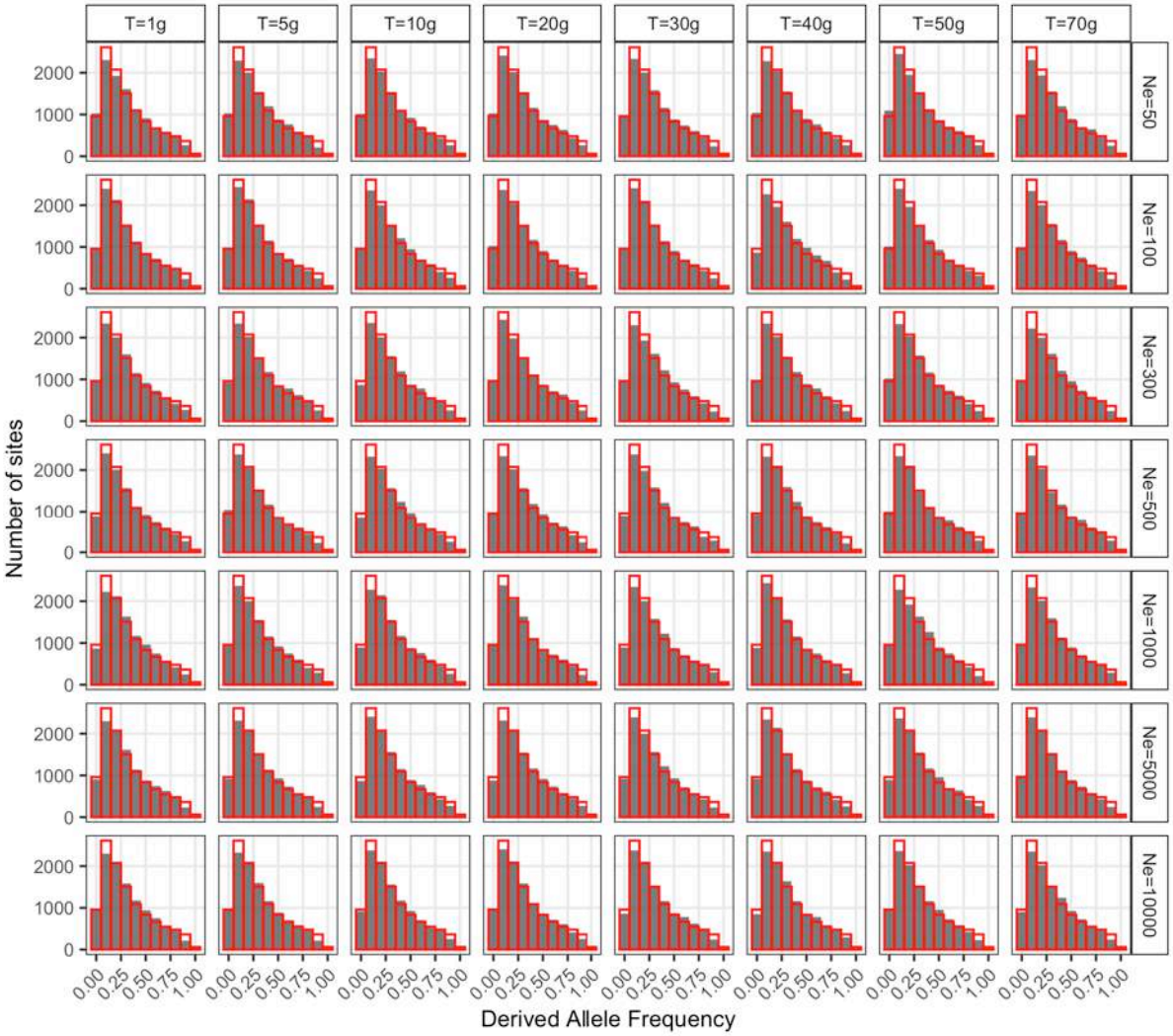
A


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

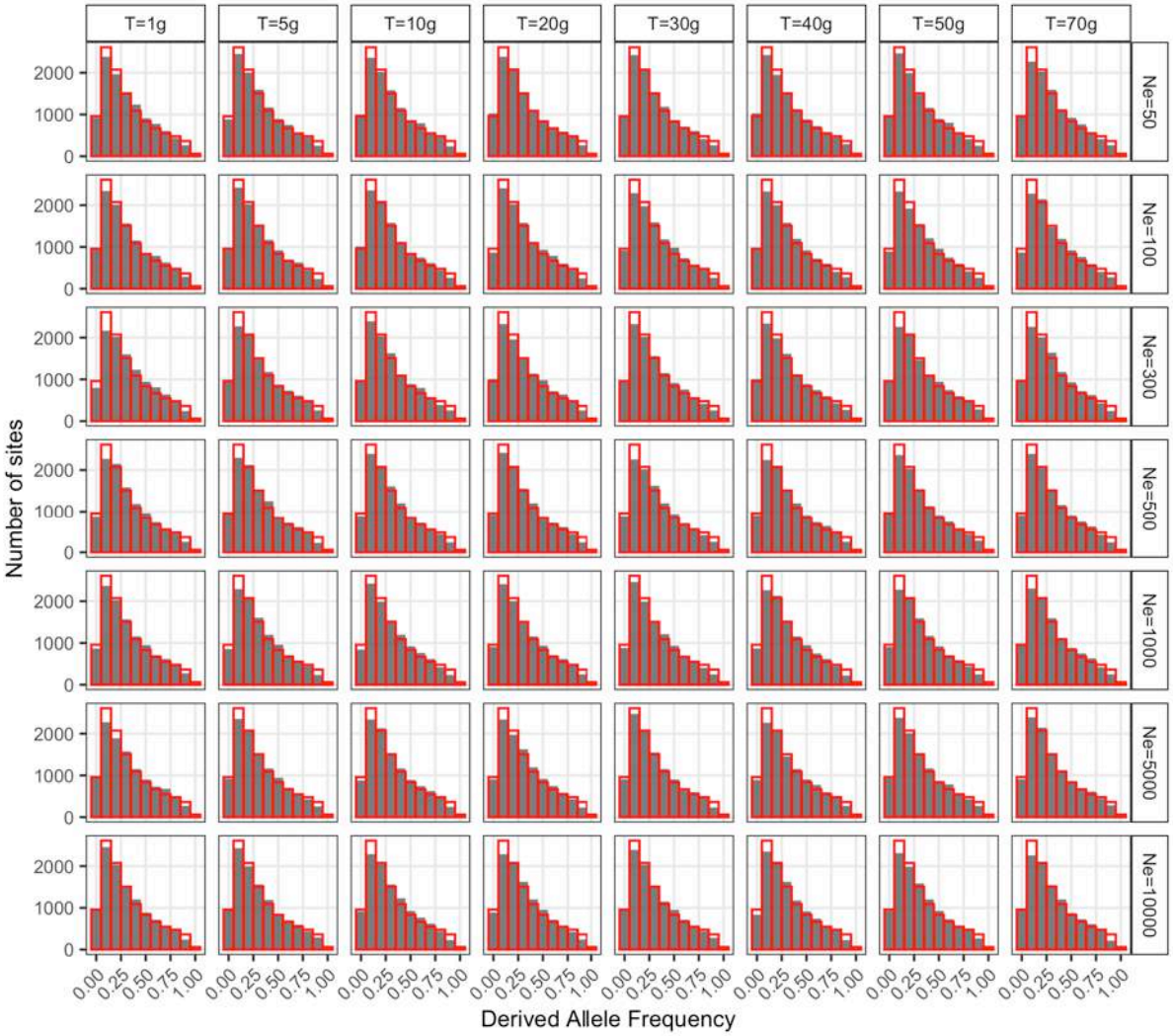
B


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

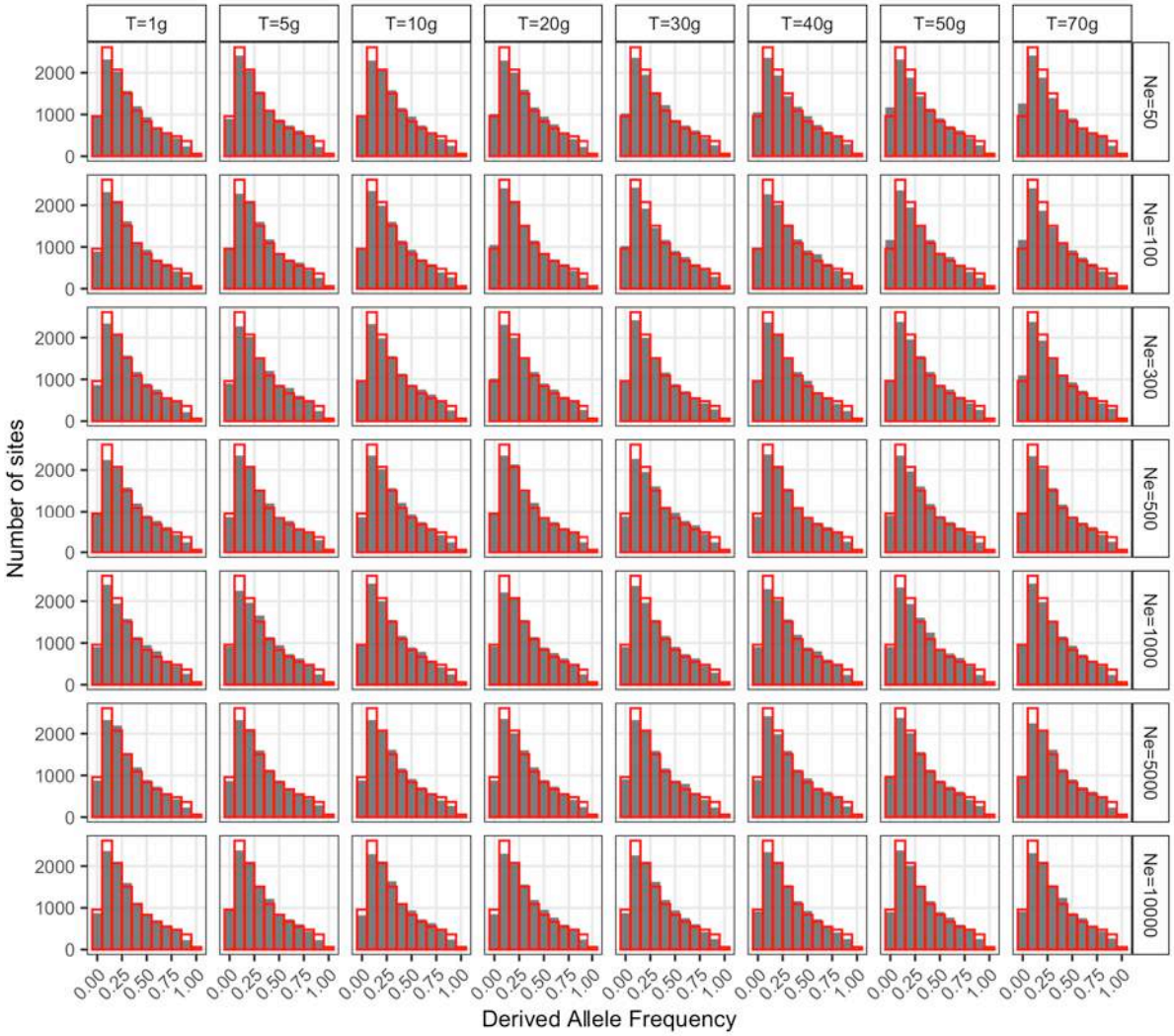
C


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

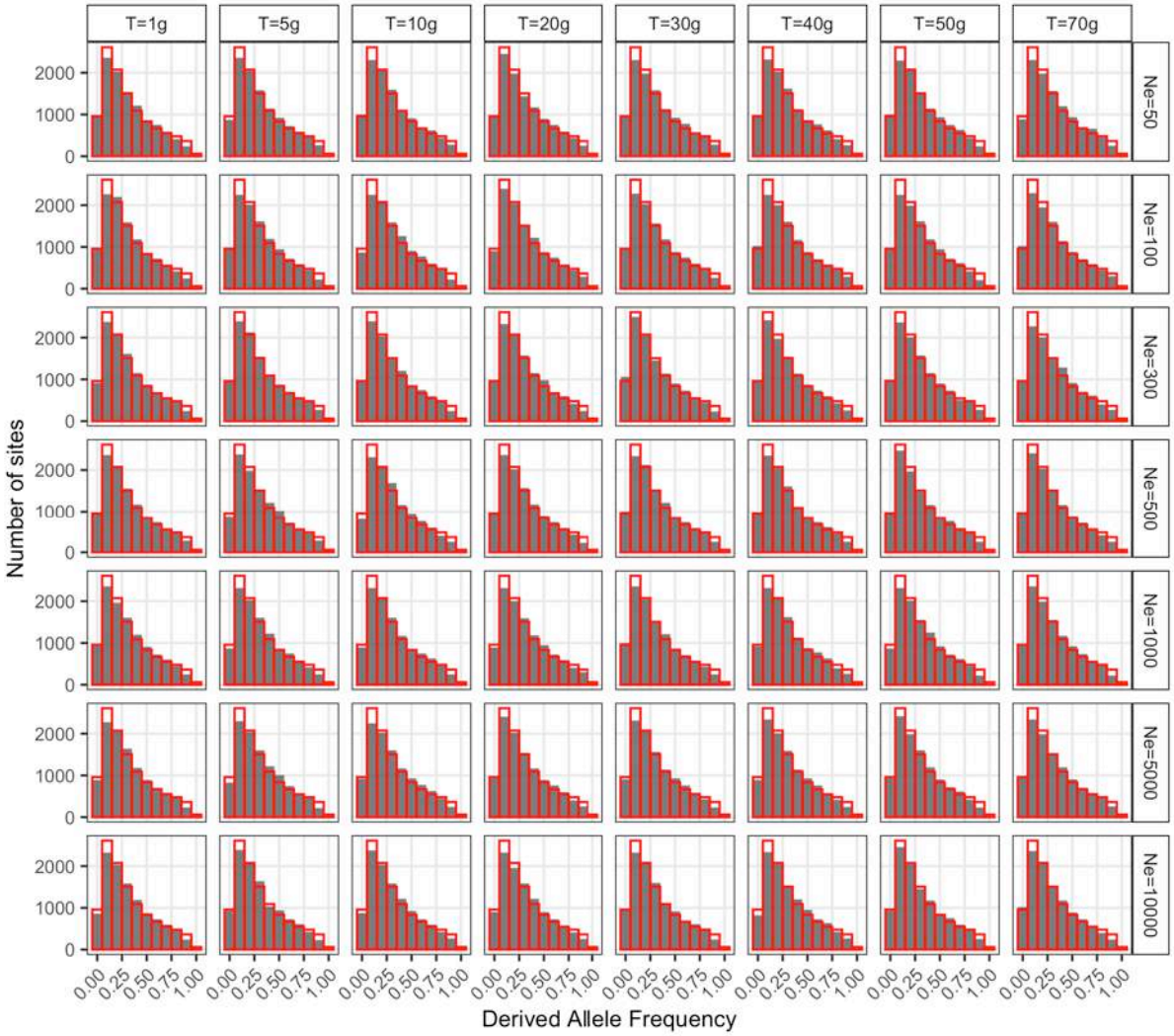
D


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

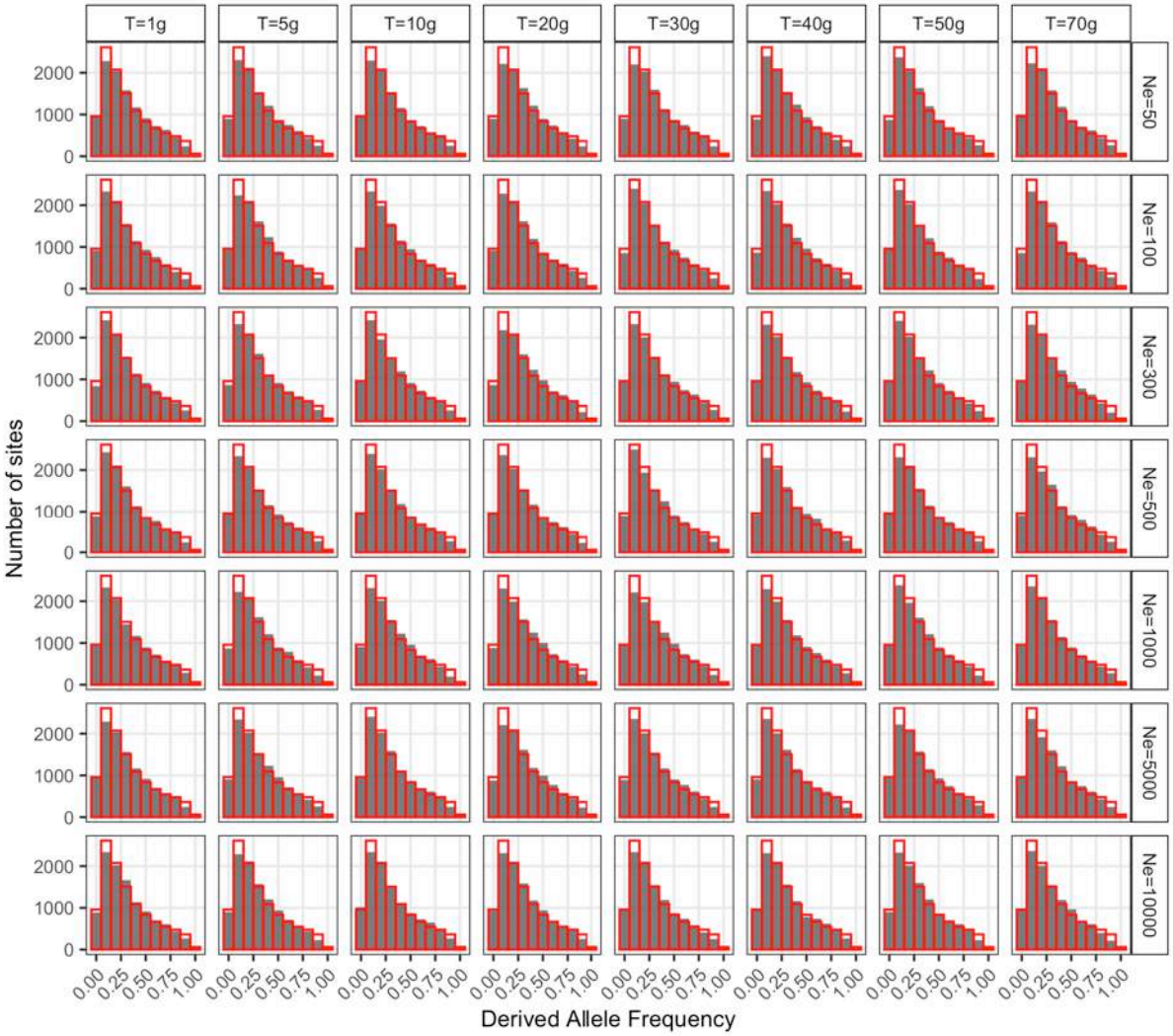
E


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

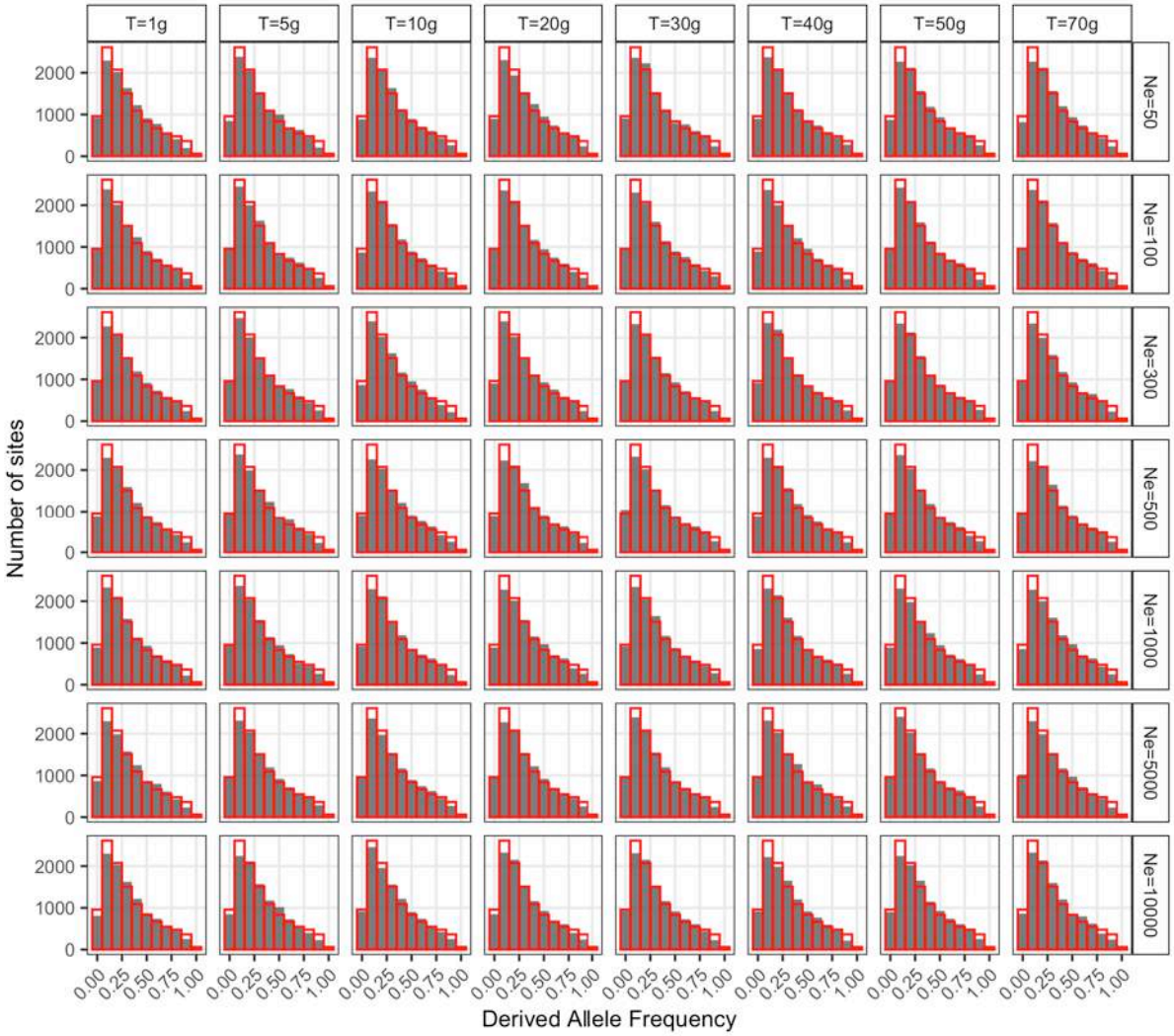
F


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

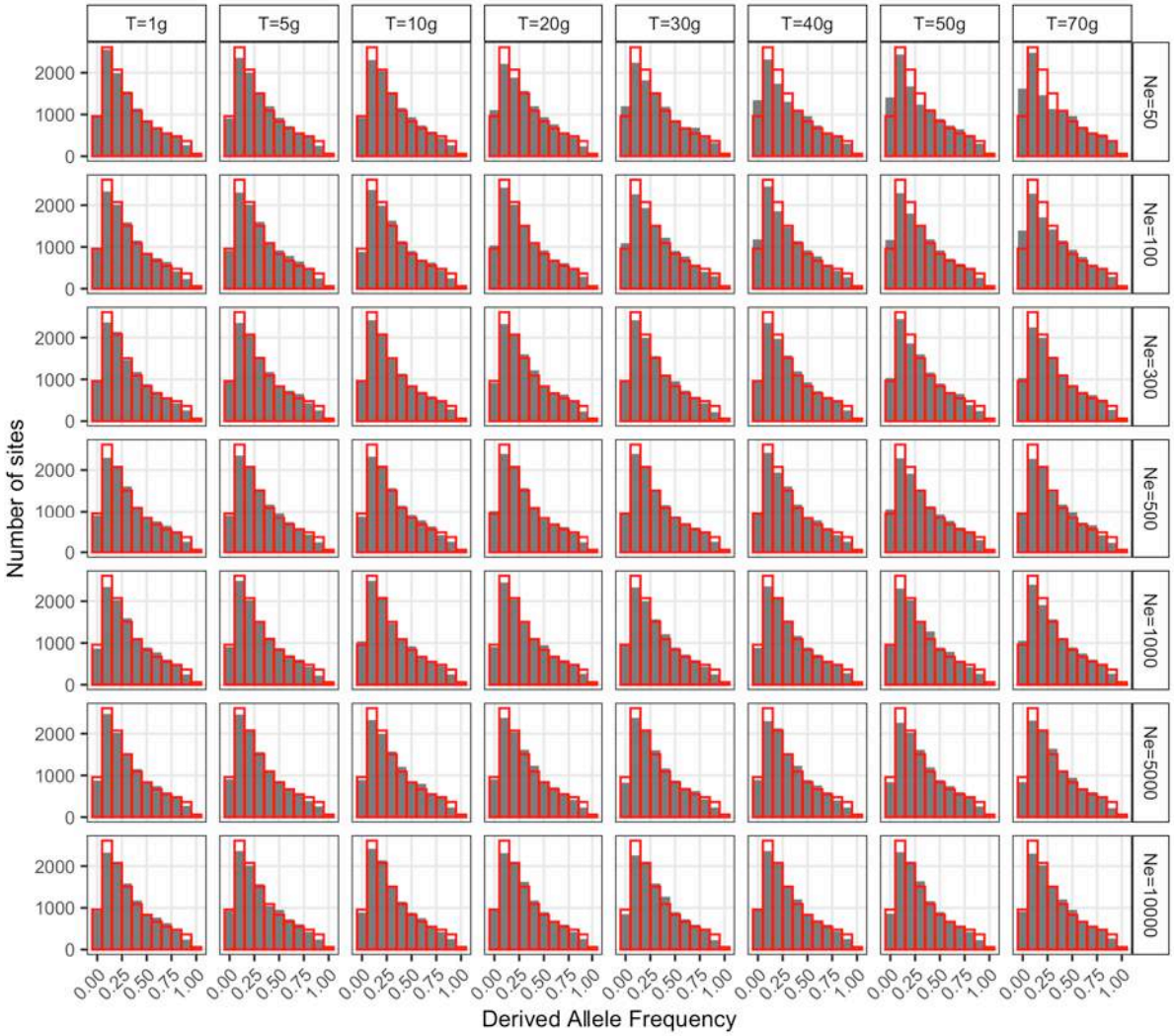
G


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

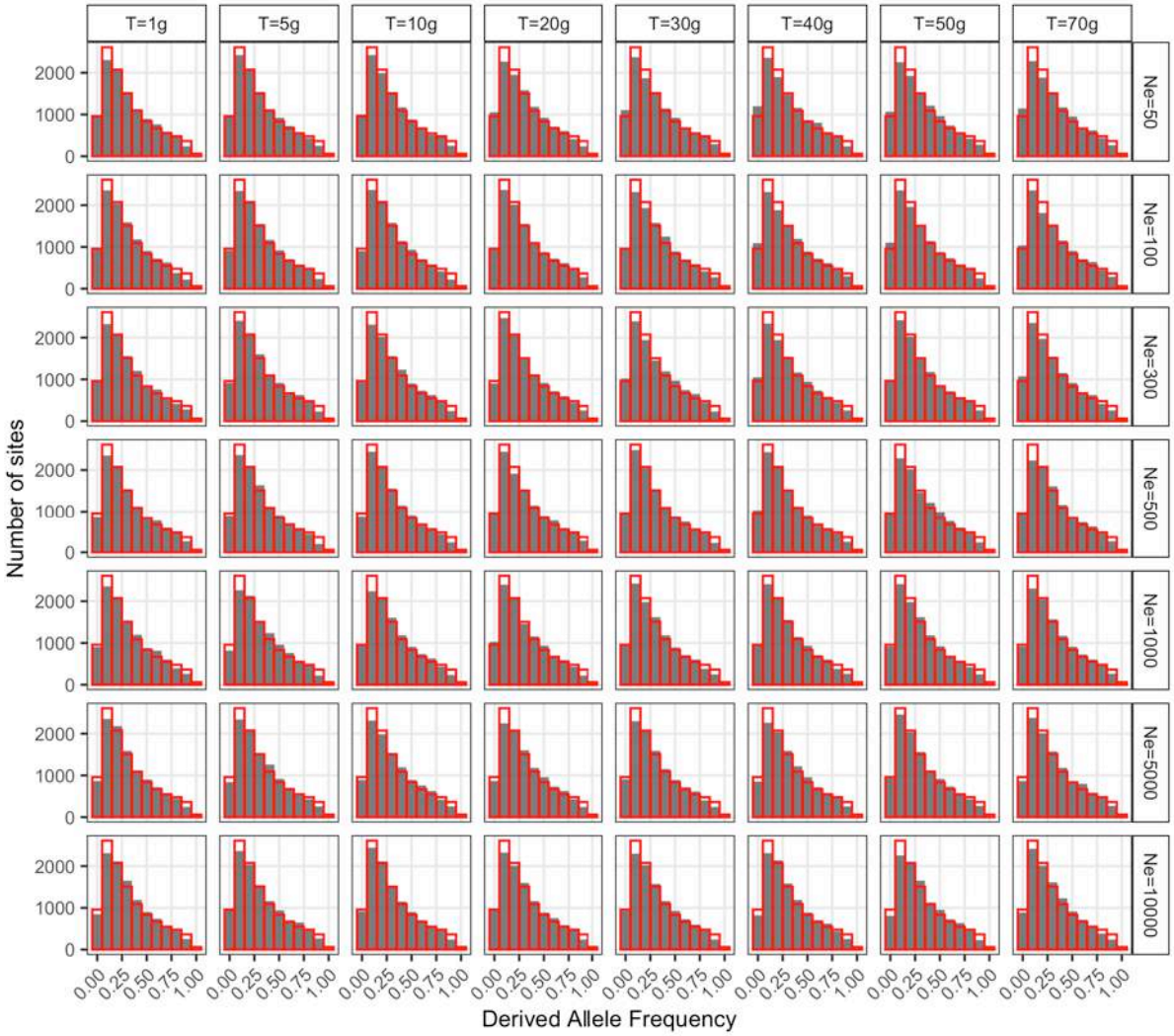
H


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

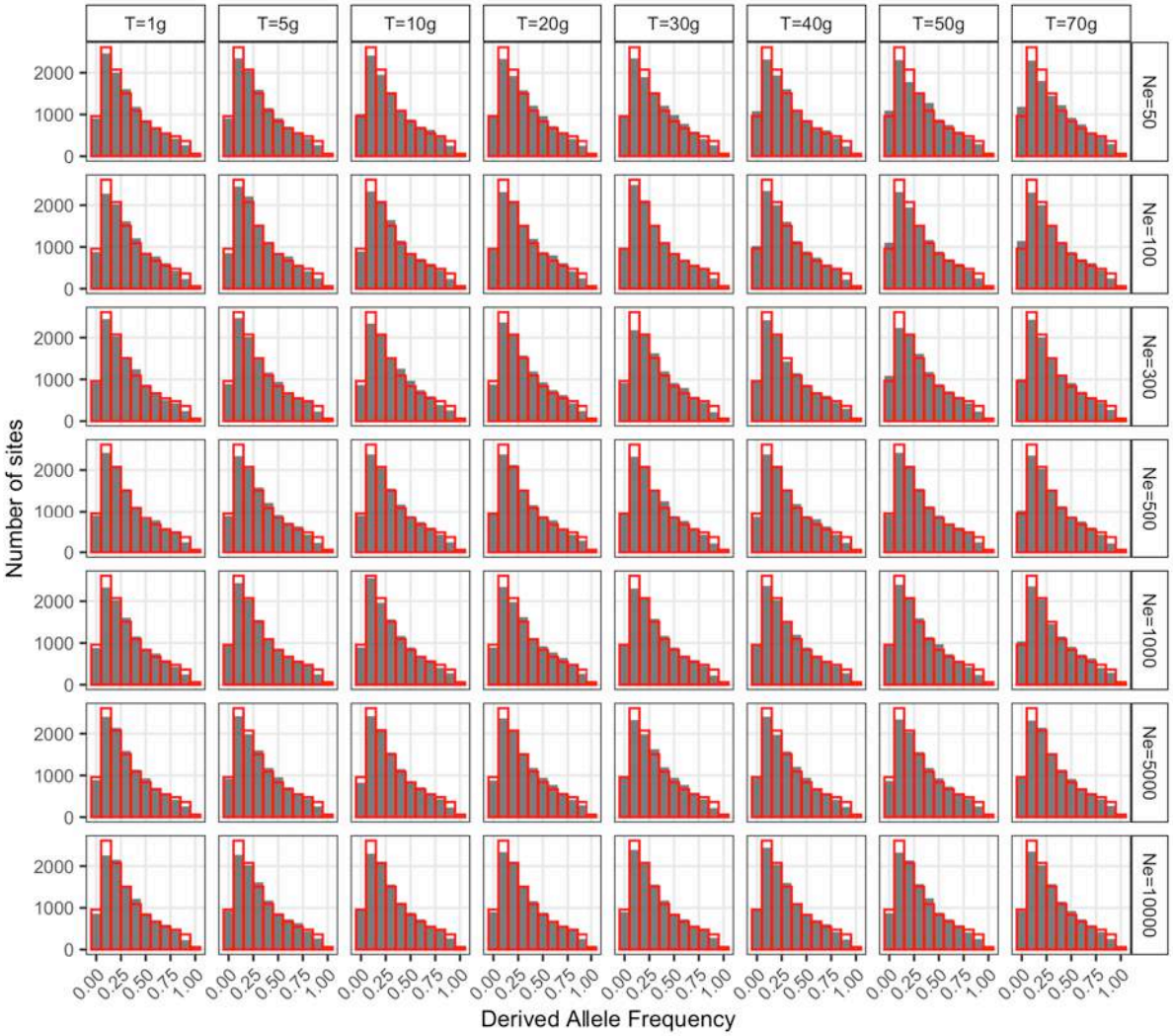
I


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

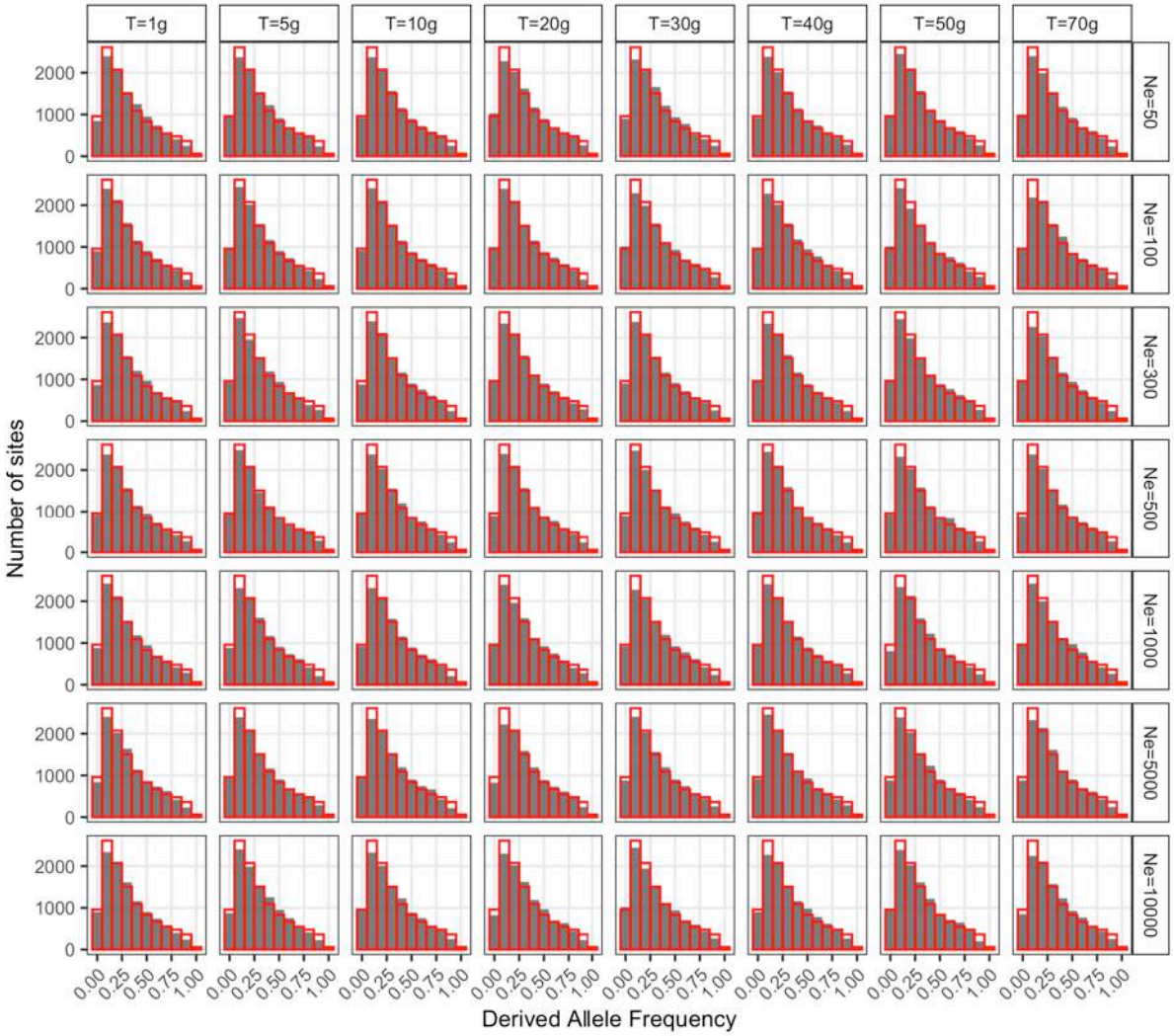
J


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

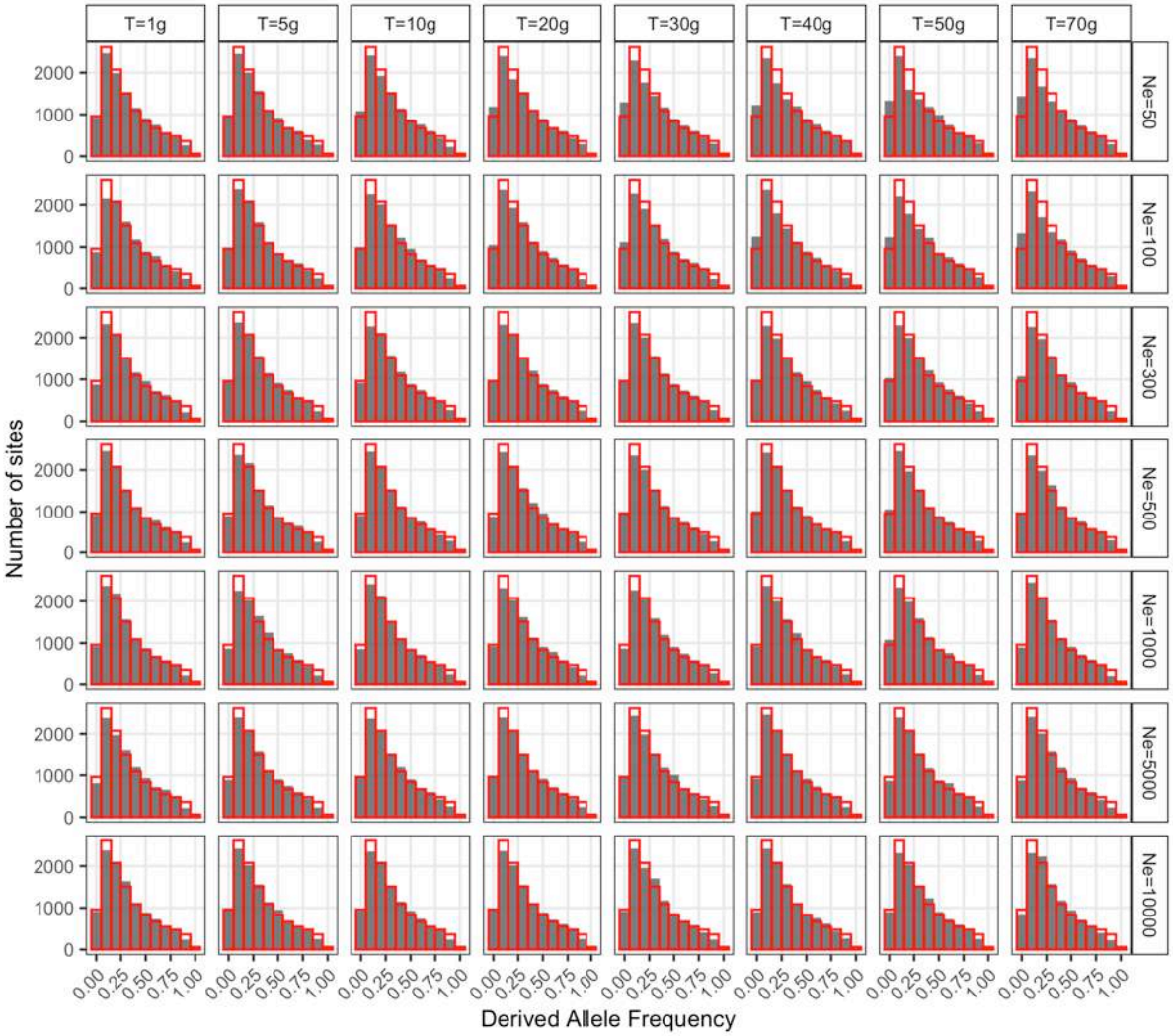
K


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

L


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

M


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of sites
0
1000
2000
0.00
0.25
0.50
0.75
1.00
Derived Allele Frequency

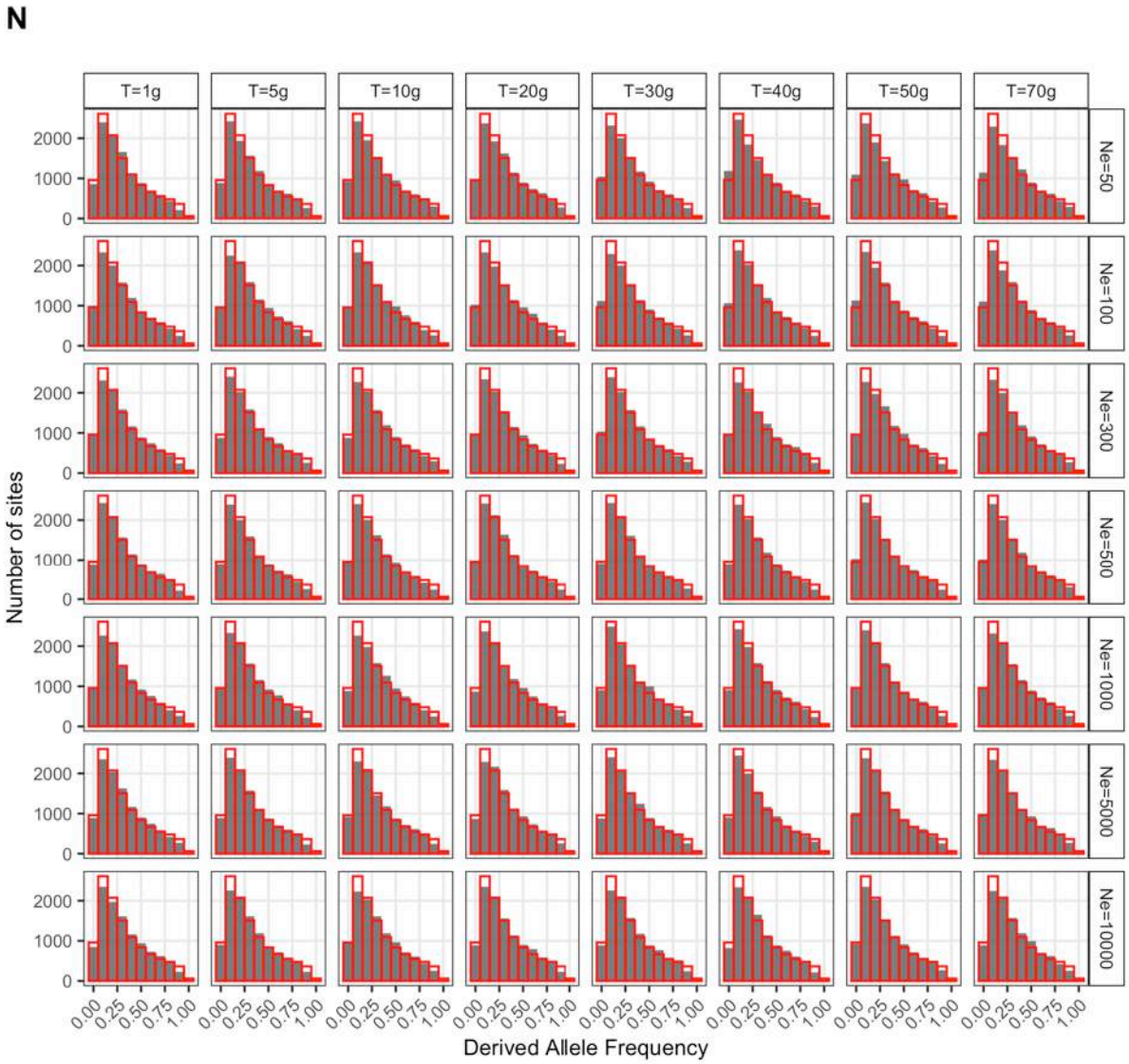
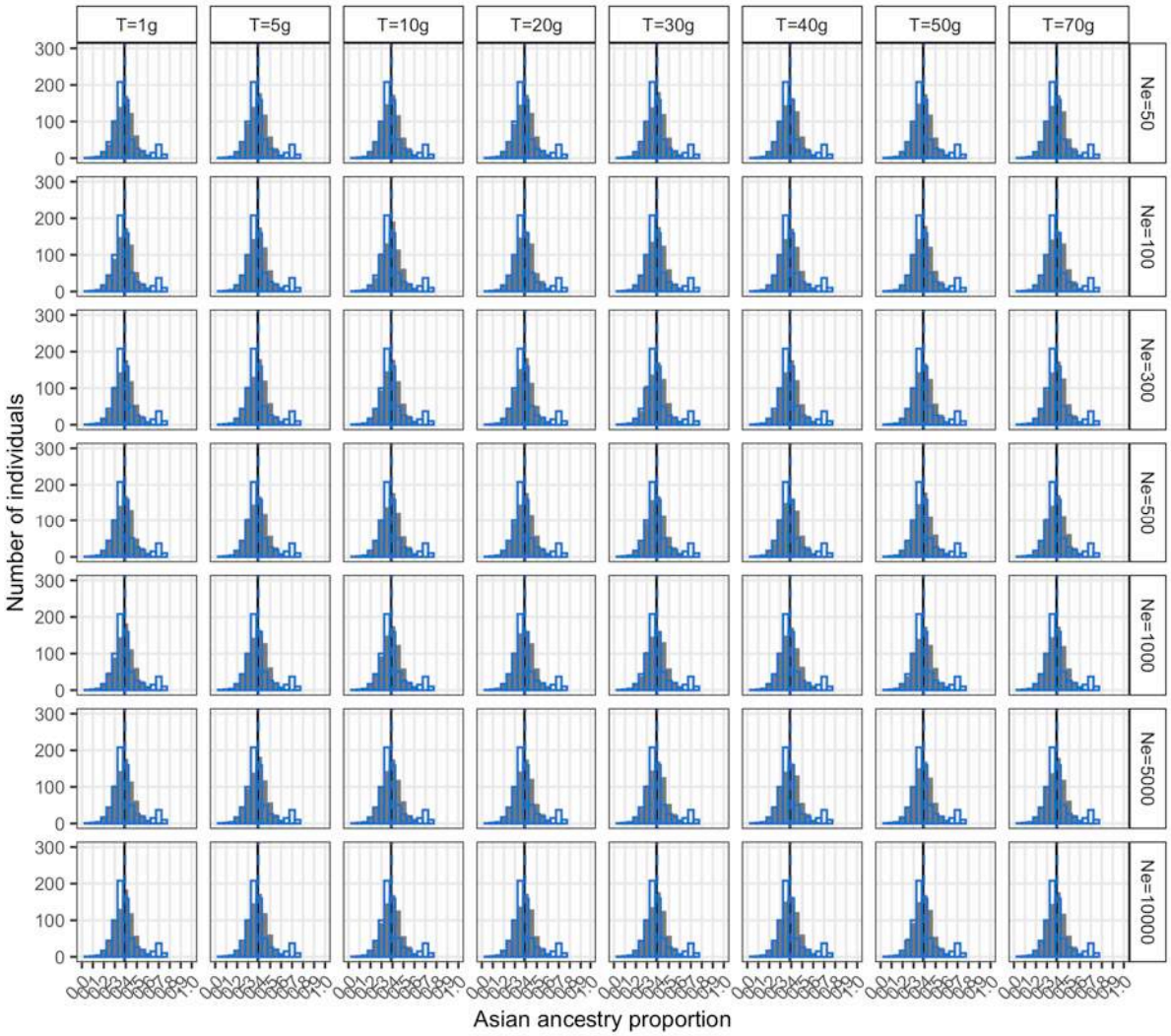
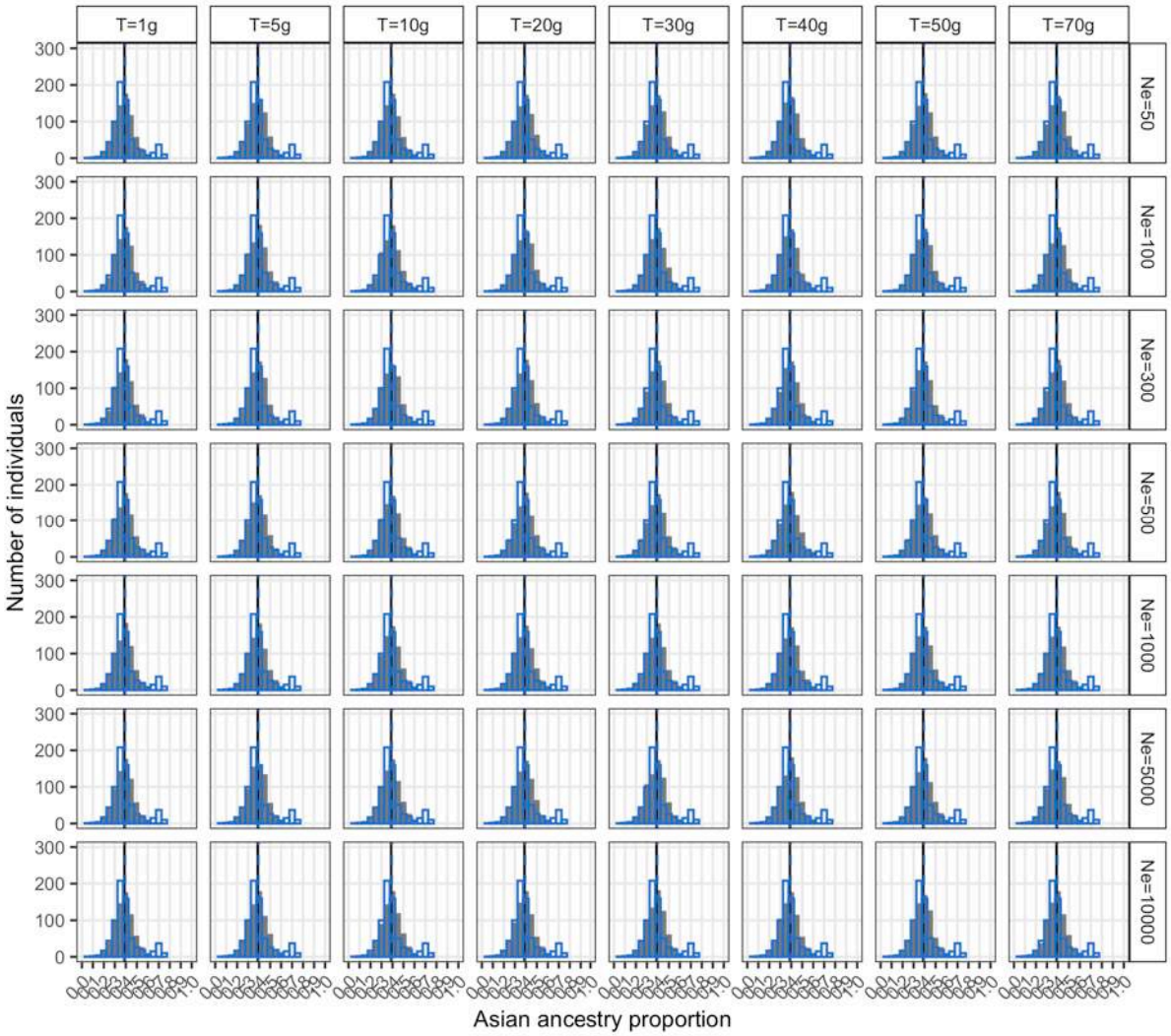


Figure S21. Effect of different demographic scenarios on the simulated global ancestry profile in Madagascar. Each panel shows the result for simulated scenarios. The details of these correspond to the legend from Fig. S19.

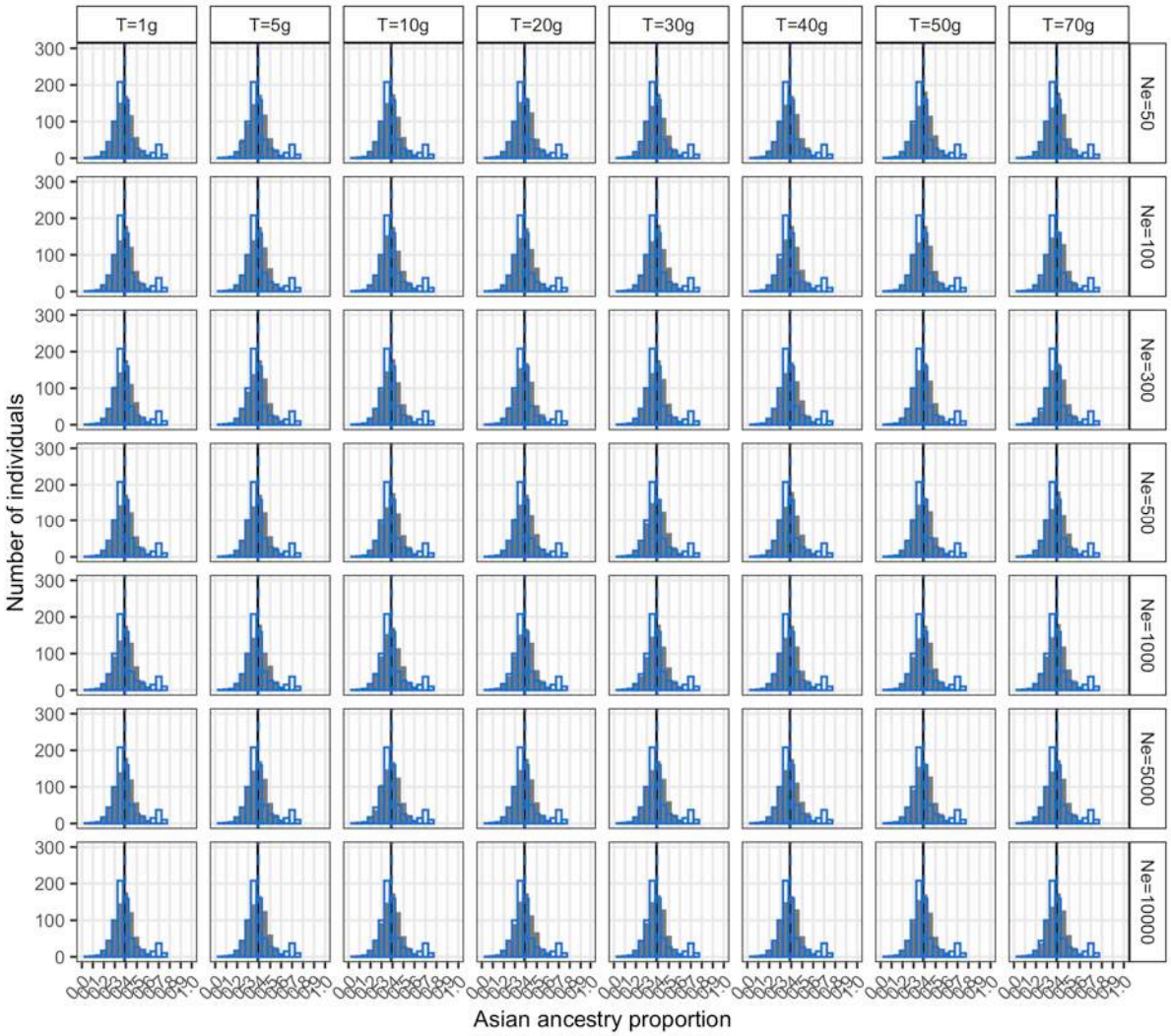
A


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

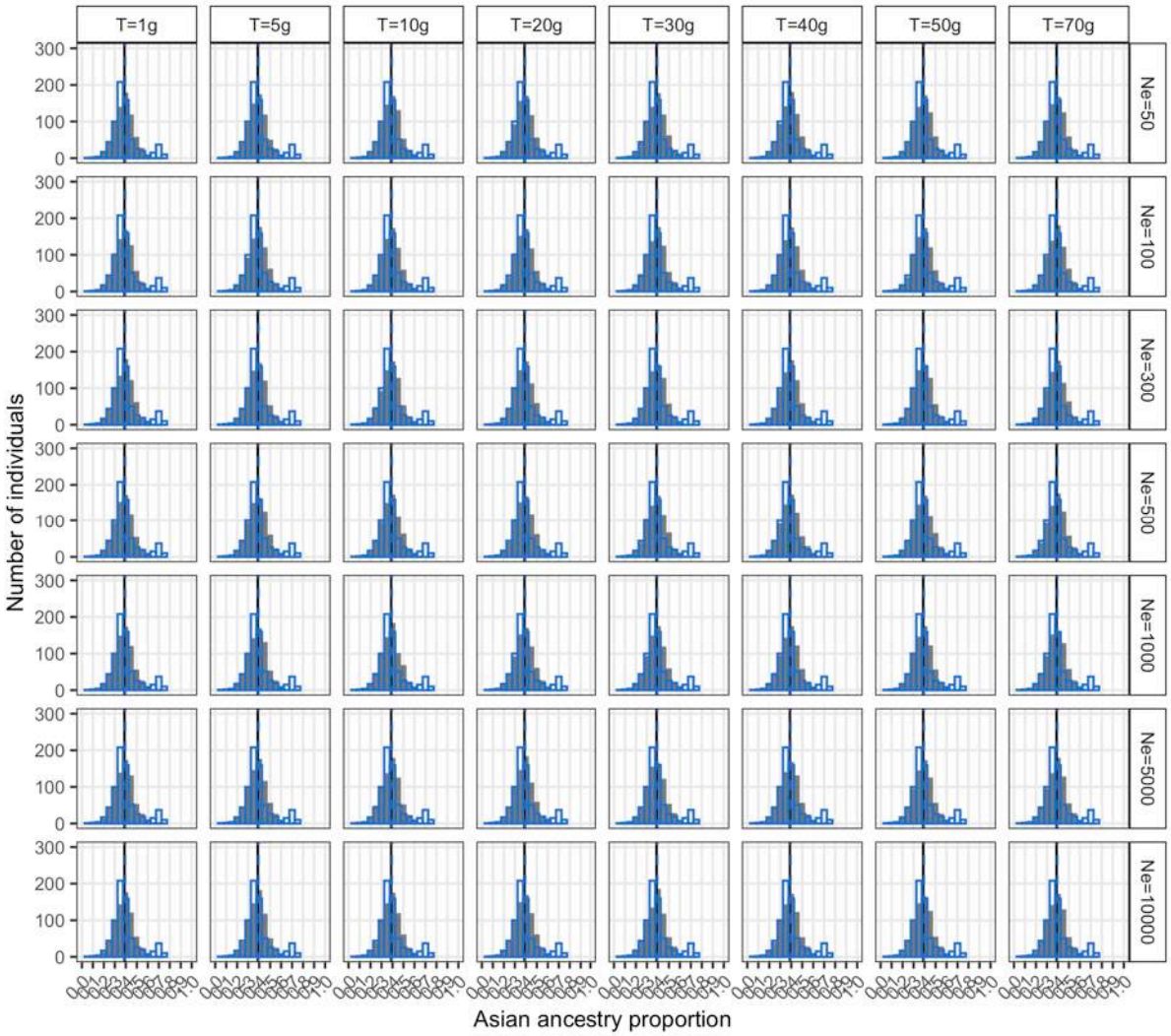
B


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

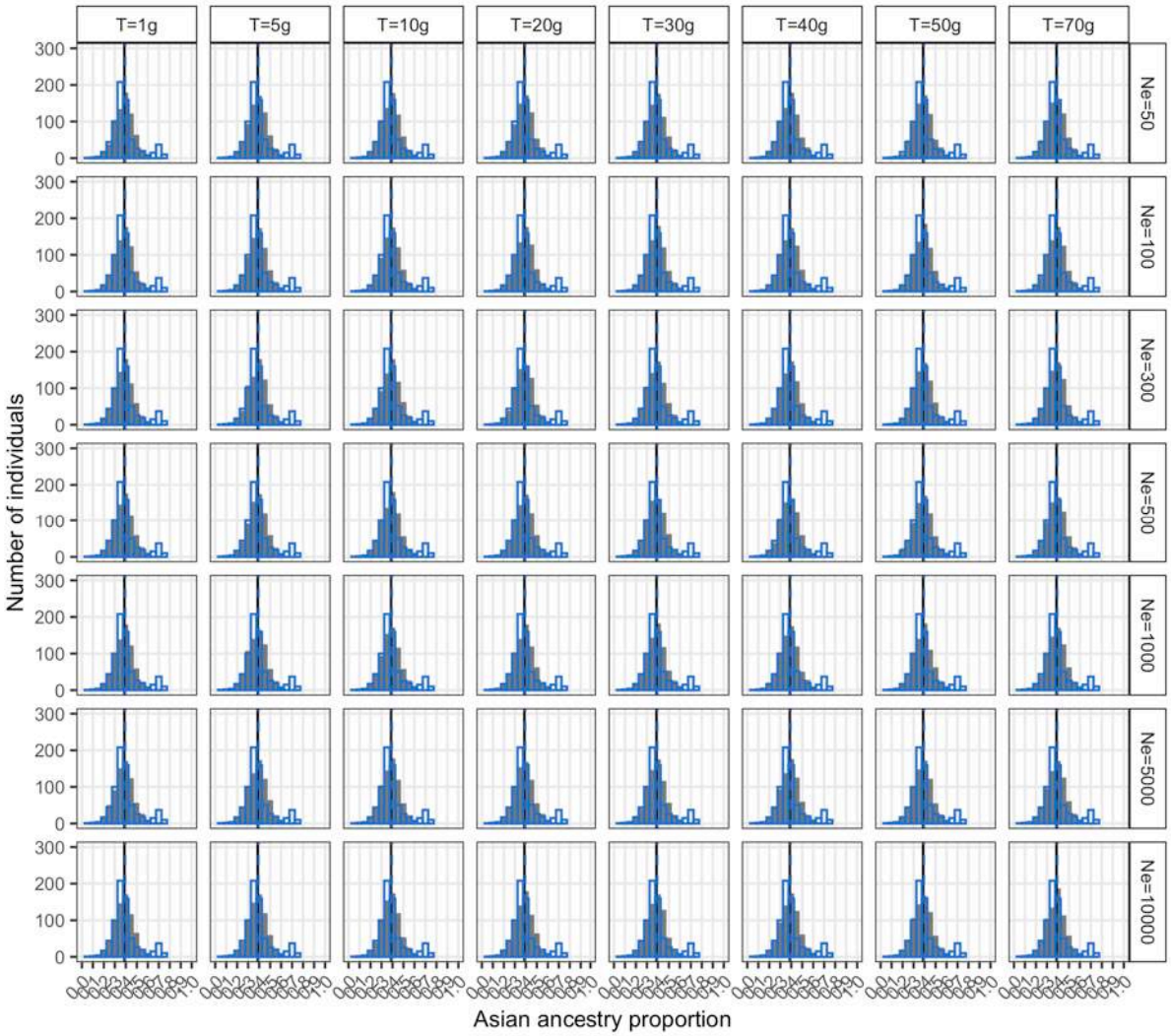
C


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
0
100
200
300
Asian ancestry proportion

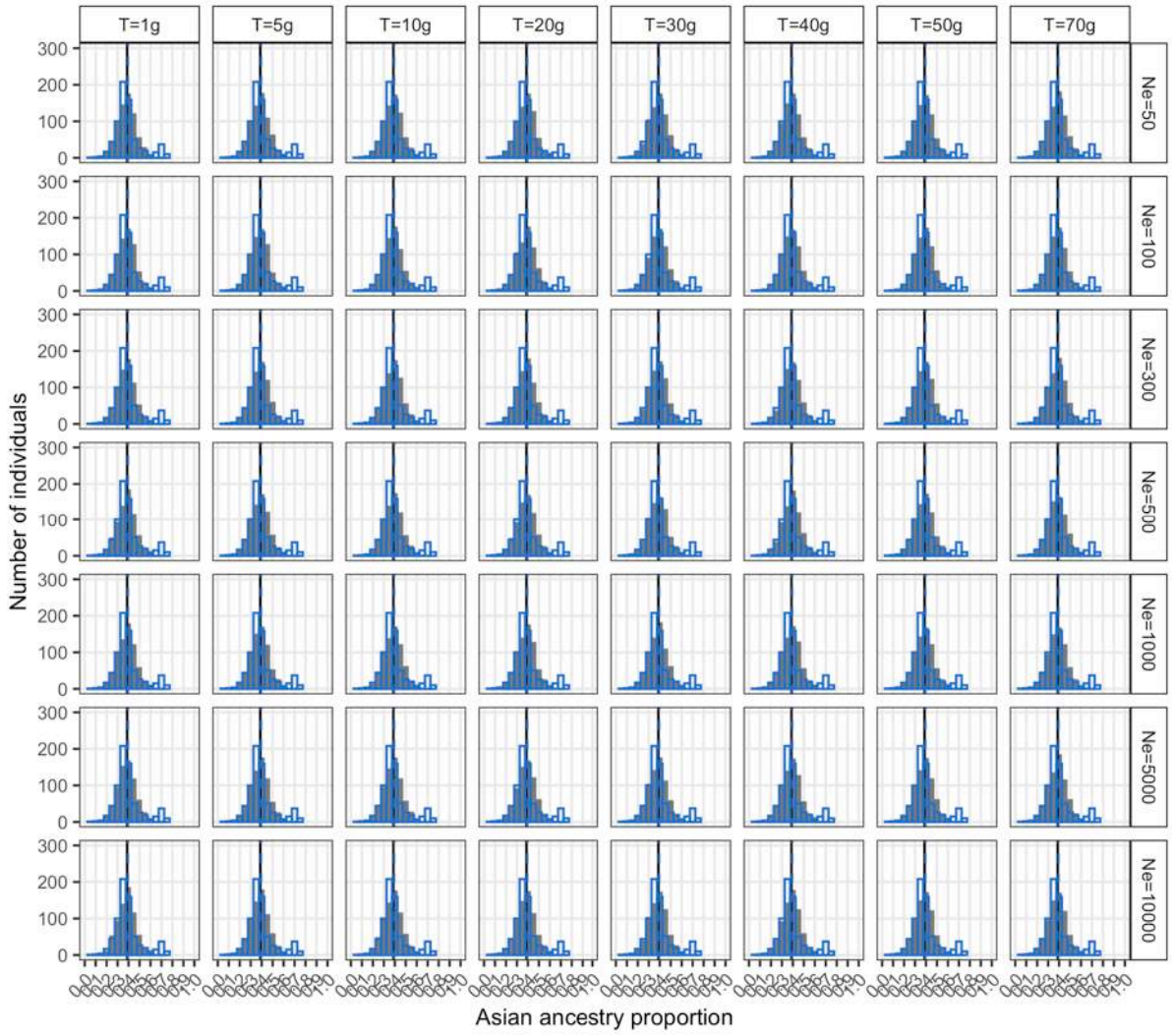
D


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

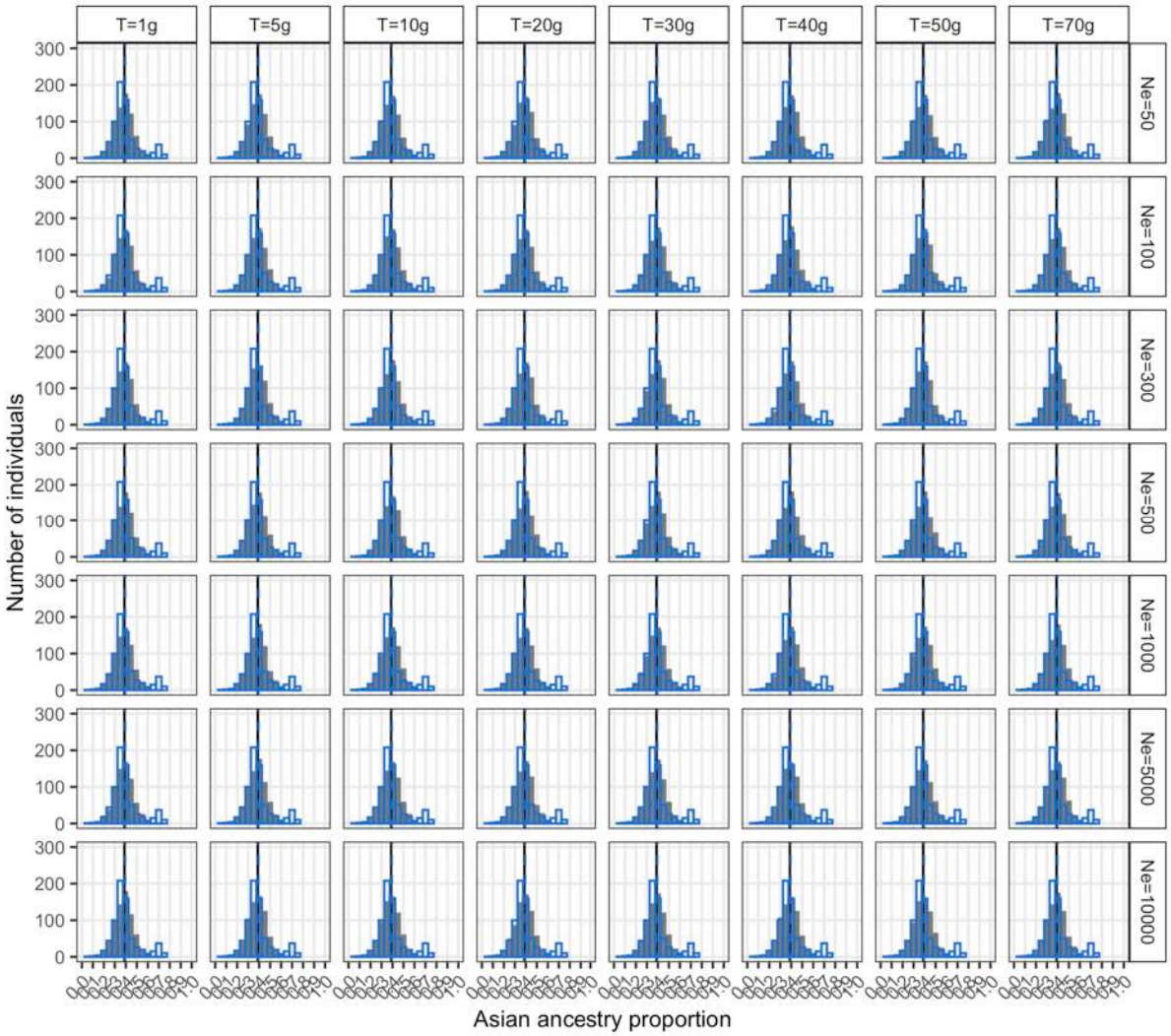
E


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

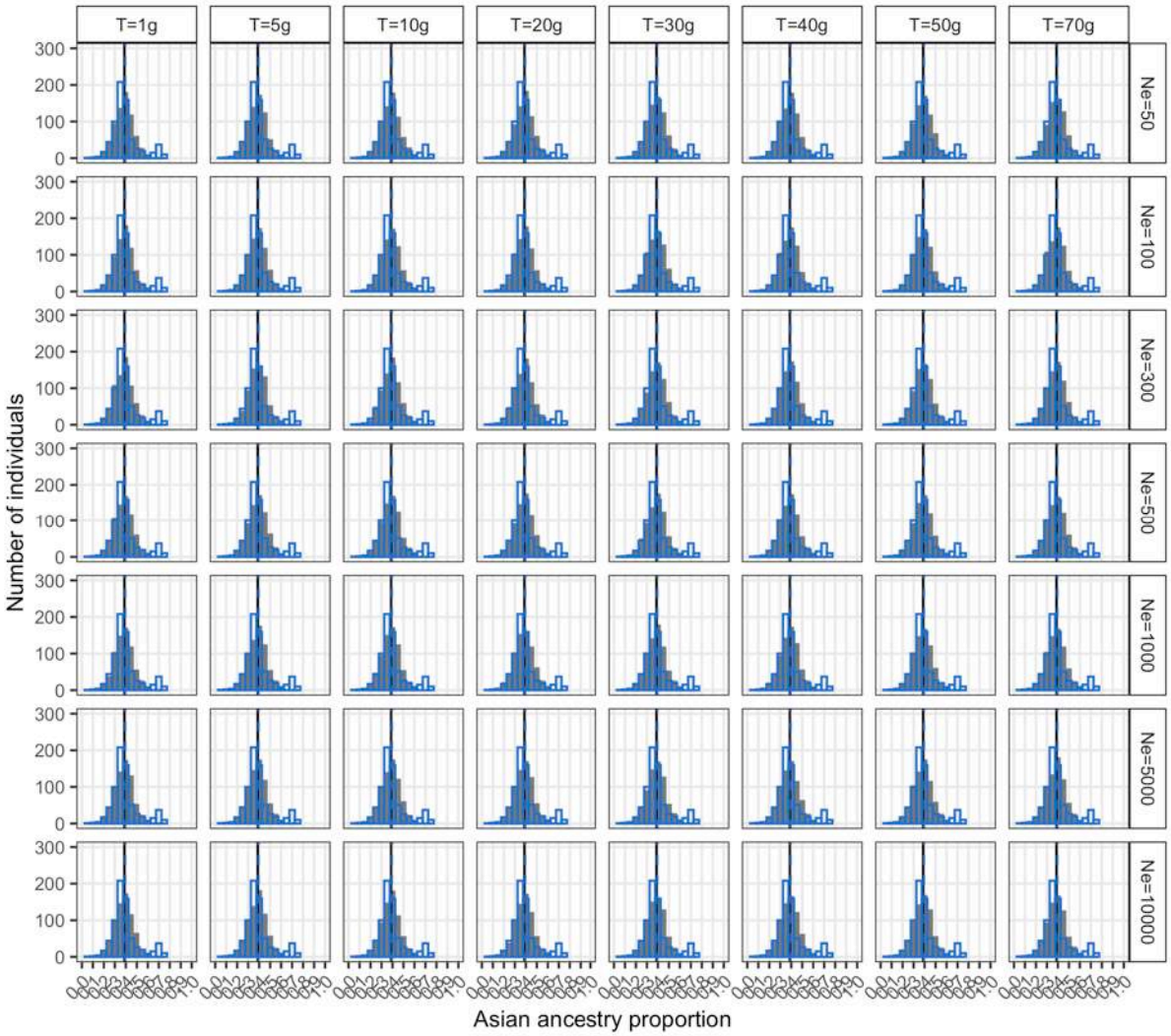
F


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

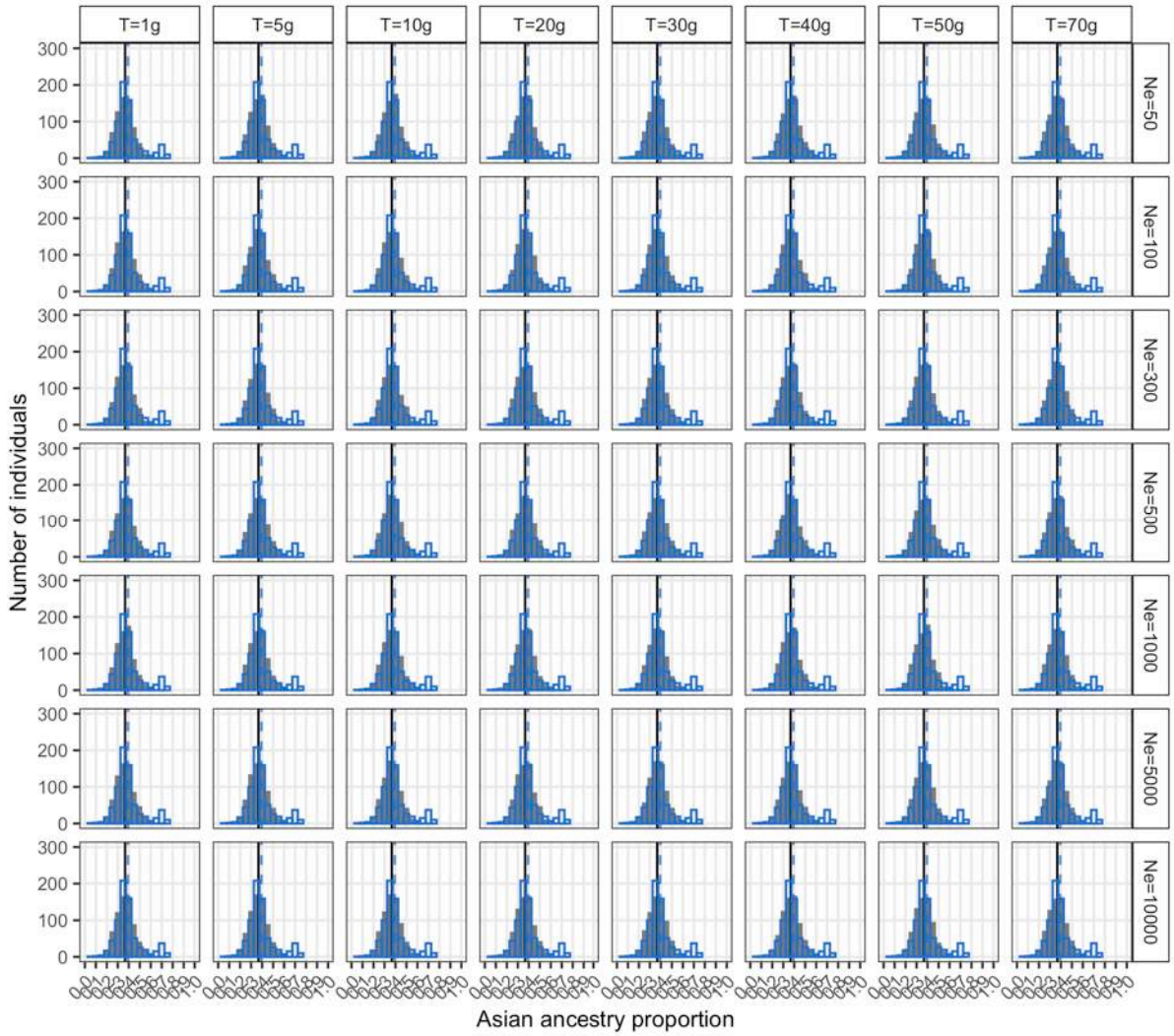
G


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
0
100
200
300
Asian ancestry proportion

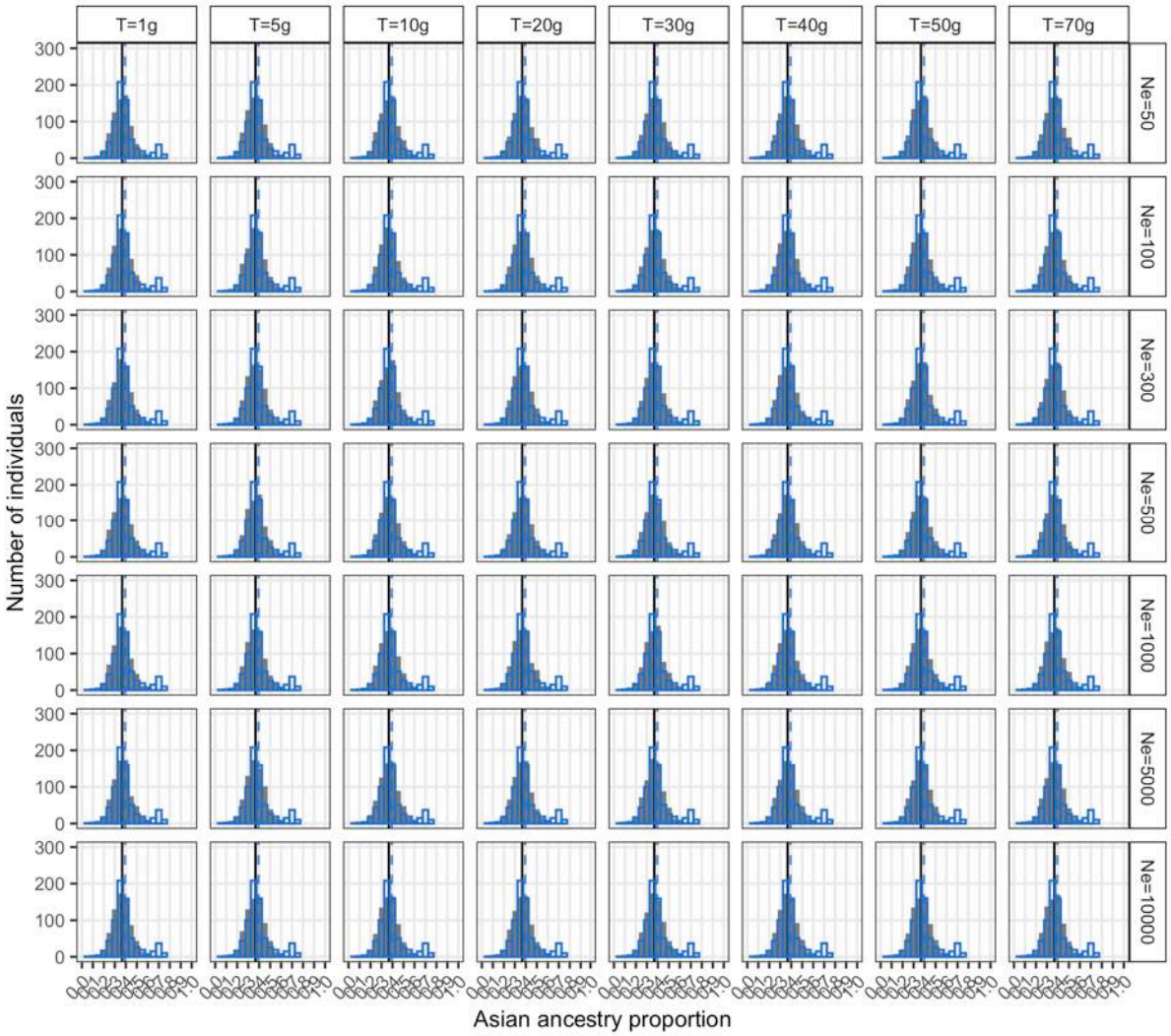
H


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

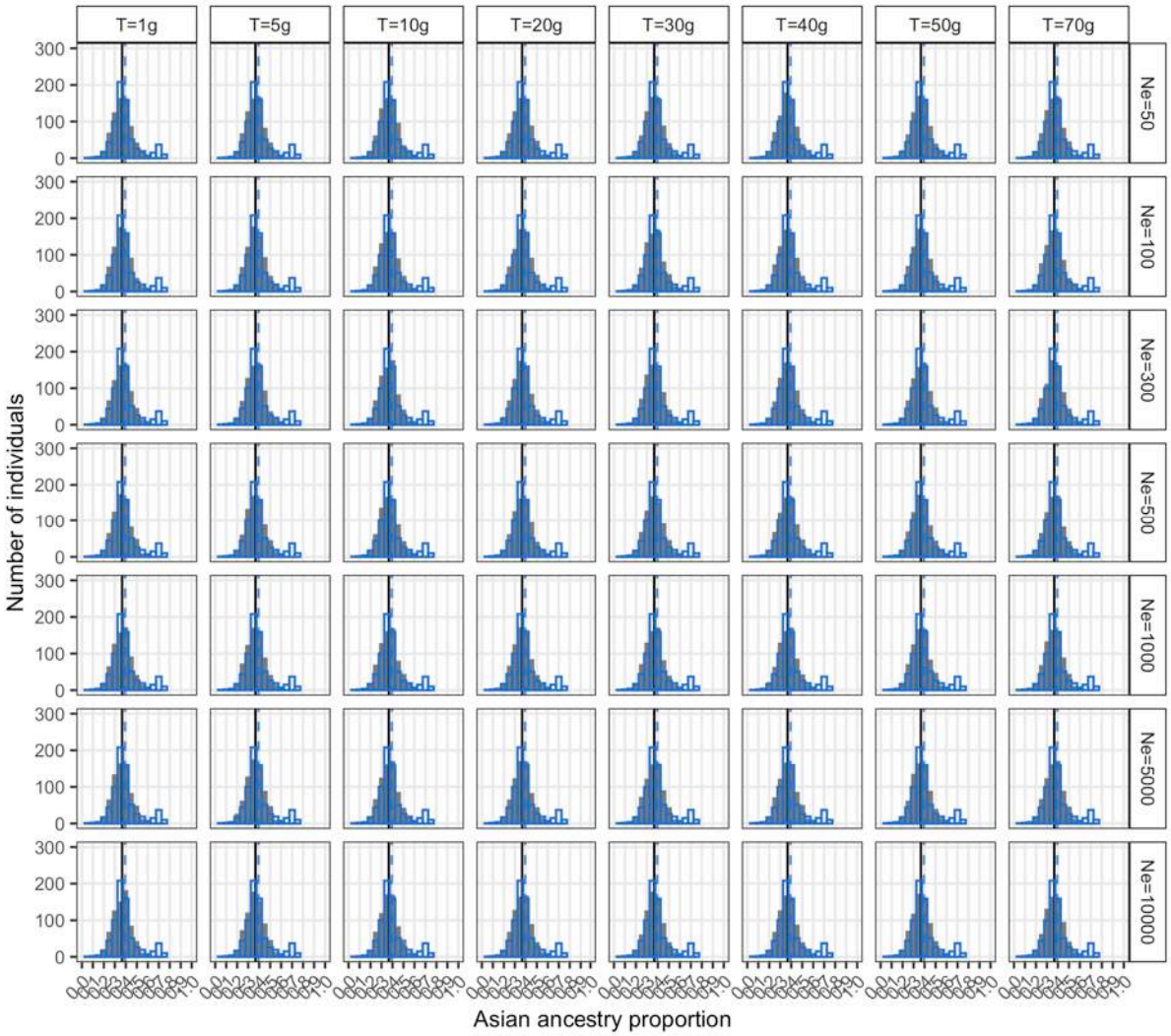
I


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

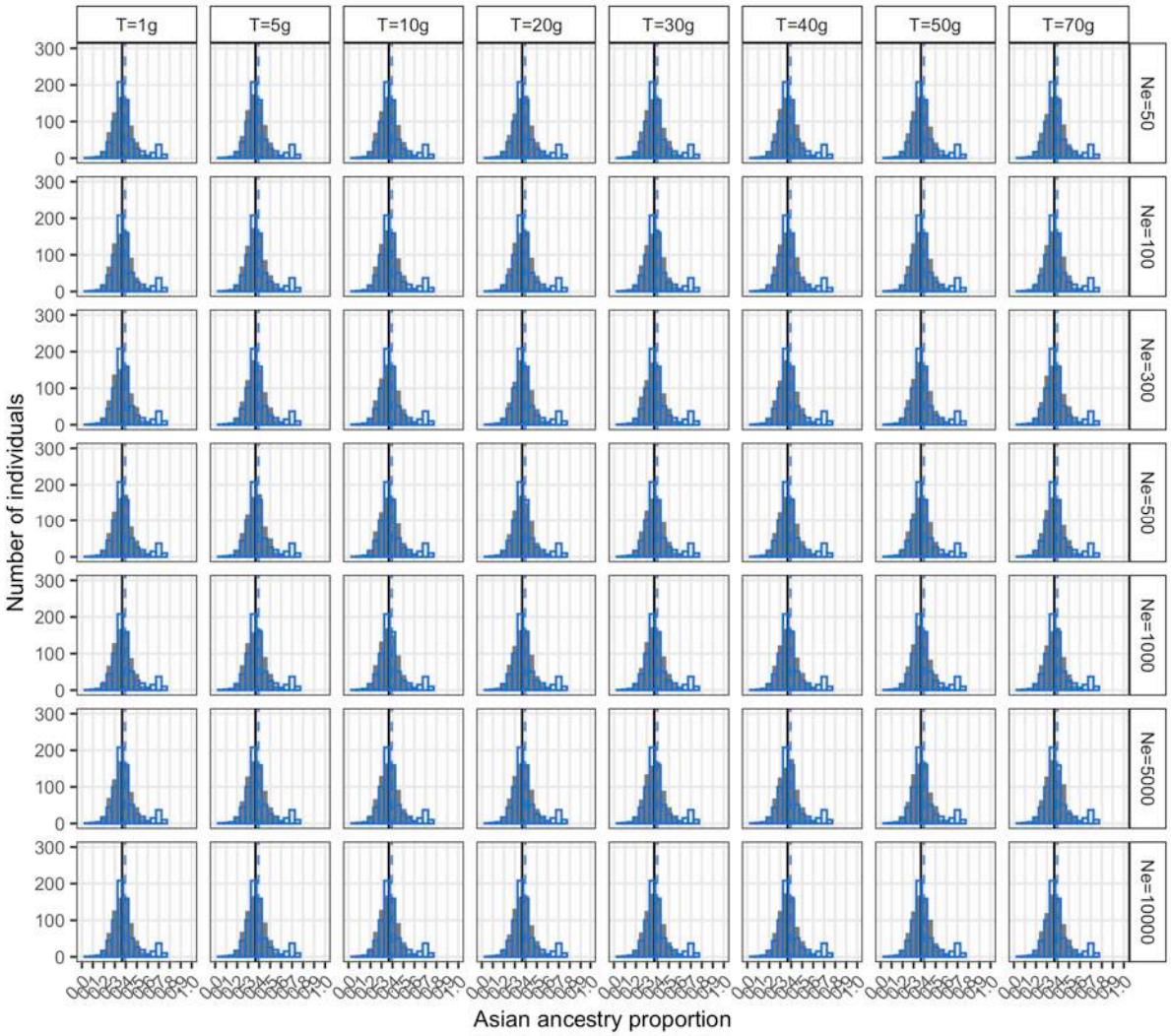
J


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

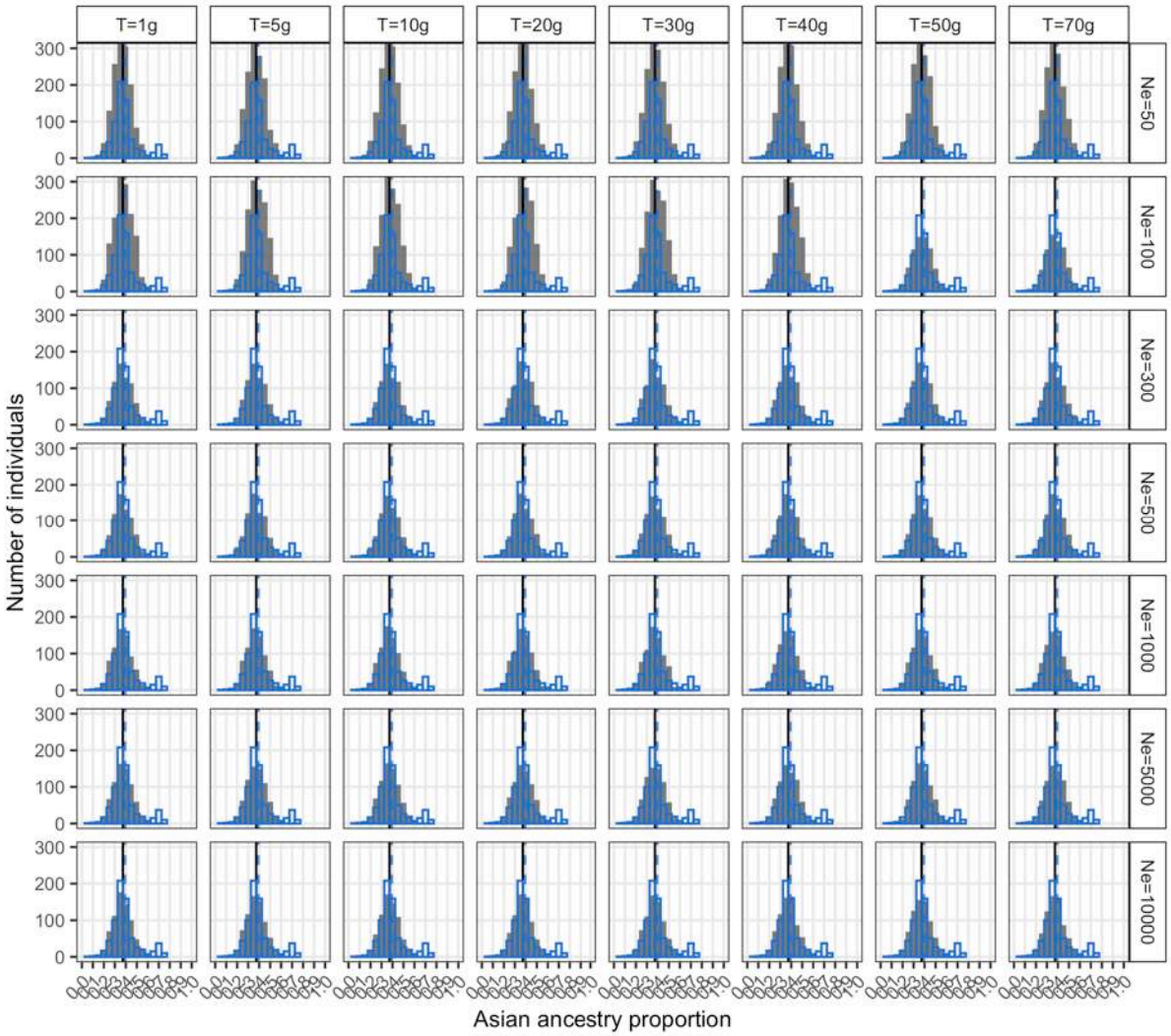
K


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

L


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

M


T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
Asian ancestry proportion

N

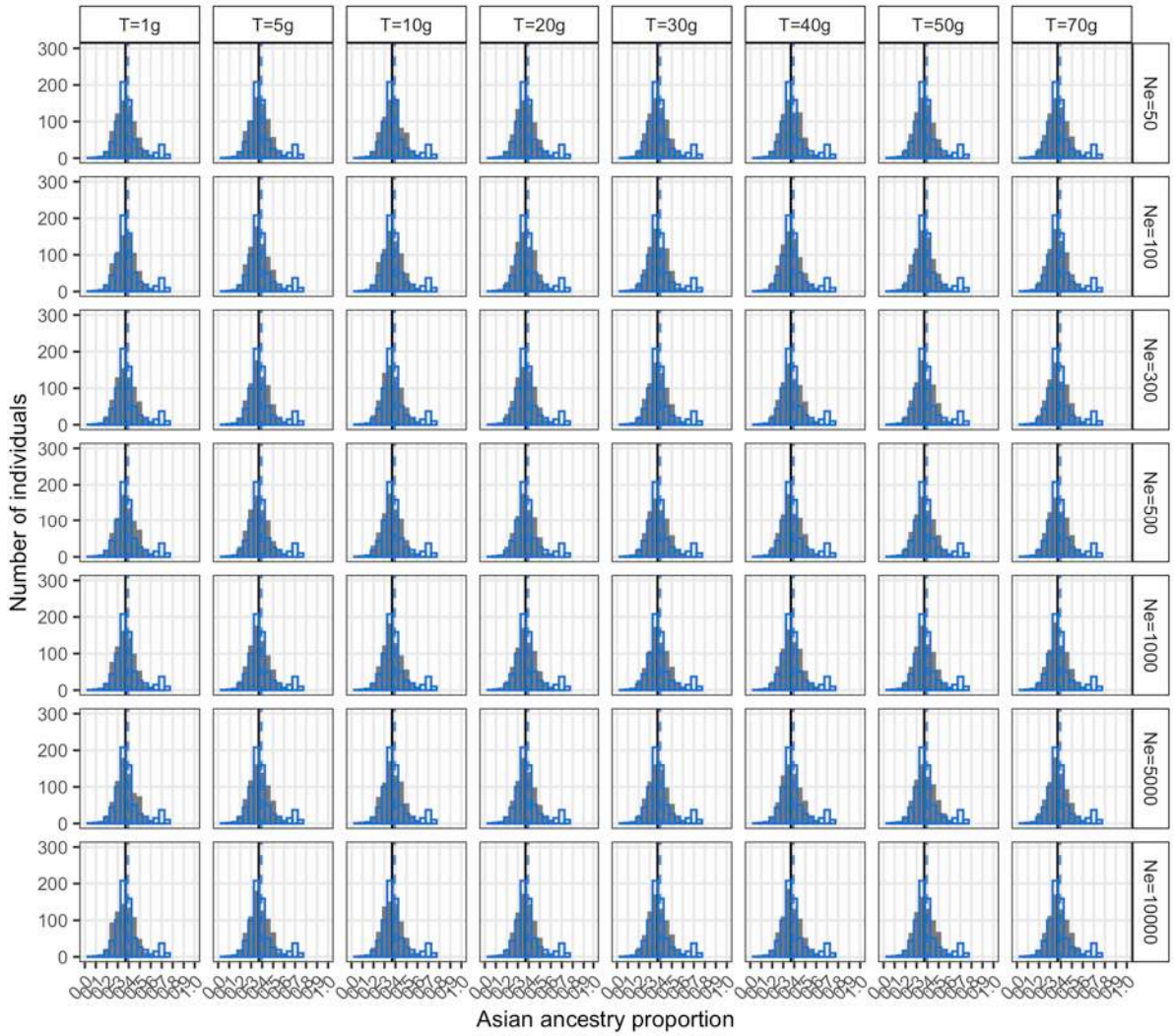

T=1g
T=5g
T=10g
T=20g
T=30g
T=40g
T=50g
T=70g
Ne=50
Ne=100
Ne=300
Ne=500
Ne=1000
Ne=5000
Ne=10000
Number of individuals
0
100
200
300
Asian ancestry proportion

Figure S22. Relationship between age of IBDs and their size detected during a long bottleneck scenario (scenario 6 of Figure S12). The limit of 5 centimorgan is marked by a red line. All IBD over 45 generations appear to be the same size, consistent with the fact that this relationship would only be correct in the case of a sufficiently large population and would cease to be correct during a bottleneck period (30-75 generations).

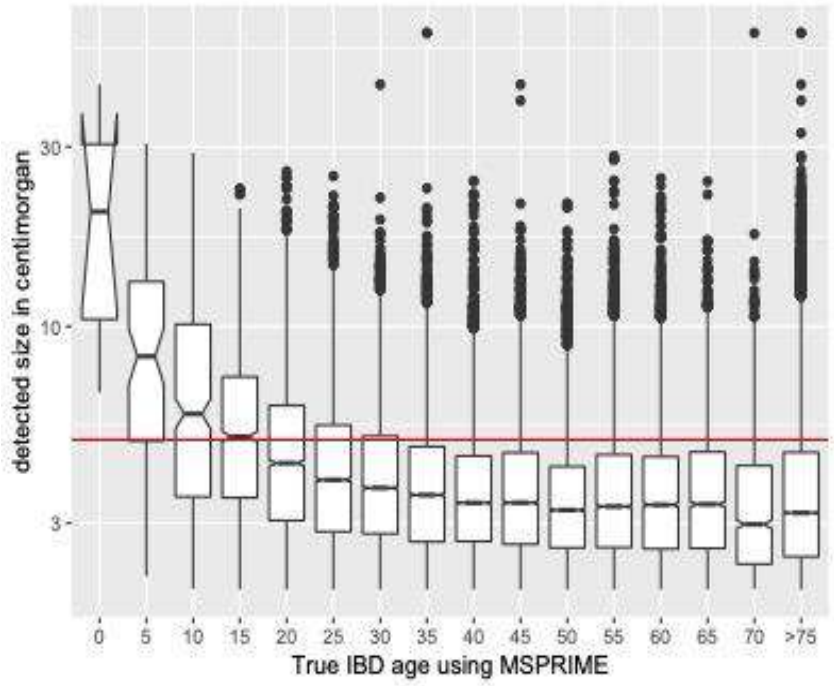


Figure S23. Effect of different simulation ancestry models on the Asian IBD-sharing distribution. We simulated the same demographic scenario (Ne_startBOT=500; Ne_endBOT=500; Tdur_BOT=40g, Mig_BOT=0.0) under the Standard Coalescent model, the discrete time Wright-Fisher model and a combination of these two (DFTW during the first 500 generations and the Standard Coalescent until the end of simulation).

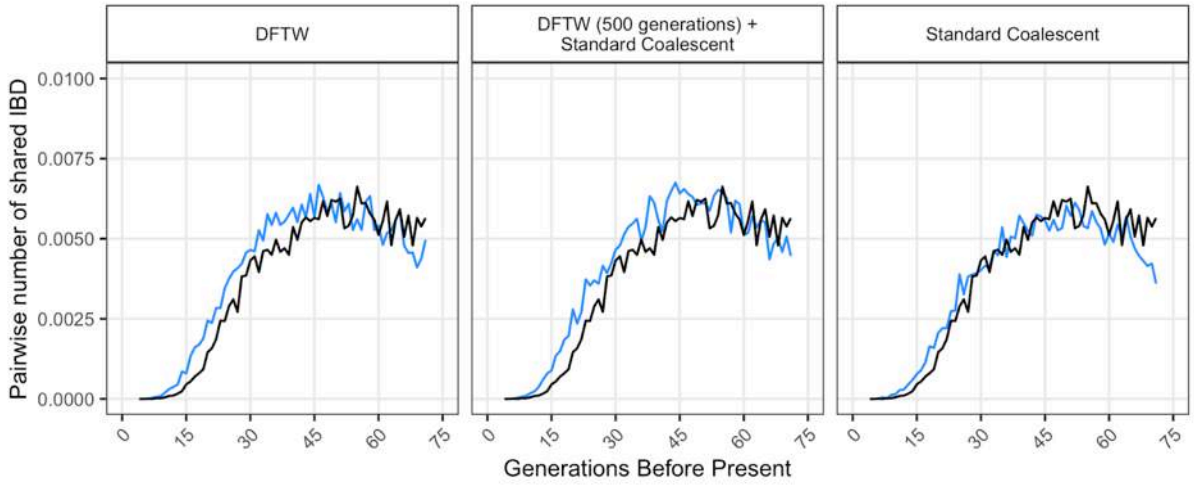
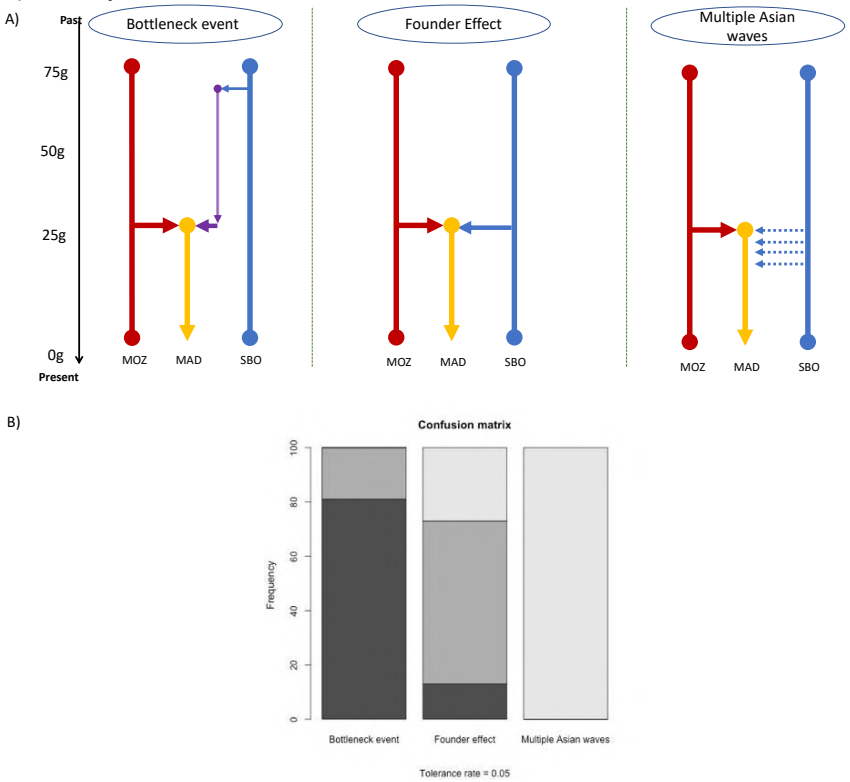
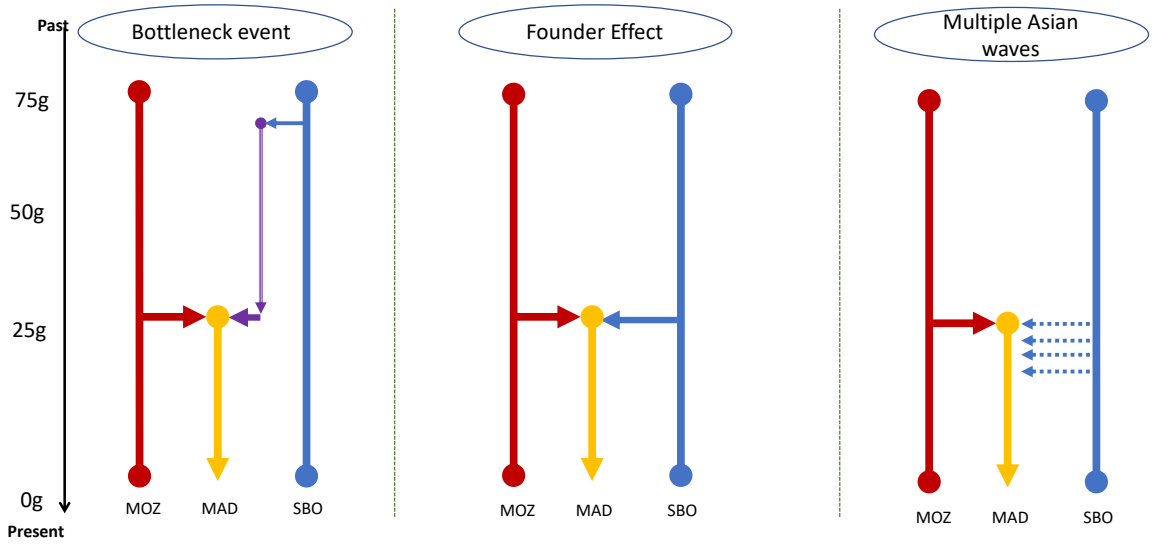


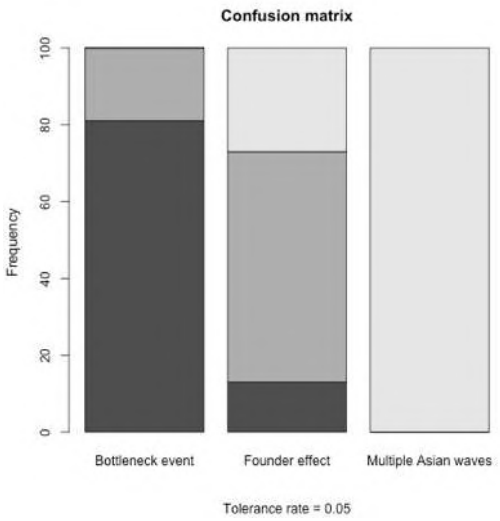
Figure S24. ABC model misclassification. A) Graphical representation of the three demographic models tested. B) Model choice step in the ABC approach. Confusion matrix for three models. The colors from dark to light gray correspond to models Bottleneck, Founder effect, Multiple Asian waves of admixture, respectively.



A)


Past
75g
50g
25g
0g
Present
Bottleneck event
MOZ
MAD
SBO
Founder Effect
MOZ
MAD
SBO
Multiple Asian waves
MOZ
MAD
SBO

B)


Confusion matrix
Frequency
0
20
40
60
80
100
Bottleneck event
Founder effect
Multiple Asian waves
Tolerance rate = 0.05

SUPPLEMENTARY TABLES

Table S1. Summary of populations used in the study and their sample sizes.

Table S2. IBDNe maximum and minimum estimated effective population sizes at the inferred date of admixture (in generations before present, according to MALDER). Growth rate was calculated using the Ne during a period of 10 generations after the admixture date.

Population	Estimated date of admixture (generations BP)	Global Estimated Ne	African ancestral pre-admixture Ne	Asian ancestral pre-admixture Ne	Growth rate
Madagascar	24.38±0.81	3610-4640	3800-4560	732-1170	0.3041
Group 01	23.70±1.08	514-625	434-523	235-278	0.1166
Group 02	22.83±1.14	2130-2840	1900-2590	583-953	0.1441
Group 03	19.99±0.95	21700-36000	15500-31900	2240-8110	0.1392
Group 04	25.94±0.88	2110-3180	1990-3090	478-616	0.0341
Group 05	24.10±1.21	2350-4040	2510-4090	503-683	0.0415
Group 06	21.82±0.97	7360-12900	4760-10400	1860-7280	0.3391
Group 07	25.53±1.60	3720-5000	3140-4240	512-837	0.0692
Group 08	29.00±1.11	1630-2240	1470-1830	376-513	0.1031
Group 09	24.70±1.04	3250-3890	2620-3380	615-989	0.2117
Group 10	28.13±1.72	2030-3090	1700-2440	407-537	0.1066

Table S3. Average reported scores from Asian IBD-sharing absolute differences between observed and simulated data.

Table S4. Description of Asian ancestral simulated demographic scenarios and Malagasy settlement.

Simulated scenarios	Simulated admixture		Simulated Asian Ancestral population			
	Admixture model	Date	Split time from South Borneo	South Borneo Ne at split time	Strength of founder event	Ne before admixture
Founder event	Single pulse of admixture	25 gBP	25 gBP	Ne=30000	Ne=900	Ne=30000
Strong founder event	Single pulse of admixture	25 gBP	26 gBP	Ne=30000	Ne=50	Ne=50
Multiple Asian waves	Multiple pulses of admixture (Asian gene flow)	20-30 gBP	30 gBP	Ne=28000	Ne=300	Ne=28000
Slow decrease of a large population	Single pulse of admixture	25 gBP	75 gBP	Ne=14000	Ne=5000	Ne=900
Bottleneck	Single pulse of admixture	25 gBP	35 gBP	Ne=26000	Ne=50	Ne=900
Long-term bottleneck	Multiple pulses of admixture (African gene flow)	25 gBP	65 gBP	Ne=16900	Ne=500	Ne=500

Table S5. Migration rates used in previous models of human demographic history.

Migration rate	Nb individuals / gen	Simulated connections		Pattern of gene flow	Distance	# Reference
7.80E-06	0.11	Africa	East Asia	Continental	~8000 km	14
5.72E-05	0.25	East Asia	Papua	Continental	~7000 km	79
2.50E-05	0.36	Africa	Europe	Continental	~5000 km	14
3.11E-05	1.06	Europe	East Asia	Continental	~6,000 km	14
1.00E-02	3	Finland	Estonia	Local (Partially isolated)	~500 km	78
2.00E-02	6	Finland	Estonia	Local	~500 km	78
3.00E-02	9				~500 km	78
4.00E-02	12				~500 km	78
5.00E-02	15	Finland	Estonia	Local (High-connected)	~500 km	78

Table S6. Inferred admixtures dates on simulated data according to different admixture models.

Admixture model	Simulated admixture date	MALDER date of admixture (generations BP)
Single pulse of admixture	20 generations BP	21.03±0.52
Single pulse of admixture	25 generations BP	26.82±0.84
Single pulse of admixture	30 generations BP	32.14±1.00
Multiple pulses of admixture (African gene flow)	20-25 generations BP	23.62±0.60
Multiple pulses of admixture (African gene flow)	20-30 generations BP	26.62±0.76
Multiple pulses of admixture (African gene flow)	20-35 generations BP	28.04±0.79
Multiple pulses of admixture (African gene flow)	20-40 generations BP	31.46±1.19

Table S7. Prior distributions of the parameters for the studied demographic models.

Model	Parameters	Variable	Minimal value	Maximal value	Shape	Notes
Founder effect	Madagascar Ne at the beggining of admixture	Ne_Mad	50	10000	Uniform	Admixture starts at 25 generations BP
Multiple pulses of admixture	Number of Asian pulses towards an African population	Num_pulses	2	15	Uniform	Admixture starts at 30 generations BP, it will finish after the sampled number of pulses. Ne_Mad is constant = 2500
Bottleneck	Ne at the beggining of bottleneck	Ne_startBOT	50	10000	Uniform	Admixture starts at 25 generations BP. Ne_Mad is constant = 2500
	Duration of bottleneck	Tdur_BOT	1	75	Uniform	
	Ne at the end of bottleneck	Ne_endBOT	50	10000	Uniform	Growth rate during the bottleneck corresponds to { log(Ne_endBOT/Ne_startBOT) / Tdur_BOT}
	Migration rate during the bottleneck	Mig_BOT	0.000001	0.02	Uniform	

Table S8. Estimated demographic parameters under the Bottleneck event model based on ABC method (tolerance rate of 5 %).

ABC algorithm	Ne_startBOT			Ne_endBOT			Tdur_BOT			Mig_BOT		
	2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Neural Network	471	487	500	365	367	368	66	66	66	0.0021	0.0021	0.0021
Local Linear Regression	109	930	3560	505	615	790	46	65	74	0.0000	0.0063	0.0197
Rejection	52	534	3608	165	1603	8457	22	52	73	0.0004	0.0071	0.0176

Table S9. Effect of minor allele frequency (MAF) filters on allele frequency spectrum of simulated African and Asian reference populations.

CODES USED IN THIS STUDY

We simulated sequence data with a mutation rate of 1.25×10^{-8} per base pair per meiosis, using the HapMap GRCh37 genetic map for each autosomal chromosome. To replicate the real data available, we sampled 700 individuals from the admixed population, 161 from the Mozambique population, 91 from the South Borneo population, 100 from the Bantu population, 100 from the Austronesian population, 400 from East-Asia, 400 from Europe and 419 from Africa.

1. MSPrime code for simulating : Split from Southern Borneo at 25 generations BP through a founder event (Ne=900), followed immediately by the admixture.

```
#!/usr/bin/env python3
#Msprime code used to simulate data - adapted from Browning 2018 IBDNe

# %% Declare constants
simulation_Ne= args.dir + "/" + args.sim + "_chr"+ args.chr +"_Ne.dat"
simulation_name= args.dir + "/" + args.sim + "_chr"+ args.chr +"_msprime.vcf"
print("Generation virtual vcf: ", simulation_name)
recomb_file=
"/Volumes/Genomique/general_data/recombination_maps/recombination_hapmapII_AdamAuto
n/genetic_map_GRCh37_chr" + args.chr + ".txt"
print("Using recombination map:", recomb_file)

# %% Define parameters
# Genomic Configuration
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8 # generate 100 Mb
# Humankind event
N0=7310 # homo sapiens initial population size
Thum=5920 # time (gens) of advent of modern humans
```

```
# Out of Africa event
Naf=14474 # size of african population
Tooa=2040 # number of generations back to Out of Africa
Nb=1861 # size of out of Africa population
mafb=1.5e-4 # migration rate Africa and Out-of-Africa
# Europe-Asia split event
Teu=920 # number generations back to Asia-Europe split
Neu=1032; Nas=554 # bottleneck population sizes
mafeu=2.5e-5; mafas=7.8e-6; meuas=3.11e-5 # mig. rates
reu=0.0038 # growth rate per generation in Europe
ras=0.0048 # growth rate per generation in Asia
# Bantu late split event
Tba= 167 # number generations bantu expansion
Nba= 2200 # initial bantu population size
rba= 0.0111 # growth rate per generation in Bantu expansion
Tmo= 60 # number generation bantu-mozambique split
Nmo= 3000 # initial mozambique population size
rmo= 0.0257 # growth rate per generation in Mozambique
mafba= 3.75e-3 # mig. rates
# Austronesian Out of Taiwan event
Tau= 166 # number generation austronesian expansion
Nau= 5000 # initial austronesian population size
rau= 0.0132 # growth rate per generation in Austronesian expansion
Tsb= 100 # number generation austronesian-southborneo split
Nsb= 5000 # initial south borneo population size
rsb=0.0219
masau= 1.60e-3 # mig. rates
# Madagascar Settlement event
Tadmix=25 # time of admixture
Nadmix= 2500 # initial size of admixed population
radmix= 0.1638 # growth rate of admixed population
seed = args.ran
print("Using Seed:")
print(seed)

# %% Population configurations
# 0 is African, 1 is European, 2 is Asian, 3 is Bantu, 4 is Austronesian, 5 is
SouthBorneo, 6 is Mozambique and 7 is Madagascar
pop_config =
[msprime.PopulationConfiguration(sample_size=838,initial_size=Naf,growth_rate=0.0),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Neu*exp(reu*Teu),
growth_rate=reu),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Nas*exp(ras*Teu),
growth_rate=ras),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nba*exp(rba*Tba),
growth_rate=rba),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nau*exp(rau*Tau),
growth_rate=rau),
     msprime.PopulationConfiguration(sample_size=182,initial_size=Nsb*exp(rsb*Tsb),
growth_rate=rsb),
     msprime.PopulationConfiguration(sample_size=322,initial_size=Nmo*exp(rmo*Tmo),
growth_rate=rmo),
     msprime.PopulationConfiguration(sample_size=1400,initial_size=Nadmix*exp(radmi
x*Tadmix),growth_rate=radmix)]
# Add migration matrix
mig_mat =
[[0,mafeu,mafas,mafba,mafas,mafas,0,0],[mafeu,0,meuas,mafeu,meuas,meuas,mafeu,0],
[mafas,mafeu,0,mafas,masau,0,mafas,0],
     [mafba,mafeu,mafas,0,mafas,mafas,mafba,0],
[mafas,meuas,masau,mafas,0,masau,mafas,0], [mafas,meuas,0,mafas,masau,0,mafas,0],
     [0,mafeu,mafas,mafba,mafas,mafas,0,0], [0,0,0,0,0,0,0,0]]

# %% Declare demographic events
# Admixture event, 5/8 Africa, 3/8 Asia
admixture_event = [
     msprime.MassMigration(time=Tadmix,source=7,destination=5,proportion=3/8),
```

```
     msprime.MassMigration(time=Tadmix+0.0001,source=7,destination=6,proportion=1.0
)
]
# Mozambique and Bantu Split
moz_event = [
     msprime.MassMigration(time=Tmo,source=6,destination=3,proportion=1.0),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,3)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(3,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,1)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(1,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,2)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(2,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,4)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,5)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(5,6))
]
# South Borneo and Austronesian Split
sbo_event = [
     msprime.MassMigration(time=Tsb,source=5,destination=2,proportion=1/3),
     msprime.MassMigration(time=Tsb+0.0001,source=5,destination=4,proportion=1.0),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,4)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(4,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,0)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(0,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,1)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(1,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,3)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,6)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(6,5))
]
# Austronesian and East Asia Split
aus_event = [
     msprime.MassMigration(time=Tau,source=4,destination=2,proportion=1.0),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,2)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(2,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,0)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(0,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,1)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(1,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(6,4))
]
# Bantu and Africa Split
ban_event = [
     msprime.MassMigration(time=Tba,source=3,destination=0,proportion=1.0),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,0)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(0,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,1)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(1,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,2)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(2,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(5,3))
]
# Asia and Europe split
eu_event = [
     msprime.MigrationRateChange(time=Teu,rate=0.0),
     msprime.MassMigration(time=Teu+0.0001,source=2,destination=1,proportion=1.0),
     msprime.PopulationParametersChange(time=Teu+0.0002,initial_size=Nb,growth_rate
=0.0,population_id=1),
```

```
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(0,1)),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(1,0))]
# Out of Africa event
ooa_event = [
     msprime.MigrationRateChange(time=Tooa,rate=0.0),
     msprime.MassMigration(time=Tooa+0.0001,source=1,destination=0,proportion=1.0)]
# initial population size
init_event = [
msprime.PopulationParametersChange(time=Thum,initial_size=N0,population_id=0)]
# Save sequence of events
events = admixture_event + moz_event + sbo_event + aus_event + ban_event + eu_event
+ ooa_event + init_event

# %%
# Use the demography debugger to print out the demographic history
# that we have just described.
dd = msprime.DemographyDebugger(
     population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events)
dd.print_history()

# %%
# Run model
# Read in the recombination map using the read_hapmap method
print(datetime.datetime.now())
infile = recomb_file
recomb_map = msprime.RecombinationMap.read_hapmap(infile)
treeseq = msprime.simulate(population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events,
     recombination_map=recomb_map,
     mutation_rate=mu,
     random_seed=seed)
print("In silico evolution has finished")
print(datetime.datetime.now())
```

2. MSPrime code for simulating : Split from Southern Borneo at 26 generation BP through a strong founder event (Ne=50), followed by admixture one generation after.

```
# %% Define parameters
# Genomic Configuration
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8 # generate 100 Mb
# Humankind event
N0=7310 # homo sapiens initial population size
Thum=5920 # time (gens) of advent of modern humans
# Out of Africa event
Naf=14474 # size of african population
Tooa=2040 # number of generations back to Out of Africa
Nb=1861 # size of out of Africa population
mafb=1.5e-4 # migration rate Africa and Out-of-Africa
# Europe-Asia split event
Teu=920 # number generations back to Asia-Europe split
Neu=1032; Nas=554 # bottleneck population sizes
mafeu=2.5e-5; mafas=7.8e-6; meuas=3.11e-5 # mig. rates
reu=0.0038 # growth rate per generation in Europe
ras=0.0048 # growth rate per generation in Asia
# Bantu late split event
Tba= 167 # number generations bantu expansion
```

```
Nba= 2200 # initial bantu population size
rba= 0.0111 # growth rate per generation in Bantu expansion
Tmo= 60 # number generation bantu-mozambique split
Nmo= 3000 # initial mozambique population size
rmo= 0.0257 # growth rate per generation in Mozambique
mafba= 3.75e-3 # mig. rates
# Austronesian Out of Taiwan event
Tau= 166 # number generation austronesian expansion
Nau= 5000 # initial austronesian population size
rau= 0.0132 # growth rate per generation in Austronesian expansion
Tsb= 100 # number generation austronesian-southborneo split
Nsb= 10000 # initial south borneo population size
rsb=0.0150
masau= 1.60e-3 # mig. rates
# Austronesian Isolation
Nis= args.eff
Sis= args.eff
Dis= args.tim
Gis= (log(Nis/Sis))/Dis
Tis= 25 + Dis
ris = 0.0
# Madagascar Settlement event
Tadmix=25 # time of admixture
Nadmix= 2500 # initial size of admixed population
radmix= 0.1638 # growth rate of admixed population
seed = args.ran
print("+DEMOGRAPHIC SCENARIO")
print("+++ADMIXTURE EVENT: start")
print(Tadmix)
print("+++ISOLATION EVENT: time, initial Ne, indv migration")
print(Tis)
print(Sis)
print("+++USING SEED:")
print(seed)

# %% Population configurations
# 0 is African, 1 is European, 2 is Asian, 3 is Bantu, 4 is Austronesian, 5 is SouthBorneo, 6 is Mozambique and 7
is Madagascar
pop_config = [msprime.PopulationConfiguration(sample_size=838,initial_size=Naf,growth_rate=0.0),
        msprime.PopulationConfiguration(sample_size=800,initial_size=Neu*exp(reu*Teu),growth_rate=reu),
        msprime.PopulationConfiguration(sample_size=800,initial_size=Nas*exp(ras*Teu),growth_rate=ras),
        msprime.PopulationConfiguration(sample_size=200,initial_size=Nba*exp(rba*Tba),growth_rate=rba),
        msprime.PopulationConfiguration(sample_size=200,initial_size=Nau*exp(rau*Tau),growth_rate=rau),
        msprime.PopulationConfiguration(sample_size=182,initial_size=Nsb*exp(rsb*Tsb),growth_rate=rsb),
        msprime.PopulationConfiguration(sample_size=322,initial_size=Nmo*exp(rmo*Tmo),growth_rate=rmo),
        msprime.PopulationConfiguration(sample_size=0,initial_size=Nis,growth_rate=ris),
        msprime.PopulationConfiguration(sample_size=1400,initial_size=Nadmix*exp(radmix*Tadmix),growth_r
ate=radmix)]
# Add migration matrix
mig_mat     =     [[0,mafeu,mafas,mafba,mafas,mafas,0,0,0],[mafeu,0,meuas,mafeu,meuas,meuas,mafeu,0,0],
[mafas,mafeu,0,mafas,masau,0,mafas,0,0],
        [mafba,mafeu,mafas,0,mafas,mafas,mafba,0,0],        [mafas,meuas,masau,mafas,0,masau,mafas,0,0],
[mafas,meuas,0,mafas,masau,0,mafas,0,0],
        [0,mafeu,mafas,mafba,mafas,mafas,0,0,0], [0,0,0,0,0,0,0,0,0], [0,0,0,0,0,0,0,0,0]]

# %% Declare demographic events
# Admixture event, 5/8 Africa, 3/8 Asia
admixture_event = [
        msprime.MassMigration(time=Tadmix,source=8,destination=7,proportion=3/8),
        msprime.MassMigration(time=Tadmix+0.0001,source=8,destination=6,proportion=1.0),
        msprime.PopulationParametersChange(time=Tadmix+0.0002,initial_size=Nis,growth_rate=Gis,populatio
n_id=7)
]
# Isolement event
iso_event= [
        msprime.MassMigration(time=Tis,source=7,destination=5,proportion=1.0),
```

```
        msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
]
# Mozambique and Bantu Split
moz_event = [
        msprime.MassMigration(time=Tmo,source=6,destination=3,proportion=1.0),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,3)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(3,6)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,1)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(1,6)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,2)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(2,6)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,4)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(4,6)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,5)),
        msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(5,6))
]
# South Borneo and Austronesian Split
sbo_event = [
        msprime.MassMigration(time=Tsb,source=5,destination=2,proportion=1/3),
        msprime.MassMigration(time=Tsb+0.0001,source=5,destination=4,proportion=1.0),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,4)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(4,5)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,0)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(0,5)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,1)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(1,5)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,3)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(3,5)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,6)),
        msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(6,5))
]
# Austronesian and East Asia Split
aus_event = [
        msprime.MassMigration(time=Tau,source=4,destination=2,proportion=1.0),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,2)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(2,4)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,0)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(0,4)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,1)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(1,4)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,3)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(3,4)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,6)),
        msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(6,4))
]
# Bantu and Africa Split
ban_event = [
        msprime.MassMigration(time=Tba,source=3,destination=0,proportion=1.0),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,0)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(0,3)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,1)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(1,3)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,2)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(2,3)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,4)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(4,3)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,5)),
        msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(5,3))
]
# Asia and Europe split
eu_event = [
        msprime.MigrationRateChange(time=Teu,rate=0.0),
        msprime.MassMigration(time=Teu+0.0001,source=2,destination=1,proportion=1.0),
        msprime.PopulationParametersChange(time=Teu+0.0002,initial_size=Nb,growth_rate=0.0,population_i
d=1),
        msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(0,1)),
        msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(1,0))]
```

```
# Out of Africa event
ooa_event = [
        msprime.MigrationRateChange(time=Tooa,rate=0.0),
        msprime.MassMigration(time=Tooa+0.0001,source=1,destination=0,proportion=1.0)]
# initial population size
init_event = [ msprime.PopulationParametersChange(time=Thum,initial_size=N0,population_id=0)]

# Save sequence of events according to splits
if Tis<=60:
        events = admixture_event + iso_event + moz_event + sbo_event + aus_event + ban_event + eu_event
+ ooa_event + init_event
        print("+++EVENTS : [adm] + [iso<=60] + [moz] + [sbo] + [aus] + [...]")
elif Tis<=100:
        events  = admixture_event + moz_event + iso_event + sbo_event + aus_event + ban_event + eu_event
+ ooa_event + init_event
        print("+++EVENTS : [adm] + [moz] + [iso<=100] + [sbo] + [aus] + [...]")
else:
        iso_event= [
    msprime.MassMigration(time=Tis,source=7,destination=4,proportion=1.0),
    msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
        ]
        events = admixture_event + moz_event + sbo_event + iso_event + aus_event + ban_event + eu_event
+ ooa_event + init_event
        print("+++EVENTS : [adm] + [moz] + [sbo<=167] + [iso] + [aus] + [...]")



# %%
# Use the demography debugger to print out the demographic history
# that we have just described.
dd = msprime.DemographyDebugger(
        population_configurations=pop_config,
        migration_matrix=mig_mat,
        demographic_events=events)
dd.print_history()

# %%
# Run model
# Read in the recombination map using the read_hapmap method
print(datetime.datetime.now())
infile = recomb_file
recomb_map = msprime.RecombinationMap.read_hapmap(infile)
treeseq = msprime.simulate(population_configurations=pop_config,
        migration_matrix=mig_mat,
        demographic_events=events,
        recombination_map=recomb_map,
        mutation_rate=mu,
        random_seed=seed)
print("In silico evolution has finished")
print(datetime.datetime.now())
```

3. MSPrime code for simulating : Admixture through multiple pulses of gene flow (Ne=70 to 300) coming from an ancestral Southern Borneo population, between 20-30 generations before present.

```
# %% Define parameters
# Genomic Configuration
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8 # generate 100 Mb
# Humankind event
N0=7310 # homo sapiens initial population size
```

```
Thum=5920 # time (gens) of advent of modern humans
# Out of Africa event
Naf=14474 # size of african population
Tooa=2040 # number of generations back to Out of Africa
Nb=1861 # size of out of Africa population
mafb=1.5e-4 # migration rate Africa and Out-of-Africa
# Europe-Asia split event
Teu=920 # number generations back to Asia-Europe split
Neu=1032; Nas=554 # bottleneck population sizes
mafeu=2.5e-5; mafas=7.8e-6; meuas=3.11e-5 # mig. rates
reu=0.0038 # growth rate per generation in Europe
ras=0.0048 # growth rate per generation in Asia
# Bantu late split event
Tba= 167 # number generations bantu expansion
Nba= 2200 # initial bantu population size
rba= 0.0111 # growth rate per generation in Bantu expansion
Tmo= 60 # number generation bantu-mozambique split
Nmo= 3000 # initial mozambique population size
rmo= 0.0257 # growth rate per generation in Mozambique
mafba= 3.75e-3 # mig. rates
# Austronesian Out of Taiwan event
Tau= 166 # number generation austronesian expansion
Nau= 5000 # initial austronesian population size
rau= 0.0132 # growth rate per generation in Austronesian expansion
Tsb= 100 # number generation austronesian-southborneo split
Nsb= 5000 # initial south borneo population size
rsb=0.0219
masau= 1.60e-3 # mig. rates
# Madagascar Settlement event
Cgf= args.flo # continuous gene flow duration
rgf= 1-exp((log(1-0.375))/Cgf) # african cgf rate
Tadmix= 20 + Cgf # time of admixture
Nadmix= 1600 # initial size of admixed population
radmix= (log(150000/Nadmix))/Tadmix # growth rate of admixed population
# Display parameters for scenario
print("+DEMOGRAPHIC SCENARIO")
print("+++ADMIXTURE EVENT: start, duration and cgf rate")
print(Tadmix)
print(Cgf)
print(rgf)
seed = args.ran
print("+++USING SEED:")
print(seed)

# %% Population configurations
# 0 is African, 1 is European, 2 is Asian, 3 is Bantu, 4 is Austronesian, 5 is
SouthBorneo, 6 is Mozambique and 7 is Madagascar
pop_config =
[msprime.PopulationConfiguration(sample_size=838,initial_size=Naf,growth_rate=0.0),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Neu*exp(reu*Teu),
growth_rate=reu),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Nas*exp(ras*Teu),
growth_rate=ras),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nba*exp(rba*Tba),
growth_rate=rba),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nau*exp(rau*Tau),
growth_rate=rau),
     msprime.PopulationConfiguration(sample_size=182,initial_size=Nsb*exp(rsb*Tsb),
growth_rate=rsb),
     msprime.PopulationConfiguration(sample_size=322,initial_size=Nmo*exp(rmo*Tmo),
growth_rate=rmo),
     msprime.PopulationConfiguration(sample_size=1400,initial_size=Nadmix*exp(radmi
x*Tadmix),growth_rate=radmix)]
# Add migration matrix
mig_mat =
[[0,mafeu,mafas,mafba,mafas,mafas,0,0],[mafeu,0,meuas,mafeu,meuas,meuas,mafeu,0],
```

```
[mafas,mafeu,0,mafas,masau,0,mafas,0],
     [mafba,mafeu,mafas,0,mafas,mafas,mafba,0],
[mafas,meuas,masau,mafas,0,masau,mafas,0], [mafas,meuas,0,mafas,masau,0,mafas,0],
     [0,mafeu,mafas,mafba,mafas,mafas,0,0], [0,0,0,0,0,0,0,0]]

# %% Declare demographic events
# Admixture event, 5/8 Africa, 3/8 Asia
admixture_event = [
     # End of admixture
     msprime.MigrationRateChange(time=20,rate=rgf,matrix_index=(7,5)),
     # Break Madagascar demography - begin asian isolated trajectory
     # Start of admixture
     msprime.MassMigration(time=Tadmix+0.0001,source=7,destination=6,proportion=1.0
),
     msprime.MigrationRateChange(time=Tadmix+1,rate=0.0,matrix_index=(7,5))
]

# Mozambique and Bantu Split
moz_event = [
     msprime.MassMigration(time=Tmo,source=6,destination=3,proportion=1.0),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,3)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(3,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,1)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(1,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,2)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(2,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,4)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,5)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(5,6))
]
# South Borneo and Austronesian Split
sbo_event = [
     msprime.MassMigration(time=Tsb+0.0001,source=5,destination=2,proportion=1/3),
     msprime.MassMigration(time=Tsb+0.0002,source=5,destination=4,proportion=1.0),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,4)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(4,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,0)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(0,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,1)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(1,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,3)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,6)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(6,5))
]
# Austronesian and East Asia Split
aus_event = [
     msprime.MassMigration(time=Tau,source=4,destination=2,proportion=1.0),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,2)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(2,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,0)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(0,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,1)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(1,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(6,4))
]
# Bantu and Africa Split
ban_event = [
     msprime.MassMigration(time=Tba,source=3,destination=0,proportion=1.0),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,0)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(0,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,1)),
```

```
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(1,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,2)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(2,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(5,3))
]
# Asia and Europe split
eu_event = [
     msprime.MigrationRateChange(time=Teu,rate=0.0),
     msprime.MassMigration(time=Teu+0.0001,source=2,destination=1,proportion=1.0),
     msprime.PopulationParametersChange(time=Teu+0.0002,initial_size=Nb,growth_rate
=0.0,population_id=1),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(0,1)),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(1,0))]
# Out of Africa event
ooa_event = [
     msprime.MigrationRateChange(time=Tooa,rate=0.0),
     msprime.MassMigration(time=Tooa+0.0001,source=1,destination=0,proportion=1.0)]
# initial population size
init_event = [
msprime.PopulationParametersChange(time=Thum,initial_size=N0,population_id=0)]

# Save sequence of events according to splits
events = admixture_event + moz_event + sbo_event + aus_event + ban_event + eu_event
+ ooa_event + init_event
print("+++EVENTS : [adm] + [moz] + [sbo] + [aus] + [...]")

# %%
# Use the demography debugger to print out the demographic history
# that we have just described.
dd = msprime.DemographyDebugger(
     population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events)
dd.print_history()

# %%
# Run model
# Read in the recombination map using the read_hapmap method
print(datetime.datetime.now())
infile = recomb_file
recomb_map = msprime.RecombinationMap.read_hapmap(infile)
treeseq = msprime.simulate(population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events,
     recombination_map=recomb_map,
     mutation_rate=mu,
     random_seed=seed)
print("In silico evolution has finished")
print(datetime.datetime.now())
```

4. MSPrime code for simulating : Split from Southern Borneo through a founder event of Ne=5,000 individuals, followed by a slow decrease of Ne during 50 generations until reaching Ne=900 (before the admixture at 25 generations BP). Also for simulating a split from Southern Borneo through a strong founder event of Ne=50, with effective population size expansion during 10 generations, reaching Ne=900 (before the admixture at 25 generations BP).

```
# %% Define parameters
# Genomic Configuration
mu=1.25e-8 # mutation rate per bp
```

```
rho=1e-8 # recombination rate per bp
nbp = 1e8 # generate 100 Mb
# Humankind event
N0=7310 # homo sapiens initial population size
Thum=5920 # time (gens) of advent of modern humans
# Out of Africa event
Naf=14474 # size of african population
Tooa=2040 # number of generations back to Out of Africa
Nb=1861 # size of out of Africa population
mafb=1.5e-4 # migration rate Africa and Out-of-Africa
# Europe-Asia split event
Teu=920 # number generations back to Asia-Europe split
Neu=1032; Nas=554 # bottleneck population sizes
mafeu=2.5e-5; mafas=7.8e-6; meuas=3.11e-5 # mig. rates
reu=0.0038 # growth rate per generation in Europe
ras=0.0048 # growth rate per generation in Asia
# Bantu late split event
Tba= 167 # number generations bantu expansion
Nba= 2200 # initial bantu population size
rba= 0.0111 # growth rate per generation in Bantu expansion
Tmo= 60 # number generation bantu-mozambique split
Nmo= 3000 # initial mozambique population size
rmo= 0.0257 # growth rate per generation in Mozambique
mafba= 3.75e-3 # mig. rates
# Austronesian Out of Taiwan event
Tau= 166 # number generation austronesian expansion
Nau= 5000 # initial austronesian population size
rau= 0.0132 # growth rate per generation in Austronesian expansion
Tsb= 100 # number generation austronesian-southborneo split
Nsb= 10000 # initial south borneo population size
rsb=0.0150
masau= 1.60e-3 # mig. rates
# Austronesian Isolation
Nis= 900
Sis= args.eff # Specify Initital Isolated Effective Size
Dis= args.tim # Specify Time of Isolation
Gis= (log(Nis/Sis))/Dis
Tis= 25 + Dis
ris = 0.0
# Madagascar Settlement event
Tadmix=25 # time of admixture
Nadmix= 2500 # initial size of admixed population
radmix= 0.1638 # growth rate of admixed population
seed = args.ran # Specify Number Seed
print("+DEMOGRAPHIC SCENARIO")
print("+++ADMIXTURE EVENT: start")
print(Tadmix)
print("+++ISOLATION EVENT: time, initial Ne, indv migration")
print(Tis)
print(Sis)
print("+++USING SEED:")
print(seed)

# %% Population configurations
# 0 is African, 1 is European, 2 is Asian, 3 is Bantu, 4 is Austronesian, 5 is
SouthBorneo, 6 is Mozambique and 7 is Madagascar
pop_config =
[msprime.PopulationConfiguration(sample_size=838,initial_size=Naf,growth_rate=0.0),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Neu*exp(reu*Teu),
growth_rate=reu),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Nas*exp(ras*Teu),
growth_rate=ras),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nba*exp(rba*Tba),
growth_rate=rba),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nau*exp(rau*Tau),
growth_rate=rau),
```

```
     msprime.PopulationConfiguration(sample_size=182,initial_size=Nsb*exp(rsb*Tsb),
growth_rate=rsb),
     msprime.PopulationConfiguration(sample_size=322,initial_size=Nmo*exp(rmo*Tmo),
growth_rate=rmo),
     msprime.PopulationConfiguration(sample_size=0,initial_size=Nis,growth_rate=ris
),
     msprime.PopulationConfiguration(sample_size=1400,initial_size=Nadmix*exp(radmi
x*Tadmix),growth_rate=radmix)]
# Add migration matrix
mig_mat =
[[0,mafeu,mafas,mafba,mafas,mafas,0,0,0],[mafeu,0,meuas,mafeu,meuas,meuas,mafeu,0,0
], [mafas,mafeu,0,mafas,masau,0,mafas,0,0],
     [mafba,mafeu,mafas,0,mafas,mafas,mafba,0,0],
[mafas,meuas,masau,mafas,0,masau,mafas,0,0],
[mafas,meuas,0,mafas,masau,0,mafas,0,0],
     [0,mafeu,mafas,mafba,mafas,mafas,0,0,0], [0,0,0,0,0,0,0,0,0],
[0,0,0,0,0,0,0,0,0]]

# %% Declare demographic events
# Admixture event, 5/8 Africa, 3/8 Asia
admixture_event = [
     msprime.MassMigration(time=Tadmix,source=8,destination=7,proportion=3/8),
     msprime.MassMigration(time=Tadmix+0.0001,source=8,destination=6,proportion=1.0
),
     msprime.PopulationParametersChange(time=Tadmix+0.0002,initial_size=Nis,growth_
rate=Gis,population_id=7)
]
# Isolement event
iso_event= [
     msprime.MassMigration(time=Tis,source=7,destination=5,proportion=1.0),
     msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
]
# Mozambique and Bantu Split
moz_event = [
     msprime.MassMigration(time=Tmo,source=6,destination=3,proportion=1.0),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,3)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(3,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,1)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(1,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,2)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(2,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,4)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,5)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(5,6))
]
# South Borneo and Austronesian Split
sbo_event = [
     msprime.MassMigration(time=Tsb,source=5,destination=2,proportion=1/3),
     msprime.MassMigration(time=Tsb+0.0001,source=5,destination=4,proportion=1.0),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,4)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(4,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,0)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(0,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,1)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(1,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,3)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(5,6)),
     msprime.MigrationRateChange(time=Tsb+0.0001,rate=0,matrix_index=(6,5))
]
# Austronesian and East Asia Split
aus_event = [
     msprime.MassMigration(time=Tau,source=4,destination=2,proportion=1.0),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,2)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(2,4)),
```

```
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,0)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(0,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,1)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(1,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(6,4))
]
# Bantu and Africa Split
ban_event = [
     msprime.MassMigration(time=Tba,source=3,destination=0,proportion=1.0),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,0)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(0,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,1)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(1,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,2)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(2,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(5,3))
]
# Asia and Europe split
eu_event = [
     msprime.MigrationRateChange(time=Teu,rate=0.0),
     msprime.MassMigration(time=Teu+0.0001,source=2,destination=1,proportion=1.0),
     msprime.PopulationParametersChange(time=Teu+0.0002,initial_size=Nb,growth_rate
=0.0,population_id=1),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(0,1)),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(1,0))]
# Out of Africa event
ooa_event = [
     msprime.MigrationRateChange(time=Tooa,rate=0.0),
     msprime.MassMigration(time=Tooa+0.0001,source=1,destination=0,proportion=1.0)]
# initial population size
init_event = [
msprime.PopulationParametersChange(time=Thum,initial_size=N0,population_id=0)]

# Save sequence of events according to splits
if Tis<=60:
     events = admixture_event + iso_event + moz_event + sbo_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [iso<=60] + [moz] + [sbo] + [aus] + [...]")
elif Tis<=100:
     events  = admixture_event + moz_event + iso_event + sbo_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [moz] + [iso<=100] + [sbo] + [aus] + [...]")
else:
     iso_event= [
        msprime.MassMigration(time=Tis,source=7,destination=4,proportion=1.0),

msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
     ]
     events = admixture_event + moz_event + sbo_event + iso_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [moz] + [sbo<=167] + [iso] + [aus] + [...]")



# %%
# Use the demography debugger to print out the demographic history
# that we have just described.
dd = msprime.DemographyDebugger(
     population_configurations=pop_config,
     migration_matrix=mig_mat,
```

```
     demographic_events=events)
dd.print_history()

# %%
# Run model
# Read in the recombination map using the read_hapmap method
print(datetime.datetime.now())
infile = recomb_file
recomb_map = msprime.RecombinationMap.read_hapmap(infile)
treeseq = msprime.simulate(population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events,
     recombination_map=recomb_map,
     mutation_rate=mu,
     random_seed=seed)
print("In silico evolution has finished")
print(datetime.datetime.now())
```

5. MSPrime code for simulating : Split from Southern Borneo through a founder event of Ne=500, with effective population size constant during 40 generations, receiving no gene flow before the admixture. It is important to notice that this was the best candidate based on the exploratory simulation analysis (Fig. S13; Table S3, S4).

```
# %% Define parameters
# Genomic Configuration
mu=1.25e-8 # mutation rate per bp
rho=1e-8 # recombination rate per bp
nbp = 1e8 # generate 100 Mb
# Humankind event
N0=7310 # homo sapiens initial population size
Thum=5920 # time (gens) of advent of modern humans
# Out of Africa event
Naf=14474 # size of african population
Tooa=2040 # number of generations back to Out of Africa
Nb=1861 # size of out of Africa population
mafb=1.5e-4 # migration rate Africa and Out-of-Africa
# Europe-Asia split event
Teu=920 # number generations back to Asia-Europe split
Neu=1032; Nas=554 # bottleneck population sizes
mafeu=2.5e-5; mafas=7.8e-6; meuas=3.11e-5 # mig. rates
reu=0.0038 # growth rate per generation in Europe
ras=0.0048 # growth rate per generation in Asia
# Bantu late split event
Tba= 167 # number generations bantu expansion
Nba= 2200 # initial bantu population size
rba= 0.0111 # growth rate per generation in Bantu expansion
Tmo= 60 # number generation bantu-mozambique split
Nmo= 3000 # initial mozambique population size
rmo= 0.0257 # growth rate per generation in Mozambique
mafba= 3.75e-3 # mig. rates
# Austronesian Out of Taiwan event
Tau= 166 # number generation austronesian expansion
Nau= 5000 # initial austronesian population size
rau= 0.0132 # growth rate per generation in Austronesian expansion
Tsb= 100 # number generation austronesian-southborneo split
Nsb= 10000 # initial south borneo population size
rsb=0.0150
masau= 1.60e-3 # mig. rates
# Madagascar Settlement event
Cgf= args.flo # continuous gene flow duration
rgf= 1-exp((log(1-0.625))/Cgf) # african cgf rate
Tadmix= 20 + Cgf # time of admixture
Nadmix= 900 # initial size of admixed population
radmix= (log(150000/Nadmix))/Tadmix # growth rate of admixed population
```

```
# Austronesian Isolation
Nis= args.eff # Specify Final Isolated Effective Size
Sis= args.eff # Specify Initital Isolated Effective Size
Dis= args.tim # Specify Time of Isolation
Gis= (log(Nis/Sis))/Dis # Growth rate
Tis= Tadmix + Dis
ris= 0.0
# Display parameters for scenario
print("+DEMOGRAPHIC SCENARIO")
print("+++ISOLATION EVENT: time and initial Ne")
print(Tis)
print(Sis)
print("+++ADMIXTURE EVENT: start, duration and cgf rate")
print(Tadmix)
print(Cgf)
print(rgf)
seed = args.ran
print("+++USING SEED:")
print(seed)

# %% Population configurations
# 0 is African, 1 is European, 2 is Asian, 3 is Bantu, 4 is Austronesian, 5 is
SouthBorneo, 6 is Mozambique and 7 is Madagascar
pop_config =
[msprime.PopulationConfiguration(sample_size=838,initial_size=Naf,growth_rate=0.0),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Neu*exp(reu*Teu),
growth_rate=reu),
     msprime.PopulationConfiguration(sample_size=800,initial_size=Nas*exp(ras*Teu),
growth_rate=ras),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nba*exp(rba*Tba),
growth_rate=rba),
     msprime.PopulationConfiguration(sample_size=200,initial_size=Nau*exp(rau*Tau),
growth_rate=rau),
     msprime.PopulationConfiguration(sample_size=182,initial_size=Nsb*exp(rsb*Tsb),
growth_rate=rsb),
     msprime.PopulationConfiguration(sample_size=322,initial_size=Nmo*exp(rmo*Tmo),
growth_rate=rmo),
     msprime.PopulationConfiguration(sample_size=1400,initial_size=Nadmix*exp(radmi
x*Tadmix),growth_rate=radmix)]
# Add migration matrix
mig_mat =
[[0,mafeu,mafas,mafba,mafas,mafas,0,0],[mafeu,0,meuas,mafeu,meuas,meuas,mafeu,0],
[mafas,mafeu,0,mafas,masau,0,mafas,0],
     [mafba,mafeu,mafas,0,mafas,mafas,mafba,0],
[mafas,meuas,masau,mafas,0,masau,mafas,0], [mafas,meuas,0,mafas,masau,0,mafas,0],
     [0,mafeu,mafas,mafba,mafas,mafas,0,0], [0,0,0,0,0,0,0,0]]

# %% Declare demographic events
# Admixture event, 5/8 Africa, 3/8 Asia
admixture_event = [
     # End of admixture
     msprime.MigrationRateChange(time=20,rate=rgf,matrix_index=(7,6)),
     # Break Madagascar demography - begin asian isolated trajectory
     msprime.PopulationParametersChange(time=Tadmix,initial_size=Nis,growth_rate=0.
0,population_id=7),
     # Start of admixture
     msprime.MigrationRateChange(time=Tadmix+1,rate=0.0,matrix_index=(7,6))
]

iso_event = [
     # Austronesian isolation
     msprime.MassMigration(time=Tis,source=7,destination=5,proportion=1.0),
     msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
]

# Mozambique and Bantu Split
```

```
moz_event = [
     msprime.MassMigration(time=Tmo,source=6,destination=3,proportion=1.0),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,3)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(3,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,1)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(1,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,2)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(2,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,4)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(6,5)),
     msprime.MigrationRateChange(time=Tmo,rate=0,matrix_index=(5,6))
]
# South Borneo and Austronesian Split
sbo_event = [
     msprime.MassMigration(time=Tsb+0.0001,source=5,destination=2,proportion=1/3),
     msprime.MassMigration(time=Tsb+0.0002,source=5,destination=4,proportion=1.0),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,4)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(4,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,0)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(0,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,1)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(1,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,3)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(5,6)),
     msprime.MigrationRateChange(time=Tsb+0.0002,rate=0,matrix_index=(6,5))
]
# Austronesian and East Asia Split
aus_event = [
     msprime.MassMigration(time=Tau,source=4,destination=2,proportion=1.0),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,2)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(2,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,0)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(0,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,1)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(1,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(4,6)),
     msprime.MigrationRateChange(time=Tau,rate=0,matrix_index=(6,4))
]
# Bantu and Africa Split
ban_event = [
     msprime.MassMigration(time=Tba,source=3,destination=0,proportion=1.0),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,0)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(0,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,1)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(1,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,2)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(2,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,4)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(4,3)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(3,5)),
     msprime.MigrationRateChange(time=Tba,rate=0,matrix_index=(5,3))
]
# Asia and Europe split
eu_event = [
     msprime.MigrationRateChange(time=Teu,rate=0.0),
     msprime.MassMigration(time=Teu+0.0001,source=2,destination=1,proportion=1.0),
     msprime.PopulationParametersChange(time=Teu+0.0002,initial_size=Nb,growth_rate
=0.0,population_id=1),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(0,1)),
     msprime.MigrationRateChange(time=Teu+0.0003,rate=mafb,matrix_index=(1,0))]
# Out of Africa event
ooa_event = [
```

```
     msprime.MigrationRateChange(time=Tooa,rate=0.0),
     msprime.MassMigration(time=Tooa+0.0001,source=1,destination=0,proportion=1.0)]
# initial population size
init_event = [
msprime.PopulationParametersChange(time=Thum,initial_size=N0,population_id=0)]

# Save sequence of events according to splits
if Tis<=60:
     events = admixture_event + iso_event + moz_event + sbo_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [iso<=60] + [moz] + [sbo] + [aus] + [...]")
elif Tis<=100:
     events  = admixture_event + moz_event + iso_event + sbo_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [moz] + [iso<=100] + [sbo] + [aus] + [...]")
else:
     iso_event = [
        # Austronesian isolation
        msprime.MassMigration(time=Tis,source=7,destination=4,proportion=1.0),

msprime.PopulationParametersChange(time=Tis,growth_rate=0.0,population_id=7)
     ]
     events = admixture_event + moz_event + sbo_event + iso_event + aus_event +
ban_event + eu_event + ooa_event + init_event
     print("+++EVENTS : [adm] + [moz] + [sbo<=167] + [iso] + [aus] + [...]")

# %%
# Use the demography debugger to print out the demographic history
# that we have just described.
dd = msprime.DemographyDebugger(
     population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events)
dd.print_history()

# %%
# Run model
# Read in the recombination map using the read_hapmap method
print(datetime.datetime.now())
infile = recomb_file
recomb_map = msprime.RecombinationMap.read_hapmap(infile)
treeseq = msprime.simulate(population_configurations=pop_config,
     migration_matrix=mig_mat,
     demographic_events=events,
     recombination_map=recomb_map,
     mutation_rate=mu,
     random_seed=seed)
print("In silico evolution has finished")
print(datetime.datetime.now())
```

ANNEXES

Annex B: Tables from Chapter III “The demographic history and mutational load of Malagasy populations”

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
8	93655802	rs115627342	6.06	C	G	0.3134	1	66	133	91	42	0.993	1	1	1	0.963	0.985	0.978							
9	3969142	rs141364611	6.17	A	G	0.3657	1	66	133	84	49	0.993	1	1	1	0.978	1	1	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
2	63220969	rs141864934	6.16	T	C	0.2985	1	66	133	93	40	0.993	1	1	1	0.963	0.978	0.993	ncbi-geneid:23301	hsa:23301	EHBP1, HPC12	EH domain-binding protein 1	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Hereditary prostate cancer
7	82059070	rs143388418	6.03	G	A	0.3881	1	66	133	81	52	0.993	1	1	1	0.985	0.985	0.985	ncbi-geneid:781	hsa:781	CACNA2D1, CACNA2	voltage-dependent calcium channel alpha-2/delta-1	Ion channels [BR:hsa04040]	Signal transduction, Endocrine system, Circulatory system, Cardiovascular disease, Protein families: signaling and cellular processes	
4	23830299	rs146691710	6.06	A	G	0.3209	1	66	133	90	43	0.993	1	1	1	0.993	0.978	0.993	ncbi-geneid:10891	hsa:10891	PPARGC1A, LEM6	alpha peroxisome proliferator-activated receptor gamma coactivator 1-alpha	Mitochondrial biogenesis [BR:hsa03029]	Signal transduction, Endocrine system, Aging, Environmental adaptation, Neurodegenerative disease, Endocrine and metabolic disease, Protein families: genetic information processing	
7	19761304	rs556495664	6.02	A	G	0.2687	1	66	133	97	36	0.993	1	1	1	1	1	1	ncbi-geneid:256130	hsa:256130	TMEM196	transmembrane protein 196		Poorly characterized	
6	98298715	rs60690715	6.08	C	A	0.291	1	66	133	94	39	0.993	1	1	1	0.948	0.918	0.948							
9	14096784	rs78364791	6.08	T	C	0.3657	1	66	133	85	48	0.993	0.955	0.985	0.985	0.993	1	0.985	ncbi-geneid:4781	hsa:4781	NFIB, CTF	nuclear factor I/B	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	
1	77088902	rs78888615	6.07	A	G	0.3284	1	66	133	90	43	0.993	1	1	1	0.993	0.985	0.985	ncbi-geneid:256435	hsa:256435	ST6GALNAC3, PRO7177	N-acetylgalactosaminide alpha-2,6-sialyltransferase (sialyltransferase 7C) [EC:2.4.3.7]	Enzymes [BR:hsa01000], Glycosyltransferases [BR:hsa01003]	Glycan biosynthesis and metabolism, Protein families: metabolism, Glycosyltransferases	
8	124988166	rs80140850	6.17	T	C	0.3657	1	66	133	85	48	0.993	1	1	1	0.993	0.97	0.97	ncbi-geneid:654463	hsa:654463	FER1L6, C8ORFK23	fer-1-like protein 6	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	

Table 1. Highly frequent extreme deleterious mutations in Malagasy genomes. GRCh37 genomic information depicts chromosome, position, rsID, GERP-RS, ancestral allele (AA), and derived/annotated allele (DA) of locus. Malagasy genetic diversity columns show results based on 67 genomes, showing the average Asian ancestry, number of derived homozygotes, number of derived alleles, and ancestry-specific count of derived alleles for each locus. Derived allele frequency is shown for 67 individuals sampled from 1,000 Genomes Panel and Malagasy populations. Functional annotation corresponds to database retrieved annotations for the gene associated to the SNV, showing the NCBI and KEGG identifier. The nomenclature, orthology, pathway, additional information and disease were annotated related to the gene from KEGG database.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
5	78964729	rs10514164	4.32	T	G	0.3731	11	0	11	6	5	0.082	0.03	0.037	0.045	0.022	0.03	0.045	ncbi-geneid:167153	hsa:167153	TENT2, APD4	poly(A) RNA polymerase GLD2 [EC:2.7.7.19]	Enzymes [BR:hsa01000], Messenger RNA biogenesis [BR:hsa03019]	Protein families: genetic information processing, Transferring phosphorus-containing groups	
1	2444470	rs114254488	5.59	C	T	0.2985	10	0	10	10	0	0.075	0.015	0.03	0.037	0.007	0.022	0.022	ncbi-geneid:55229	hsa:55229	PANK4, CTRCT49	bifunctional damage-control phosphatase, subfamily II, fusion protein		Unclassified: metabolism	Cataract
5	106016334	rs12514358	6.06	C	T	0.4254	9	0	9	3	6	0.067	0.037	0.045	0.037	0.015	0.007	0.007							
16	7223641	rs17142766	5.73	A	G	0.3955	12	0	12	7	5	0.09	0.03	0.037	0.007	0.022	0.007	0.015	ncbi-geneid:54715	hsa:54715	RBFOX1, 2BP1	RNA binding protein fox-1	Spliceosome [BR:hsa03041]	Protein families: genetic information processing	
3	114379193	rs17681207	6.16	C	T	0.3134	5	0	5	4	1	0.037	0.03	0.045	0.007	0.037	0.037	0.022	ncbi-geneid:26137	hsa:26137	ZBTB20, DPZF	zinc finger and BTB domain-containing protein 20	Transcription factors [BR:hsa03000], Ubiquitin system [BR:hsa04121]	Protein families: genetic information processing	Primrose syndrome
5	137031775	rs2905612	4.16	C	T	0.2612	14	0	14	7	7	0.104	0.007	0.007	0.015	0.045	0.007	0.022	ncbi-geneid:26249	hsa:26249	KLHL3, PHA2D	kelch-like protein 2/3	Ubiquitin system [BR:hsa04121]	Protein families: genetic information processing	Hyperkalemic distal renal tubular acidosis (RTA type 4),Renal tubular acidosis
2	146219068	rs56673819	6.02	T	C	0.403	5	0	5	1	4	0.037	0.045	0.015	0.015	0.007	0.037	0.037							
6	38729511	rs61748600	6	T	C	0.306	5	0	5	2	3	0.037	0.03	0.015	0.015	0.045	0.037	0.045	ncbi-geneid:1769	hsa:1769	DNAH8, ATPase	dynein axonemal heavy chain	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Neurodegenerative disease, Protein families: signaling and cellular processes	Spermatogenic failure
1	107289830	rs61788201	4.76	A	T	0.2537	8	1	10	10	0	0.075	0.007	0.022	0.007	0.022	0.015	0.015							
22	33199362	rs62232903	4.62	G	A	0.3507	16	2	20	10	10	0.149	0.03	0.015	0.03	0.037	0.022	0.007	ncbi-geneid:8224	hsa:8224	SYN3	synapsin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
22	33199362	rs62232903	4.62	G	A	0.3507	16	2	20	10	10	0.149	0.03	0.015	0.03	0.037	0.022	0.007	ncbi-geneid:7078	hsa:7078	TIMP3, HSMRK222	metalloproteinase inhibitor 3	Peptidases and inhibitors [BR:hsa01002]	Cancer: overview, Protein families: metabolism	Sorsby fundus dystrophy
11	103624943	rs71482405	5.06	A	G	0.3284	12	0	12	7	5	0.09	0.022	0.045	0.015	0.022	0.045	0.022							
16	50263762	rs7203951	6.05	A	G	0.291	8	0	8	4	4	0.06	0.015	0.045	0.015	0.03	0.015	0.007	ncbi-geneid:64282	hsa:64282	TENT4B, PAPD5	non-canonical poly(A) RNA polymerase PAPD5/7 [EC:2.7.7.19]	Enzymes [BR:hsa01000], Messenger RNA biogenesis [BR:hsa03019], Transfer RNA biogenesis [BR:hsa03016]	Folding, sorting and degradation, Protein families: genetic information processing, Transferring phosphorus-containing groups	
8	21984765	rs73549523	5.17	C	T	0.3284	13	0	13	6	7	0.097	0.007	0.007	0.037	0.03	0.022	0.007	ncbi-geneid:55806	hsa:55806	HR, ALUNC	[histone H3]-dimethyl-L-lysine9 demethylase [EC:1.14.11.65]	Enzymes [BR:hsa01000], Chromosome and associated proteins [BR:hsa03036]	Protein families: genetic information processing, Acting on paired donors, with incorporation or reduction of molecular oxygen	Atrichia with papular lesions,Hypotrichosis,Marie-Unna hereditary hypotrichosis,Alopecia universalis
15	93536173	rs76621355	5.25	C	T	0.3582	12	0	12	1	11	0.09	0.015	0.022	0.015	0.022	0.045	0.022	ncbi-geneid:1106	hsa:1106	CHD2, DEE94	chromodomain-helicase-DNA-binding protein 2 [EC:5.6.2.-]	Enzymes [BR:hsa01000], Chromosome and associated proteins [BR:hsa03036]	Protein families: genetic information processing, Isomerases altering macromolecular conformation	Early infantile epileptic encephalopathy,Epileptic encephalopathy, childhood-onset
2	99483990	rs77944740	4.05	A	C	0.3881	9	1	11	1	10	0.082	0.03	0.045	0.045	0.03	0.007	0.007	ncbi-geneid:343990	hsa:343990	CRACDL, C2orf55				

Table 2. Variants present in Malagasy genomes and abroad (DAF < 0.05) predicted to have a large and extreme deleterious effect. Legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity							Derived allele frequency (DAF)							Functional annotation to the mapped gene					
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	HWE	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
6	257186	rs1011327	2.11	T	C	0.2985	1.28E-12	61	1	63	46	17	0.47	0.5	0.59	0.5	0.418	0.455	0.448							
6	32487462	rs115001626	2.61	G	T	0.2985	3.99E-09	7	15	37	29	8	0.276	0.448	0.373	0.187	0.194	0.231	0.284	ncbi-geneid:3127	hsa:3127	HLA-DRB5, HLA-DRB5*	MHC class II antigen	Exosome [BR:hsa04147]	Signaling molecules and interaction, Transport and catabolism, Immune system, Infectious disease: viral, Infectious disease: bacterial, Infectious disease: parasitic, Immune disease, Cardiovascular disease, Endocrine and metabolic disease, Protein families: signaling and cellular processes	Primary central nervous system lymphoma
9	29095016	rs1342251	2.66	T	A	0.3433	1.59E-07	12	33	78	58	20	0.582	0.784	0.836	0.776	0.642	0.537	0.687							
6	294268	rs1511159	2.78	T	C	0.2985	4.58E-11	59	1	61	44	17	0.455	0.53	0.575	0.545	0.507	0.463	0.433	ncbi-geneid:56940	hsa:56940	DUSP22, JKAP	atypical dual specificity phosphatase [EC:3.1.3.16 3.1.3.48]	Enzymes [BR:hsa01000], Protein phosphatases and associated proteins [BR:hsa01009]	Protein families: metabolism, Acting on ester bonds	Anaplastic large-cell lymphoma,Peripheral T cell lymphoma
1	61288285	rs1996168	2.92	C	T	0.4776	2.35E-09	55	0	55	32	23	0.41	0.209	0.246	0.254	0.455	0.44	0.47							
8	11247075	rs2572442	2.26	A	T	0.291	7.44E-07	4	56	116	77	39	0.866	1	1	1	0.552	0.731	0.627							
6	57290485	rs35444204	3.66	T	C	0.3358	9.85E-20	67	0	67	45	22	0.5	0.515	0.493	0.493	0.5	0.493	0.522	ncbi-geneid:5558	hsa:5558	PRIM2, PRIM2A	DNA primase large subunit	DNA replication proteins [BR:hsa03032]	Replication and repair, Protein families: genetic information processing	
6	349343	rs3778605	2.27	T	A	0.2985	6.97E-16	64	3	70	49	21	0.522	0.515	0.478	0.507	0.507	0.507	0.522	ncbi-geneid:56940	hsa:56940	DUSP22, JKAP	atypical dual specificity phosphatase [EC:3.1.3.16 3.1.3.48]	Enzymes [BR:hsa01000], Protein phosphatases and associated proteins [BR:hsa01009]	Protein families: metabolism, Acting on ester bonds	Anaplastic large-cell lymphoma,Peripheral T cell lymphoma
1	61288266	rs3936928	2.12	A	G	0.4776	2.35E-09	55	12	79	37	42	0.59	0.791	0.769	0.746	0.552	0.56	0.53							
13	39934649	rs55958075	2.01	C	T	0.3433	7.25E-08	10	40	90	57	33	0.672	0.903	0.948	0.881	0.828	0.866	0.91	ncbi-geneid:10186	hsa:10186	LHFPL6, LHFP	LHFPL tetraspan subfamily member protein	Transporters [BR:hsa02000], Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	
8	25068604	rs60821775	2.37	G	C	0.3731	9.97E-08	11	37	85	50	35	0.634	0.903	0.94	0.963	0.373	0.336	0.336	ncbi-geneid:80005	hsa:80005	DOCK5	dedicator of cytokinesis protein 5	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
6	22053672	rs6927366	2.1	A	G	0.2836	1.59E-07	11	38	87	62	25	0.649	0.172	0.209	0.201	0.679	0.493	0.604							
20	52480410	rs73912566	2.61	T	C	0.3955	3.36E-18	66	0	66	41	25	0.493	0.328	0.358	0.358	0.478	0.5	0.485							
6	57295212	rs75428242	2.16	A	T	0.3358	9.85E-20	67	0	67	43	24	0.5	0.522	0.5	0.5	0.53	0.5	0.493	ncbi-geneid:5558	hsa:5558	PRIM2, PRIM2A	DNA primase large subunit	DNA replication proteins [BR:hsa03032]	Replication and repair, Protein families: genetic information processing	
10	47107666	rs9423080	2.93	G	T	0.3433	5.84E-17	65	0	65	40	25	0.485	0.47	0.455	0.44	0.448	0.47	0.47							

Table 3. Variants with predicted deleterious effects that significantly deviate from Hardy-Weinberg equilibrium. Legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
15	71654056	rs10775203	2.3	T	C	0.3209	23	0	23	7	16	0.172	0.522	0.507	0.537	0.03	0.052	0.022	ncbi-geneid:79875	hsa:79875	THSD4, AAT12	thrombospondin type-1 domain-containing protein 4	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Protein families: signaling and cellular processes	Familial thoracic aortic aneurysm and dissection
21	32659284	rs11702036	2.38	T	C	0.291	20	0	20	8	12	0.149	0.448	0.515	0.478	0.03	0.037	0.022	ncbi-geneid:7074	hsa:7074	TIAM1, NEDLDS	T-lymphoma invasion and metastasis-inducing protein 1	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Cellular community - eukaryotes, Cell motility, Immune system, Cancer: overview, Protein families: signaling and cellular processes	
7	147677544	rs12532921	2.07	C	A	0.2985	17	0	17	4	13	0.127	0.597	0.552	0.56	0.045	0.007	0.037	ncbi-geneid:26047	hsa:26047	CNTNAP2, AUTS15	contactin associated protein-like 2	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signaling molecules and interaction, Protein families: signaling and cellular processes	Pitt-Hopkins syndrome,Autism
11	21178116	rs16907565	2.03	A	T	0.3731	22	0	22	5	17	0.164	0.515	0.478	0.478	0.075	0.097	0.112	ncbi-geneid:4745	hsa:4745	NELL1, IDH3GL	protein kinase C-binding protein NELL	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
19	4312103	rs3745989	2.84	C	A	0.2985	18	0	18	6	12	0.134	0.53	0.552	0.612	0.09	0.037	0.075	ncbi-geneid:79187	hsa:79187	FSD1, GLFND	fibronectin type III and SPRY domain-containing protein 1	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
15	71642403	rs3784384	3.86	C	A	0.3209	23	0	23	7	16	0.172	0.537	0.485	0.537	0.022	0.045	0.022	ncbi-geneid:79875	hsa:79875	THSD4, AAT12	thrombospondin type-1 domain-containing protein 4	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Signal transduction, Protein families: signaling and cellular processes	Familial thoracic aortic aneurysm and dissection
3	185484257	rs4686388	3.77	A	G	0.3284	27	0	27	10	17	0.201	0.612	0.612	0.59	0.067	0.045	0.112	ncbi-geneid:10644	hsa:10644	IGF2BP2, IMP-2	insulin-like growth factor 2 mRNA-binding protein 2	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	Type 2 diabetes mellitus
4	21315062	rs6812694	2.87	C	T	0.2836	16	0	16	8	8	0.119	0.448	0.463	0.448	0.03	0.022	0.045	ncbi-geneid:80333	hsa:80333	KCNIP4, CALP	Kv channel-interacting protein	Domain-containing proteins not elsewhere classified [BR:hsa04990]	Protein families: signaling and cellular processes	
18	59763727	rs10432259	2.97	T	G	0.291	23	0	23	11	12	0.172	0.582	0.575	0.619	0.112	0.082	0.067	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
18	59791520	rs10503065	2.16	A	G	0.291	23	0	23	9	14	0.172	0.537	0.552	0.582	0.119	0.082	0.06	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
18	59779231	rs12607624	3.17	G	A	0.291	23	0	23	10	13	0.172	0.59	0.575	0.619	0.112	0.075	0.06	ncbi-geneid:23556	hsa:23556	PIGN, MCAHS	GPI ethanolamine phosphate transferase 1 [EC:2.7.-.-]	Enzymes [BR:hsa01000]	Glycan biosynthesis and metabolism, Transferring phosphorus-containing groups	Multiple congenital anomalies-hypotonia-seizures syndrome,Inherited glycosylphosphatidylinositol deficiencies
4	47589203	rs1316874	3.13	G	A	0.2985	11	0	11	2	9	0.082	0.56	0.478	0.478	0.03	0.007	0.037	ncbi-geneid:57205	hsa:57205	ATP10D, ATPVD	phospholipid-translocating ATPase [EC:7.6.2.1]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Catalysing the translocation of other compounds	
10	90216485	rs2312610	3.63	C	T	0.3582	25	0	25	7	18	0.187	0.515	0.5	0.507	0.03	0.052	0.06	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
10	90257225	rs2576171	2.87	T	C	0.3582	22	0	22	6	16	0.164	0.448	0.448	0.455	0.007	0.022	0.03	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
14	74538706	rs4646861	5.28	A	G	0.3806	16	0	16	3	13	0.119	0.582	0.567	0.5	0.037	0.007	0.007	ncbi-geneid:4329	hsa:4329	ALDH6A1, MMSADHA	malonate-semialdehyde dehydrogenase (acetylating) / methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.18 1.2.1.27]	Enzymes [BR:hsa01000]	Carbohydrate metabolism, Amino acid metabolism, Metabolism of other amino acids, Acting on the aldehyde or oxo group of donors	Methylmalonate semialdehyde dehydrogenase deficiency
10	90216054	rs4934409	2.63	T	A	0.3582	25	0	25	7	18	0.187	0.515	0.5	0.507	0.03	0.052	0.06	ncbi-geneid:55328	hsa:55328	RNLS, C10orf59	renalase [EC:1.6.3.5]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on NADH or NADPH	
2	64804495	rs10180097	2.89	A	G	0.3358	13	0	13	3	10	0.097	0.493	0.53	0.478	0.045	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
3	32024251	rs12629793	2.18	C	G	0.2985	23	0	23	7	16	0.172	0.44	0.522	0.463	0.007	0.015	0.037	ncbi-geneid:114884	hsa:114884	OSBPL10, ORP10	oxysterol-binding protein-related protein 9/10/11	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64803774	rs12713520	2.38	A	G	0.3358	13	0	13	3	10	0.097	0.493	0.53	0.478	0.06	0.052	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64773779	rs1978404	2.12	A	G	0.3358	12	0	12	4	8	0.09	0.478	0.53	0.478	0.045	0.06	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
21	35934476	rs2251817	3.31	C	T	0.3134	22	0	22	6	16	0.164	0.627	0.597	0.612	0.067	0.104	0.104	ncbi-geneid:1827	hsa:1827	RCAN1, ADAPT78	calcipressin-1	Membrane trafficking [BR:hsa04131]	Endocrine system, Infectious disease: viral, Protein families: genetic information processing	Down syndrome
2	64783870	rs7562829	2.14	T	G	0.3358	13	0	13	4	9	0.097	0.493	0.522	0.478	0.052	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	64752872	rs7585497	2.03	G	A	0.3358	12	0	12	3	9	0.09	0.47	0.522	0.463	0.045	0.045	0.06	ncbi-geneid:54812	hsa:54812	AFTPH, Nbla10388	aftiphilin	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
2	144400860	rs78141710	2.42	C	A	0.3657	18	0	18	4	14	0.134	0.448	0.381	0.418	0.015	0.082	0.022	ncbi-geneid:55843	hsa:55843	ARHGAP15, BM046	Rho GTPase-activating protein 15	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	
21	35938024	rs8130808	2.29	G	T	0.3134	22	0	22	6	16	0.164	0.604	0.582	0.612	0.067	0.104	0.104	ncbi-geneid:1827	hsa:1827	RCAN1, ADAPT78	calcipressin-1	Membrane trafficking [BR:hsa04131]	Endocrine system, Infectious disease: viral, Protein families: genetic information processing	Down syndrome
13	24701318	rs9553191	2.59	A	G	0.3731	22	0	22	11	11	0.164	0.463	0.396	0.433	0.03	0.045	0.037	ncbi-geneid:221178	hsa:221178	SPATA13, ARHGEF29	Rho guanine nucleotide exchange factor 4/29	Membrane trafficking [BR:hsa04131]	Cell motility, Protein families: genetic information processing	
14	65555471	rs10143198	2.55	T	C	0.3284	22	0	22	6	16	0.164	0.649	0.604	0.575	0.097	0.06	0.097	ncbi-geneid:4149	hsa:4149	MAX, bHLHd4	Max protein	Transcription factors [BR:hsa03000]	Signal transduction, Cancer: overview, Cancer: specific types, Protein families: genetic information processing	Malignant paraganglioma
18	53170201	rs1037430	2.1	T	C	0.3358	16	0	16	7	9	0.119	0.425	0.41	0.493	0.03	0.015	0.015	ncbi-geneid:6925	hsa:6925	TCF4, CDG2T	transcription factor 4/12	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Pitt-Hopkins syndrome,Fuchs corneal dystrophy
3	32024251	rs12629793	2.18	C	G	0.2985	23	0	23	7	16	0.172	0.44	0.522	0.463	0.007	0.015	0.037	ncbi-geneid:344787	hsa:344787	ZNF860,NA	KRAB domain-containing zinc finger protein	Transcription factors [BR:hsa03000]	Infectious disease: viral, Protein families: genetic information processing	
10	3821561	rs17731	3.18	G	A	0.403	17	0	17	3	14	0.127	0.493	0.44	0.47	0.03	0.022	0.015	ncbi-geneid:1316	hsa:1316	KLF6, BCD1	krueppel-like factor 6/7	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Prostate cancer
7	42067356	rs2237425	5.15	C	G	0.2463	13	0	13	7	6	0.097	0.41	0.328	0.418	0.007	0.007	0.007	ncbi-geneid:2737	hsa:2737	GLI3, ACLS	zinc finger protein GLI3	Transcription factors [BR:hsa03000]	Signal transduction, Cancer: overview, Cancer: specific types, Protein families: genetic information processing	Pallister-Hall syndrome,Polysyndactyly,Postaxial polydactyly,Greig cephalopolysyndactyly syndrome,Preaxial polydactyly
6	45545074	rs3899143	2.72	C	T	0.306	15	0	15	4	11	0.112	0.463	0.418	0.425	0.015	0.015	0.007	ncbi-geneid:860	hsa:860	RUNX2, AML3	runt-related transcription factor 2	Transcription factors [BR:hsa03000]	Endocrine system, Cancer: overview, Protein families: genetic information processing	Cleidocranial dysplasia
9	4169816	rs4740753	2.16	T	C	0.3209	26	0	26	10	16	0.194	0.664	0.694	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4170020	rs4740755	2.55	A	C	0.3209	25	0	25	9	16	0.187	0.664	0.687	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4169621	rs7045178	3.48	T	A	0.3209	26	0	26	10	16	0.194	0.664	0.687	0.627	0.06	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
9	4170457	rs7851070	2.04	A	G	0.3209	26	0	26	9	17	0.194	0.664	0.687	0.627	0.067	0.104	0.082	ncbi-geneid:169792	hsa:169792	GLIS3, NDH	zinc finger protein GLIS1/3	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	Permanent neonatal diabetes mellitus
19	31829903	rs924150	3.3	A	C	0.3358	20	0	20	7	13	0.149	0.485	0.463	0.53	0.06	0.022	0.045	ncbi-geneid:57616	hsa:57616	TSHZ3, TSH3	teashirt	Transcription factors [BR:hsa03000]	Protein families: genetic information processing	

Table 4. Variants showing highly differentiated allele frequencies between African and Asian populations. Legend is the same described in Table 1.

GRCh37 genomic information						Malagasy genetic diversity						Derived allele frequency (DAF)							Functional annotation to the mapped gene						
CHR	POS	rsID	GERP-RS	AA	DA	EAS_anc	Heterozygote	Homozygote Derived	Derived	African Derived	Asian Derived	MGY	CHB	KHV	CDX	LWK	ESN	YRI	NCBI_id	KEGG_id	GENE NAME	ORTHOLOGY	PATHWAY	KEGG_info	DISEASE
8	124154697	rs10101626	5.36	G	T	0.3806	28	2	32	24	8	0.239	0.209	0.142	0.112	0.343	0.321	0.261	ncbi-geneid:93594	hsa:93594	TBC1D31, Gm85				
19	1534042	rs10853954	2.29	T	C	0.291	37	10	57	27	30	0.425	0.575	0.575	0.694	0.269	0.373	0.313	ncbi-geneid:126520	hsa:126520	PLK5, PLK-5	inactive serine/threonine-protein kinase PLK5	DNA repair and recombination proteins [BR:hsa03400]	Protein families: genetic information processing	
8	143922642	rs114099899	2.55	A	G	0.4328	2	65	132	74	58	0.985	1	1	1	0.97	0.978	0.97	ncbi-geneid:2765	hsa:2765	GML, LY6DL				
14	57948380	rs1152522	2.75	T	C	0.2612	29	12	53	23	30	0.396	0.91	0.993	0.978	0.187	0.172	0.209	ncbi-geneid:55195	hsa:55195	CCDC198, C14orf105				
1	40773150	rs12077871	3.92	G	A	0.3358	14	0	14	5	9	0.104	0.119	0.09	0.082	0.127	0.104	0.097	ncbi-geneid:1298	hsa:1298	COL9A2, DJ39G22.4	collagen type IX alpha	Proteoglycans [BR:hsa00535]	Signal transduction, Signaling molecules and interaction, Cellular community - eukaryotes, Digestive system, Infectious disease: viral, Protein families: signaling and cellular processes	Multiple epiphyseal dysplasia,Stickler syndrome
1	5935162	rs1287637	4.93	T	A	0.2985	11	2	15	13	2	0.112	0.224	0.231	0.276	0.127	0.097	0.067	ncbi-geneid:261734	hsa:261734	NPHP4, POC10	nephrocystin-4	Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	Nephronophthisis,Senior-Loken syndrome
14	24470138	rs1811890	2.84	C	T	0.3881	19	3	25	10	15	0.187	0.269	0.209	0.134	0.142	0.022	0.03	ncbi-geneid:10901	hsa:10901	DHRS4, CR	dehydrogenase/reductase SDR family member 4 [EC:1.1.-.-]	Enzymes [BR:hsa01000]	Metabolism of cofactors and vitamins, Transport and catabolism, Acting on the CH-OH group of donors	
14	24470138	rs1811890	2.84	C	T	0.3881	19	3	25	10	15	0.187	0.269	0.209	0.134	0.142	0.022	0.03	ncbi-geneid:317749	hsa:317749	DHRS4L2, SDR25C3	dehydrogenase/reductase SDR family member 4-like protein 2 [EC:1.1.-.-]	Enzymes [BR:hsa01000]	Metabolism of cofactors and vitamins, Acting on the CH-OH group of donors	
18	25616451	rs1944294	3.2	A	T	0.3134	22	5	32	13	19	0.239	0.254	0.351	0.321	0.037	0.067	0.082	ncbi-geneid:1000	hsa:1000	CDH2, ACOGS	cadherin 2, type 1, N-cadherin	Chromosome and associated proteins [BR:hsa03036], Cell adhesion molecules [BR:hsa04515], CD molecules [BR:hsa04090]	Signaling molecules and interaction, Cardiovascular disease, Protein families: genetic information processing, Protein families: signaling and cellular processes, (CDH)	Arrhythmogenic right ventricular cardiomyopathy,Agenesis of corpus callosum, cardiac, ocular, and genital syndrome
1	179858444	rs2245425	5.26	G	A	0.291	30	23	76	46	30	0.567	0.776	0.731	0.761	0.328	0.425	0.507	ncbi-geneid:26092	hsa:26092	TOR1AIP1, LAP1	torsin-1A-interacting protein	Chaperones and folding catalysts [BR:hsa03110]	Protein families: genetic information processing	Limb-girdle muscular dystrophy
1	236706300	rs2273865	2.59	T	A	0.3507	16	3	22	17	5	0.164	0.149	0.201	0.216	0.172	0.216	0.261	ncbi-geneid:3964	hsa:3964	LGALS8, Gal-8	galectin-8	Membrane trafficking [BR:hsa04131], Lectins [BR:hsa04091]	Protein families: genetic information processing, Protein families: signaling and cellular processes, (LGALS)	
4	111542154	rs2278782	2.67	G	A	0.2761	15	2	19	14	5	0.142	0.172	0.104	0.201	0.112	0.179	0.127	ncbi-geneid:5308	hsa:5308	PITX2, ARP1	paired-like homeodomain transcription factor 2	Transcription factors [BR:hsa03000]	Signal transduction, Protein families: genetic information processing	Axenfeld-Rieger syndrome,Ring dermoid of cornea,Anterior segment dysgenesis
14	88862529	rs3179969	2.45	A	G	0.2761	36	20	76	58	18	0.567	0.664	0.537	0.59	0.343	0.485	0.425	ncbi-geneid:55812	hsa:55812	SPATA7, HEL-S-296	spermatogenesis-associated protein 7	Cilium and associated proteins [BR:hsa03037]	Protein families: signaling and cellular processes	Leber congenital amaurosis
5	74965122	rs34358	4.88	G	A	0.4104	34	11	56	37	19	0.418	0.396	0.455	0.418	0.597	0.515	0.493	ncbi-geneid:728780	hsa:728780	ANKDD1B,NA				
8	99205612	rs3735887	3.65	T	C	0.291	27	19	65	48	17	0.485	0.485	0.328	0.433	0.522	0.545	0.493	ncbi-geneid:79815	hsa:79815	NIPAL2, NPAL2	magnesium transporter	Transporters [BR:hsa02000]	Protein families: signaling and cellular processes	
19	51535130	rs3745540	4.27	A	G	0.3209	34	9	52	26	26	0.388	0.567	0.657	0.687	0.358	0.306	0.321	ncbi-geneid:43849	hsa:43849	KLK12, KLK-L5	kallikrein 12 [EC:3.4.21.-]	Enzymes [BR:hsa01000], Peptidases and inhibitors [BR:hsa01002]	Protein families: metabolism, Acting on peptide bonds (peptidases)	
15	31294714	rs3784589	5.05	C	A	0.2985	4	0	4	3	1	0.03	0.007	0.007	0.007	0.112	0.045	0.037	ncbi-geneid:4308	hsa:4308	TRPM1, CSNB1C	transient receptor potential cation channel subfamily M member 1	Ion channels [BR:hsa04040]	Protein families: signaling and cellular processes	Congenital stationary night blindness,Chromosome 15q13.3 microdeletion syndrome
1	228469903	rs3795786	4.21	A	T	0.3134	12	1	14	3	11	0.104	0.269	0.313	0.254	0.022	0.007	0.037	ncbi-geneid:84033	hsa:84033	OBSCN, ARHGEF30	obscurin-RhoGEF [EC:2.7.11.1]	Enzymes [BR:hsa01000], Protein kinases [BR:hsa01001], Membrane trafficking [BR:hsa04131]	Protein families: metabolism, Protein families: genetic information processing, Transferring phosphorus-containing groups	
1	67242087	rs3816989	5.92	G	A	0.3731	14	2	18	12	6	0.134	0.112	0.157	0.134	0.097	0.052	0.037	ncbi-geneid:200132	hsa:200132	DYNLT5, TCTEX1D1	dynein light chain Tctex-type 5	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Protein families: signaling and cellular processes	
7	156468559	rs3823617	5.22	T	C	0.4179	10	0	10	9	1	0.075	0.127	0.112	0.097	0.097	0.149	0.127	ncbi-geneid:140545	hsa:140545	RNF32, FKSG33				
7	156468559	rs3823617	5.22	T	C	0.4179	10	0	10	9	1	0.075	0.127	0.112	0.097	0.097	0.149	0.127	ncbi-geneid:64327	hsa:64327	LMBR1, ACHP	limb region 1 protein	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Acheiropodia,Triphalangeal thumb-polysyndactyly syndrome,Syndactyly,Preaxial polydactyly,Laurin-Sandrow syndrome
1	247719769	rs56043070	3	G	A	0.4254	2	0	2	2	0	0.015	0.022	0.015	0.022	0.022	0.045	0.022	ncbi-geneid:148823	hsa:148823	GCSAML, C1orf150				
1	223285200	rs5744168	4.67	G	A	0.3507	2	0	2	2	0	0.015	0.045	0.045	0.06	0.022	0.015	0.03	ncbi-geneid:7100	hsa:7100	TLR5, MELIOS	toll-like receptor 5	Pattern recognition receptors [BR:hsa04054]	Immune system, Infectious disease: bacterial, Immune disease, Protein families: signaling and cellular processes	Systemic lupus erythematosus
15	55722882	rs57809907	4.68	C	A	0.403	25	6	37	34	3	0.276	0.007	0.022	0.007	0.463	0.396	0.619	ncbi-geneid:161582	hsa:161582	DNAAF4, CILD25	dynein axonemal assembly factor 4	Cilium and associated proteins [BR:hsa03037], Cytoskeleton proteins [BR:hsa04812]	Protein families: signaling and cellular processes	Primary ciliary dyskinesia,Dyslexia
1	47080679	rs6671527	2.2	A	G	0.3582	30	3	36	31	5	0.269	0.06	0.075	0.075	0.44	0.5	0.448	ncbi-geneid:8569	hsa:8569	MKNK1, MNK1	MAP kinase interacting serine/threonine kinase [EC:2.7.11.1]	Enzymes [BR:hsa01000], Protein kinases [BR:hsa01001]	Signal transduction, Endocrine system, Protein families: metabolism, Transferring phosphorus-containing groups	
1	47080679	rs6671527	2.2	A	G	0.3582	30	3	36	31	5	0.269	0.06	0.075	0.075	0.44	0.5	0.448	ncbi-geneid:148932	hsa:148932	MOB3C, MOB1E				
8	100133706	rs7460625	3.32	G	T	0.291	27	8	43	30	13	0.321	0.5	0.463	0.455	0.396	0.187	0.313	ncbi-geneid:157680	hsa:157680	VPS13B, BLTP5B	vacuolar protein sorting-associated protein 13B	Membrane trafficking [BR:hsa04131]	Protein families: genetic information processing	Cohen syndrome
10	90354423	rs74777687	3.87	G	C	0.3582	8	0	8	6	2	0.06	0.007	0.022	0.037	0.037	0.067	0.03	ncbi-geneid:142910	hsa:142910	LIPJ, LIPL1	lipase member J [EC:3.1.1.-]	Enzymes [BR:hsa01000]	Unclassified: metabolism, Acting on ester bonds	
1	158549492	rs863362	4.18	C	T	0.0821	33	24	81	73	8	0.604	0.634	0.515	0.604	0.552	0.627	0.604	ncbi-geneid:128367	hsa:128367	OR10X1, OR1-13	olfactory receptor	G protein-coupled receptors [BR:hsa04030]	Sensory system, Protein families: signaling and cellular processes	
21	31744127	rs877346	2.31	A	T	0.3209	24	10	44	21	23	0.328	0.246	0.366	0.336	0.09	0.142	0.179	ncbi-geneid:337959	hsa:337959	KRTAP13-2, KAP13-2				

Table 5. Potentially deleterious variants based on three different annotation software. GRCh37 genomic information depicts

chromosome, position, rsID, GERP-RS, annotated effect allele (EA) according to snpEFF, ancestral allele (AA), and derived allele (DA) of locus. The rest of the legend is the same described in Table 1.