Histories of Admixture

Paul Verdu

MUSEUM NATIONAL D'HISTOIRE NATURELLE

Année 2022

HABILITATION A DIRIGER DES RECHERCHES

du

MUSEUM NATIONAL D'HISTOIRE NATURELLE

Spécialité : Anthropologie et génétique des populations

Présentée et soutenue publiquement par

Paul Verdu

2022

Histoires de Métissages

Histories of Admixture

JURY :

Mme. Anouk BARBEROUSSE	**Professeure à Sorbonne Université**	**Examinatrice**
Mr. Simon BOITARD	**Chargé de Recherche à l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement**	**Examinateur**
Mme. Catherine BOURGAIN	**Directrice de recherche à l'Institut National de la Santé et de la Recherche Médicale**	**Rapportrice**
Mr. Simon GRAVEL	**Professeur à l'Université McGill**	**Rapporteur**
Mme. Emmanuelle PORCHER	**Professeure au Muséum National d'Histoire Naturelle**	**Examinatrice**
Mr. Alexandre ROBERT	**Professeur au Muséum National d'Histoire Naturelle**	**Rapporteur**

A Marie-France Mifune,

Grâce à qui tout est possible.

Table of Contents

Introduction

Histories of Admixture

Introduction. Histories of Admixture

The object of study at the core of any population genetics investigation of biological evolution processes is genetic variation, observed, or imagined, among individuals within and between populations and species. In a classical synthetic theory of evolution framework, neutral and neo-Darwinian, illustrated in the schematic figure below, genetic variation is influenced by major evolutionary forces. Mutation is the only force generative, randomly, of genetic variation across individuals from one generation to the next. Demographic forces (or rather "demo-genetic" forces) then shape the distribution of this diversity across individuals and populations, randomly via genetic drift, or through reproductive isolation and, conversely, genetic admixture or migration. Finally, Darwinian natural selection influences indirectly genetic variation via the selection of those particular phenotypes related to reproductive fitness that have a genetic determination.

As I was implicitly taught classically during most of my studies before my PhD, genetic variation and natural selection pertained to the "natural", "innate", world, while mutation and demography were largely considered as formal forces often disjointed from any specific environmental considerations as they mechanistically influenced genetic variation "blindly". The Neutral Theory of evolution (Kimura 1968, 1983), which ultimately lead to this synthetic theory of genetic evolution, made a tremendous impression on me personally: only a small proportion of genetic variation was determining directly phenotypic variation across individuals.

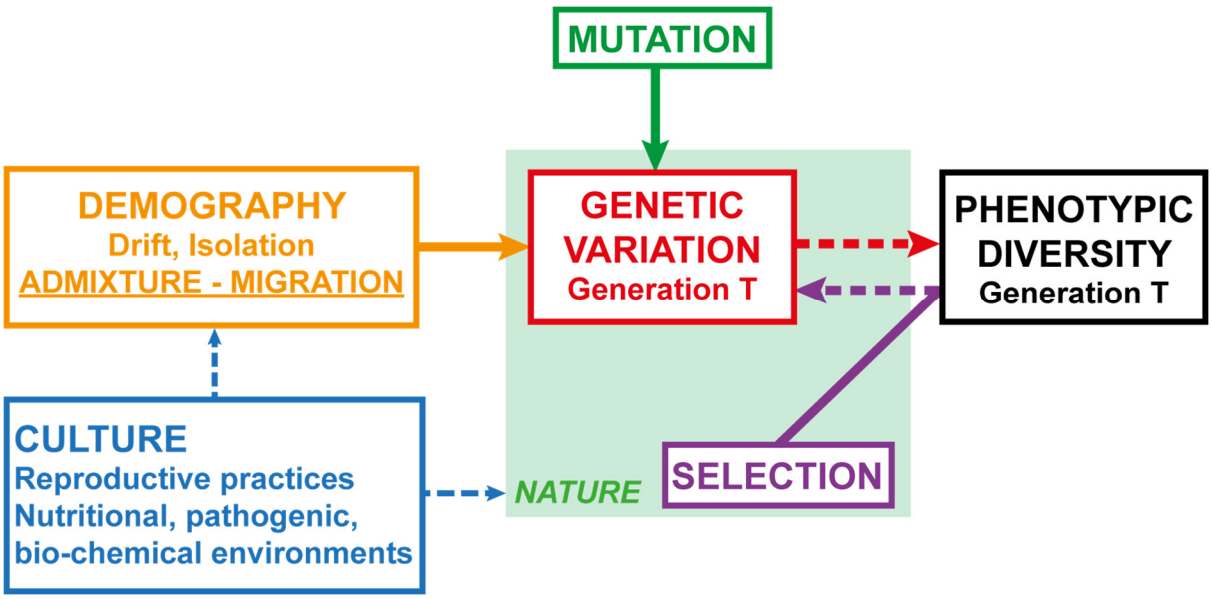


Figure Introduction

Schematic representation of the influence of evolutionary forces (mutation, demography, selection) on genetic diversity and partially associated phenotypic diversity, in the classical neutral, neo-Darwinian, synthetic theory of evolution.

Culture, defined as all practices and behaviors transmitted from one generation to the next in a non-genetic way, indirectly influences genetic diversity patterns of all species that practice it. Admixture and migrations are mainly demographic forces known to influence genetic diversity patterns. Note that, the Neutral Theory of Evolution (Kimura 1968, 1983), showed that only a very limited part of genetic diversity determines directly phenotypic diversity.

Culture, which I define comfortably here as all practices and behaviors transmitted among individuals from one generation to the next in a non-genetic way, was recognized as a possible force influencing the distribution of genetic variation of any species which practiced it, including humans but not limited to them, by far. However, in humans, while cultural practices were known to influence reproductive practices and/or nutritional and pathogenic environments, thus indirectly shaping demographic and natural selection forces, evolutionary geneticists very often considered that this force was too versatile and changing to influence significantly the evolution of our species. From a cultural anthropology perspective, the raging "Nature-Culture" (or "Nature-Nurture") debate often resulted in separating human biology from the realm of cultural behaviors, creating different, hermetically sealed, objects of study. On one hand, anthropological biologists would focus on the natural part of our evolution and the strict equivalence between genetic diversity and phenotypic diversity, hereby often completely missing the teachings of the Neutral Theory revolution. On the other hand, cultural anthropologists would focus on the cultural practices and representations disconnected from biological considerations, and relatively often disconnected even from natural environments in which humans lived.

In this context, genetic admixture and migrations represented essential demographic mechanisms of human biological evolution, as for any other species investigated by population geneticists. A large body of seminal theoretical approaches, built since the very beginning of population genetics at the turn of the 20th century, provided methodological statistical tools to identify admixture and migration signatures in the observed genetic data, but populations were first conceived as isolated and non-admixed before admixture and migration may influence their evolution. Furthermore, how admixture or migration had occurred in human genetic evolution was often not directly investigated until the end of the 20th century.

The object of this dissertation is to recapitulate advances in anthropological genetics and human population genetics since the early 2000's having furthered our understanding on how admixture and migration influenced genetic evolution. In particular, I illustrate these developments with my personal research endeavors since 2005 and the beginning of my PhD. I hereby intend to show that socio-cultural processes massively determine admixture and migration across human populations and communities, in complex ways over time and space. This shows that Culture is far from being a weak force in human evolution, and further illustrates how the opposition of Nature and Culture is often not operational to understand human genetic variation and evolution. Altogether, I strongly plea for the necessity to embrace different disciplinary paradigms and methods to investigate the multi-faceted history of human evolution.

I believe that my research work stands at the crossroads of certain cultural anthropology disciplines, mainly ethnology and historical-, computational-, and socio-linguistics, with population genetics. In particular, most of my research aim at reconstructing the past demographic and evolutionary history of human populations from genetic variation patterns observed within and across populations today, and using population genetics approaches. I thus consider my work to be either anthropological genetics or human population genetics work, two disciplines sharing the same population genetics paradigms and most methods, and that I differentiate, somehow a bit arbitrarily and cosmetically I reckon, only based on their finality: anthropological genetics is mainly interested in the history of human populations and groups, while human population genetics is mainly interested in the genetic evolution of our species (Cavalli-Sforza and Feldman 2003).

In practice, I first elaborate anthropological questions about the peopling and evolutionary history of human populations and about the influence of socio-cultural behaviors having shaped the biological and cultural diversity of populations. I then design and conduct fieldwork in order to collect cultural anthropology, genetic, and linguistic data from extant populations, mainly in Central and Western Africa,

so far. I then generate the genetics data on the collected samples using a variety of molecular genetics technics. In parallel, I elaborate certain theoretical, mathematized, population genetics frameworks for analyzing population genetics data, when existing approaches do not allow me to directly investigate my specific questions of interest. I also develop the methodological framework and associated statistical and computational tools to use these new frameworks for the analysis of observed genetic data. I then apply these methods to the newly generated data, as well as any other existing population genetics statistical approaches relevant to address my anthropological questions of interest. More recently, my research developed to jointly investigate genetic and linguistic diversity within populations and languages, which required developing a novel framework, defining novel objects of study, and the accompanying population genetics and linguistics methods for investigating them. Therefore, my research is both of an anthropological and a population genetics nature.

In this dissertation, I aimed at recapitulating my past research, with a particular focus on explaining how did I obtain results and, thus, on how my research was elaborated and conducted. Indeed, I am mostly interested in how scientific results are obtained, often vastly more than in the obtained results themselves. Therefore, I was not interested in producing here a summary of all my results, big or small, supposedly influential or confidential, without trying to explain how I got to ask myself these questions and how I decided to try to answer them. In particular, several of my former and current PhD students repeatedly asked me how to elaborate a research project from scratch, and how did I proceed for my previous published works. While I try to teach them exactly this by doing research, as any research supervisor I believe, I thought it might be interesting to them and my future students to have a written document presenting numerous conceptual and practical aspects of my previous research endeavors, concerning the elaboration of anthropological questions of interest, fieldwork design, ethics and deontology, theories and methodologies, molecular data generation, and statistical analyses.

Therefore, I tried to narrate as much as possible, sometimes with personal comments, how I conducted my research, how I made mistakes and restarted, how I sometimes succeeded, and, most of all, how I did not do things alone, far from it… Indeed, my work is essentially collaborative and pluri-disciplinary, sometimes even interdisciplinary, and I would be incapable of pursuing scientific research without the help and support of numerous researchers, including all the students I have had the chance and honor to supervise over the years.

Hopefully, and sincerely, this dissertation will also be of some interest to other more senior researchers in cultural anthropology, evolutionary biology, and population genetics, more generally and among others.

The **first Chapter** of this dissertation presents how scientific and non-scientific categories and categorization processes of human individuals and populations shape the possible questions of interest to different disciplines of anthropology and evolutionary biology, often naïvely and sometime prejudicially, and also shape all the downstream scientific dialectics deployed to address them. I introduce here the genesis of my principal research projects conducted over the years in Central Africa and Cabo Verde in West Africa, the building of the anthropological questions that I tried to address, and the design of fieldwork sampling strategies and methods. I provide a practical guide concerning categorization issues in anthropological genetics and human population genetics, a major subject in my daily research practice. I then discuss specific ethical and deontological issues in the practice of human population genetics research, in particular focusing on data privacy protection and data sharing in the Open Science era. I also provide a practical guide about these issues. Finally, I develop perspectives about categorization issues in the paleo-genomics era, that I believe to be at the root of difficult relationships between archaeologists and geneticists.

Note that the first two sections of this chapter also largely serve as detailed introductions to **Chapter 2** and **5**, respectively. Finally, this chapter is the only largely original production of this dissertation.

Chapter 2 focuses on summarizing the dialectics, methodologies, and results we have generated for addressing multiple questions about the peopling and evolutionary histories of Central African populations, and on how they were profoundly shaped by complex socio-cultural behavior determining admixture processes among populations. Altogether, this work, conducted since 2005, highlighted, for me, how complex admixture processes could be a major mechanism of human biological evolution, which determined all my other research endeavors.

Chapter 3 focuses specifically on the theoretical, mathematical, framework we have elaborated for investing highly complex admixture histories using genetic data.

Chapter 4 presents the methodological approach we developed for reconstructing such complex admixture histories from observed genetic data, and the sets of statistical and computational tools we produced to achieve this goal.

Chapter 5, finally, summarizes our investigations of the genetic and linguistic diversity of Cabo Verde, and how complex admixture processes, determined by changing socio-historical contexts over the last 400 years, may have shaped them. Ultimately, I present here the novel joint population genetics and population linguistics framework that we are currently developing for investigating the diversity of human biological and cultural histories.

Note that, instead of listing perspectives of future research in the conclusion, each chapter presents specific “ongoing research” and “perspectives” sub-sections were I elaborate on some of my current endeavors and future prospects concerning each subject separately.

Finally, I **conclude** this dissertation with general perspectives echoing the first part of this introduction, and with future challenges of disciplinary and inter-disciplinary anthropology research. I finish with a personal recommendation to vastly increase the efforts for mediating and diffusing the Scientific Methods towards specialists and non-specialists alike, a major academic interdisciplinary necessity as well as a crucial challenge for scientific research in the society.

Chapter 1

Who should I sample?

Nditam at dawn.
Pays Tikar, Région Centre, Cameroun, 2016

Chapter 1. Who should I sample?

At the heart of every population genetics investigation lies the genetic variation that emerges from the comparison of germinal DNA sequences (or genotypes) among individuals. Contrarily to, for instance, numerous fungi, algae or vegetal multicellular organisms, *Homo sapiens* individuals are most often easily identified as spatially separated biological organisms. Investigating human genetic variation is thus a priori unequivocal as genetic variation itself stems directly from the comparison of DNA sequences unambiguously gathered from different individuals. Fundamental scientific categorization issues, with major ethical, deontological, methodological, and dialectical consequences for almost every aspect of anthropological genetics research, arise only a single small step ahead of this trivial and common-place consideration, when researchers group individuals' DNA into groups and populations.

Anthropological geneticists and human population geneticists are interested in reconstructing the evolutionary history of predefined groups of individuals or *populations*. In population genetics, most fundamental theoretical expectations, statistical reasoning, hypothesis testing approaches, and methodologies in general rely on the allelic composition and frequencies estimated on sets of individuals grouped a priori into *populations*. Thus, they inherently depend on which individuals are grouped in which *populations*. In other words, if one groups individuals differently from another, estimates of allelic composition and frequencies will possibly be different; and all subsequent results of statistical inferences and their interpretations may also differ. Hence, explicitly providing the criteria used to group individuals into *populations* is essential to understand and interpret the outcome of population genetics statistical descriptions and historical inference reconstructions.

This liminal problem of each and every statistical-based scientific method becomes very practical for population geneticists involved in the primary sampling of individual DNA in the field, including human population geneticists. Indeed, "*Who should I sample?*" is the single question that determines a priori both the anthropological genetics questions that researchers may or may not address and which population genetics methods may or may not be deployed for addressing them.

In this chapter, I will present the categorization challenges and sampling methodologies we developed in order to reconstruct the peopling and evolutionary histories of Central African populations (**section 1.1**), and the genetic and linguistic admixture histories of Cabo Verde (**section 1.2**). Note that both sections will also largely serve as introductions to **Chapters 2** and **5** respectively.

In **section 1.3**, I conclude based on these experiences by providing practical guidelines for thoroughly informing and discussing categorization criteria of individuals and populations in human population genetics and anthropological genetics research.

In **section 1.4**, I elaborate on the ethical and deontological aspects of human population genetics research specifically related to these categorization issues and, more generally, on the sharing of genetic, anthropological, and linguistic data.

In **section 1.5**, I conclude this latter ethics and deontology section with practical recommendations to researchers and students.

Finally, in **section 1.6**, I develop on the perspectives of these issues in the recently emerged and massively developing field of human paleo-genetics.

1.1. Sampling individuals in Central Africa to reconstruct the genetic origins and evolutionary history of "Pygmy" and neighboring "non-Pygmy" populations

Back in 2005, the PhD project I started at the UMR7206 Eco-anthropology (UMR5145 back then) provided me with the rare and ambitious opportunity to elaborate an anthropological genetics project starting with the collection of DNA samples in the field. The project designed initially by Pr. Evelyne Heyer (my official PhD advisor) and Pr. Serge Bahuchet (director of the lab at the time and unofficial PhD advisor) aimed at gathering DNA samples from numerous Congo Basin so-called *"Pygmy"* populations in order to characterize their genetic diversity and descriptively compare genetic differentiation patterns with linguistic variation from the same populations. To do so, the population genetics side of the project could capitalize on the massive expertise of the diverse ethnologists from the laboratory historically specialized in the study of Central African Pygmy populations' cultures and ways-of-life from Cameroon, Gabon, Central-African Republic (CAR), Congo, Democratic Republic of Congo (DRC), Uganda, Burundi, and Rwanda (**Figure F1.1.a**).

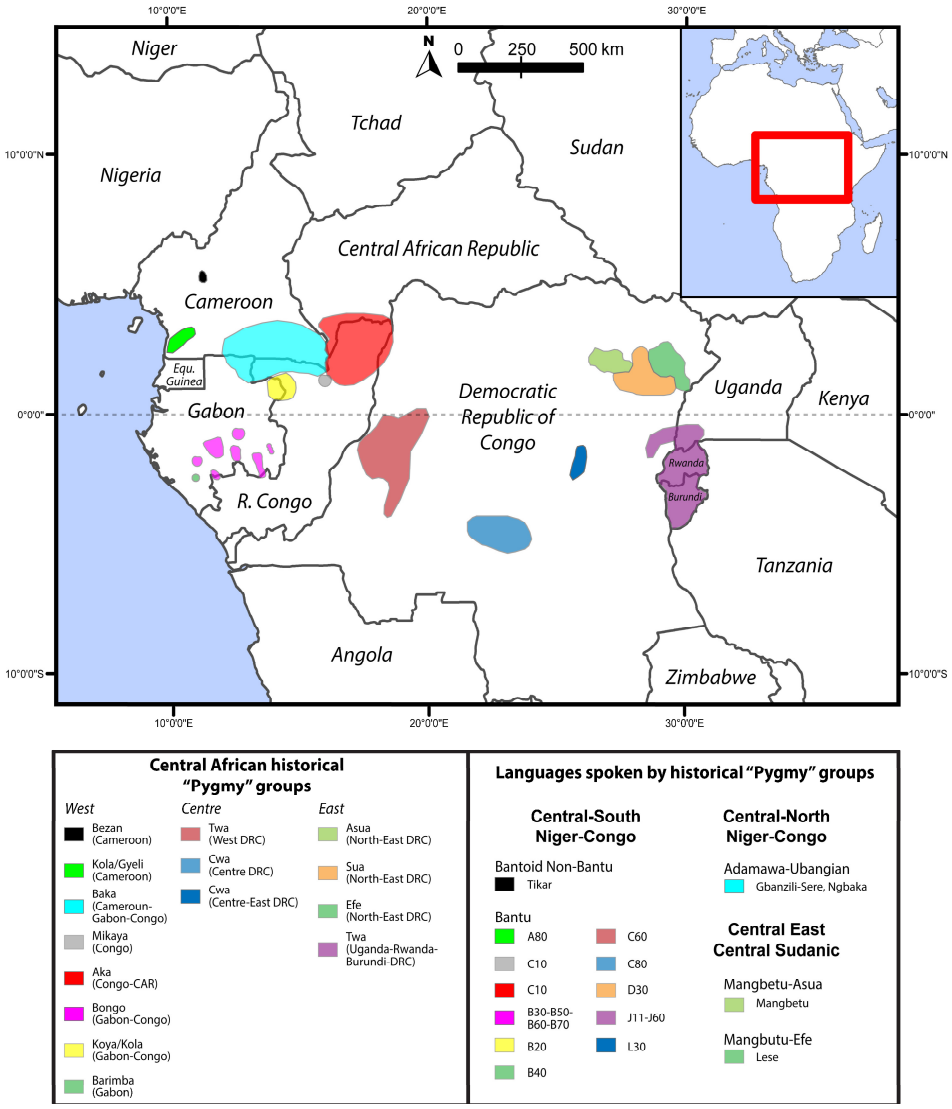


Figure F1.1.a.
Geographical map of historical "Pygmy" populations in Central Africa, inferred from ethnological surveys by S. Bahuchet, A. Froment, B. Hewlett, S. Le Bomin, H. Pagézy, P. Verdu and colleagues. Figure originally published in Verdu P. *Current Biology* 2016

Therefore, as a nascent population geneticist willing to learn about every aspect of an anthropological genetics study, from sampling in the field to population genetics data analyses via molecular genetics data generation, I had the reassuring luxury to focus on historically emblematic populations from a largely under-investigated region of the world in a scientific environment built since the 1960's by most of the main worldwide specialists of this geo-cultural region. Theoretically, I thus just had to go to Central African Pygmy villages long studied by the ethnologists of the team (together with them in a pluri-disciplinary framework), gather DNA samples from volunteer participants, come-back to France, and do some population genetics magic.

1.1.a. Who are the Central African Pygmies?

But who are the Pygmies? How am I to know who should I sample and who should I not sample on a field I had never been to myself? In the anthropological genetics' realm, Central African Pygmy populations had been the focus of famous studies initiated mainly by L. L. Cavalli-Sforza et al. in the 1980s (Cavalli-Sforza 1986), and further explored in seminal work by Destro-Bisol et al. (2004). Some DNA samples from mainly three populations, called by these authors the Biaka and Mbenzele from CAR, and the Mbuti from DRC, had been gathered and various aspects of their genetic diversity explored. They were presented as hunter-gatherer populations specialized in equatorial forest activities, living separated from other agriculturalist sedentary populations in the region, and, therefore, thought to have retained a nomadic hunter-gatherer way-of-life in the forest exemplifying most of human evolutionary history before the Neolithic agricultural revolutions. However, none of these population genetics studies detailed the inclusion criteria of individuals nor the categorization criteria of their grouping in each population and further grouping of populations into the "Pygmy" category. They were named and labelled as if the nature of these categorial groupings was obvious and innate. As a matter of fact, and most importantly, these different populations were *de facto* assumed to have a common biological origin differing from the origin of other populations in the region: they were the famous "Pygmies" and everyone was supposed to know who we were talking about…

In this context, the ethnologists involved in my PhD project had a very different perspective, substantiated by numerous publications since the 1950's and summarized in, to my views, one of the most important works on the subject: "L'Invention des Pygmées" by Serge Bahuchet (1993). With the leading expertise of Pr. Bahuchet and colleagues, I rapidly learned that, in fact, "the Pygmies do not exist". Indeed, the word Pygmy itself is derived from ancient Greek and refers to a cubit, thus a measure of short size (or the small volume of a clenched fist in alternative translations). It was first explicitly used by Homer in his Iliad (Song 3, v. 1-8), in a single instance, where the Trojan warriors fighting against the Achaeans are compared to a flock of migrating cranes pouncing on the "Pygmy" population with whom they are at war, somewhere on the other side of the seas. And that's it... Pygmies are thus, classically, anthropomorphic individuals of short stature (the word itself), living somewhere far and unknown, and at war with big migrating birds.

Why on earth did 20th century human population geneticists grouped numerous human populations from the entire Congo Basin area under such blanket term? Well, since Homer, Western cultures have been swarming with myths elaborating on Homer's few words, whether representing Homer's Pygmies on ancient vases as short-stature individuals at war with birds (**Figure F1.1.b**), or medieval dissertations and poems about the alleged particularities of the Pygmy people, and even cinematographic representations in the very first Tarzan movie in the 20th century, among numerous other instances (Bahuchet 1993; Verdu 2009).

Figure F1.1.b.
Two ancient vases figuring classical representations of "Pygmies" fighting cranes in reference to Homer's Iliad.

Left vase: Attic red-figure *chous* (oinochoe, type 3), dated around 430–420 BCE; National Archaeological Museum (Madrid) -Wikimedia commons. https://upload.wikimedia.org/wikipedia/commons/thumb/2/25/Fight_Pygmy_crane_MAN.jpg/640px-Fight_Pygmy_crane_MAN.jpg.

Right vase: Attic black-figure *hydria*, Etruria (?), dated around 530-520 BC; Roma, Museo nazionale etrusco di Villa Giulia 50425 / 438 - Beazley Archive Pottery Database 43377 - Wikimedia commons. https://upload.wikimedia.org/wikipedia/commons/thumb/3/36/Attic_black-figure_hydria_-_ABV_extra_-_Pygmies_and_cranes_-_orgiastic_komos_-_Roma_MNEVG_50425_-_04.jpg/640px-Attic_black-figure_hydria_-_ABV_extra_-_Pygmies_and_cranes_-_orgiastic_komos_-_Roma_MNEVG_50425_-_04.jpg.

Furthermore, Western explorers, including Marco Polo, travelling the world from the 14th century and on, often searched for Homer's Pygmies, obviously not finding populations at war with birds. Far from questioning the reality of Homer's depiction, they, instead, always moved the putative location of existence of the famous Pygmies to the other side of the edge of the known world, thus somewhere yet unexplored.

We thus had to wait, as Bahuchet said, more than 20 centuries after Homer, for two Western explorers independently travelling the Congo Basin during the mid-19th century (Paul Du Chaillu from West to East, and Georg Schweinfürth from East to West), who claimed to have finally found the famous Pygmies from Homer (Schweinfurth 1873; Du Chaillu 1892). Of course, they were not at war with birds… but they were apparently of relatively short stature, lived a hunting-gathering way of life in the Central African equatorial forest, and were known as different populations specialized in the forest realm by other Congo Basin agriculturalist populations with whom Westerners already used to interact since at least the 15th century. From then on, the Central African Pygmies were born, and more than 20 populations throughout the Congo Basin (**Figure F1.1**) met by various Western explorers and adventurers, including the famous journalist Stanley for instance, were categorized historically under the blanket-term "Pygmy" based on being apparently of relatively short stature (in fact, so-designated "Pygmies" were not often measured before the mid-20th century, Froment 1993), and on being recognized by outsiders as different populations specialized in forest activities.

Throughout the 20th century, ethnologists and cultural anthropologists extensively described and characterized cultures and ways-of-life in numerous (but not all) populations historically labelled as Pygmies, in turn substantiating the grouping of these populations into this category or, instead, revealing their massive cultural variation. Ultimately, they showed that numerous non-mutually exclusive cultural criteria could be informed and intersected to group populations into "Pygmies" in reference to the historical

category used by Westerners (Bahuchet 1992a, 1992b; Bahuchet and Guillaume 1992; Hewlett 1996; Fürniss 2011; Verdu and Destro-Bisol 2012; Hewlett 2014; Verdu 2014, 2016).

1) Pygmy populations would self-identify as different communities with different ethno-names than outsiders, and would symmetrically be recognized as different communities by outsiders.

2) Pygmy populations would be recognized by outsiders as specialists of hunting, fishing, and gathering subsistence activities, as well as specialists of medical and magical knowledge of forest environments.

3) They would live a mobile way-of-life in their environment, with semi-permanent habitat, as opposed to more sedentary neighboring agriculturalist populations.

4) They would practice specific musical and dancing activities, for ritual or entertainment purposes, recognized as specific and different by outsiders.

Importantly, and somewhat apparently contradictory, ethnologists and cultural anthropologists systematically described exceptions to at least one of the second, third, and fourth inclusion criteria in a given historically labelled "Pygmy" population or another throughout the Congo Basin[1].

First, while Pygmy populations are indeed often specialized in foraging activities in the forest, and recognized as such by outsiders, they all know about, and practice, agriculture and horticulture, whether in their own fields or gardens, or in their neighbors' fields and gardens; and the neighboring agriculturalists very often hunt, fish, and gather in the rain-forest themselves. Furthermore, some Pygmy populations, such as the Twa from Rwanda and Burundi (**Figure F1.1.a**), are recognized as skilled potters or weavers rather than specialized hunters and gatherers.

Second, some Pygmy populations, such as the Bongo from Gabon or the Twa from Rwanda (**Figure F1.1.a**), in fact live a very sedentary way-of-life in villages much resembling those of their agriculturalist's neighbors, and only conduct their foraging activities for limited specific periods of time rather than frequently move across semi-permanent forest-camps such as some communities of Aka from CAR and Baka from Cameroon (**Figure F1.1.a**). Furthermore, it is not uncommon that Pygmy communities and their neighbors live in the very same villages, sometimes in separate neighborhoods but sometimes not.

Third, some Pygmy populations live in other environments than the equatorial rainforest, such as the Bezan from Cameroon who live in clear-forest environments, or the Twa from Western DRC who live in forest swamps (**Figure F1.1.a**).

Fourth, some Pygmy populations indeed have very specific musical practices, with different vocal techniques and instruments than their neighbors, such as the emblematic practice of polyphonies and yodel in Aka from CAR and Baka from Cameroon or flutes orchestras in the Efe from Eastern DRC, but numerous other historical Pygmy populations do not have specific musical practices fundamentally differing from those of their neighbors.

Finally, Pygmy populations do not have a specific "Pygmy" language; "unfortunately" as the linguistic difference criteria is often used by geneticists for grouping individuals into different and separate populations. Instead, Pygmy populations all practice a language most often practiced by some of the neighboring populations not categorized as Pygmies, sometimes even from different linguistic families (Bahuchet 2012; Verdu and Destro-Bisol 2012). Thus, languages spoken by the different Pygmy populations, even geographically close from one-another, are sometimes massively differing and not mutually intelligible without learning (**Figure F1.1.a**).

[1] The first criteria always held throughout the Congo Basin, but it could not be used on its own to group populations into the Pygmy category, as there is no equivalent to a "Pygmy" word shared by all populations in the Congo Basin and as this criterion is essential for identifying any community separate from any other everywhere around the world.

1.1.b. Categorizing Central African populations into a binary category: Pygmies and non-Pygmies

In this context, grouping populations under the blanket-term Pygmy could be achieved by intersecting the majority of the above criteria of categorization with another non-exclusive and complex criteria, the only one shared by all Congo Basin populations historically labelled "Pygmies". Indeed, ethnologists and cultural anthropologists showed that all Central African Pygmy populations shared specific complex socio-economic interactions with at least one of the neighboring populations not labelled historically as Pygmies. These interactions often involved exchanges of goods, such as forest products and game, or medical and magical knowledge specific to the Pygmy groups, against money, agricultural produces, or iron products not produced by the Pygmy groups. Most importantly, these interactions are not limited to exchanges of products or knowledge, but also involve complex systems of Pygmy vassalage and infeudation to their dominant neighbors, including labor-force employment and socio-marital segregation and discrimination against Pygmies (Kazadi 1981; Bahuchet 1992a; Bahuchet and Guillaume 1992; Joiris 2003; Hewlett 2014; Verdu 2014).

As a result, it became obvious to me that I could not rely on the historical categorization of populations into Pygmies without gathering the detailed information of each population I was to visit to inform these criteria and, only then, label individuals into populations and "super-group" Pygmies. Furthermore, a classical ethnographic categorization method, relying on a nucleus of cultural traits and practices shared by members of a group excluding all other individuals, would also prove difficult in the Central African context of complex socio-economic interactions and multilingualism described above. Therefore, I decided to rely on the anthropological framework proposed by Frederik Barth which favors, instead, informing the dynamic and shifting borders and frontiers among cultural features shared across different groups in socio-cultural interactions locally (Barth 1969). Finally, ethnologists have had a long-standing debate about how to group, or not, Central African populations that were not historically grouped into the Pygmy category. They have been often designated as "Villagers", albeit as mentioned above, some Pygmy populations are also living in villages; "Agriculturalists", albeit all Pygmy populations practice agriculture, sometimes to a large extent; or even, in particular by geneticists, designated as "Bantus", a vast linguistic family encompassing more than 250 languages spoken from Central to Austral Africa, and, again, very problematic as numerous Pygmy populations themselves practice Bantu languages largely undifferentiated from those of their neighbors (**Figure F1.1.a**).

Facing these fundamental categorization issue, I decided to introduce a binary categorization system that would at least allow me to maintain separate categories systematically based on the same numerous and complex criteria described above and informed indiscriminately in all populations locally: the Pygmy/non-Pygmy binary categories (Verdu et al. 2009). We thus designed interview questionnaires heavily relying on ethnologists' expertise at a local scale that would allow us to formally inform cultural practices in all visited communities with respect to i) endogenous systems of self-identification and of identification of other than self; ii) subsistence strategies and specializations; iii) life-history traits related to spatial mobility; iv) musical practices; v) socio-economic and cultural relationships with immediate neighboring populations including intermarriages and discriminations. Intersecting *a posteriori* these cultural criteria informed in various communities locally would thus allow us to categorize samples into populations with different ethno-names, categorize every such population into Pygmy or non-Pygmy populations, and further provide key information about detailed ways-of-life for all individuals investigated and sampled, whichever the population category. Indeed, possessing all this anthropological information

from each population for which we had collected DNA samples would further allow us *a posteriori* to group the same sets of individuals into different categories, in order to formally evaluate how cultural differences or similarities may, or may not, translate into genetic differentiation patterns.

Important warning: Throughout my work, I have explicitly decided to keep the ambiguous "Pygmy" label (and its counter-part "non-Pygmy") for naming the groups of individuals based on the above detailed criteria of categorization. It is important to emphasize here that whether it is relevant to continue to use this label is still an ongoing debate in the cultural and biological anthropology communities; a debate to which I have myself participated (Hewlett 2014). A practical problem of using this somehow archaic term is due to the fact, that, in some parts of Central Africa, the word "pygmy" itself is used derogatorily by non-Pygmy neighbors. However, I also found it to be ambivalent in the field. Indeed, certain Pygmy groups or individuals choose to claim this name in front of outsiders, governmental or non-governmental outsiders, nationals or foreign, to emphasize their socio-cultural particularity and advocate, in their local power play, against the socio-economic discriminations they suffer from their neighbors. Finally, at least one tentative "re-naming" of the four historical Cameroonian "Pygmy" populations (the Bezan, the Baka, the Ba.Ghieli, and the Ba.Koya), into the "Quatre B" (the "four Bs"), by Cameroonian civil society in the 2000's, also resulted in this new label being used derogatorily as well, at least as I last witnessed in 2016…

In this complex context, I decided to first define precisely in my publications the categorization criteria used for grouping individuals into the binary category explained throughout this chapter, and decided to keep the word "Pygmy" as a label, making almost always sure to mention the potential derogatory usage of the term. I decided to use the original historical term for two reasons: as an exogenous term from ancient Greek, I believe it is more prone to ask the naïve reader the question "who are we talking about?", which I think is essential for reflexive critical thinking about the results and study design. Second, I do it in reference to the history of anthropology, and to the works of my giant predecessors.

However, as long as criteria of categorization are maintained, I really have no problems whatsoever to use another "label", as this will not change anything to the population genetics statistical descriptions and inferences performed, nor their discussions based on said criteria rather than only on the label "Pygmy". We just haven't reached a consensus on a name, yet. Human population geneticists often prefer to transform the "Pygmy" and "non-Pygmy" categories into "Rainforest Hunter Gatherers" and "Agriculturalists", as I did myself in certain collaborative publications. I believe this is very far from ideal and, in fact, even more problematic than the classical exogenous and literary "Pygmy" and "non-Pygmy". Indeed, this particular denomination for genetics research formally essentializes ways-of-life and economic practices and strictly separates them albeit, as explained above, the Central African context is not as strictly dichotomic regarding these cultural traits, far from it. Furthermore, this essentialization, as it is highly meaningful for anthropological geneticists and human population geneticists, might be in fact misleading in that it oversimplifies hypotheses and the methods used to test them into a "case-control" approach that may inevitably fail to capture the known more complex reality of human adaptation to varying ways-of-life.

For all these reasons, and bearing all these warnings in mind, I will henceforth keep using the "Pygmy" and "non-Pygmy" labels throughout this dissertation.

1.1.c. Population genetics questions about Central Africans categorized as Pygmies and non-Pygmies

Most importantly, investigating this classical anthropology categorization problem of Central African populations, and further experimenting it during numerous pluri-disciplinary fieldwork in Gabon, Cameroon and Uganda, highlighted two fundamental questions that remained, in fact, largely unexplored.

First, do Central African Pygmy populations share a common genetic origin more recently than they share a common genetic origin with neighboring non-Pygmy populations throughout the Congo Basin? In other words, do populations categorized as Pygmies all share a common biological evolution distinct from that of neighboring non-Pygmy populations after their divergences from the common ancestor, or not?

Indeed, ethnological approaches are most often synchronic by nature and explore shared and distinct cultural features among populations at a given point in time, without formally testing whether these features stem originally from a shared origin, from converging but independent cultural constructions, or from ancient exchanges among otherwise distinct communities. Furthermore, with respect to ethnological endogenous representations of the origins of Pygmy populations, note that the various Pygmy populations very rarely know the existence of other communities also labelled "Pygmies" by Westerners, even when geographically close, and that they do not share, with other such communities, a common cosmogony nor myths of origin[2]. Therefore, ethnologists repeatedly discussed the nature of the observed similarities and profound differences among the various Pygmy populations and with neighboring non-Pygmy populations, leaving most often explicitly open the question of their ancient common or independent origins. Surprisingly, before the 2005's, human population geneticists virtually never formally considered this question, probably due to the fact that the "Pygmy" category itself *de facto* assumed a common origin of all these populations (and see third point below). Nevertheless, the nature of population genetics questioning and statistical-based inference methods represented an ideal framework to formally test the common or independent biological origin of numerous populations categorized, based on cultural criteria only, into Pygmies and neighboring non-Pygmies, provided that these criteria had been systematically informed adequately in a first place.

Second, are Central African Pygmy populations reproductively isolated from their non-Pygmy neighbors, or do they form distinct cultural communities nevertheless commonly intermarrying? In other words, do Pygmy/non-Pygmy categories and associated discriminations against intermarriages indeed translated into populations' genetic isolation or, instead, were these cultural behaviors not prevalent and normative enough to avoid substantial admixture events or even random matting among communities?

Again, population geneticists interested in Central African Pygmy populations very rarely explicitly investigated population migrations and admixture at a local scale, considering a priori Pygmy populations as isolated and remote from contacts with other populations. This was consistent with common popular representations of remote and isolated Central African Pygmies in the Western cultures, despite extensive ethnologists' work describing a very different reality. This is true to the point that, most of the time, while some historical Pygmy populations were sampled for DNA in the Congo Basin, no non-Pygmy neighboring populations from the region were investigated, propagating implicitly in the human population genetics community the wrong idea that Pygmy populations were the sole inhabitants of the rain-forest...

[2] Note that several non-Pygmy neighboring populations do share common myths of origins involving "Pygmy" populations in highly similar ways, but these myths are not shared by the Pygmy populations themselves (Klieman 2003).

Finally, what about height, the alleged morphological characteristics for which the word Pygmy was used in a first place? Most interestingly, the categorization system described above does not rely on morphological criteria. Nevertheless, intersecting these cultural categories with individual height, biological anthropologists identified during the 20th century that, indeed, Pygmy populations were among the shortest worldwide. However, in a seminal work (Froment 1993), biological anthropologist Alain Froment showed the massive variation of average non-pathological adult height across the various Pygmy populations throughout the Congo Basin, ranging from 143cm high in DRC Efe males to 161cm in Twa males from Western DRC. Moreover, and most importantly, Froment *et al.* further showed that numerous neighboring non-Pygmy populations were also of relatively short stature, such as the Ba.Konjo from Western Uganda measuring on average less than 158cm high. Therefore, using only a single threshold of non-pathologic adult height in the Congo Basin cannot be sufficient to categorize individuals into the Pygmy and non-Pygmy categories overlapping the other cultural criteria described above; quite the opposite in fact as it would result in reversing the cultural and historical categorization for numerous Pygmy as well as non-Pygmy populations. Nevertheless, before the 2010's, numerous biological anthropologists and population geneticists alike considered this short stature criteria as a proof of common origin and shared evolutionary history, without ever formally testing it, and most often not correcting for inter-population variation of height. Also, importantly, note that in an almost ubiquitous way in the biological anthropology and genetics literature before the 2010's, researchers most often tried to identify why Pygmies were shorter (Perry and Dominy 2009), but never explicitly focused on why non-Pygmy neighbors were taller, nor on the evolutionary relevance of height variation among groups.

In this context, we decided to conduct state-of-the-art morphological measurements of individual height in, hopefully, all individuals sampled for DNA and interviewed extensively as described above during our project. However, we never used height as a criterion of inclusion of individuals in our study or of categorization into "Pygmy" and "non-Pygmy". Instead, we collected height in order to formally test whether individual Central African non-pathological adult height variation could be due to genetic factors or solely explained by environmental plasticity, a still vigorously debated question in the biological anthropologist community at the beginning of the 21st century. Finally, in this framework, note that during all of my yearly fieldwork interviews in Central Africa between 2006 and 2016, height differences were never invoked as an endogenous criterion for distinguishing communities and individuals. Pygmy groups never referred to height as a shared common feature of their community distinct from that of other, taller, non-Pygmy communities, and, vice versa, non-Pygmy neighbors never referred to height as identifying an individual as Pygmy or different than self.

Based on these premises, I conducted, between 2006 and 2016, 9 fieldworks (totaling 8 months residing in sampled communities) in Gabon, Cameroon, and Uganda, in order to collect more than 800 DNA samples from 14 communities categorized *a posteriori* in Pygmy and neighboring non-Pygmy based on the categorization criteria and sampling scheme presented above (**Figure F1.1.c**). Notably, I could not sample myself numerous other populations of interest to the peopling history of Central Africa, and relied for these on sampling performed by other researchers with whom I could interact during my project (**Figure F1.1.c**). I could thus obtain the various information needed to inform my categorization criteria directly from the samplers themselves, which allowed me to exclude from analyses samples for which all of my categorization criteria were not sufficiently informed.

This is often out of reach of population genetics studies conducted on data previously collected and accessible in genomics databases (see **section 1.4**). As exemplified above, this poses vast difficulties as detailed categorization criteria used for the inclusion of DNA samples in a population are often not readily

provided to other researchers. In turn, this lack of information often limits the spectrum of questions that can be addressed using population genetics methods about the genetic diversity of human populations and demographic and evolutionary processes involved at its origins. Furthermore, it is plausible that, in the lack of sufficient categorization information, population geneticists would presume that sampling strategies and inclusion criteria might be comparable across studies interested in similar questions and geo-cultural areas. Based on this assumption they may spend substantial amount of time reconciling apparently discrepant results and interpretations, or alternatively, overinterpreting similarities and homologies. However, it is plausible that categories and inclusion criteria may differ in a first place, making results from different studies hardly comparable without detailed formal discussions, themselves rendered impossible by the initial lack of information about the samples investigated.

I further develop and exemplify these aspects about another research setting part of my long-term research project in Cabo Verde, in the following section.

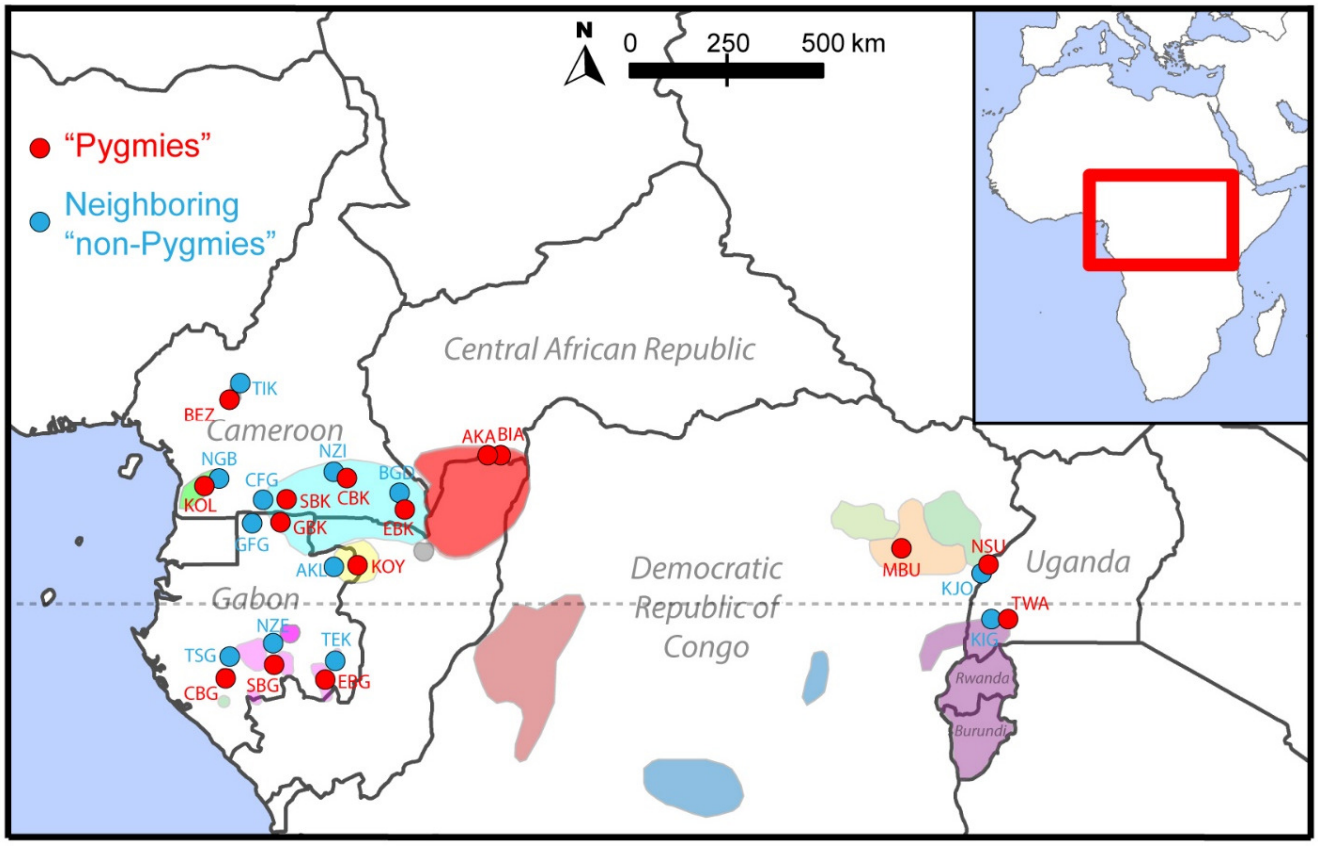


Figure F1.1.c.

	Population Name	Code	Country	Linguistic Family (Guthrie 1971)	Sampling
Pygmies	Bezan	BEZ	Cameroon	Bantoïd Non-Bantu, Tikar	P. Verdu
	Kola	KOL	Cameroon	Bantu, A80	A. Froment
	Central Baka	CBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment, P. Verdu
	Eastern Baka	EBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment
	Southern Baka	SBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment, P. Verdu
	Gabonese Baka	GBK	Gabon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	J-M Hombert, L. Van Der Veen
	Kova	KOY	Gabon	Bantu, B20	S. Le Bomin
	Central Bongo	CBG	Gabon	Bantu, B30	P. Verdu, S. Le Bomin
	Eastern Bongo	EBG	Gabon	Bantu, B70	P. Verdu, S. Le Bomin
	Southern Bongo	SBG	Gabon	Bantu, B30-B50	P. Verdu, S. Le Bomin
	Aka	AKA	CAR	Bantu, C10	B. Hewlett
	Biaka	BIA	CAR	Bantu C10	L.L. Cavalli-Sforza, B. Hewlett
	Nsua	NSU	Uganda	Sudanic, Manybetu-Efe	P. Verdu, M-F Mitune
	Mbuti	MBU	DRC	*nd*	L.L. Cavalli-Sforza, B. Hewlett
	Twa	TWA	Uganda	*nd*	G.H. Perry, L. Barreiro
Non-Pygmies	Tikar	TIK	Cameroon	Bantoïde Non-Bantu, Tikar	P. Verdu
	Nzime	NZI	Cameroon	Bantu, A80	A. Froment, P. Verdu
	Bangando	BGD	Cameroon	Adamawa-Ubangian, Bangandu, Gbaya	A. Froment
	Ngumba	NGB	Cameroon	Bantu, A80	A. Froment
	Fang	CFG	Cameroon	Bantu, A70	A. Froment
	Fang	GFG	Gabon	Bantu, A70	J-M Hombert, L. Van Der Veen
	Akele (Bongomo)	AKL	Gabon	Bantu, B20	S. Le Bomin
	Teke	TEK	Gabon	Bantu, B70	J-M Hombert, L. Van Der Veen
	Nzebi	NZE	Gabon	Bantu, B50	J-M Hombert, L. Van Der Veen
	Tsogho	TSG	Gabon	Bantu, B30	J-M Hombert, L. Van Der Veen
	Konjo	KJO	Uganda	Bantu, D30	P. Verdu, M-F Mitune
	Kiga	KIG	Uganda	*nd*	G.H. Perry, L. Barreiro

Figure and table adapted from previous publication in Verdu P. *in Hunter-Gatherers from the Congo Basin, B. Hewlett Eds* 2014.
Color legends for peopling areas can be found in Figure F1.1.a above.

1.2. Sampling individuals in Cabo Verde to reconstruct the genetic and linguistic admixture histories of the archipelago

In 2009, at the beginning of my post-doctorate research on theoretical approaches to complex admixture processes under the supervision of Pr. Noah A. Rosenberg at the University of Michigan, I had the chance to be associated to initiate a new project about the peopling history of Cabo Verde, which eventually became one of my long-term research endeavors (see **Chapter 5**). With Pr. Rosenberg, we were initially contacted by Marlyse Baptista, linguistics' professor at the University of Michigan, who hoped to get insights from genetics about the peopling history of the Cabo Verde archipelago.

Cabo Verde is a small Atlantic-ocean archipelago 500 km West of Senegal in Western Africa (**Figure F1.2.a**), discovered in the 1460's presumably un-inhabited (based on largely consensual historical records) by Antonio De Noli, an Italian navigator sailing the Atlantic coast of Africa on behalf of the Portuguese crown (Albuquerque et al. 1991; Carreira 2000). Soon after, Portuguese migrants supported by the Portuguese Crown settled its main island, Santiago, aiming to benefit from the ideal geo-strategic location of Cabo Verde in the developing commercial routes between Europe and Africa. Equally rapidly, enslaved Africans were forcibly removed from continental Africa and deported to the archipelago to participate in the building of the colony, the first European peopling-colony in Sub-Saharan Africa during this era. By the end of the 15th century, missionaries and adventurers reported the emergence of a novel language, the Kriolu language, spoken by Cabo Verdeans and built from Portuguese and a variety of African languages, de facto making it the first Atlantic Creole[3] language from the era (Quint 2000; Baptista 2002, 2015; Lang 2009).

During the following 400 years, Cabo Verde served as a major slave-trade platform between Africa, Europe and the Americas, in particular during the massive expansion of the Trans-Atlantic Slave Trade induced by the expansion of the plantation economic system in the Caribbean and the Americas, between the 1630's, and the abolition of the TAST and that of slavery in the mid-19th century (Eltis and Richardson 2015). During this Plantation Economy era, historical records unequivocally show that thousands of enslaved Africans were forcibly deported to Cabo Verde from all over the Atlantic coast of the continent, thus bringing together numerous populations and languages that contributed to shape Cabo Verdean society. In parallel to these forced African migrations, diverse European populations also migrated to Cabo Verde, attracted by novel economic opportunities (including blooming piracy during the 16th and 17th centuries), exiled due to religious persecutions in Europe, or prisoners being deported to the archipelago (Albuquerque et al. 1991). On top of this already complex historical peopling, the nine inhabited islands of the archipelago were successively settled over the course of more than 350 years since the founding of Santiago colony, due to the hot and dry Sahelian climate of the archipelago with few easily accessible water resources on most islands, shifting political and commercial opportunities, including temporary shifts of dominion among European colonial empires, intense piracy disrupting the fragile settlements, and recurring pandemic outbreaks and famines (Albuquerque et al. 1991; Patterson 1988; Carreira 2000; Soares 2011)…

[3] Notably, "Créole" is the French derivative of the Portuguese word "crioulo" (from which Kriolu stems), itself deriving from the latin "creāre" meaning in general "to create", "to feed", or "to raise", and more specifically in our context here "servant fed at home".

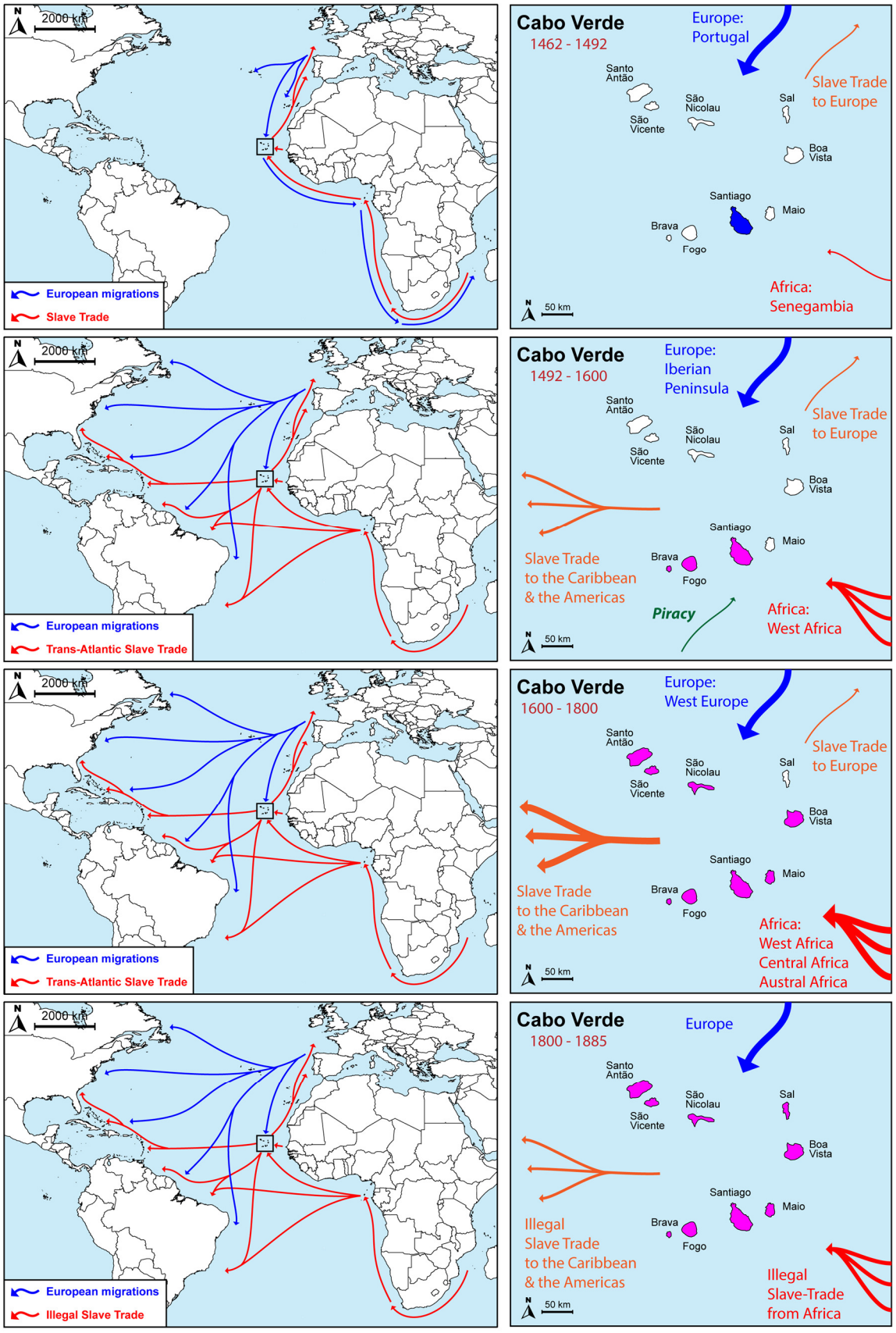


Figure F1.2.a

Schematic representation of the successive peopling of Cabo Verde islands in the TAST context since the 1460's until the abolition of the TAST and of Slavery in the 19th century, inferred from historical records only.

In this context, Pr. Baptista had been working extensively on the linguistic variation of Cabo Verdean Kriolu and was interested in better understanding the origins of the people having historically given birth to Kriolu speakers today. Genetics would thus possibly, in her views at the time, further inform which African and European languages may have been involved in the construction of Cabo Verdean Kriolu.

Indeed, while substantial historical records attested to the multiple origins of people forcibly deported or voluntarily immigrated to Cabo Verde, they very often lacked precision due to partial and biased geographical, linguistic, and anthropological knowledge of slave-traders, slave-owners, and administrators; or due to voluntary alterations and destructions of records during and after the TAST[4]. In this context, historical and comparative linguistics already had attested that European Portuguese, French, and English languages, and African Mandingo, Wolof, Temne, and Kimbundu languages, in particular, contributed to the lexical and syntactic diversity of Cabo Verdean Kriolu today (Quint 2000; Baptista 2002; Lang 2009). However, when, during the peopling history of Cabo Verde, did these different European and African languages contribute to Kriolu remained often unclear. Furthermore, whether these linguistic borrowings were accompanied by individual migrations and long-term residency (forced or voluntary) in the archipelago, or not, also remained largely unelucidated. Finally, it is important to stress that the origins of numerous linguistic traits, such as lexical roots or morphological and grammatical constructs of this contact language, remain unknown to linguists.

Therefore, any information allowing to determine which enslaved-African communities remained only temporarily in Cabo Verde before being re-deported elsewhere, and which remained for longer periods of time or even indefinitely, would be of major interest to historians and linguists, as well as every anthropologist interested in the TAST era (Albuquerque et al. 1991; Carreira 2000; Quint 2000; Baptista 2015). Similarly, which European populations remained in Cabo Verde or only transited there is also largely unknown, although it would massively benefit to our understanding of the socio-cultural and biological construction of this new population, the first to be born from the TAST (Soares 2011). Most importantly, note that these questions concern in fact all enslaved-African descendant populations on either side of the Atlantic, and studying Cabo Verde in this context would provide major information about the influence of the TAST on cultural and biological diversity of admixed populations since the very beginning of this era in the 15th century, before Columbus.

As a result, during the construction of the project with Pr. Baptista, we decided that we would try to answer the following questions in a pluri-disciplinary framework: which European and African populations contributed to the linguistic and genetic diversities observed today throughout the Cabo Verdean archipelago, and when did they do so?

In this context, the fundamental question of this chapter, *"who should I sample?"*, had, in all appearances, a much more trivial answer for this project than for the previous Central African context: we should sample genetic data from "Cabo Verdean individuals" that "speak Kriolu". I thus had, *a priori*, no major categorization challenges to design and prepare field-work for this project… Nevertheless, against my somewhat candid optimism, two major partly-overlapping categorization and interdisciplinarity issues rapidly arose about: *i)* comparing our future genetics results with the rest of the human population genetics literature about other enslaved-African descendant admixed populations, and therefore determining in part who should I sample in Cabo Verde; and *ii)* recruiting participants on the field for the purpose of joint anthropological, genetic, and linguistic sampling.

[4] These historical knowledge issues are similar for most if not all populations descending from enslaved-Africans forcibly removed from the continent during the TAST.

1.2.a. Participants' inclusion in population genetics studies related to the TAST: who did they sample?

Since the beginning, in the early 20th century, of the mathematical formalization of Mendel's genetic inheritance laws at the root of the emergence of population genetics as a discipline, extensive research has been conducted to characterize and elucidate the genetic diversity and evolutionary histories of TAST enslaved-African descendant populations in the Americas. Indeed, since the first statistical formalization of expected genetic diversity patterns in the context of admixture among previously isolated human populations by Bernstein (1931), the genetic diversity of genetically admixed enslaved-African descendant communities in the USA, known as Afro-American or African American communities today (see below), has been continuously and massively investigated to understand the major influence of admixture processes on human recent biological evolution[5].

Jumping forward in time to the beginning of the 21st century, with the major advances of DNA sequencing and genotyping technologies[6], human population genetics investigation of USA African American communities continued to attract major attention in human genetics. Furthermore, they have been enriched by similar investigations conducted on other TAST-related genetically admixed populations in the Caribbean, North, Central, and South America, as well as Africa and Europe albeit massively less frequently there (e.g. Baharian et al. 2016; Mathias et al. 2016; Ongaro et al. 2019; Micheletti et al. 2020; Fortes-Lima and Verdu 2021; **Chapter 5**).

Despite massive amounts of datasets and results available today about enslaved-African descendant populations mainly in the Americas, as we illustrated in a recent review article with my former post-doctorate student Cesar Fortes-Lima, all these investigations did not necessarily sample individuals based on comparable categorization criteria (Fortes-Lima and Verdu 2021). While medical population genetics studies included participants based on explicit medical criteria, forensic, biological anthropology, and historical demography inference studies relied instead on data collected potentially with very different criteria of inclusion of participants. It is therefore often difficult to compare the results obtained on genetic admixture patterns and the inferred admixture histories across studies, as well as with the expected results for our own new Cabo Verde project.

For instance, numerous human population genetics studies use a criterion of inclusion of participants based on self-reported or self-identified African descendancy, whether for medical genetics or other forensic or anthropological genetics questions of interest. However, it is often unclear in publications if individuals self-reported African origins based on known genealogical information or based on alleged origins stemming from community identity constructions and representations without explicit genealogical knowledge. In turn, this may result in "populations" grouping individuals with ancient admixture histories due to TAST migrations and slavery, together with individuals whose African ancestors are genealogically known in the last couple generations and thus not related to the historical TAST. This is the case in the USA for individuals not descending from enslaved-Africans deported to the USA during the TAST, but descending, for instance, from African migrants after the 1960's when such recent migrations have brought more than 10 times more Africans to the USA in 30 years than the slave-trade did during the 17th and 18th century altogether. These individuals of recent continental African origins might be identified or might self-

[5] To a point where several thousands of articles have been published on the subject, rendering, I believe (but perhaps I am wrong of course), the task of exhaustive academic bibliographic census-work improbable.

[6] Born, in particular, from the 1980's invention of Polymerase Chain Reaction and from the completion of the first assemblage of the human genome in the 2000's.

identify as "African American" although they do not share a common demographic, migratory, and social-segregation history with individuals also self-identifying in the same category but descending from enslaved-Africans during the TAST.

This is often the case in studies targeting individuals from specific communities recognized in national demographic census protocols and historically linked to enslaved-African forced migrations. The historical and social construction and recognition of these national, or census-recognized, communities is complex and diverse across countries and across time within-countries, as we illustrated in Fortes-Lima and Verdu (2021) on page R83:

> "Among 23 continental American countries, the last census questionnaires record up to four questions about self-reported indigenous, ethnic, linguistic, or racial identity (only four countries use explicitly the word 'race' or its literal translation). Apart from French Guyana where no self-categorizations are recorded, the census questionnaires' categories are based on self-reported imprecise skin-colour criteria, genealogical ancestry and cultural affinities, linguistic practices, or mixtures of these criteria. 14 countries out of 22 propose a category based on self-reported African descent, seven of which propose the word 'Black' (or its translation) as an alternative naming for the same category, and three of 14 which distinguish the 'Afro-descendant' category from the 'Black' category. Additionally, four countries among the 22 consider only a 'Black' category without explicit reference to 'Africa'. Furthermore, eight countries out of the 22 propose a separate category for 'explicitly' admixed individuals of African descent ('Mulata/o', 'Creole' or 'Mixed'), two countries have a separate 'Brown' category, and three countries gather some or all these labels under the broader 'Afro-descendant' category. In this complex categorization patchwork, population geneticists showed that self-perception of ancestry and self-reported identities overlapping national census categories are often at odds with individuals' genomic ancestries and admixture patterns. This is due to the fact that self-constructed cultural identities emerge from multiple familial and societal experiences superimposing self-perceived phenotypic features. This process is thus much more complex than simple mendelian inheritance which only reflects a very specific (genetic) part of an individual's history, furthermore only very partially translated into complex phenotypic features such as skin-colour."

For an additional example, the USA's NIH defined[7] "racial" categories (namely in 2022: "American Indian" or "Alaska Native", "Asian", "Black" or "African American", "Native Hawaiian" or "Other Pacific Islander", and "White") juxtaposed with two mutually-exclusive "ethnic not-racial" categories ("Hispanic or Latino" and "Not Hispanic or Latino"). Conversely, USA census categories include often many more "boxes" (e.g. for the 2010 US Census: 15 explicit racial category boxes and 4 explicit categories for the "ethnic not racial Hispanic or Latino" categories), which change with every national census every ten years, sometimes massively from one census to the next. Interestingly, both NIH categories and Census Bureau categories are legally required to follow the same definitions and criteria of self-identification since 1997[8],

[7] See the official NIH 2015 grant web-page which provides very interesting definitions of these categories: https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html

[8] Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. https://obamawhitehouse.archives.gov/omb/fedreg_1997standards

but the number of possible categories and sub-categories available for individuals' self-identification nevertheless may differ albeit based on the same legal requirements. In this context, an individual identified or self-identifying as African American at some point in time, might not necessarily self-identify in the same category after the change of its official definition.

The problems here suggested do not stem from the diversity of inclusion criteria and recruitment and sampling methods themselves. Ultimately, we expect different enslaved-African descendant admixed populations to have experienced different histories of admixture in different, changing, socio-historical contexts. In fact, this is what anthropological geneticists are generally interested in and what I am personally very interested in for my research. The problems instead stem from the fact that inclusion criteria, sampling methods, and questions asked by recruiters to potential participants are not often explicitly exposed and discussed in the human population genetics literature (including some of the work I participated in, I reckon). Similarly, when categorizing individuals into different communities or populations with a given label, categorization processes are not always clarified by the researchers.

As a first consequence in anthropological genetics, when using previously published datasets or discussing novel results in the light of previous work, it is often very difficult to know "who are we talking about?", and to derive expectations and reconcile results about observed admixture patterns or about the admixture histories inferred from these patterns. As a second major and critical consequence, we may work to test hypotheses on a given dataset that we would never have thought reasonable to test in a first place, if we have had access to the detailed information about "who did they sample?", instead of drawing wrong assumptions of a population's evolutionary history based on their sole "label" and sampling location provided by the original authors.

For a last, frequent, example with major theoretical and applied issues, certain studies considered individuals recruited in the general population for other purposes, such as medical genetics. After the initial population genetics description of admixture patterns in these samples, authors decided to group sub-samples of individuals based on arbitrary threshold levels of genetic admixture from a given source population rather than other self-reported information prior to genetic testing. They therefore use a genetics criterion for selecting individuals that they will then consider as a "population", whose genetic evolutionary history they will, then, infer using the very same genetic data they used to categorize them in a first place.

Beyond the risks of dialectical circularity and those, complex, of double use of the data for statistical inferences that such approaches may critically face (Devezer et al. 2021), this raises the question of how to define "a genetically admixed population" in a population genetics perspective. Indeed, I find surprisingly unclear, in both the human and non-human population genetics literatures, whether the admixed or hybrid population of interest is composed only of genetically admixed individuals[9], or whether the admixed population is a population where admixture is possible but not necessary (see **Chapter 3**). In other words, should we consider as the admixed population of interest the group of individuals possibly reproducing with one another (the classical definition of a population in population genetics) before knowing whether they are actually genetically admixed or not? Or should we consider instead as the admixed population the sub-sample of individuals within this group that are indeed genetically admixed and only them? In the former, we expect proportions of admixture from each source population to be between 0 and 1, included, while in the latter, such admixture proportions will be bounded by the thresholds chosen by the authors. In turn, allelic composition and frequencies, admixture fraction distributions, theoretically-derived expectations (see **Chapter 3**), appropriate and adapted statistical methods (see **Chapter 4**), and population

[9] Individuals who have at least two-gene-pools from different, previously isolated, populations of origin contributing to their genome

genetics inferences' results will differ between the two definitions although the same larger group of individuals was initially sampled and investigated.

1.2.b. Sampling design for population genetics purposes in Cabo Verde: who did I sample?

In this context, relatively extensive historical information about Cabo Verde peopling showed that these population-genetics and anthropological categorization issues could be adequately circumvented (but never fully overcome nonetheless) by a sampling scheme anchored in classical anthropological family questionnaires about mobility behaviors.

First, as opposed to continental Africa, the Caribbean, and the Americas, it is highly likely that Cabo Verde was not permanently settled at the time of its colonization by the Portuguese crown. Indeed, despite few recent archeological traces of possible temporary and limited peopling likely due to shipwreck survivors, no historical records attest to pre-established colonies in the archipelago at the time of its exploration by De Noli; although this question nevertheless remains investigated by historians and archaeologists (Albuquerque et al. 1991).

Second, after a long and complex history of socio-economic and socio-marital segregation between enslaved and non-enslaved communities during the colonial era, as well as between newcomers and pre-established communities, Cabo Verde does not recognize, among its national citizens, communities based on self-identified European or African or other descent[10], since its independence in 1975. It is clear that Cabo Verdean society allows the formation of communities and civil associations among individuals self-identifying to a given descent, but democratic representation and governmental redistribution, segregation, and discriminations of any form for individuals or groups of individuals based on these criteria are constitutionally prohibited[11].

Finally, historical records clearly show that admixture events between Portuguese settlers and free or enslaved Africans occurred since the very beginning of Santiago settlement in the 1470's (Albuquerque et al. 1991; Carreira 2000). Indeed, there are numerous commercial, marital and notarial records mentioning the genetically admixed origin of individuals in Cabo Verde (using Portuguese words such as "criolo", "mulato", or "mestiço" among others), as well as enslaved Africans freed by their owners willing to legally transmit to them their financial and real estate possessions. From my interpretation of historical records and investigations, one could imagine that the Cabo Verdean founding peopling had been particularly difficult due to limited access to water resources in the dry Sahelian climate of the archipelago and distance to the European homeland. Therefore, the Portuguese crown rapidly and repeatedly released laws and very liberal authorizations of commerce to incite Portuguese settlers to colonize Santiago for its geo-strategic position in the developing commercial routes between Europe and Africa. The initial settlers were thus, apparently in historical investigations, largely young Portuguese males seeking fortune or even common law prisoners deported to the archipelago, and very few Portuguese females seemed to have voluntarily migrated permanently during the first 50 years of the colony. In this context, freeing enslaved African females or genetically admixed offspring provided opportunities for single males to legally transmit their inheritance to their descendance. Conversely during the very beginning of the colony at least, males married to Portuguese females most often seemed to have sought to ultimately return to Portugal after having socially and economically succeeded in the colony, whether their wives actually came to Cabo Verde or remained

[10] file:///C:/Users/Paul/AppData/Local/Temp/relatorio_metodologico_rgph-2010.pdf - p76-79.

[11] https://www.governo.cv/governo/constituicao/ - Article 1.2

in Portugal. Finally, at the other end of colonial history, since the end of the TAST and of Slavery in the 19th century and up until today, numerous continental European or African populations migrated to Cabo Verde whether individually seeking economic opportunities or via work-force migrations organized across countries throughout the former Lusophone empire or beyond; an economic migratory process still active today.

It was therefore clear to me that genetic admixture *might have been at least possible* between populations of continental European and African origins voluntarily migrated or forcibly deported to the archipelago throughout the entire history of Cabo Verde, since its very founding settlement in the second half of the 15th century up until today. I was thus extremely interested to unravel when and how did admixture occurred in Cabo Verde giving birth to the observed genetic diversity patterns today, and to possibly link the inferred events with the variable socio-historical and demographic peopling contexts for each island in the archipelago. The numerous peopling questions I sought to answer at the time, after the deciphering of the classical questions of continental European and African origins of the Cabo Verde gene-pool of initial interest to Pr. Baptista at the root of the project, were:

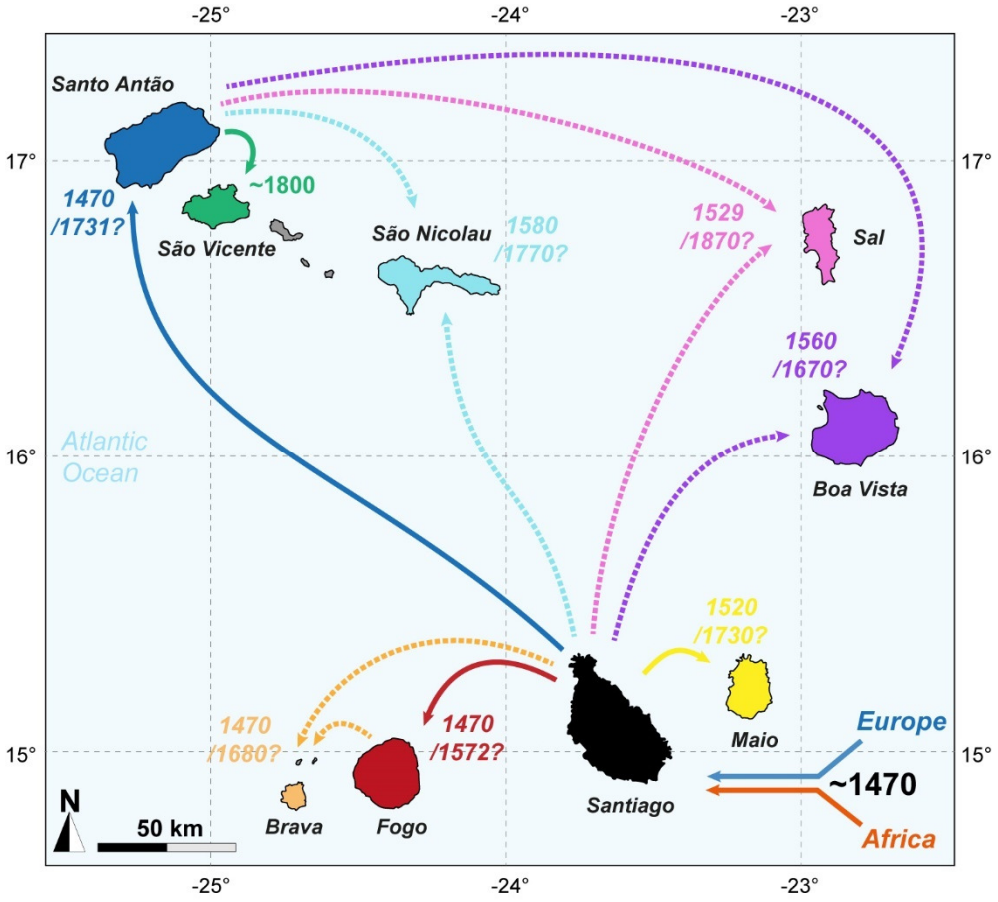


Figure F1.2.b

Schematic representation of founding and migration movements of populations among Cabo Verde islands between founding in the 1460's and the end of the 19th century inferred from historical records

i) What are the genetic diversity and genetic admixture patterns within Cabo Verde at reduced geographical scale? Are their islands or group of islands relatively more or less reproductively isolated than others due to serial founding events and isolation triggered by political changes over the 400 years of the peopling of the archipelago? Or, instead, are genetic diversity and admixture patterns evenly distributed throughout the archipelago as suggested by numerous population movements across islands throughout history? (**Figure F1.2.b**)

ii) When and where did the admixture events that built genetic diversity and admixture patterns today occur? Were there more admixture events identifiable during the massive population movements triggered by the expansion of Plantation Economy between the mid-17th and the early 19th century during the TAST? Did the abolition of the TAST reduce the opportunity for admixture events to shape extant genetic patterns? Did the abolition of slavery and the major sociological changes that followed increased the opportunity for admixture events to occur? Did recent 20th century work-related migrations result in novel, recent, admixture events? (**Figure F1.2.a**)

Altogether, I was thus interested in Cabo Verdean admixture history in general, and should thus recruit participants without distinction of self-reported descent, and whether admixture in their genealogical histories, if any at all, was recent or not, and I was not interested in foreigners having very recently migrated to Cabo Verde. I therefore decided to consider a single recruitment criterion: people should be Cabo Verdean citizens (and healthy adults only, as usual). I would then conduct detailed familial anthropology interviews with each volunteer to collect detailed information about her/his mobility history and the same information self-reported concerning her/his family. I thus collected geographical locations, main residence locations (at the time of the interview), birth-places, and the same information, when known, concerning all biological parents and grand-parents. I also collected information about the volunteer's history of mobility and the context and duration of these mobility events (for professional, familial, leisure, military, religious reasons for some frequent instances), first within Cabo Verde across islands, second outside of Cabo Verde. From previous experiences in Central Africa, I expected these interviews to require between 1 and 2 hours with each participant (depending on how much they would be talkative about their life history), compared to the 2 minutes usually required for providing a saliva sample for DNA… However, this extensive time would allow me to better intuit the reality of sampled individual's mobility behavior throughout Cabo Verde nowadays, and, most importantly, potentially qualitatively and quantitatively measure it systematically for all the sampled individuals. In turn, this would, *a posteriori*, allow me to formally compare genetics patterns in particular with interview, residence, and birth place locations and those self-reported for biological parents and grand-parents.

Finally, I had also prepared a classical interview section about individuals' mobility in marriages (residence prior and after marriage, prior and after the birth of each offspring) to be able to work on extant philopatry practices. Although I conducted systematically these parts of the interview for everyone interviewed during the project, it soon became clear that this part of my field-work design was profoundly un-adapted to Cabo Verdean society. I indeed rapidly found out that official marriages were in fact rare in Cabo Verdeans born after the 1950's (the vast majority of my participants), which I could have stumbled upon if I had conducted my bibliographical research more thoroughly, as I later discovered at the end of the extended three volumes of the General History of Cabo Verde (Albuquerque et al. 1991)... It also became clear on the field that social organization with respect to marital not-married life was complex and

that high levels of intimacy with participants would be required to access this information as politely and respectfully as required. Indeed, ethnographic semi-directed interviews on these matters require numerous questions to be asked, each of a highly personal nature, and thus much more intrusive compared to official publicly recorded marriage and divorce statements. Facing this setback, I decided that a novel sampling scheme would be required to investigate in detail philopatry and marital choices, together with a novel protocol and accompanying ethical authorizations and informed consents procedures, and decided to accept that these questions of major interest were out of direct reach of this project, and would need to be explored separately (see **Chapter 5**).

Importantly, as detailed above (section **1.2.a**), the unique inclusion criteria on Cabo Verdean citizenship here chosen, put me at odds with numerous previous studies investigating the admixture history of enslaved-African descendants in the Caribbean and the Americas, as samples were often gathered with different criteria in different historical and social contexts, and for sometimes different population genetics aims. However, I would at least have different criteria of categorization explicitly and thoroughly informed for each participant that I could use at will to analyze and re-analyze the patterns of Cabo Verdean genetic diversity. I would thus be able to provide to the population genetics community extensive and detailed categorization information and let them decide whether it could be of any use to them, as well as have the material for informed discussion on the comparability of previous results with my own, in turn highlighting the complex processes and caveats described above.

1.2.c. Genesis of a multidisciplinary sampling-design for joint genetic-linguistics investigations.

Numerous human population geneticists have been working with computational linguists (e.g. Cavalli-Sforza et al. 1992; Cavalli-Sforza 1997, 2001; Creanza et al. 2015), and compared the history of populations' genetic divergences and migrations (reconstructed from observed genetic diversity patterns), with the history of languages' divergences and migrations (reconstructed from observed linguistic variation patterns). Indeed, computational and historical linguistics are interested in part in the origins of linguistic diversity and the history of linguistic changes over time that produced observed languages; questions that strongly echo fundamental questions of human evolutionary biologists since Darwin himself in The Descent of Man (1871), and therefore triggered the mutual interest for intersecting results obtained separately.

However, numerous linguistic disciplines are interested in "Languages" as symbolic and performative meta-constructions, while only certain socio-linguistic, cognitive, and natural-languages disciplines are interested in the diversity of individual realizations of said languages in different contexts and at different times. As a result, computational and historical linguists, *lato sensu*, reconstruct the phylo-linguistic relationships among extant (or extinct) languages most often based on lexical or syntactic diversities across languages; languages which are considered as such meta-constructions and which do not explicitly and systematically consider inter-individual or within-individual linguistic variation. Here lies one of the major paradigmatic discrepancies between genetics and linguistics in pluri-disciplinary approaches trying to contrast human history of origins and migrations from either approach: the former stems from observed individual genetic variation while the latter stems from variation across languages, languages here being already complex objects built by linguistic expertise rather than inter-individual linguistic variation (a.k.a. "no one speaks the dictionary, but every one speak parts of versions of it").

Therefore, as I also witnessed personally in the field in Central Africa, linguists gather extensive data from written material and/or specific linguistic interviews conducted with a minimal number of focal informants chosen by the linguists based on specific criteria. For instance, after random discussions and a series of preliminary interviews, linguists may identify individuals as particularly expert in certain aspects of the language they speak, and privilege detailed interviews with them in particular. In all cases, they very rarely perform systematic interviews with a vast number of individuals, recruited a priori based on other, non-linguistic, criteria. They may, for a schematic instance, conduct word-list interviews with several individuals of roughly the same age and born in a given location, but they rarely conduct the exact same interview with more than 20 such individuals. They then transcribe and synthesize the information obtained using extensive linguistic expertise and methods in order to build a single word-list for a given "language". As a result, linguistic word-lists, often used by computational and comparative linguists to reconstruct the historical phylo-linguistic relationships among languages, are unique for a given language (e.g. Gray and Jordan 2000; Gray and Atkinson 2003; Pagel 2009). Data about linguistic variation across individuals, even not systematically recorded, form parts of the linguist's corpus, but are rarely provided to the community nor the direct object of their investigations. These concepts and this fundamental difference in classical historical inferences from genetic and linguistic data are schematized in **Figure F1.2.c**, below. Note, importantly, that I matured this schematic synthesis in collaboration with Valentin Thouzeau during his PhD about joint genetic and linguistic inferences, that I co-supervised (with Dr. Frédéric Austerlitz) between 2015 and 2018, thus several years after the beginning of our Cabo Verde project (see **Chapter 5**).

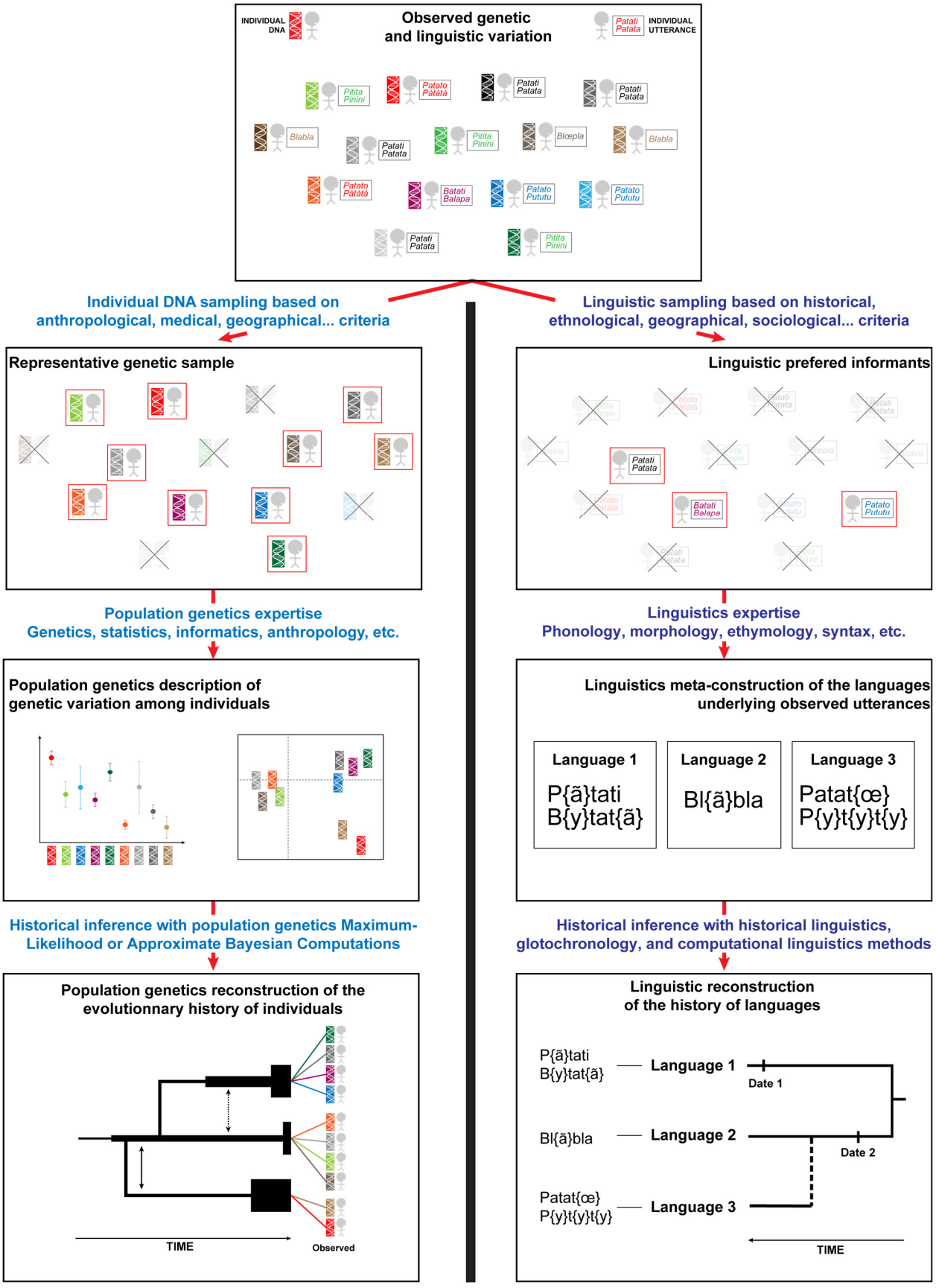


Figure F1.2.c

Schematic representation of paradigmatic and methodological differences between population genetics and computational and historical linguistics inferences aiming at reconstructing populations and languages demographic histories of origins and migrations or borrowings.

In this context, Pr. Baptista initially came to ask us about our population genetics expertise for inferring which populations of continental African and European origins contributed to the genetic landscape of Cabo Verde today. She would then use the results of these investigations to discuss, independently, her linguistics objects of study within the methodological framework of her linguistic discipline paradigms. Therefore, the project was not ignited from a multidisciplinary perspective, let alone from a fully integrated interdisciplinary approach. Nevertheless, Pr. Baptista taught Pr Rosenberg and me about the linguistic diversity of Cabo Verdean Kriolu, and we discovered, as population geneticists, that lexical, phonetic, and syntactic diversity, "substantially"[12] differentiated the manners of performing the Cabo Verdean Kriolu language across certain islands of the archipelago (Quint 2000; Baptista 2002, 2015; Lang 2009). Based on my previous experiences in Central Africa (see **section 1.1** above and **Chapter 2**), I became very interested in deploying classical genetic-linguistic comparisons as described above, but at the within language-scale, across Cabo Verdean islands, rather than across languages as most often conducted in previous studies. Following previous joint genetics-linguistics work, I thus hoped to be able not only to reconstruct the genetic peopling history of Cabo Verde, but also to be able to compare it with the linguistic history of this archipelago reconstructed separately with computational linguistics methods. This endeavor had never been conducted, to my knowledge at the time, at the within-language scale and considering a contact language such as Kriolu (see **Chapter 5**).

To do so, we needed to sample individuals' DNA across islands as explained above (**section 1.2.b**), and obtain, separately, linguistic data, such as word-lists, already compiled for the same islands. However, such linguistic information was not readily available nor homogeneously and systematically reported by linguists for the different Cabo Verdean Kriolu variants they had identified based on numerous other criteria than "just" lexical. Well, we could do it ourselves, couldn't we? And even more! Indeed, I then asked Pr. Baptista whether we could consider individual linguistic sampling: we would collect familial anthropology and saliva samples as planned above, and for the same participants, all of them, also collect linguistic data. This would allow us to conduct at least descriptive quantitative analyses of the various manners of speaking Kriolu across individuals at the same scale and among the same individuals for which we would investigate familial anthropology and genetic data.

At that point, the choice of which linguistic data and which protocol became critical. Indeed, I had in mind conducting "simple" word-lists, such as the famous Swadesh list (Swadesh 1971), for every participant, similar to what I had participated to in Central Africa but for all participants rather than for some specific individuals selected by linguistic expertise. Among the numerous classical and well-known methodological difficulties of the Swadesh protocol (see **Chapter 5**), one was outstanding for the purpose of capturing within language variation in a Kriolu language. Swadesh word-list was initially built mainly to obtain lexical roots and phonetical transcriptions of highly stable words within languages (mainly of European origins); such as words describing body parts (hands, fingers, nose…), numerals (1, 2, 3, 10, 100…), or basic action verbs (to see, to eat, to drink…). Indeed, Swadesh designed a list of lexical items recognized by linguists as being very likely shared by individuals within a language, and as little polysemic and/or ambiguous as possible. This was not a priori an ideal list for recording possible differences in manners of speaking Kriolu, as we did not expect major diversity in ways to say, for instance, the number "4" across individuals and islands… We were wrong here, as we indirectly found out during our first field-works and as Valentin Thouzeau brilliantly further demonstrated during his PhD (see **Chapter 5**)… Nevertheless, we opted at first for another strategy.

[12] Qualitatively from a linguistic perspective and expertise, not quantitatively measured.

Pr. Baptista suggested instead to shunt another classical semi-directed interview protocol sometimes used in cognitive linguistic studies: *The Pear Story* movie (Chafe 1980). We decided to show volunteer participants this 6-minutes speech-less (but not silent) movie, *The Pear Story*, which roughly showed the story of the adventures of a kid on a bicycle stealing (or at least surreptitiously borrowing…) a basket of pears from a farmer[13]. We would then only ask individuals to narrate the story they just saw, in their own way of speaking Kriolu, the Kriolu they would "speak every day at home". We would then record their entire speech without interruption and fully transcribe it for linguistic diversity analyses across participants. Note that we were interested in whichever semi-spontaneous speech hereby pronounced by participants, whether specifically narrating the movie itself or talking about something completely different (which did happen sometimes in real life, albeit relatively rarely…). Indeed, based on previous results from socio-linguistics and *N-gram* analyses about idiolectal diversity of manners of realizing a language in a given context (Labov 1972), we thought that any sufficient amount of speech would allow linguists to identify specific individual signatures of their manners of speaking Kriolu in the specific context of our interviews (that of researchers showing them a movie and recording what they had to say and how they would say it, with us as their only audience), whichever the subject of their speech.

Therefore, we set out to collect a very specific type of linguistic variation among individual realizations within a language, often referred to as "utterances" in linguistics (Croft 1991, p 107): *"a particular instance of actually-occurring language as it is pronounced, grammatically structured, and semantically interpreted in its context"*.

Obviously, while I could show the movie and record the speech, I could neither transcribe the speech, nor conduct linguistic analyses downstream, as this required linguistic expertise that I did not have as a biological anthropologist. Therefore, we needed to conduct the fieldwork together, Pr. Baptista and myself, at least, for the entire project and for collecting all the anthropological, genetic, and linguistic data for our entire corpus and for all Cabo Verdean islands. Despite the apparently difficult organizational and logistics challenges that this choice posed, it is important to say that, in fact, it increased our enthusiasm for the project: it would allow us to discover, in deep and first-hand, the practices and methods deployed in both disciplines on the field and beyond, and therefore better understand our respective challenges, intuitions, technics, and most importantly, what could be and what could not be analyzed and tested by either discipline based on the data we would collect jointly.

Therefore, we set out to collect such anthropological, genetics, and linguistic data from volunteer participants of Cabo Verdean nationality (**section 1.2.b**), and that self-reported speaking Kriolu. Indeed, in the context described above, we were interested in idiolectal variation, variation of manners of speaking Kriolu, in the given context of our project setting, whether individuals were deemed "Kriolu-specialists" or "barely speaking Kriolu" by expert linguists. While Pr. Baptista was at first not overly enthusiastic about this choice at odds with linguistic traditional approaches, she also recognized that it could not be detrimental to her work, as it would not prevent her from identifying key informants for future work among our participants. On the other hand, it was crucial for the genetics part of the project, in order to avoid spurious unexpected sampling biases, as illustrated above in the previous **sections 1.2.a** and **b.**, that might have resulted in extended investigations of the genetic admixture history of Kriolu specialists only, rather than that of Cabo Verdeans more generally.

Nevertheless, from a socio-linguistics perspective, this choice would in turn require us to record at least some basic additional information about the participants' birth-places, residences, and mobility

[13] This movie was initially produced and used by linguists in a semi-directed protocol to inform language and narration construction by participants after they have seen it in a controlled environment.

behaviors, which we already were interested in (see **section 1.2.b**), self-reported age and gender, main languages spoken or practiced regularly and their contexts of usage, as well as that of the parents and grand-parents, education duration (number of cumulated years of schooling, of higher education, and of temporary professional training), as well as employment situation (Labov 1972)… Indeed, all these factors were known by linguists to possibly influence manners of speaking a language. Note that, despite the additional interview time needed to collect this information in a controlled way (another 1 hour per individual), our goal was never to conduct a full-blown socio-linguistic investigation here. Instead, we simply aimed to inform these aspects which would help us to contrast observed idiolectal variation patterns with possible known covariables, similarly to how life-history mobility questionnaires were acquired mainly to contrast possible genetic diversity patterns.

Ultimately, we conducted these interviews for 261 Cabo Verdean participants to our study (**Chapter 5**), originating from all nine islands of Cabo Verde, and interviewed in more than 60 different locations across 7 islands (**Figure F1.2.d**). Time constraints for Pr. Baptista due to other academic and personal commitments strongly restrained the duration of each one of our fieldworks to roughly 20 days per year. The first two fieldwork we conducted together, in 2010 and 2011, empirically showed that we would need backup, in particular for advertising our project and goals, and reaching out to possible participants interested in hearing us out and, hopefully, participating. Therefore, I first started to rigorously optimize field-work organization and logistics as much as possible which is, in fact, contradictory with the very nature of field work where, usually, a lot of room has to be made for uncertainty and the unexpected. Second, we had the chance to meet and hire for four fieldworks between 2014 and 2018, Sergio Costa, one of Pr. Baptista's former student long before our project and a permanent translator (Portuguese – Kriolu - Portuguese) at the National Parliament of Cabo Verde. Sergio Costa was not only a skilled linguist, but also deeply passionate about Kriolu language and soon legitimately became a key scientific collaborator to our project, for successfully and efficiently conducting field work and in particular recruiting participants by word of mouth, for systematically and rigorously transcribing Kriolu in ALUPEC (the morpho-phonetic orthographic norm for Cabo Verdean Kriolu), as well as for discussing sampling strategies, downstream analyses and results, and participating to scientific publications as a co-author.

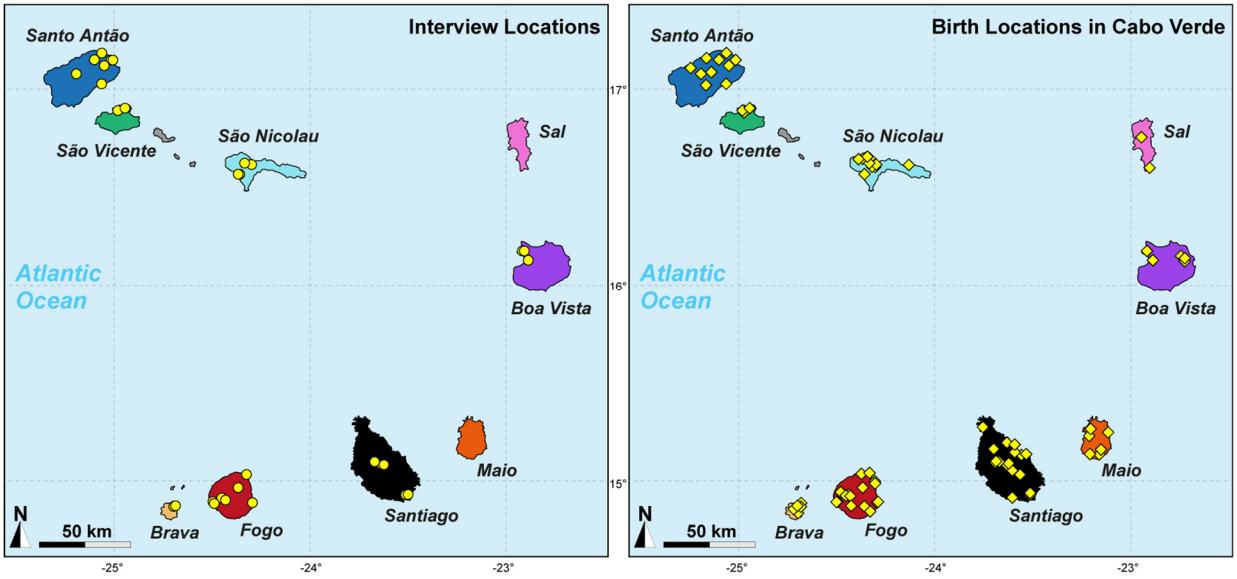


Figure F1.2.d

Sampling and interview locations for Cabo Verde participants to our project between 2010 and 2018, as well as participants' self-reported birth-places in Cabo Verde. GPS coordinates for each visited or reported location were recorded on site by myself, and complemented by geographical maps and Google Earth™ for a (minority) number of locations I did not have the chance to visit, in particular in the islands of Sal and Maio which we did not visit during our six fieldworks.

1.3. Concluding remarks to sections 1.1 and 1.2.

For concluding remarks concerning the above **sections 1.1** and **1.2**, note that the sampling schemes and categorization procedures we chose for our Central African and Cabo Verdean projects are never "ideal" nor should they be considered as "definitive references" for these projects or even beyond them; they are at the very best *ad hoc*, and I intended to demonstrate here that such issues should always be intended in an *ad hoc* manner for the questions of interest to a given project. Indeed, the main point I intended to make here, besides the scientific introduction to our research work and results synthesized in the following chapters of this work, is to show how these sampling and categorization questions are at the root of every downstream human population genetics and anthropological genetics research. In this context, the overall lack of detailed information about inclusion criteria, sampling methodologies, and categorization choices in the literature (including in some of the works I was implicated in, I reckon), from my perspective, strongly diminishes the range of other questions and discussions that the community may further ask about these datasets, when made available, and about contrasting novel results with previous publications. I intended to show here that this has not only consequences on general discussions and perspectives, but most importantly may affect the applicability of statistical testing and inference methods, similarly to whether a dataset fits the prerequisite or not of a parametric statistical test, and therefore whether it can be used or not. Indeed, equations can very often be mechanically applied to a dataset, they are just pluses and minuses after all. However, can the results obtained be interpreted as theoretically expected and intended for these equations? This may very well not be the case.

1.3.a. A practical toolbox for anthropological categorization issues in human population genetics research.

For a practical toolbox for human population geneticists about anthropological categorization issues, I suggest the following protocol aimed at informing *"who are we talking about?"* when collecting anew DNA samples in human groups or when using previously published genetic data collected by others, specifically for questions of evolutionary anthropology and population genetics analyses and inferences. These recommendations stem from my personal experiences described in part in the previous sections, backed up by theoretical and empirical classical cultural anthropology approaches about categorization issues.

1. *Population names and labels:* investigate the etymology of the name used or self-reported to designate a group of individuals and corresponding DNA samples or sequences. Is it self-reported or used by outsiders of the group? Is it an endogenous ethno-name or another type of classifier? If explicitly meaningful endogenously, what type of lexical category does this name refer to? Since when is it used to designate this group of individuals? Did the designated group of individuals change over time, incorporating or excluding individuals based on which criteria?

2. *Criteria for categorization in "self" and "others"*: what are the criteria by which individuals self-identify to a given group, different from another group geographically close? Of which nature are these criteria (e.g. linguistic, geographical, theological/religious, genealogical, ways-of-life, socio-economic, etc.)? Are these self-identification criteria shared uniformly by everyone in the group as well as recognized by outsiders? Do outsiders use other criteria for designating the target group than the criteria used endogenously? If yes, do they intersect or are they completely separated?

3. *Relationships between "self" and "other", in particular regarding possible reproductive events*: what are the marital relationships and rules between geographically-close groups? Are there social rules or recommendations for mate choices within and across groups (e.g. marital segregation or favored marriages)? Are they prevalent and normative, and what are the rules for dealing with the transgressions of these norms? This point is in fact crucial for population geneticists: it is not the social rule and endogenous representation *per se* that may influence reproduction patterns, and thus genetic diversity patterns, but their prevalence in the society and over time (see **Chapter 2** and **5**). What is the mobility of spouses in marriages and divorces (philopatry)? What is the mobility of children in their parents' marriages and divorces? Since when these rules apply?

4. *Systematic information related to individual samples*: do we have actual information about individuals' language, birth-place[14], residence location, as well as interview or sampling location? This information was self-reported by individuals or inferred by a third party? If the latter, who did it, and based on which criteria? If the former, how questions were asked to participants?

Those are, to my views, the fundamental questions that human population geneticists and anthropological geneticists should at least consider before using previously published data, as well as when collecting novel samples. It is clear that answering those questions can require a life-time of detailed ethnological or familial anthropology work. If this work has been conducted before, it is for the best for population geneticists, but it is likely not going to be the case in most cases for the specific group of DNA samples under investigation. Nevertheless, trying to identify which information is available and which is not will allow the researcher to reframe research questions to be more adequate to the investigated samples, and also tune statistical inference methods used, as well as reframe discussion points. Ultimately, it shall help evaluating the cost/benefit ratio of conducting tedious dataset merging and population genetics analyses. Finally, for human population geneticists conducting DNA sampling on the field, it is crucial to try to systematically record this information for each individual, which requires to design *ad hoc* questionnaires. Note that, such questionnaires of an ethnological nature will always benefit (I am willing to say "require") from inputs from ethnologists and cultural anthropologists. Indeed, these disciplines have extensive relevant expertise in how to ask a question, which is often not as trivial as biologists may sincerely think. Finally, making this information available to the community is crucial as I advocated in the previous sections. It nevertheless faces major ethical and deontological questions further discussed in the following **section 1.4**.

1.3.b. Miscellaneous conclusion.

In my personal experience, cultural anthropologists are very well familiarized and often conscious about anthropological categorization issues, as these often constrain and explain the cultural diversity they are interested in in a first place. In mirror, I often find human population geneticists and human geneticists surprisingly candid about the importance of these categorization issues on their own paradigms, methodological tools and approaches, as well as results' interpretations. Intuitively to my views, this may

[14] Note that population geneticists sometimes rebut birth-places information with comments like: *"people are now always born in the nearby hospital. Therefore, birth-places do not mean much in terms of genetics."* First, I think instead that it means exactly what it means about reproductive behaviors and is, as such, of compelling importance to understand the genetic patterns observed. Birth-places are evidently directly related to reproduction, while sampling locations are most of the time conjectural to the research project... Second, birth place is an easy and "clean" categorization criterion: one is only born once and people often know where exactly. It can thus be recorded in a relatively unambiguous manner for all individuals. I think that this is perhaps one of the least ambiguous question of such anthropological questionnaire.

possibly stem from the very different epistemological backgrounds experienced by the two communities during their academic training. Indeed, human population geneticists most often come from medicine, biology, ecology, mathematics, informatics and/or statistics backgrounds, rather than from cultural anthropology and ethnology. As a consequence, it seems to me that, for certain human population geneticists, knowledge can directly stem from DNA samples or sequences only, even when very limited anthropological context is available. While I likewise think that numerous fundamental and essential knowledge can indeed be acquired from these DNA samples or sequences only, I also think that knowledge of an evolutionary anthropology nature cannot be rigorously acquired if detailed sampling and categorization issues are not systematically discussed.

This is, to my views, highly similar to classical naturalist categorization issues, between taxonomic and phylogenetic classifications for instance, or to ecology, demography, and conservation biology questions. In these disciplines, *"who did we sample?"* and *"how did we sample them?"* are Material and Methods questions that, despite also facing major methodological difficulties, are very often extensively explained in the literature. I see no dialectical reasons why this should not be equally the case in human population genetics and anthropological genetics studies.

Of major importance, I am very well aware that there are, in the human population genetics literature, numerous glaring exceptions to the above critics. They will keep on inspiring me for the necessary changes I need to bring to my research practices in the future.

1.4. Ethics, deontology, laws, and scientific research administration: *"what information can I share about whom I sampled?"*

Disclaimer:
Ethics and deontology have represented and still represent a significant part of my work-time. However, they are not, per se, my objects of research nor my scientific discipline. My experience in these questions stems mainly from field sampling and data generating and sharing, and are almost exclusively empirical from my repeated interactions with various National Ethics Committees and academic Institutional Review Board, and from fieldwork experience with donors over the years in Central Africa and Cabo Verde. As a result, starting in 2014, I have been extensively working with CNRS and MNHN legal practitioners and Ethics committees to create (and validate) a novel package of CNIL (the French national commission for informatics and personal data protection) National Methodologies[15] *suited for anthropological genetics data sampling and usage, conducted in my research specifically but also tuned for almost all other anthropological genetics projects at the UMR7206 Eco-Anthropology. These methodologies and recommendations are now applicable for numerous other human sciences and biological anthropology projects conducted by academic researchers in France. Based on these previous empirical experiences, I aimed, in this section, to share with fellow researchers and students, some of my thoughts about major empirical ethical and deontological issues for anthropological genetics and human population genetics projects in the genome-wide era. This section thus strongly lacks philosophy and judicial academic rigor, in fact both disciplines outside my field of scientific expertise.*

In the previous sections, I have been trying to answer the crucial question *"who should I sample?"* that I faced in my previous projects concerning anthropological, genetic, and linguistic data sampling on the field. I have strongly advocated for the need, in scientific publications and communications, to present and discuss anthropological categories for sampled individuals, as well as details about the sampling methodologies themselves. However, this raises another difficult ethical and deontological question, especially, but not only, for researchers conducting the fieldwork itself: *"what information can I share about whom I sampled?"*.

The Declaration of Helsinki, initiated in 1964 and of which the latest of 7 revised versions dated in 2013, provides "ethical principles for medical research involving human subjects"[16]. In fact, I never conducted "medical research" in the strict sense of this declaration. Nevertheless, the ethical and deontological principles here presented readily applied to "human genetics research" in general. Furthermore, a substantial number of items in the declaration, including the most essential core points *research scientific finality and appropriate dimensionality*, *justice to access the study and focus groups*, *identifiability, personal privacy and vulnerable subjects*, *informed consents*, *risks, burdens, and benefits*, and *data sharing*, could easily be translated for the anthropological genetic purposes described in the previous sections[17]. Indeed, simply replacing the words "medical research" and "physician", by the words "anthropological research" and "anthropologist" in both the extended and synthetic versions of the declaration, provides a very useful framework and applicable guidelines for Informed Consent procurement during anthropological genetics field sampling.

---

[15] https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037187386
https://www.cnil.fr/fr/declaration/mr-001-recherches-dans-le-domaine-de-la-sante-avec-recueil-du-consentement
https://www.cnil.fr/fr/declaration/mr-003-recherches-dans-le-domaine-de-la-sante-sans-recueil-du-consentement

[16] https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

[17] In fact, virtually all essential aspects of the declaration beyond "clinical testing", "health assistance", and "invasive procedures" (this later point being easily avoided for me as I considered only non-invasive procedures for DNA sampling with saliva spits provided by donors, rather than blood samples or even buccal swabs).

In this context, over the years and probably as everyone else in the community conducting sampling in the field, I modeled my "Informed Consent" procedures to encompass all items suggested in the Declaration of Helsinki and applicable to my work. I then proposed these to my colleagues and Ethics Committees and Institutional Review Boards from the country where my labs were at the time, as well as those of the countries where the sampling was taking place; received (sometimes laboriously after substantial amount of time) comments; then corrected and resubmitted in order to eventually become approved to conduct the proposed research.

I will not dwell here on all the aspects of these ethical and deontological guidelines. I previously published a chapter in a book edited by Bonnie Lynn Hewlett, (Hewlett 2019), aiming at educating cultural and biological anthropology students about fieldwork and "what we never tell to our students about fieldwork, but we all know that we really really should". Briefly, my chapter herein, entitled *"Do you consent to participate in the research study?"* (Verdu 2019), approached the, I believe, massive scientific and methodological benefits of preparing anthropological genetics fieldwork in the light of the Declaration of Helsinki's Informed Consent requirements. I also extensively exemplify important pitfalls, and their possible consequences, due to the discrepancy between ethically approved Informed Consent procedures and the reality I have faced in some of my Central African fieldwork experiences. Specifically, I developed this chapter on pitfall examples concerning the *format of Informed Consent* procurement on the field, the complexity of establishing *"voluntariness" of volunteers*, and troubled *benefits* and *compensations*.

Here, I would like to elaborate on the specific issues of data-sharing that arise between scientific reproducibility and Open Access, on one side, and the respect of participants' privacy and personal data, on the other.

As developed in the previous sections, I think that it is critical for human population geneticists and anthropological geneticists to provide detailed categorization criteria and associated anthropological information gathered on the field, as well as inclusion and sampling processes, aiming at improving scientific reproducibility and discussions. While numerous aspects of fieldwork processes and data can readily be provided to the community in publications, some information are harder to share due, first, to their format unfit for articles (genetic data, individual anthropological information tables…), and, second, to the fact that they can represent identifiable privacy-protected information about participants, or both. Furthermore, some of the data may be part of intellectual property and ownership policies of funding agencies and institutions. In this context, note that such ethical and legal constraints may vary between the countries where sampling is conducted (the first and foremost law that applies in our case), and countries hosting academic research laboratories where data treatment and analyses are conducted, hence further complexifying the matter.

1.4.a. Human individual genetic data is "identifiable sensitive data".

Human genetic data is uniformly considered as "identifiable sensitive data" in all countries where my previous research projects have been conducted, whether for sampling or for downstream analyses. Indeed, a human DNA sample is very largely uniquely identifying his/her provider and may further carry information that could be exploited for forensic and judicial purposes, for fundamental or medical and pharmaceutical research purposes, or for commercial recreational purposes more recently. This is why there are, to my knowledge, no country worldwide that has not already codified and legislated about the conditions of sampling and usage of human genetic data in these three contexts (judicial, scientific research, commercial). This is also why "data usage and sharing" are an important section of Informed Consent

procurement recommended by the Declaration of Helsinki, and, from my experience, one of the most discussed sections of this procedure with Ethical committees, IRBs, and donors alike.

While some jurisdictions may allow that individual donors may specifically agree and allow by contract that researchers can release his/her personal genetic data freely and openly (nevertheless without providing their name associated with the study and a specific DNA sample or sequence), this is never the case in the French bio-ethics and research conduct and deontology laws under which I operate (see Note 15 above). I can thus never openly release human genetic data such as, for example, the 1000 Genome Project where data can be downloaded freely by anyone with a serious internet connection and a suitable hard drive. Furthermore, France also still completely prohibits commercial or recreational DNA sampling and testing, both for consumers and private firms, aiming at protecting the ill-usage of sensitive personal data for commercial or a variety of judiciary purposes (see below).

However, for ensuring scientific reproducibility of my work, and, separately, for sharing this genetic data (acquired within a public academic research framework and funding) with the scientific community in order to fuel novel projects and novel knowledge beyond my specific projects, other laws and provisional procedures apply. In all cases, I can only legally share genetic data with legally responsible and identified scientific research institutions and, namely, identified researchers, and if the proposed research, its means, and its proposed Data Management Plan fit what was agreed by donors upon procurement of the Informed Consent on the field. In agreement with the Declaration of Helsinki, no financial or commercial usage and patents can ever be conducted using the samples and data I collected in my research projects. Thus, other re-usage of the data for other research finalities, means, and DMPs can only be considered by re-obtaining a revised Informed Consent from the donors, including for my projects.

Based on these premises, often shared by academic researchers worldwide in my field, data is often "made available to interested researchers in agreement with Informed Consent and Ethical requirements, upon request to the corresponding author", as also extensively explained in Informed Consent procedures to potential donors. More recently, the European Union created and maintains a data repository accessible online, the European Genome-Phenome Archive (eGA[18]) which complies to the above French laws and allows me to deposit data and control their access by other interested researchers on a case-by-case basis, while providing me with secured storage and (to a small but significant extent) the administrative support needed for this crucial task.

This apparent classical (in the population genetics community) solution to a fundamental problem due to the nature of human genetic data nevertheless becomes, to my views, more challenging every day. This is due to massive technical and methodological advances in the field over the past 20 years, to the increasing diversity of complex tasks combined in any project which requires more researchers and labs involved (and therefore more shared data), and to recent societal changes introduced in part by commercial/recreational[19] broad-audience DNA testing in certain countries. Finally, the massive increase in different types and

[18] https://ega-archive.org/

[19] "Recreational" DNA-testing was the term sometimes used even by the international private corporations proposing them to customers, at least when their businesses started roughly 10 years ago, for technologies providing data and analyses at the genome-wide scale as by 23andMe™, ancestry.com™, MyHeritage™, or ScotlandDNA™ for instances. "Commercial" or "Customer" DNA testing has since then been preferred by these actors and by consumers and genealogical companies lobbying for the legalization of such practices in countries where it was, or still is, forbidden (such as France). This re-branding was deemed necessary by these actors, as they considered that the word "recreational" could be used derogatorily by opponents, as well as conveyed a sense of frivolity that ill fitted the needs and interests of certain categories of consumers resorting to these private DNA testing (namely people in deep, sometimes medically pathological, search of genealogical roots or identity such as, among others, individuals with unknown biological parenthood or individuals with a complex uninformed genealogical history due to the trauma of the Trans-Atlantic Slave Trade).

volumes of genetic data generated on DNA samples in the past 20 years, complexifies the already difficult question of legal transmission of responsibility of the data when the researcher who originally collected and/or generated them retires, quits academic research, or dies.

1.4.b. Updating Informed Consents for novel population-genetics research projects

Indeed, technical and methodological advances in population genetics since the first "complete" human DNA sequences have largely increased, and continue to increase every day, the ranges of possible applications beyond those explicitly planned for the initial research project at the root of the sampling and initial data generation. It is therefore harder and harder to evaluate whether the newly proposed research still fits the finality and scientific objectives on which relied the original Informed Consent procurement, in turn justifying to provide the data to the requester, or not. As a logical result, ethics committees and IRBs are legitimately more and more attentive to the finality of the proposed research project: "studying the genetic evolution of human populations" is no longer a plausible general finality accepted by these institutions; one has to be more precise and, therefore, constrain the possible research topics and methods used to achieve them, in turn de facto constraining the field of possible public academic scientific research that could be conducted without re-procuring Informed Consents.

While 10 years ago, I had little requests from other teams to access the data and I thus mostly re-procured Informed Consent for new applications largely outside the scope of the initial study for my projects only, it is now frequent that I have to refuse sharing the data for reasons of incompatibility with initial Informed Consents. Even for my own projects, re-contacting previous volunteer participants is often impossible (e.g. Central African populations that I worked with still rarely have electricity, let alone private phones nor internet nor email access) without organizing a novel fieldwork; and, even then, some previous participants may often be hard to find again several years after their first participation, as they might have moved or changed names since. In turn, these legitimate and legal ethical constraints have also been accompanied by an increasing amount of bureaucratic and administrative hurdles, very often at complete odds with the reality of fieldwork sampling (Verdu 2019). Altogether, these developments may impair the opportunities and enthusiasm of researchers willing to initiate novel fieldwork or procure updated Informed Consents from previous donors. These tasks have very often become simply too big, too complex, and with too little legal and administrative support for considering launching new projects, in particular for small-scale projects requiring relatively modest means for limited amounts of time.

1.4.c. When genetic data escape researchers for scientific necessities

The increase in the number of researchers (*lato sensu*, including students, engineers, technicians, and faculty) from different institutions involved in a given project highly complexified the traceability of data and that of scientific productions (publications, conferences, thesis and memoires) generated using them. The increase itself in the number of interesting scientific methods to try on the data is perfectly expected and is simply due to the amount of work and seminal discoveries of our giant predecessors in the field, combined with the increasing number of students they trained in societies where access to advanced academic training has been strongly democratized in the last 70 years roughly[20].

[20] This exponential phenomenon is by no means unique to human population genetics. It is happening in all active scientific fields of research.

As a result, specialization has increased for complex tasks that require expertise that can desirably be brought to the project by novel collaborations. In turn, the need for data sharing is greater than it used to be. The number of researchers putting their hands on sensitive genetic and anthropological data within a project increases, the number of students, destined to move elsewhere possibly with said data stored somewhere on their portable hard drives, increases, and the number of other labs in which data are stored and used, increases. It has therefore become virtually impossible for a single researcher responsible of Informed Consents from donors, and thus downstream data usage, to keep track of all possible unlawful or ill-advised data sharing and usage that may happen within only five years of the original Informed Consent procurement, let alone on the longer term.

While laws, research institutions, and institutional databases (such as eGA), significantly support and protect individual researchers from deliberate abuses or simple inadvertences, the consequences of data escaping researchers remains ethically and even psychologically extremely heavy to bear for any researcher involved in primary field-sampling. Indeed, field sampling relies almost exclusively on the trust, respect, and honesty given by donors to researchers. I, as others researchers who shared their thoughts on the matter with me, deeply feel the need to honor these. Witnessing even the sheer possibility that data may escape us and be ill-used with respect to what was consented by donors is thus, literally, heart-breaking.

1.4.d. Returning individual genetic data to their biological owners

The ethical and deontological necessity to make the genetic data available to their donors is far from new in the medical-genetics community, for medical-ethics reasons, transparency, and respect of the patient. It is technically relatively easy to provide patients with the results of their genetic testing for a small number of short targeted genetic regions or mutations: such result fit on one or two sheets of paper, and patients are usually in regular contact with their physician hence facilitating communication. These two technical aspects are, in fact, crucial for feasible "genetic data return to their biological owner", and are often impossible in the case of anthropological genetics fundamental research projects[21], in particular in the genomic era.

Indeed, while I could provide sheets of paper with an individual's genotyping results for 50 microsatellites 20 years ago, it is today impossible to provide individuals with a paper version of their entire genome. After a rapid computer-based simulation using Microsoft Office Word™, I calculated that a single almost-complete 6 billion bases (A, T, G, or C) diploid genome written in Times New Roman font size 10 on default-margins A4 format would span 1,139,817 one-sided pages, or an A4 paper-roll long of 338,5km, either format requiring roughly 2,590 trees each 12m high and 20cm wide for producing the paper[22]… This is not possible logistically and ecologically ill-advised. Of course, I could provide individuals with a portable hard-drive or high-capacity flash drive with said sequence in a compressed format, also disregarding the ecological cost of doing that for several hundreds of individuals often involved in a single project. However, I would then also ethically need to provide individuals with the computer means to actually access the data, train them to use these devices, and would further need to regularly provide them

[21] Note that finding the individuals to give them back their genetic results is also far from trivial in numerous communities with little to no regular access to medical services and facilities, whether in developing countries or in developed countries.

[22] Despite what is here an attempt at humor, in my experience, broad audience public as well as legal practitioners involved in ethic committees rarely grasp the size of a genome and of genomic data in general, hence producing fundamentally unachievable procedures despite legitimate ethical concerns and requests. In practice, I found that this empirical exemplification helps them intuit what is cognitively really not that trivial to understand, even for specialists.

with updated versions of the hardware and software against the rapid obsolescence of such material. Ultimately, and to be thorough, as I worked essentially in Central and Western African locations with more than difficult access to electricity, I would somehow also need to ensure access to stable sources of electricity for donors receiving such numerical version of their genome.

Fortunately, beyond these practical impossibilities, in my experience, participants were never interested in their raw genetic data. They were mostly interested in copies of the anthropological interviews I conducted with them to "keep a material archive for their offspring", and in the knowledge that I would, with my (to them) magical scientific method, extract from their genetic data itself. Therefore, and in agreement with ethical committees providing research authorizations, I simply photocopied the raw anthropological interviews conducted on the field to give them back to participants, and prepared synthetic non-specialized versions of my published scientific findings that I presented to the communities in person during the following visit (see **Figure F1.4**), or for specific "return of the results" missions, and via postal mail on a field I visited only once in Uganda[23]. Also note that obtaining the funding for such return to participants on the field, as it requires similar logistics and associated costs as the original field sampling in the case of my low intensity fieldwork methodologies (**sections 1.1** and **1.2**), has massively improved, albeit still not being trivial, over my more than 15 years practice. Indeed, while funding agencies accepted but rarely required return to participants work-packages in research-proposals, and even sometimes proposed reducing the overall budget by cutting on these specific costs, they now strongly require that such return to participants plan be readily and thoroughly prepared and budgeted in the original proposals.

However, in my most recent fieldworks, things have changed at least for some people in the general public I try to recruit for my research projects, namely the younger and/or those with frequent personal access to internet. Due to the recent strong development of commercial DNA testing aimed at the general public in the last five years, and massive advertisement campaigns for these products on internet and TV, it is now not uncommon that some participants to my research studies ask me for their complete DNA data. They explained to me that they thus hoped to obtain their personal DNA testing results similar to what these commercial products propose.

In such cases, I would clarify an important point, in fact a complex population-genetics methodological point, already explicated during Informed Consent procurement before sampling, but apparently not sufficiently understood by these individuals (and I can't and do not blame them…). The aims of my projects are to reconstruct past histories of genetic migrations and origins for groups of individuals rather than reconstructing individually-centered specific genealogies. I thus work with population genetics paradigms and methods that apply to groups of individual genetic data rather than establish individual diagnostics similar to what is provided by medical genetics or these commercial tests (those at least who perform a decent job, which is very often not the case…). Therefore, the analyses I am conducting are fundamentally (mathematically) not the same than those conducted for medical genetics or these commercial tests, and the type of genetic data that I generate may also differ. In other, statistical, words, I use inference statistical approaches rather than diagnostic and/or predictive statistical approaches.

[23] Note that the Covid19 pandemia impaired me to conduct the final return mission to Cabo Verde, initially planned in the summer 2020 and which I will organize, hopefully in 2023. In parallel, we also think about conducting an online return event, which in all cases will not be sufficient as perhaps 80% of participants to this study do not have computer or internet access.

Figure F1.4
Photographs (courtesy of Evelyne Heyer) of Paul Verdu during a return-to-participants fieldwork in Bezan communities of Ngombe in the Tikar Country in Central Cammeroon, in 2011. Individuals provided their consent for the non-commercial diffusion of their image for educational and pedagogical purposes only.

In my last fieldwork in 2018, when this happened a couple of times, I proposed to participants to provide them access to their genetic data upon request to me and via secured internet servers (and reminding them that they would need sufficient secured disk space and internet bandwidth to download and store said personal data, disclaiming responsibility for data loss after transfer), but cannot provide them with the results these companies claim they can provide as I do not do this. The answer has yet always been that they understood well the difference, that participating in my project was very interesting nonetheless, and that they retracted their request without a problem as, really, said raw data was of no interest to them: participating for the pleasure of participating and eventually hearing about the outcome of my research was completely sufficient for them to accept to participate.

The specialized informed reader may see how my answer to them is becoming more and more ambiguous: while it is still true that numerous population genetics inference methods rely on allelic and haplotypic frequencies only, thus not directly using individually inferred genealogies nor using individual-centered diagnostic or predictive approaches, several very interesting new methods now operate directly on individual genetic landscape and patterns rather than just population frequencies. Indeed, for one glaring example, certain local-ancestry approaches map genomic segments within individuals, and the genetic history of a population is, in fact in these specific methods, the marginal product of all the individual genetic

histories reconstructed for each individual separately; as opposed, again, to more classic inferences where the specific genetic history of a given individual is not inferred and could only be assessed as a statistical possible realization of the population-wide inferred history.

In this context, there are increasing societal and political needs for improving the diffusion of the scientific method and culture to non-specialized audiences as well as reshaping scientific projects and initiatives to be more intricated socially. As a result, ethics committees and commissions protecting personal sensitive data now expect Data Management Plans to explicitly facilitate access to personal genetic data for their original biological owners. They still hear that in most of my cases, this is not logistically and practically feasible as explained above, including in terms of the lack of the truly massive informatic investments and associated human resources that it would require from public academic institutions and labs for such ambitious task. But for how long ethics institutions and volunteer participants will keep on accepting our still legitimate derogations? I expect not long, which in turn readily clearly poses a major problem that most fundamental academic research laboratories, equivalent to the one I work in now, are not even close to be able to overcome.

And of course, the problem of traceability and protection of sensitive data for researchers exposed above would then literally explode. I have been very frequently solicited in the past five years by TV, radio, internet, and paper journalists, as well as by various civil-society associations for the defense of minorities or of migrant communities, of adopted children, or genealogical societies, as well as by learned societies, requiring my expertise about various aspects of genetic commercial testing ([24]; see curriculum vitae for an extended list). I systematically took these opportunities to try to educate the audience about what is genetic data, what is the population genetics methods and dialectics deployed for reaching results, and what it can or cannot tell. In all cases, I hereby tried to explain why personal genetic data are universally considered as sensitive data that are always protected by specific laws (more or less restrictive in reality, but still); and that every person should at least be careful and understand why sharing openly one's genetic data, or giving them out to a private foreign corporation, is never trivial nor free of possible major backlash, individually or societally[25]. This, in turn, also explains why, contrarily to these commercial enterprise, academic research ethically approved protocols for collecting and using genetic data have to deploy numerous procedures and heavy protocols meant in part to ensure privacy and protect donors from these risks.

This last comment has another perhaps imaginary but nevertheless possibly dreadful consequence: how can public research, aimed in general at improving the "common good", hope to justify public-money expenditures for increasingly complex and time consuming (and thus expensive) procurement of ethical research authorizations and Informed Consent and high quality genetic data from volunteer participants, when certain private corporations obtain such data much more easily and cheaply (as a matter of fact, people pay them for this) with much lighter ethical and legal constraints?

---

[24] https://www.francetvinfo.fr/sante/biologie-genetique/tests-genetiques/c-est-important-il-me-manque-une-partie-de-moi-interdits-et-pourtant-largement-pratiques-faut-il-legaliser-les-tests-adn-genealogiques-en-france_3986789.html

[25] Note that the dystopian anticipation movie *Gattaca*, by Andrew Nichols and released in 1998, was not only extremely accurate on numerous human genetics methodological aspects, but also profoundly accurately illustrated how a eugenic society could expand on the availability of massive genetic data. As a consequence, I recommend to use it for general public debates and for population genetics university classes. My experiences are that it works great both to explain the scientific method involved in human genetics and ethical and deontological issues in the field and, beyond that, society in general.

1.4.e. Who gets to use the genetic data once I am gone?

Finally concerning specifically human genetic data in this section, we are now the first generation of population geneticists for whom the archival and continuity for future re-usage of massive human genomic data becomes a crucial question. Indeed, what happens to such sensitive personal data and the evaluation of their future sharing and re-usage when the researcher who received the original Informed Consent from donors retires, quits, or dies? To my knowledge, and after having interviewed legal practitioners at the CNRS and the MNHN on this matter, this is at best a grey zone in French legal regulations, and completely left out of the Declaration of Helsinki guidelines. While there are numerous laws and regulations concerning the transmission and re-usage of personal medical data, and while there are provisions about the transmission and archival of research data including biological samples in general within public academic research institutions such as the CNRS and the MNHN, they do not explicitly indicate how human genetic sensitive and personal data should be handled, and most importantly whether they can be re-used in the absence of the researcher responsible for evaluating that data-access requests appropriately suit the original Informed Consents.

I, unfortunately, recently experienced first-hand this issue; an unexpected example of what will eventually always occur and may further illustrate how unprepared I and our institutions are. Together with a fellow researcher, Dr. Trevor Pemberton, launching his lab in Manitoba State University (Canada), I started in 2014 an ambitious project that required extensive anthropometrical and DNA sampling in 400 individuals from 4 "Pygmy" and neighboring "non-Pygmy" populations (see **section 1.1**) from Central and South-East Cameroon. After two years of obtaining an NSRC Canadian grant, building the research protocol and overcoming the ethical, deontological, and administrative hurdles, we were approved by Canadian IRBs and ethic committees, as well as by the IRD (French public research Institute for Research and Development) in Cameroon who, at the time, had a research agreement with Cameroonian government to authorize research involving the human subject that applied to our specific project. We therefore set out in the spring of 2016 to conduct two months of intense fieldwork in villages I had conducted numerous researches in the past, together with our CNRS colleague Dr. Fernando Ramirez-Rossi, our IRD colleague Dr. Alain Froment, and Canadian PhD student Dr. Alexandra Blant. Of major importance, I was never to keep, even temporarily, any original DNA sample and anthropological data in my lab in France: all material was to be stored and not shared, at least before the original publication, in Canada, and was directly sent from Cameroon to Canada, without ever entering in my lab in France.

Everything went well, and DNA samples for this unique database (by far the first of its kind for Central African populations) were extracted in the Canadian lab and ready to be genotyped and whole-genome sequenced. At that point in mid-2018, Dr. Pemberton unexpectedly and brutally completely retired from academic research for reasons I will not explain here. The department head of his lab at the time, Pr. Barbara Triggs-Raine, was left, from one day to the next, with an inextricable situation regarding, in particular, samples and data.

After a very long administrative and legal procedure, in 2019, the newly appointed dean of the Canadian university requested that we destroy the samples, a decision that apparently was validated by the university's ethics committee. Fortunately, Pr. Triggs-Raine opposed this decision and contacted me in an emergency: did we have legal provisions in the original IRBs that would prevent such thing from happening and was I able to legally be given the care of these samples? Fortunately, following the Declaration of Helsinki guidelines and downstream discussion in various national bio-ethics committees, we included in our Informed Consent and ethical applications the general provision that human DNA samples are sensitive

human biological samples that cannot be destroyed without the explicit consent or request of their original donors, unless they pose a biohazard imminent threat (which was not the case for us). Such provision was explicated to participants and agreed upon in my Informed Consent procedures under specific sections about "rights and procedures for retracting from the project".

Furthermore, the Canadian IRBs had asked us to include a non-standard section about what we should do if the Canadian PI "died" (sic), which I had in all honesty totally forgotten about at that point. Indeed, reparsing the documents in my possession, we had simply made a provision that all material and responsibilities should be transferred to me and my lab, which had been accepted by ethics committees.

Together with Pr. Triggs-Raine, we thus launched an official appeal to the university's ethics committee, pointing towards these two provisions that they had themselves endorsed initially. They then requested proof that I was involved first-hand in the original Informed Consents procurements to donors, even if my name and all contact information were explicitly included in Informed Consent forms provided as a paper archive to each participant in Cameroon. I have thus had to justify that I was in fact the one who procured Informed Consents from all participants during a fieldwork I had largely organized and conducted, with documents and photographs and video recordings[26]. Fortunately, our appeal was accepted, and we could launch the Material Transfer Agreement for permanent transfer of all material left in the freezers and drawers of my collaborator. This took another year and half, after extensive legal discussions between the Canadian university and the CNRS, mediated by Pr. Triggs-Raine and myself. Finally, during the fall of 2020, I received officially the material and right to store it as agreed upon between the Canadian university and the CNRS.

That would sound like the end of a long and tedious story, but no. As far as I have the right to keep and care for all this material, I do not have the right to exploit them. Indeed, the Cameroonian research authorization provided to my colleague could not be transferred to me in his absence. As of today, I am still trying to obtain this transfer from Cameroonian authorities. In fact, after a moving of the archives of the ethics committee, they seem to have lost our original file, and thus cannot legally amend it. I am now in the process of re-applying, which is not trivial as I need retro-activity for 2016, when the sampling happened. This is far from standard for ethics approvals to conduct research in general, and obviously not a priority for Cameroonian research ethics committees in particular…

This was an unexpected and accidental experience, but what will happen to, for instance, my eGA deposits and the Data Access Committees I pilot, if I quit or die, or when I retire? I do not have the answer to this question... yet!

[26] In fact, I had established a video-recorded protocol for Informed Consent procurement during my PhD, aiming at legally overcoming the multifactorial problem of having illiterate participants sign a lengthy classical Informed Consent form (Verdu 2019). During my post doctorate, after extensive discussion and audits, I got this protocol to be included in IRBs standard operating procedures of the University of Michigan and, later, I did the same with the CNIL in France. Video-recorded Informed Consents were now acceptable in my research cases following a precise protocol, and could only be used for legal and ethics committees' procedures and requirements. I could not use them myself for any other purposes, even research (visual anthropology for instance, as I have had to reject requests of data access from researchers from this field in several occasions), although I have to ensure their storage and availability.

1.4.f. What about anthropological and linguistic data?

The previous sections concerned specifically archival and sharing of human genetic data, as they are intrinsically identifying data of a highly sensitive nature due to their vast judicial, scientific research, and commercial potential. But, as I introduced in **sections 1.1 to 1.3**, what about the individual anthropological (cultural or biological) and linguistic data I also collected during my fieldwork? Indeed, I strongly advocated for releasing to the scientific community numerous information about individual's self-identification, residence and birth locations, marital rules, and numerous other such data useful for individual categorization in groups and populations. Furthermore, in my Cabo Verde project in particular, I also investigated individual speech patterns to evaluate jointly the genetic and linguistic history having given birth to the observed biological and cultural patterns of diversity throughout Cabo Verde (see **section 1.2**); and obviously advocated for releasing such data together with my publications in a similar way to genetic data, for deontological purposes regarding reproducibility of scientific results as well as for promoting future research endeavors.

i) Linguistic recordings.

After discussion with legal practitioners, we agreed to recognize that they could be used in existing legal forensics procedures similar to existing cases when voice recordings are used as evidence in a trial. My public research linguistic recordings were thus *de facto* identifying sensitive data... Furthermore, I had legal practitioners realize that such recordings could also be of interest, beyond judicial and police institutions, to scientific research laboratories and private companies alike working with natural languages and automated voice recognition, a very active public and private Research and Development field today. Altogether, we agreed with legal practitioners that it would be easier to build a methodology about this data using as a template what we already had for human genetic samples and data; despite the lower risk-potential of such data as they had, to our views, no obvious medical and pharmaceutical nor direct individual-targeted commercial interests that we could imagine, yet… I thus included them under the exact same Informed Consent procedures as the genetic samples and data, nevertheless precisely describing that the nature and risks associated with either type of data strongly differed. This process has thus given birth to a specific CNIL methodology usable for any research protocol archiving and analyzing individual linguistic recordings, and my Cabo Verde protocol is since cited as a case-example of applicability for this methodology by the CNRS ethics committee for Data Management Plans in particular towards interested researchers in Human Social Sciences[27].

ii) Anthropological interviews, self-reported data.

The French law, I reckon, is very specific and restrictive on personal data sharing as compared to the Anglo-Saxon laws I have been in contact with both as a post doctorate fellow and, later, for sharing data with USA fellow collaborators for my projects. In the French law and CNIL recommendations, any information or batch of information allowing to identify an individual is subject, by default, to the right to be protected, even if not "sensitive" as defined by them similarly as for genetic and linguistic data. In this context, while individual birth-places cannot on their own allow to identify a given participant to my study, the vector of age, sex, residence place, birth-place, and parental and grand parental birth and residence-

[27] https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf ; which was regularly enthusiastically received by some members of this vast community, but not always, in particular, by those who thought, until then, that their research topics and methods were absolutely not concerned by such ethical and deontological matters…

places can be used to do so in principle. Furthermore, in particular in Central Africa but not only, while the risk associated with such hypothetical identification is probably low, it is not reasonable to completely exclude that the political powers in place locally now or in the future will not use this information to act one way or another specifically towards individuals that participated to my research and/or the communities they belong to; physical retaliation being one of the worst-case scenarios.

Therefore, with the legal practitioners that assisted me in these matters, we decided to create a specific category for this part of my data in general, and to model it upon the existing methodologies concerning sensitive personal data protection about religious and political opinions, and marital and sexual orientations and opinions. As a matter of fact, I also had to explain to them why human population genetics research projects often gathered religious and marital data, even if only indirectly via their familial anthropology interviews: these are of major interest to certain human population geneticists (including me) as they are often key factors determining who can reproduce with who locally, and further may be part of the social frameworks determining marriage rules and norms and mate-choice recommendations, hence of obvious possible influence on genetic diversity distribution within and among communities. As a consequence, I cannot, in fact, freely and openly share such individual data as I advocated for previously in favor of Open Science and reproducibility, due, here again, to personal data protection and respect of privacy.

iii) Transmitting population inclusion and categorization criteria to future generations

This poses also the same problem of transmission of this personal information about categorization criteria than for the DNA sample and genetic data itself as discussed above in **section 1.4.e**. As a last example of this issue, and to further exemplify here deontological problems of transmission of information about the categorization issues discussed in the previous **sections 1.1 to 1.3**, let us consider in some details the Human Genome Diversity Panel – Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) DNA panel[28]. This panel has been used over the past 20 years to massively increase our understanding of human genetic diversity and evolutionary history. Indeed, more than 200 scientific publications directly investigated this panel, and, furthermore, samples from the panel were largely provided to other international human genetics initiatives of fundamental importance such as the HapMap Project or the 1000 Genomes Project among many others.

The HGDP-CEPH panel itself, provides 1048 individual DNA samples (and associated genotyped markers using various technologies and whole-genome sequences) from 58 worldwide populations (**Table T1.4**). Most importantly, very little information about the sampling and recording of the categorization-criteria for individuals are readily provided to researchers willing to use this panel, and, in fact, very little information about the recruitment of DNA donors and sampling methodologies on the field are made available or explicitly referenced. Nevertheless, the HGDP-CEPH panel readily provides population-information tables for which the population labels leave very little doubts about the fact that panel-inclusion criteria of DNA donors and population categorization-criteria were highly variable from one population to the other, some possibly involving self-report "populations" but some others very evidently not so.

Indeed (**Table T1.4**, first column "Population"), six population labels indicate UN-recognized states' nationalities ("Japanese", "French", "Cambodian", "Colombian", "Palestinian"[29], "Russian"). One population label indicates a vague geographic region within a state not evidently linked to any other ethno-geographical or ethno-linguistic group: the "North Italian" population, differentiated from the "Tuscan"

[28] https://cephb.fr/hgdp_panel.php

[29] Note that Palestine was not yet recognized as a state by the UN when the HGDP-CEPH panel was made available to the scientific community.

and "Sardinian" which are other Italian *populations* labeled with a name referring to official administrative regions of Italy, and possibly also referring to ethnic and ethnolinguistic regional groups. Seven populations' labels start with the word "Bantu", which is a linguistic family containing more than 250 different languages from all over Sub-Saharan Africa (Guthrie 1958), associated with three different geographical codes ("North East", "South East" and "South West"), and six ethnic or ethnolinguistic collated labels ("Pedi", "Sotho", "Herero", "Tswana", "Ovambo", "Zulu") for the South African samples only, while the "Banthu North East" from Kenya is not further named in more details. Of particular interest to me (see **section 1.1**), note that the "Biaka Pygmy" population are also Bantu speakers but are not labeled as such. Indeed, only two African populations are named according to an ethno-name associated with the word "Pygmy" ("Biaka" and "Mbuti"), which, let me remind here, is an exogenous (European) extremely complex historical category with an ancient Greek etymology.

Finally, all other populations are labeled with a name indicating what can only be assumed to be regional (geography), ethnic, ethno-geographic, and/or ethnolinguistic groups, albeit no definition of the "ethnicity" criteria are provided here, with associated nationality in the sole case of the "French Basque". Note that, rather surprisingly, the "Population table" is associated with another "Individual population information table" from the same HGDP-CEPH reference. The latter indicates that "Colombian" population is in fact composed by "Piapocco and Curripacco" individuals, two ethnic and ethnolinguistic groups of Colombia, a class of categorization naming that is only specified in this individual information table, but not in the population table...

Furthermore, the same "individual information" table for HGDP-CEPH individuals readily provides a grouping of individuals into so-called "Geographic origin" (**Table T1.4**, column 2), another grouping of individuals into so-called "Regions" (**Table T1.4**, column 3), and a third grouping of individuals into so-called "Pop7Groups" (**Table T1.4**, column 4), without providing any form of references or explanations for the choices hereby made. Nevertheless, I strongly suspect that this Pop7Groups comes from the Analysis of Molecular Variance (AMOVA) results published in the very famous and indeed seminal Rosenberg et al. *Science* 2002 population genetics article on genome-wide microsatellite genotyping and analysis of this panel; but it is not reference as such in this table provided directly on the HGDP-CEPH panel resources. Thus, some of the HGDP populations are apparently categorized by the panelists based on alleged or pre-assumed geographic origins, possibly using the results of genetic clustering analyses, and other inexplicit historical and genealogical data…

Table T1.4: HGDP-CEPH population table (https://cephb.fr/hgdp_panel.php)

HGDP-CEPH Population	Geographic origin	Region	Pop7Groups
Surui	Brazil	America	America
Karitiana	Brazil	America	America
Colombian	Colombia	America	America
Maya	Mexico	America	America
Pima	Mexico	America	America
Cambodian	Cambodia	Asia	East_Asia
Han	China	Asia	East_Asia
Tujia	China	Asia	East_Asia
Yizu	China	Asia	East_Asia
Miaozu	China	Asia	East_Asia
Miaozu	China	Asia	East_Asia
Oroqen	China	Asia	East_Asia
Daur	China	Asia	East_Asia
Mongola	China	Asia	East_Asia
Hezhen	China	Asia	East_Asia
Xibo	China	Asia	East_Asia
Uygur	China	Asia	Central_South_Asia
Dai	China	Asia	East_Asia
Lahu	China	Asia	East_Asia
She	China	Asia	East_Asia
Naxi	China	Asia	East_Asia
Tu	China	Asia	East_Asia
Japanese	Japan	Asia	East_Asia
Brahui	Pakistan	Asia	Central_South_Asia
Balochi	Pakistan	Asia	Central_South_Asia
Hazara	Pakistan	Asia	Central_South_Asia
Makrani	Pakistan	Asia	Central_South_Asia
Sindhi	Pakistan	Asia	Central_South_Asia
Pathan	Pakistan	Asia	Central_South_Asia
Kalash	Pakistan	Asia	Central_South_Asia
Burusho	Pakistan	Asia	Central_South_Asia
Yakut	Siberia	Asia	East_Asia
French	France	Europe	Europe
French_Basque	France	Europe	Europe
Sardinian	Italy	Europe	Europe
Tuscan	Italy	Europe	Europe
North_Italian	Italy (Bergamo)	Europe	Europe
Orcadian	Orkney Islands	Europe	Europe
Russian	Russia	Europe	Europe
Adygei	Russia Caucasus	Europe	Europe
Druze	Israel (Carmel)	Middle East	Middle_East
Palestinian	Israel (Central)	Middle East	Middle_East
Bedouin	Israel (Negev)	Middle East	Middle_East
Mozabite	Algeria (Mzab)	North Africa	Middle_East
NAN_Melanesian	Bougainville	Oceania	Oceania
Papuan	New Guinea	Oceania	Oceania
Biaka_Pygmy	Central African Republic	Subsaharan Africa	Africa
Mbuti_Pygmy	Democratic Republic of Congo	Subsaharan Africa	Africa
Bantu_N.E.	Kenya	Subsaharan Africa	Africa
San	Namibia	Subsaharan Africa	Africa
Yoruba	Nigeria	Subsaharan Africa	Africa
Mandenka	Senegal	Subsaharan Africa	Africa
Bantu_S.E._Pedi	South Africa	Subsaharan Africa	Africa
Bantu_S.E._S.Sotho	South Africa	Subsaharan Africa	Africa
Bantu_S.W._Herero	South Africa	Subsaharan Africa	Africa
Bantu_S.E._Tswana	South Africa	Subsaharan Africa	Africa
Bantu_S.W._Ovambo	South Africa	Subsaharan Africa	Africa
Bantu_S.E._Zulu	South Africa	Subsaharan Africa	Africa

Finally, and of importance in the literature, the HapMap panel, also vastly used in human population genetics and genomics and further part of the 1000 Genomes Project, considers the "CEU" population simply described as "Utah residents with Northern and Western European ancestry from the HGDP-CEPH collection". However, the HGDP-CEPH collection itself refers to said families as:

> *« - 10 familles Françaises qui avaient participé aux travaux de recherche du Professeur Jean Dausset sur le complexe majeur d'histocompatibilité*
> *- 48 familles de Mormons d'Utah, collectées par le Professeur Ray White*
> *- 1 famille d'Amish de Pennsylvanie, collectée par le Professeur Ken Kidd*
> *- 2 familles du Venezuela, collectées par le Professeur Nancy Wexler »* [30]

We may thus assume that the HapMap CEU who are said to have been sampled in Utah (USA), comprised the self-reported Mormon individuals from the HGDP-CEPH. They are often categorized in human population genetics studies, without discussion whatsoever, as "European" populations, and even sometimes mapped somewhere in Germany rather than in Utah, exactly as if these samples had been in fact sampled in non-Mormon communities from "Central Europe". This categorization "approximation" neglects the recent history of European migrations to the USA and accompanying demographic founder events, as well as the specific socio-cultural marital behavior of Mormon populations allowing and even sometimes favoring polygyny and religious endogamy[31], which very possibly influenced their resulting genetic diversity compared to extant Central European populations (at least one should test it formally before assuming).

Altogether for the HGDP-CEPH, note that since the recent passing of Luca Cavalli-Sforza and Howard Cann, leaders of the HGDP-CEPH initiative and direct major witnesses of the DNA sample inclusion in the panel, the detailed information about individual sampling has significantly been lost forever, as we cannot even directly ask them to recall what was the genesis of this table that, to my views, exemplifies largely beyond my specific research projects, what is faced by the human population genetics and anthropological genetics community regarding deontological practices on anthropological categorization issues and reproductible research.

[30] https://cephb.fr/familles_CEPH.php

[31] Polygamy was officially legally banned in Utah by the Morrill Act Federal law in 1862 (https://www.uen.org/utah_history_encyclopedia/p/POLYGAMY.shtml), but more recently, the ban was weakened successively in 2013 and recently in 2020 (https://www.nytimes.com/2020/05/13/us/utah-bigamy-law.html) as enacted in 2022 (https://le.utah.gov/xcode/title76/chapter7/76-7-s101.html)

1.5. Concluding remarks to section 1.4, and personal recommendations to researchers and students.

Altogether, for data sharing ethical and deontological issues discussed above in **section 1.4** concerning the reproducibility and associated advanced discussion of human population genetics and anthropological genetics investigation relying on complex categorization issues, as discussed in **sections 1.1 to 1.3**, and concerning self-reported anthropological data collected from interviews with participants,

I recommend to:

1. readily provide as many elements as possible in scientific publications and communications, within the ethical and deontological constraints stated above regarding individual privacy and personal data protection.
2. provide data at the level of "populations" or "communities", sufficiently wide to avoid personal identification within these larger social or geographical groups of individuals, for types of individual-level data that could allow indirect identification of participants.
3. retain identifying and sensitive anthropological data at the individual scale, and share them exclusively within the frame of close collaborations or at least backed up by appropriate legally binding Material Transfer Agreement, or equivalent. For other research projects conducted outside the involvement of the researcher who originally procured Informed Consents, bind data sharing by MTA, or the equivalent, based only on explicit compliance to Informed Consents procured by donors.

Finally, for sharing genetic samples and data, and for linguistic recordings and systematic transcripts,

I recommend to consider them as systematically identifying highly sensitive data. As such, I recommend to retain them and only share them based on judicial or medical research requests from authorized parties, and share them with the scientific community based only on explicit compliance to Informed Consents procured from donors, backed by legally-binding MTA, or equivalent.

Altogether, the issues discussed in the above **section 1.4** emerge from the fundamental contradiction between protecting sensitive private data, such as human genetic data, and using such data for the advance of science, technology, knowledge, and society. They are exacerbated by the recent possibilities of generating rapidly massive amounts of genetic data, and of sharing them ubiquitously virtually instantaneously.

The solutions however are not to be expected from ethical changes alone, I think and from my personal experiences. Indeed, I strongly believe that ethically justified deontological and legal constraints on the practice of anthropological genetics sampling and data sharing can improve research practices and, as exemplified in this chapter, strongly improve reproducibility and productive discussion for future work in the field. However, they highlight major gordian knots, sometimes impossible to break, between data protection and Open Science, which in turn may strongly limit academic freedom and the advances of research projects and endeavors.

The solutions to this conundrum are, I think, mainly socio-political, both nationally and internationally. To overcome, in practice, the ethical and deontological challenges I described here, researchers worldwide need more administrative and legal support from their institutions; countries need more administrative and

legal support devoted to appropriately evaluate the increasing number and complexity of requests submitted by researchers; and we all need to harmonize administrative requirements and procedures as well as conduct frequent ethical debates at an international scale, even broader than that of the Declaration of Helsinki.

This will only be achieved by a renewed effort and mobilization from the scientific community to:

i) engage and participate in the public debate concerning bio-ethics and deontological research practices, including concerning a reflexive critic of their own past practices;

ii) massively increase the efforts of diffusion of the scientific method itself towards civil non-specialized audiences in general, including specifically towards legal practitioners and policy-makers.

Indeed, I believe there is a crucial need for diffusing methods, dialectics, and, in general, the scientific production of evidence, rather than communication and advertisement of shiny results obtained without ever explaining how they were obtained. I believe that we lack massive understanding in the society today about how science operates in principle and in practice, which strongly impairs, to my views, the impact of the scientific community on societal decisions and policies (including that directly concerning scientific research). This is, I believe, the true "*Science in the Society"* challenge we face as a community in the proximal future.

1.6. Perspectives: Population genetics categorization issues in the paleo-genomics era and interdisciplinarity necessities between genetics and anthropologies.

Disclaimer:

I have not personally conducted paleo-genomics research projects in the past. However, my personal experience about this topic and the points of view exposed in this section rely on the building from scratch of the paleo-genomics molecular platform P2GM at the MNHN on the site of the Musée de l'Homme which started in 2011 (there were literally no walls in the lab section of the Musée de l'Homme when I started this project in 2011 during the complete makeover of the building). Since its opening in 2015, I have been co-directing the platform with Dr. Céline Bon (a real paleo-geneticists), coordinating the work of researchers and engineers specialized in this discipline for their projects conducted on P2GM. I have thus been in operational contact with more than 40 paleo-genomics projects conducted since then on the platform, concerning numerous scientific problematics and interested in numerous types of archaeological materials and species, including hominids. As a result, I have been largely involved, since 2019, in a human paleo-genomics project piloted by Dr. Marie-Claude Marsolier-Kergoat and Dr. Céline Bon on reconstructing the paleo-genetics history of the Paris Basin since the Mesolithic until now. My scientific participation in the project has gone beyond the operational co-coordination of molecular genetics experiments on P2GM since 2021 (an already substantial task in itself…), as I am now carrying a work-package of the project newly funded by the French ANR (Project ParisAncientDNA) about developing ABC inference methods suited for investigating specifically complex admixture processes using paleo-genomics data (see ***Chapter 4.3.b****).*

Finally, and as a result, I have been invited to participate in 2020 to a national workshop on the regulation of invasive and destructive scientific procedures conducted on human remains from archaeological material excavated in France and under the supervision of the Ministry of Culture. This workshop, initiated by the Minister of Culture overseeing all archaeological excavations and preservation in France, has put together the extensive national community of archaeology practitioners (academic researchers, INRAP, private preventive-archaeology enterprises…) and human remains collections curators (national, regional, institutional, academic researchers…), together with scientific representatives of the four main paleo-genomics public-research platforms, active in France at the time of this workshop. Based on the work of this taskforce, comprising more than 50 individuals, the Minister of Culture produced and released in June 2022 novel norms, regulations, and procedures for access to human remains archaeological material for conducting invasive research (paleo-genetics and isotopic research mainly), with the main goals to ensure deontological good-practices across involved parties, the integrity of research-conduct, and the maximal preservation of the resource for future generations.

Therefore, while not practicing it with my own hands, beyond paleo-genetics literature reading as any human population genetics, the thoughts presented below result from numerous extended interactions, and sometimes mediation, between the archaeologist and paleo-anthropological communities and that of human population genetics, as well as my interdisciplinary practice within the MNHN and Musée de l'Homme over the years.

However, the thoughts below are just food-for-thoughts, and will need to be formally elaborated and extensively substantiated to reach the academically acceptable level that they lack at that point.

For the past 20 years, and particularly since 2010, human paleo-genetics molecular techniques aiming at extracting and sequencing DNA from ancient human skeletal remains have massively improved. They fueled the adaptation and deployment of novel population genetics methods for the study of ancient genetic diversity patterns, which have profoundly re-drawn our knowledge of human evolutionary history, and which continue to do so almost on a daily basis. Today, paleo-genetics studies can ambition (technically and financially) to investigate "*populations*", by which I mean here a statistically substantial number (>20) of individuals from the same geographic region and which have died relatively at the same time, rather than seldom individuals dispersed over gigantic geographic and temporal spaces.

These novel technological and methodological developments, together with those of environmental, archaeo-zoological, and archaeo-botanical paleo-genetics, provide population geneticists with what is almost a time-machine for travelling in the real history of genetic evolution. In practice, they bring novel almost direct information into genetic patterns of populations in the past and therefore allow researchers to formally test genetic continuity, or the absence of genetic continuity, across populations over time and until today. Furthermore, and importantly, they also allow to empirically test the power and accuracy of virtually all population genetics inference models and methods developed since the birth of population-genetics at the turn of the 20th century and, exclusively until now, aimed at reconstructing and inferring the past from the observed present. This allowed recently to revisit numerous previous results about human genetic evolution (Schlebusch and Jakobsson 2018; Bergström et al. 2021).

To be honest, we are not completely there yet, mainly due to the nature of ancient and degraded genetic data which differs strongly from that of "fresh" genetic data, and therefore requires very specific statistical treatments, in turn requiring adapting existing or developing new models and expectations, and therefore not always strictly comparable to "all previous population genetics methods". Nevertheless, this emerging novel paradigm rapidly builds a novel scientific discipline which has roots and strong conceptual and methodological ties with "classical" population genetics. In return and as a consequence, it has already deeply impacted human population genetics and anthropological genetics to the point that it is impossible today for these practitioners to work, or be trained, without considering and discussing paleo-genetics methods and results. This is thus, I believe, a revolution for these latter disciplines.

In this context, I will next address, as a specific perspective, categorization issues for population genetics as raised in a novel manner due to paleo-genetics. Second, I will try to evaluate whether paleo-genetics is also a revolution for paleo-anthropologists and archaeologists, as very often claimed by geneticists and the journalists echoing them in the non-specialized audience. Ultimately, I will conclude these perspectives about the broader issues of the practice of interdisciplinary research between paleo-genetics and archaeology, echoing very similar issues discussed in the previous sections about interdisciplinary practices between cultural anthropology, human population genetics, and linguistics.

1.6.a. Archaeological cultures are not population-genetics' populations

Paleo-geneticists require skeletal material as their primary object of investigation. The discipline itself is already strongly specialized in two, heavily intertwined, sub-fields of specialization: molecular genetics and population genetics. Molecular paleo-geneticists are primarily focusing on extracting and sequencing ancient and degraded DNA from skeletal remains, a major molecular-chemistry challenge in itself that only needs specific inputs from archeological knowledge about the remains themselves and the context from which they were unearthed. Namely, while molecular paleo-geneticists can work blindly (*"give me a bone, and I'll extract and sequence it's DNA"*), they often much prefer having additional information about the material itself, as it can crucially orientate their approaches and methods. Indeed, DNA preservation, and therefore the extraction and sequencing-preparation molecular techniques to be deployed, strongly depends on the age and the type and osteological structure of the remain, as well as on the environmental (temperature, pluviometry, humidity, type of vegetation coverage…) and geo-chemical (pH, soil chemical composition) context from which remains were unearthed by archaeologists. However, note that if this information is not fully available, this does not prevent molecular paleo-geneticists to exploit the material itself…

For population genetics paleo-genetics investigation, once the DNA sequences have been generated, the categorization and labelling of skeletal remains into "populations" is as methodologically crucial as it is for any population genetics investigation, as exemplified and discussed throughout this Chapter. This information stems primarily from archeological work and *cannot* be filled-in anew by paleo-geneticists alone.

This is precisely where major problems of categorizations emerge in the practice of paleo-genetics, with some strong analogies with the problems previously described between cultural anthropology and human population genetics as well as novel issues.

Indeed, paleo-geneticists tag the skeletal material they investigate most often based on categories provided to them directly by archaeologists or based on archaeological collection databases and metadata associated with the remains. Ultimately, they label their samples with a "population tag", preferably further informed with an "archaeological date", but not always so. This "population tag" is based on archaeological work of an extremely complex nature very rarely explicitly discussed in paleo-genetics work, and often apparently mis-understood by geneticists, very unfortunately. As a result, except for more recent paleo-genetics articles making a significant effort to integrate archeological knowledge at the root of their investigation in sincere interdisciplinary efforts, a substantial amount of the population-genetics paleo-genetics investigations are conducted in a way that may be strongly at odds with archaeology, albeit recklessly feeding on this discipline.

In my views, archaeologists integrate multiple knowledge brought from varied disciplines or subdisciplines such as geo-chronology, stratigraphy, sedimentology, and paleo-palynology, art-history, ethnology and material-culture studies, biological anthropology and thanatology, etc. They do so in order to propose a complex meta-constructed knowledge about the archeological sites and objects they investigate, and about the societies that produced them.

They build "archaeological cultures" based on i) a single precisely defined *archaeological criterion* (or a very limited number of such criteria), such as a particular ceramic manufacture, ii) observed throughout sites at a *relatively reduced geographical scale* (regional), and, iii), spanning *relatively short periods of time* (seldom more than one thousand years). For a random instance, the "Corded Ware culture" refers to an archeological category of sites and practices based on a specific ceramic style and manufacture that is found between roughly 3000 and 2000 years before the common era from North-Central Europe to Central Asia. Most importantly, note that it is very frequent that two different archeological cultures, coexisting in time and space, are, respectively, defined based on an archaeological criterion of a very different nature. Some may rely on ceramic manufacture while another may stem from shared specific funerary practices. Therefore, "archaeological cultures" can overlap in time and space, or not, and be produced by biological individuals from different societies, or not, or in biological genealogical relationships with one another, or not.

Once such criterion "archaeological culture" is built, and relatively consensually agreed upon within the archaeologist community, more in deep analyses of archaeological sites inform said "culture" by identifying other sets of archaeological phenomena and practices more or less variable among sites within the culture and not necessary to build the initial categorization. This is how, for instance, funerary practices may vary largely across sites within culture, and even be structured spatially and/or temporally, without re-defining the culture initially defined on another criterion than funerary practices.

In this context, archaeologists may talk about an "archaeological population" most often when historical linguistic records further inform the archaeological site directly or indirectly; hence several intense debates as to whether certain recent pre-historic archaeological cultures should be considered as populations or not.

Therefore, they explicitly add another stringent criterion for such categorization into a population compared to the categorization into an archaeological culture.

Next, archaeologists build "archaeological periods" corresponding to much less specific archaeological phenomenon (such as significant changes in lithic industries between the "Mesolithic period", literally the "*middle stone* period", and the "Neolithic period", literally the "*new stone* period", or metal industry such as the shift from the "Bronze Age" to the "Iron Age"), found at massive geographical scales, spanning much larger periods of time, and relatively less variable across sites and within a period. Importantly, as a consequence of such categorization process, the same archeological period often does not span the same absolute chronological dates in the different regions where it is identified; e.g. the "Bronze Age" or the "Neolithic" do not have the same dates and durations, respectively, in different regions. Furthermore, there may be temporal periods during these "archaeological periods" in certain geographic areas, where no "archaeological cultures" have been defined and recorded albeit archaeological sites have been found and investigated.

Finally, archaeologists and biological anthropologists often further characterize given archaeological periods, cultures, or populations with specific features about the ways of life of the individuals who produced said cultures. For example, among many others, an archaeological site categorized as ancient Mesolithic is often being considered as the product of a hunting-gathering way of life, while, for Neolithic sites, their producers are sometimes characterized as agriculturalists or herders depending on varied archaeological markers such as the presence of remains from domesticated farmed crops or animals. Note, however, that once such way of life has been characterized for certain sites in a given archaeological culture, it may be applied to all other sites in this culture, even in the absence of the same explicit markers locally, *de facto* conducting a dialectical induction without formally testing it and thus possibly wrongly reducing the presumed cultural variation within said archaeological culture.

Most importantly in this context, archaeological cultures were produced by individuals and groups of individuals that indeed existed, but for whom we do not necessarily have skeletal remains. For instance, in certain Celtic archaeological populations, individuals were not buried but most often cremated, leaving little occurrences of human skeletal remains within given sites associated with otherwise rich material cultures and archeological material remains at the origins of the "Celtic" labelling of said sites. Furthermore, and in all cases, even when skeletal remains are associated with cultural artifacts and funerary practices, it is perhaps probable but cannot be formally and scientifically verified that said skeletal remains pertained to an individual having directly participated in the production of this culture, as opposed to a recent migrant conjecturally buried here.

Based on such elaborations, archaeologists may propose models and hypotheses about the varied processes involved in archaeological cultures' or populations' emergences, expansions, diffusions, and disappearances over time and space. Intuitively, the archaeological scientific dialectics resulting in naming an archeological culture, population, period, or, simply, phenomena or trait, from varied information extracted from an archaeological site, may be seen as analogous to the linguistics dialectics building a "Language" from observed communication practices as explained in **section 1.2** above[32]. Indeed, an archaeological culture, population, or period is not a simple object directly observable with our own senses, nor stemming from complex manipulations borrowed from experimental Popperian sciences only; an archaeological culture, an archaeological population, and an archeological period are, in fact, complex

[32] with all academic cautions needed when doing such analogy.

disciplinary interpretations, each non-necessarily overlapping, which are all the subject of extensive discussions and debates within the scientific discipline.

Therefore, archaeological categories are, by their very construction, nowhere close to the population-genetics' definition of a population, i.e. a group of individuals more likely to reproduce with one another than they are with other groups of individuals.

Nevertheless, paleo-geneticists often tend to use, blindly, these archeological labels as population tags for the DNA they extracted and sequenced from skeletal remain. They thus de facto assimilate sets of DNA sequences within an archaeological culture or population or period often independently of the sometimes very large temporal scales the original remains spanned, and further associate these *de novo* paleo-genetics "populations" with sets of ways-of-life practices frequently investigated by human population geneticists such as "agricultural", "hunter-gatherer", "farmer", "herder", etc. Once this is done, they then investigate these groups of DNA sequences with the same paradigms used to investigate "populations" in population genetics. Based on their DNA sequences, they calculate allelic and haplotypic frequencies across these false populations but real sets of biological individuals, and infer their origins and migration and admixture histories across time and space.

Expectedly at that point, paleo-geneticists often synthesize their results and discussions using these borrowed and altered "population" labels and associated traits in ways that are very often inappropriate as they end up mixing geographical criteria, archeological cultures, populations and/or periods names, absolute chronological dating, and presumed or archaeologically substantiated ways of life…

As an illustrative example, which I find to be representative of many other famous published articles in paleo-genetics, one can read in an article published in Nature in 2015 and already cited more than 700 times (Allentoft et al. 2015):

"Abstract:

> The Bronze Age of Eurasia (around 3000–1000 BC) was a period of major cultural changes. However, there is debate about whether these changes resulted from the circulation of ideas or from human migrations, potentially also facilitating the spread of languages and certain phenotypic traits. We investigated this by using new, improved methods to sequence low-coverage genomes from 101 ancient **humans** from across Eurasia. We show that the Bronze Age was a highly dynamic period involving large-scale **population** migrations and replacements, responsible for shaping major parts of present-day demographic structure in both Europe and Asia. Our findings are consistent with the hypothesized spread of Indo-European languages during the Early Bronze Age. We also demonstrate that light skin pigmentation in Europeans was already present at high frequency in the Bronze Age, but not lactose tolerance, indicating a more recent onset of positive selection on lactose tolerance than previously thought."

Which therefore readily refers to:

- archaeological periods (in bright red),
- geographical regions (in dark blue),
- archaeological dating (in bright green),
- linguistics (in orange),
- intertwined with purely paleo-genetics information (in purple)

- and a population-genetics' biological definition of "a population" and of "an individual" (in **black**), probably, as the way they use these lexical terms do not fit the above definition of an archaeological population.

After the body of the article deploys, the last concluding paragraph reads:

"Implications:

> It has been debated for decades if the major cultural changes that occurred during the Bronze Age resulted from the circulation of **people** or ideas and whether the expansion of Indo-European languages was concomitant with these shifts or occurred with the earlier spread of agriculture. Our findings show that these transformations involved migrations, but of a different nature than previously suggested: the Yamnaya/Afanasievo movement was directional into Central Asia and the Altai-Sayan region and probably without much local infiltration, whereas the resulting Corded Ware culture in Europe was the result of admixture with the local Neolithic **people**. The enigmatic Sintashta culture near the Urals bears genetic resemblance to Corded Ware and was therefore likely to be an eastward migration into Asia. As this culture spread towards Altai it evolved into the Andronovo culture (Fig. 1.), which was then gradually admixed and replaced by East Asian **peoples** that appear in the later cultures (Mezhovskaya and Karasuk). Our analyses support that migrations during the Early Bronze Age is a probable scenario for the spread of Indo-European languages, in line with reconstructions based on some archaeological and historical linguistic data. In the light of our results, the existence of the Afanasievo culture near Altai around 3000 bc could also provide an explanation for the mysterious presence of one of the oldest Indo-European languages, Tocharian in the Tarim basin in China. It seems plausible that Afanasievo, with their genetic western (Yamnaya) origin, spoke an Indo-European language and could have introduced this southward to Xinjang and Tarim. Importantly, however, although our results support a correspondence between cultural changes, migrations, and linguistic patterns, we caution that such relationships cannot always be expected but must be demonstrated case by case."

Which therefore refers to:

- archaeological periods (in bright red),
- archaeological cultures (in dark red),
- geographical regions (in dark blue),
- archaeological dating (in bright green),
- ways of life (in dark green)
- historical linguistics (in orange),
- intertwined with paleo-genetics and population genetics inference results (in purple)
- and a population-genetics' biological definition of "a population" and of "an individual" (in **black**), probably, as the way they use these lexical terms do not fit the above definition of an archaeological population.

Despite the academic conditional tense rigorously used and the important final provision, the mixing of categories from very different disciplines and with very different definitions here proposed is not discussed. Furthermore, the archaeological cultures and periods here used as genetic population labels are

not themselves discussed -- this is probably left by paleo-geneticists for archaeologists to discuss in the future… Symmetrically, the implications for population genetics of the archaeological categorizations here used are not discussed, and this is probably left by paleo-geneticists to anthropological geneticists and population geneticists to discuss in the future…

As a result of the collision of complex categories from different disciplines that are seldom explained and even less discussed, as an anthropological geneticist and as explained throughout this chapter, I have honestly a hard time to understand *"who are they talking about and who are they conducting their population genetics inferences on?"*, not without substantial efforts of re-reading and disentangling categories often outside my areas of expertise. Furthermore, I honestly understand that some archaeologists and biological anthropologists feel that their work and discipline are mutilated by paleo-geneticists (Callaway 2018)…

As a final anecdote, a famous archaeologist that I only briefly encountered at conferences and that knew I was very sensitive about issues of categorization in human population genetics research, directly contacted me to convince me to write with him a rebuttal of a very recently published paleo-genetics paper. He wanted to re-analyze and re-interpret the paleo-genetics data after more careful categorization of samples. He did so the day after the publication of the paper in *Nature*. And he was a co-author of this paleo-genetics paper he wanted to refute the day after its publication[33]…

Ultimately, I find the same categorization pitfalls with paleo-genetics as I did with human population geneticists. Both communities apply and use population labels drawn from other disciplinary productions without sufficient caution nor discussion. Technically, as previously said iteratively above, I am not immune to these issues myself, I have done such mistakes myself in previous publications and will probably find myself doing them again in the future… Without excuses, one has to recognize that those are very complex and difficult problems in practice, both for the collaborative and interdisciplinary scientific work itself, as well as, pragmatically, when it comes down to mimetic publication opportunism and editorial constraints in a challenging academic world…

1.6.b. Is there really a paleo-genetics revolution, or at least its' possibility, in paleo-anthropology and archaeology?

I believe that it is possible that, for paleo-anthropology, the development of paleo-genetics indeed represents a scientific revolution, at least an embryo of revolution as of today. Indeed, in the future, the genetic determination of complex anthropometrical and phenotypic traits of interests and their evolution over time in interaction with environmental and developmental constrains will possibly be unraveled by human quantitative and population genetics (it is the explicit goal of countless researchers and projects since the birth of human genetics). In turn and expanding these findings to the study of ancient remains, paleo-anthropology will then be able to build synthetic and mechanistic models of morphological transmission with changes through time anchored on evolutionary genetics mechanisms. Such knowledge and models will thus integrate paleo-anthropology and paleo-genomics to fundamentally reform the former discipline and associated scientific practices, as they currently only rely on their own descriptive ad hoc

[33] It also denotes the very ambiguous relationship that some archaeologists have with paleo-geneticists in the challenging academic times we experience, here at odds with deontology of research practice: one does not put his/her name on an article he/she fundamentally disagrees with… For the record, I refused to work on his rebuttal project.

models which lack most detailed knowledge about underlying molecular, cellular, and biological development mechanisms.

For instance, the classical models of morphological evolution from "robust" to "gracile" forms are currently *ad hoc* and exclusively based on descriptive investigation of bone-records. They remain actively largely debated and regularly revisited in paleo-anthropology. I believe that they cannot reach a scientific consensus as they cannot explicit the underlying environmental and biological plasticity and evolutionary biology mechanisms needed to generalize a synthetic theory of human morphological evolution rooted in the synthetic theory of evolution. Paleo-genetics could thus indeed revolutionize paleo-anthropology, provided that human genetics indeed provide the needed understanding of the genetic and environmental determination of complex morphological traits, which still very largely remains to be discovered today (e.g. with the genetic determination of adult non-pathologic height in **Chapter 2**, introduced above in **section 1.1.c**). Furthermore, note that while paleo-genetics may revolutionize paleo-anthropology as a whole, not all paleo-anthropologists are directly, in practice, concerned by paleo-genetics. Indeed, it is not yet possible for paleo-genetics to investigate completely fossilized skeletal material: DNA is an organic molecule and its extraction and sequencing require that skeletal remains still comprise organic matter.

Conversely, I believe that paleo-genetics provides novel very interesting tools for archaeologists without fundamentally changing their disciplinary paradigms and scientific questioning. Instead, it seems to me that these novel genetics disciplines offer, to archaeologists, sets of novel technical means to address long-standing questions and formally test certain aspects of long-standing hypotheses that could not be addressed in numerous pre-historical and historical contexts. To some extent, from my perspective, this "non-revolution" may nevertheless be as important to these fields than the "non-revolution" brought, starting in the 1940's and still ongoing today, by the deployment of isotopic characterization for archaeological purposes[34]. In other words, without being a revolution (at least in the radical sense I employ here), paleo-genetics might represent a "very very important" milestone in the long history of archaeological disciplines...

Indeed, paleo-genetics approaches can provide, and have already provided in numerous instances, key information to archaeologists about, among existing examples, the determination of biological sex (in particular when bio-anthropometrical approaches are inoperable or inconclusive, and/or when only social gender can be assessed by archaeologists), the genealogical biological relationships among remains (in the lack of explicit such records which would, in any case, only provide familial genealogical information rather than biological genealogies), and more generally about genetic migrations and admixture histories, as provided by phylogenetics and population genetics demographic inference approaches.

Thus, with paleo-genetics in the archaeological tool-box, it becomes possible for archaeologists to formally test hypotheses about the links between genealogical and social structures and organizations in the past, including biologically familial or non-familial funerary practices. Furthermore, it can allow to formally test whether past cultural diffusions where accompanied, or not, by demic migrations and/or population admixture. Note importantly here, that this latter point is of major interest to virtually any anthropological geneticists, but of major interest to some archaeologists only.

[34] Indeed, it seems to me that this latter technological development, derived from fundamental and applied quantum physics, did not change the fundamental paradigms underlying paleo-anthropological and archaeological scientific questioning, but nevertheless provided means to inform archaeological sites about numerous aspects and hypotheses that were largely out of reach of researchers before. They involve, but are not limited to, absolute dating beyond classical relative dating using geo-chronology and sedimentology, understanding human biological and skeletal development, reconstructing past alimentation, past migrations, and, in general, past relationships between humans and their environments.

Indeed, archaeologists are often interested in the diversity of past cultural practices, the conditions and processes underlying their emergence, expansion, diffusion, and disappearance, whether they are accompanied by reproduction events leaving genes behind or not. Therefore, when past cultural representations and practices themselves are the objects of study, answering anthropological genetics questions about genetic migrations and admixture is often of very limited interest, if not completely anecdotical. This is the case for numerous archaeologists in practice, including those who otherwise may be interested in the other types of information brought by paleo-geneticists, such as determining biological gender as exemplified above.

Finally, while I do not believe that paleo-genetics represents a fundamental and radical revolution for archaeology, I also believe that it is true that the expansion of paleo-genetics, whether of interest or not to archaeologists, already influenced changes in the technical practices of this discipline at the source of the skeletal material investigated by paleo-geneticists. In particular, archaeologists are now almost all implementing protocols trying to limit human-DNA contamination of the skeletal remains they discover in the field or investigate in collections; and if they do not implement them systematically enough as per paleo-geneticists, they are at least less and less oblivious of these novel practices. Furthermore, they try to implement novel constrains for remains' preservation in collection environments, in order to ensure specifically a better conservation of ancient-DNA for future requests even when not interested in paleo-genetics studies themselves. However, I believe that these technical changes do not fundamentally change archaeological paradigms, methods and aims.

1.6.c. "The interdisciplinary requirement" for human population genetics and paleo-anthropologists

In this context, one can see that there is a fundamental asymmetry in the interdisciplinary relationships between paleo-genetics and archaeology: paleo-geneticists imperatively need archeologists as primary providers of skeletal remains and associated information, their object of study, while archaeologists do not always require inputs from paleo-geneticists in their disciplinary practice. This is to some extent similarly asymmetrical between anthropological geneticists and cultural anthropologists or linguists, as explained earlier in this Chapter.

I believe it is important, in such pluri-disciplinary work, to recognize and acknowledge the contours of these asymmetries, as not all projects require true inter-disciplinary integration and their massive level of investment from either discipline to understand each other. In all cases, for pluri-disciplinary asymmetric projects as well as for inter-disciplinary integrated projects, there is a massive need for exchanges and dialogue between disciplines, relying on the efforts made by all parties to truly try to understand and accept the other paradigms, constraints, and focus of interest. It is only then that trust can be gained by both parties, and productive scientific work conducted; a trust that it seems has largely been lost between paleo-geneticists and archaeologists, in part for the varied reasons explained in this section, in turn resulting in defiance if not rupture between research practitioners from either disciplinary field.

To restore this trust and elaborate and conduct future productive projects between paleo-genetics and archaeology, time spent in dialogue is of the essence. It is however a major challenge as both disciplines work at different speed. Nevertheless, I think paleo-geneticists need to devote much more time speaking with archaeologists and trying to better understand their disciplinary paradigms, for the sake of their projects and that of the future advances that they can bring to either field. This comment is also valid for human population geneticists, cultural anthropologists and linguists, as I think I have shown throughout the

previous sections. Frank Alvarez-Pereyre named his seminal book on interdisciplinary work *"L'Exigence interdisciplinaire"* (2003); I think the population genetics disciplines need to embrace this *"exigence"* in their much-needed relationships with other disciplines.

Ultimately, investigating ancient DNA, modern DNA, and human evolution is, very obviously by definition, of major interest to paleo-geneticists, anthropological geneticists, and human population geneticists. However, these academic communities should often realize that their findings, major as well as minor, may not speak to other disciplines as eloquently as they presume. Indeed, while cultural anthropologists and archaeologists are always, respectively, interested in any knowledge acquired on material and/or cultures they explore, it is crucial to understand that their respective paradigms and methods participate to elaborate their objects of study, as for any scientific discipline and as in genetics. This may sound trivial, but has deep consequences that I think are far too often overlooked, or at least not enough actively mobilized in results' interpretations and discussions. Indeed, each one of these disciplines, while all focusing on the same human individuals or groups, and the symbolic or material cultures they perhaps perform and produce, in fact talk about different "populations".

Geneticists reconstruct the genetic diversity of populations, and sometimes try to reconstruct the history of genetic evolution of groups of DNA sequences carried by biological individuals. Linguists reconstruct the linguistic diversity of populations, that is the diversity of Languages or of manners of performing languages among biological individuals or groups of individuals, sometimes (but not always) trying to reconstruct the history of changes of these different objects. Ethnologists reconstruct the diversity of cultural practices performed by biological individuals and groups of individuals, and sometimes (but not always) try to understand their dynamics and possibly changes over time. Archaeologists reconstruct the diversity of past cultural practices necessarily performed at the time by biological individuals, but not necessarily focusing on said individual performances, and try to reconstruct the history of cultural changes and diffusions, not necessarily interested in the biological individuals involved in these processes *per se*.

All these disciplines use as a central object of study and interpretation the "population" or "group". However, these terms are very often defined very differently, which introduces, almost systematically in my experience, major mis-understanding during cross disciplinary talk, in turn irremediably impairing fruitful scientific interdisciplinarity[35]. A "linguistic population", an "archaeological population", an "ethnological population", and a "genetic population" represent very different populations, which may ideally refer to the exact same group of biological individuals, but which very rarely, if ever, do so as they in fact involve very different categorization construction processes.

I hope that I have convincingly illustrated this fact in this chapter, which finally leads me to propose that any such interdisciplinary, or pluri-disciplinary, endeavor should imperatively start with a precise and extended discussion to try to explicit for each discipline and each study-case: ***"What is the human population under study?"***. First, it is, I believe, a scientific necessity to build a truly interdisciplinary project by building a common object of study. Second, when and if such object is impossible to conjure due to impossible reconciliation of definitions, addressing this question would at least allow to investigate the different aspects of an object, only similar in appearances and in the lexicon, from different pluri-disciplinary perspectives.

[35] Note that, the problem of not making explicit the definition of what is named "a population" also sometimes lead to major misunderstanding in within-disciplinary talk, however less frequent and dramatic I think, as the common disciplinary jargon and largely shared paradigms often allows easier understanding despite polysemy.

Ultimately, the History of Humans is composite and complex, each discipline can bring different insights exploring different facets of this kaleidoscopic object from different perspectives. I think that these facets do not necessarily always have to tell the exact same story or to "always be reconciled", they altogether exist, whether different or analogous.

Chapter 2

Admixture in the demographic history and biological evolution of Central African "Pygmy" and neighboring "non-Pygmy" populations

Breakfast in Mbonde
Pays Tikar, Région Centre, Cameroun, 2011
©Paul Verdu

Chapter 2. Admixture in the demographic history and biological evolution of Central African "Pygmy" and neighboring "non-Pygmy" populations

***Important warning repeated from Chapter 1.1.b*:**

"Throughout my work, I have explicitly decided to keep the ambiguous "Pygmy" label (and its counter-part "non-Pygmy") for naming the groups of individuals based on the criteria of categorization de tailed in ***Chapter 1.1****. It is important to emphasize here that whether it is relevant to continue to use this label is still an ongoing debate in the cultural and biological anthropology communities; a debate to which I have myself participated. A practical problem of using this somehow archaic term is due to the fact, that, in some parts of Central Africa, the word "pygmy" itself is used derogatorily by non-Pygmy neighbors. However, I also found it to be ambivalent in the field. Indeed, certain Pygmy groups or individuals choose to claim this name in front of outsiders, governmental or non-governmental, to emphasize their socio-cultural particularity and advocate, in their local power play, against the socio-economic discriminations they suffer from their neighbors. Finally, at least one tentative "re-naming" of the four historical Cameroonian "Pygmy" populations (the Bezan, the Baka, the Ba.Ghieli, and the Ba.Koya), into the "Quatre B" (the "four Bs)" by Cameroonian civil society in the 2000's, also resulted in this new label being used derogatorily as well…*

In this complex context, I decided to first define precisely in my publications the categorization criteria used for grouping individuals into the binary category explained throughout this chapter, and decided to keep the word "Pygmy" as a label, making almost always sure to mention the potential derogatory usage of the term. I decided to use the original historical term for two reasons: as an exogenous term from ancient Greek, I believe it is more prone to ask the naïve reader the question "who are we talking about?", which I think is essential for reflexive critical thinking about the results and study design, as explained throughout this dissertation. Second, I do it in reference to the history of anthropology, and to the work of my giant predecessors.

However, as long as criteria of categorization are maintained, I really have no problems whatsoever to use another "label", as this will not change anything to the population genetics statistical descriptions and inferences performed, nor their discussions based on said criteria. We just haven't found a consensus on a name, yet. Human population geneticists often prefer to transform the "Pygmy" and "non-Pygmy" categories into "Rainforest Hunter Gatherers" and "Agriculturalists", as I did myself in certain collaborative publications. I believe this is very far from ideal and, in fact, even more problematic than the classical exogenous and literary "Pygmy" and "non-Pygmy". Indeed, this particular denomination for genetics research formally essentializes ways-of-life and economic practices and strictly separates them albeit, as explained above, the Central African context is not as strictly dichotomic regarding these cultural traits, far from it. Furthermore, this essentialization, as it is highly meaningful for anthropological geneticists and human population geneticists, might be in fact misleading in that it oversimplifies hypotheses and the methods used to test them into a "case-control" approach that may inevitably fail to capture a known more complex reality of human adaptation to varying ways-of-life.

For all these reasons, and bearing all these warnings in mind, I will henceforth keep using the "Pygmy" and "non-Pygmy" labels throughout this dissertation, in the hopes that readers understand that this is never, in my mouth, used derogatorily."

I detailed in **Chapter 1-section 1.1** the genesis, fundamental questions, and sampling protocols that presided my research endeavors since the beginning of my PhD, and until now, concerning the reconstruction, from genetic data, of the genetic history and evolution of Central African populations categorized in "Pygmy" and "non-Pygmy" populations based on historical and cultural criteria. In this **Chapter 2**, I will summarize and discuss several previous publications reporting the key-results we obtained over the years concerning this project.

As mentioned previously, the socio-historical construction of the "Pygmy" category in Western cultures resulted in more than 20 populations throughout the Congo Basin being labelled as such by Europeans. Cultural anthropologists and ethnologists during the 20th century showed that numerous cultural criteria could be informed to group these populations into a single category and used the historical word "Pygmy" to designate this super-group of populations differing from all other neighboring populations categorized into the mirror group "non-Pygmies". We also saw in this **Chapter 1-section 1.1** that anthropologists and ethnologists could not determine whether these populations had a common biological origin or not, as such question is often beyond their focus of interest, and contradictory in essence with their disciplinary paradigms and methodological means.

However, for biologists, such question is of interest to better understand human evolution and the patterns of genetic diversity observed today within and among populations in the region. In this context, the historical and widely-spread use of the blanket-term Pygmy resulted for numerous biologists in *de facto* considering these historical Pygmy populations as sharing a common origin more recent than the common origin shared between them and neighboring non-Pygmy populations, albeit without formally testing this hypothesis. Furthermore, archaeological work extensively showed the ancient occupation of the Congo Basin for the last 60,000 years at least (Phillipson 2005). However, due in part to the overall lack of ancient human remains in the acidic soils of the rainforest and to methodological limitations inherent to the discipline (see **section 1.6**), they could not link the archaeological occupation of the region with either Pygmy or non-Pygmy extant populations. Furthermore, the lack of knowledge about extensive previous anthropological and ethnological work among population geneticists also widely led to believe that Pygmy populations lived genetically isolated from geographically neighboring populations, without formally testing such prejudice. In this context, it is essential to better understand the demographic and migratory genetic history of these populations before one can even hope to identify the genetic determination and evolutionary history of phenotypic differences across populations and groups of populations in the region.

First (**section 2.1**), we described genetic diversity patterns obtained for numerous populations throughout the Congo Basin, categorized in Pygmies and non-Pygmy neighbors based on numerous cultural criteria.

Second (**section 2.2**), we formally tested whether Pygmy populations shared a common or an independent origin with neighboring non-Pygmy populations.

Third (**section 2.3**), we reconstructed the influence of complex socio-cultural behavior regarding inter-community marriages on genetic diversity patterns, effective populations sizes, and asymmetric gender-biased admixture patterns.

Fourth (**section 2.4**), we synthesize here these fundamental results about the demographic history of the Congo Basin peopling in a schematic way.

Fifth (**section 2.5**), we synthesize our work about the genetic evolution and adaptation signatures in Congo Basin populations, with a specific focus on the genetic determination of non-pathological adult height differentiation across individuals and populations.

Finally (**section 2.6**), we propose ongoing and future perspectives about investigating the "Neutral" demographic history of Central African extant populations.

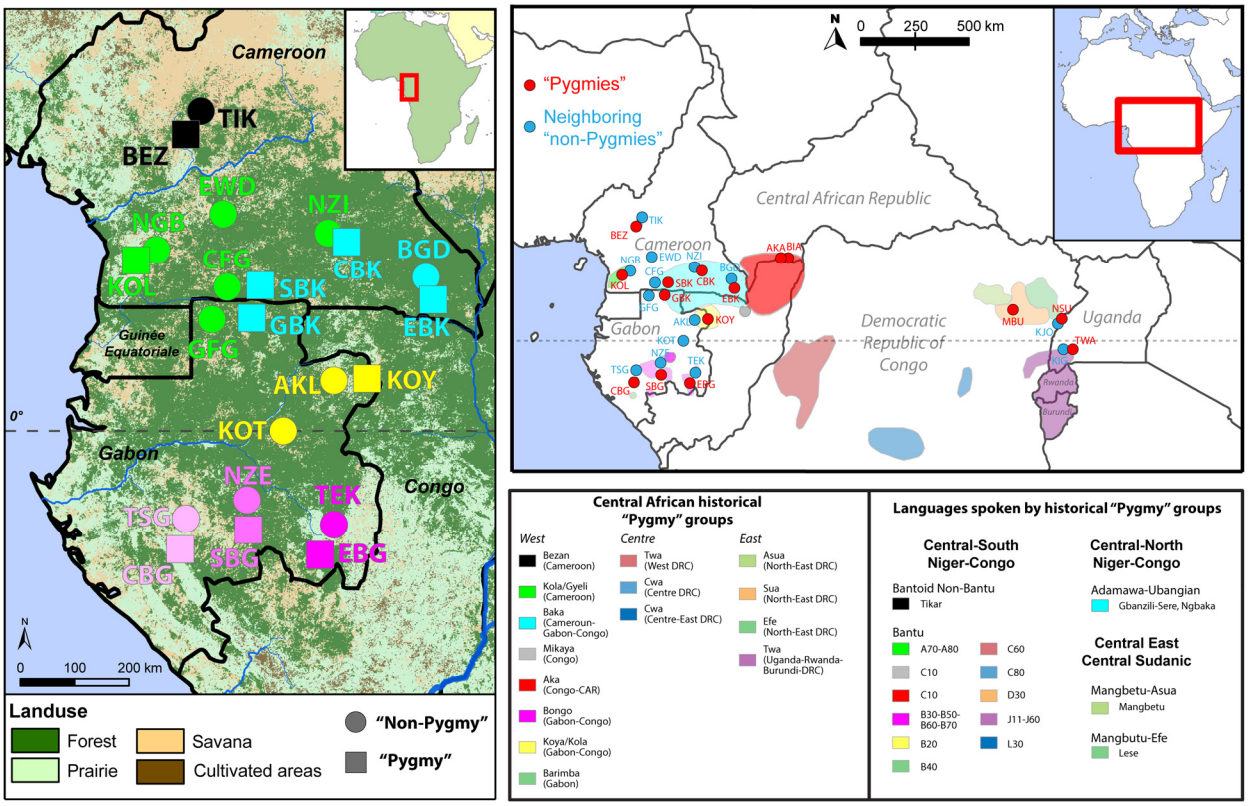


Figure F2.1.a.

	Population Name	Code	Country	Linguistic Family (Guthrie 1971)	Sampling
Pygmies	Bezan	BEZ	Cameroon	Bantoïd Non-Bantu, Tikar	P. Verdu
	Kola	KOL	Cameroon	Bantu, A80	A. Froment
	Central Baka	CBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment, P. Verdu
	Eastern Baka	EBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment
	Southern Baka	SBK	Cameroon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	A. Froment, P. Verdu
	Gabonese Baka	GBK	Gabon	Adamawa-Ubangian, Gbanzili-Sere, Ngbaka	J-M Hombert, L. Van Der Veen
	Koya	KOY	Gabon	Bantu, B20	S. Le Bomin
	Central Bongo	CBG	Gabon	Bantu, B30	P. Verdu, S. Le Bomin
	Eastern Bongo	EBG	Gabon	Bantu, B70	P. Verdu, S. Le Bomin
	Southern Bongo	SBG	Gabon	Bantu, B30-B50	P. Verdu, S. Le Bomin
	Aka	AKA	CAR	Bantu, C10	B. Hewlett
	Biaka	BIA	CAR	Bantu C10	L.L. Cavalli-Sforza, B. Hewlett
	Nsua	NSU	Uganda	Sudanic, Mangbetu-Efe	P. Verdu, M-F Mirune
	Mbuti	MBU	DRC	*nd*	L.L. Cavalli-Sforza, B. Hewlett
	Twa	TWA	Uganda	*nd*	G.H. Perry, L. Barreiro
Non-Pygmies	Tikar	TIK	Cameroon	Bantoïde Non-Bantu, Tikar	P. Verdu
	Ewondo	EWD	Cameroon	Bantu, A70	A. Froment
	Nzime	NZI	Cameroon	Bantu, A80	A. Froment, P. Verdu
	Bangando	BGD	Cameroon	Adamawa-Ubangian, Bangandu, Gbaya	A. Froment
	Ngumba	NGB	Cameroon	Bantu, A80	A. Froment
	Fang	CFG	Cameroon	Bantu, A70	A. Froment
	Fang	GFG	Gabon	Bantu, A70	J-M Hombert, L. Van Der Veen
	Akele (Bongomo)	AKL	Gabon	Bantu, B20	S. Le Bomin
	Kota	KOT	Gabon	Bantu, B20	J-M Hombert, L. Van Der Veen
	Teke	TEK	Gabon	Bantu, B70	J-M Hombert, L. Van Der Veen
	Nzebi	NZE	Gabon	Bantu, B50	J-M Hombert, L. Van Der Veen
	Tsogho	TSG	Gabon	Bantu, B30	J-M Hombert, L. Van Der Veen
	Konjo	KJO	Uganda	Bantu, D30	P. Verdu, M-F Mirune
	Kiga	KIG	Uganda	*nd*	G.H. Perry, L. Barreiro

Population samples used in genetic analyses in **Chapter 2** see detailed Materials and Methods in references cited therein. **See Chapter 1.1** for sampling and anthropological categorization procedures in Pygmies and non-Pygmies based on complex cultural criteria only.
Figures and tables adapted from previous publications:
Verdu P et al. *Current Biology* 2009 and Verdu P. *in Hunter-Gatherers from the Congo Basin, B. Hewlett Eds* 2014.

2.1. Genetic variation patterns among Central African populations with respect to different anthropological categories

After several years of DNA and anthropological data sampling throughout Central Africa conducted by myself and other colleagues (see **Chapter 1.1** above), we investigated, first between 2006 and 2009, autosomal microsatellite variation for 28 tetranucleotide markers dispersed throughout the genome (and thus genetically independent with very low Linkage Disequilibrium), that I genotyped for more than 800 individuals (**Figure 2.1**), with help from Myriam Georges at the nascent[36] molecular laboratory of the UMR7206 at the time of my PhD (Verdu 2009; Verdu et al. 2009).

As for any statistical analysis project, one first has to describe statistically the data, before trying to infer the mechanisms (genetic evolution in our case) that gave birth to the observed pattern. Note importantly, however, that inference and evolutionary hypotheses testing should imperatively be formally elaborated so as to avoid circularity and double-use of the data (Devezer et al. 2021). In particular, any population genetics statistical testing should be careful not to test hypotheses circularly based on observed genetic patterns.

In our case, individuals and populations could be categorized in "Pygmies" and "non-Pygmy neighbors" based on numerous cultural criteria described extensively in **Chapter 1.1**, without using genetic or phenotypic criteria such as adult standing height. Alternatively, we could also categorize individuals and populations linguistically based on Guthrie (1967) categorizations of Central African languages, a primary interest of ethnologists in the initial interdisciplinary project I was part of. We therefore set out to investigate genetic variation patterns within and across populations and groups of populations using classical hierarchized or un-hierarchized AMOVA approaches (Excoffier et al. 1992), in order to evaluate how cultural anthropology categories translated into genetic variation distribution, or not.

We first considered only genetically family unrelated individuals at the 2nd genealogical degree (within and across populations), and focusing on roughly 600 individuals born in at relatively small geographical scale in the Western part of the Congo Basin only (**Figure 2.1.a left panel**), from 21 populations that could be categorized in nine "Pygmy" and 12 neighboring "non-Pygmy" populations. We found that a significant proportion of the total genetic variation differentiated across the 21 populations overall (un-hierarchized AMOVA, F_{ST} = 0.013, 10,000 permutation $p < 0.001$). Furthermore, we found that the nine populations categorized as Pygmies a priori where considerably more differentiated (F_{ST} = 0.019, 10,000 permutation $p < 0.001$), than, separately also in un-hierarchized AMOVA, the 12 neighboring non-Pygmy populations (F_{ST} = 0.004, 10,000 permutation $p < 0.001$). Interestingly, we found a much smaller proportion of the total variance explained by linguistic categories as we found a non-significant F_{CT} = 0.0025 (10,000 permutation $p = 0.042$) in a hierarchized AMOVA grouping all 21 populations into 7 linguistic groups (**Figure 2.1.a left panel colors**), without distinction in Pygmies and non-Pygmies.

Furthermore, investigating pairwise-F_{ST} values (Weir and Cockerham 1984) across all pairs of 21 populations, and setting non-significant (10,000 permutations) such values to 0, we found that all Pygmy populations where significantly differentiated from one another except the two South Cameroon Baka SBK and CBK populations. Conversely, we found that 68.2% of pairwise F_{ST} values across non-Pygmy populations were not significant. **Figure 2.1.b** shows Principal Coordinates Analyses (PCoA) projections

[36] The molecular lab had been built prior to my PhD in the Musée de l'Homme by Dr. Patricia Balaresque under the supervision of Pr Evelyne Heyer and, together with Myriam Georges, we inaugurated the first DNA extractions and PCRs conducted herein in 2006.

of this population-pairwise F_{ST} matrix further visually illustrating that most genetic variation in our dataset was accounted for by major differentiation across "Pygmy"-labelled populations and that, conversely, non-Pygmy populations were much more closely resembling one another. In these plots, the greatest genetic differentiation between pairs of populations on average along the first PCoA axis occured between, roughly, the four Baka Pygmy populations and the group of non-Pygmy populations, while the second and third PCoA axes both distinguished populations labelled as Pygmies. Importantly here, note that some Pygmy populations, such as the Gabonese Bongo populations, were more resembling non-Pygmy, and that other populations such as the Kola and the Bezan were found at intermediate distances between the Baka and the Bongo along the first PCoA axis. It is important to note in this context that pairwise-F_{ST} patterns here observed did not stem from vast differentiation in loci's average heterozygosities (Nei 1978) between Pygmy and non-Pygmy populations (*He* = 0.736 SD = 0.012 across 9 populations in Pygmies, and *He* = 0.741 SD = 0.007 across 12 neighboring non-Pygmy populations, Wilcoxon two-sided rank sum test p = 0.30), albeit these latter results further showed greater *He* variation across Pygmy populations than across non-Pygmy neighbors. Finally, as expected from the linguistic AMOVA, we see no correlation between population genetic differentiation patterns and linguistic groupings: linguistic categories do not translate into genetic variation differentiation across populations in Western Central Africa.

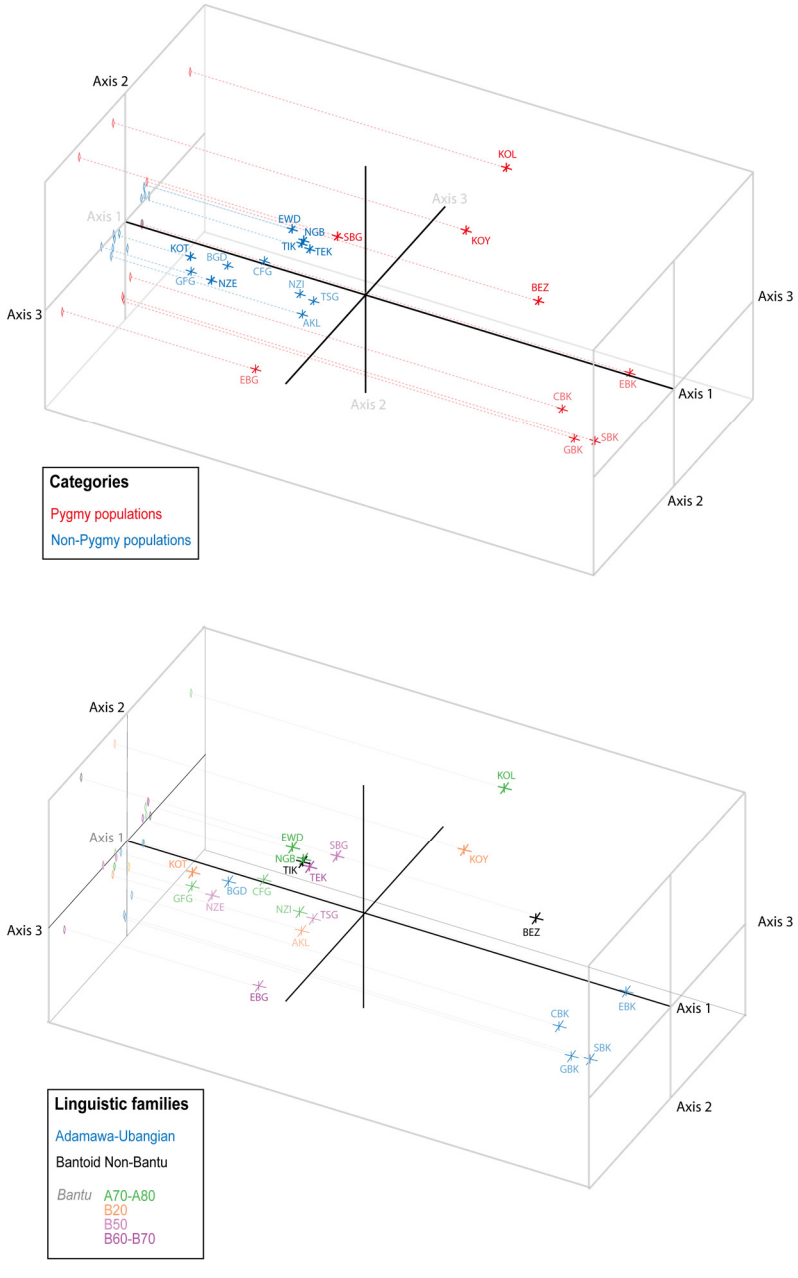


Figure F2.1.b.
Principle Coordinates Analyses based on population pairwise F_{ST} (Weir and Cockerham 1984) calculated for 28 independent autosomal tetranucleotide microsatellites. Population categorization in Pygmies and non-Pygmies are based on numerous historical and cultural criteria without considering phenotypic information (see **Chapter 1.1**).
Top panel was previously published in Verdu et al. *Current Biology* (2009).
Bottom panel is the same genetic differentiation PCoA projection colored in linguistic affiliations instead. Note that this latter panel was, in essence, originally published in a slightly different analysis in Verdu and Destro-Bisol *Human Biology* 2012.
Colors and population labels can be found in Figure F2.1.a above.

Finally (**Figure 2.1.c**), note that these results could be qualitatively expanded at the scale of the entire Congo Basin when considering three additional Aka Pygmy populations (BIA, AKA, AKM) from Central African Republic, Mbuti (MBU) Pygmies from Eastern Democratic Republic of Congo, and Nsua Efe (NSU) Pygmies from Western Uganda (see Map in **Figure 2.1.a right panel**). Importantly, considering interindividual pairwise Allele Sharing Dissimilarity (Bowcock et al. 1994) calculated from 153,798 autosomal SNPs in another, later (Pemberton, Verdu et al. 2018), study using an overlapping sample set, rather than the microsatellite markers used above, we found very similar patterns, also consistently found in other studies considering neutral short sequences dispersed across the genome (Patin et al. 2009), or whole exome sequences (Lopez et al. 2018, 2019). Note, here, that not only certain Pygmy populations are genetically more resembling non-Pygmy groups, but this result expands to inter-individual genetic variation where some Pygmy individuals resemble more some non-Pygmy individuals, even at reduced geographical scales in Western Central Africa only.

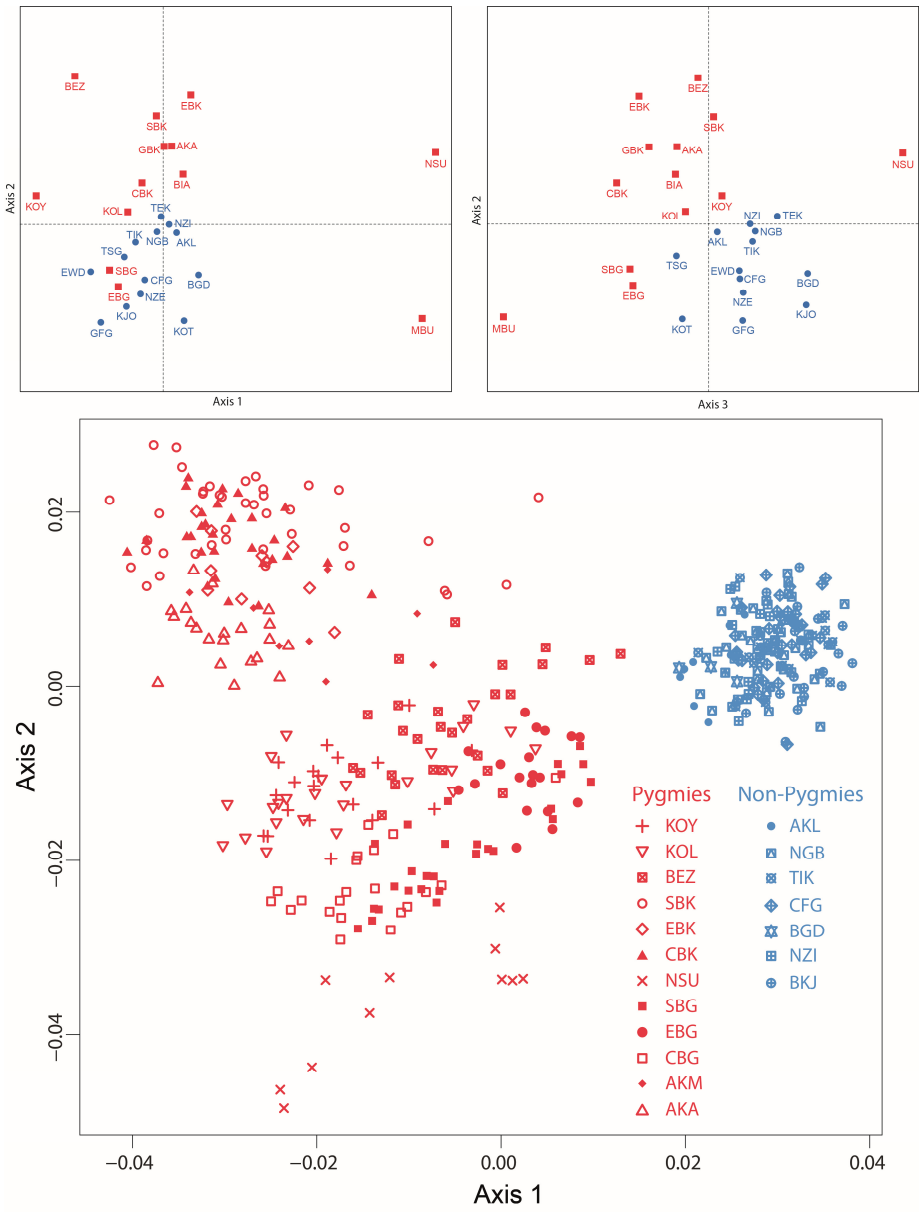


Figure F2.1.c.
Top panels: Multi-Dimensional Scaling based on population pairwise F_{ST} (Weir and Cockerham 1984) calculated for 28 independent autosomal tetranucleotide microsatellites, data here re-analyzed based on previous analyses in Verdu P. PhD Thesis (Université Pierre et Marie Curie – MNHN, Paris, France, 2009). ***Bottom panel***: Multi-Dimensional Scaling based on individual pairwise Allele Sharing Dissimilarities (Bowcock et al. 1994) calculated using 153,798 autosomal SNPs genome-wide from Pemberton, Verdu et al. *Human Genetics* 2018. Population categorization in Pygmies and non-Pygmies are based on numerous historical and cultural criteria without considering phenotypic information (see **Chapter 1.1**).

We then further explored inter-individual genetic variation using STRUCTURE unsupervised Bayesian clustering algorithms (Pritchard et al. 2000; Falush et al. 2003), whether using autosomal microsatellites or SNPs, and found interesting patterns of shared genetic resemblance across groups of individuals (**Figure 2.1.d**). Indeed, in all cases, this powerful MCMC-based clustering algorithm systematically found that individuals categorized *a posteriori*[37] as Pygmies had, on average, higher levels of genotypes' membership to one (red) virtual genetic cluster, while non-Pygmy individuals *a posteriori* clustered more substantially to the alternative (blue) cluster. However, and most importantly, neither red nor blue clusters were represented close to 100% in large groups of individuals. Instead, a substantial amount of blue can be found in all individuals categorized as Pygmies, and a substantial amount of red can be found in all individuals categorized as non-Pygmies. Furthermore, while membership to the red cluster is relatively low and little variable across individuals within each non-Pygmy-labelled population, the amount of blue in most Pygmy individuals is substantially higher on average; and, interestingly, on average variable across the various Pygmy populations: Baka Pygmies have on average lower genetic memberships to the blue cluster while Bongo Pygmies have on average much higher such memberships. This pattern would be expected in the case of asymmetric and heterogeneous admixture across Pygmy and neighboring non-Pygmy populations, where Pygmy populations would be more introgressed by the non-Pygmy gene-pool than the opposite, and where each Pygmy population would be introgressed to a different extent on average. However, such apparent STRUCTURE patterns could also be explained by other scenarios where each Pygmy population would have diverged from their neighboring non-Pygmy population at different times in the recent past and experienced different genetic drift since then. Therefore, whether such patterns stem from differential admixture or differential origins and drift remained, at that point of our descriptions of autosomal genetic patterns, to be assessed.

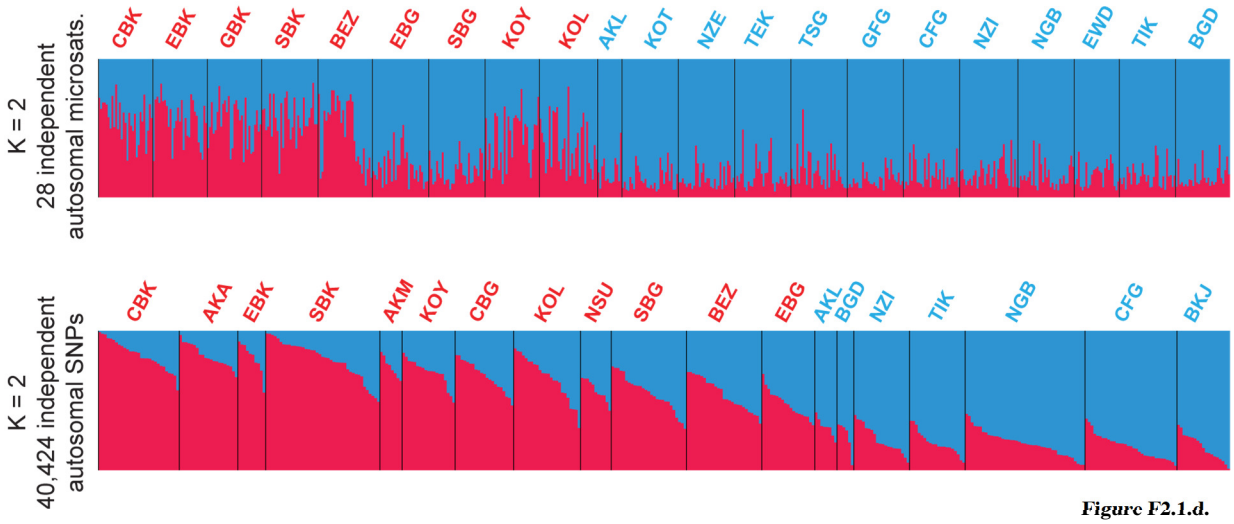


Figure F2.1.d.

Top panel: STRUCTURE analysis obtained for 28 autosomal tetranucleotide markers genotyped in 607 individuals from Western Central Africa only (Figure F2.1.a right panel), previously published in Verdu et al. *Current Biology* 2009.
Bottom panel: STRUCTURE analysis obtained for 40,424 independent (low LD) genome-wide SNPs genotyped in 406 individuals from Western to Eastern Central Africa only (Figure F2.1.a left panel), previously published in Pemberton, Verdu et al. *Human Genetics* 2018.
Each individual is represented by a single vertical line divided into 2 colors (red and blue) corresponding to the relative proportion of genotypes assigned to either virtual cluster inferred for *K*=2 with this software based only on allele frequencies calculated across all individuals. Population categorization in Pygmies (labelled in red) and non-Pygmies (labelled in blue) are based on numerous historical and cultural criteria without considering phenotypic information (see **Chapter 1.1**).

[37] As a reminder, note that unsupervised clustering methods implemented in STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) or ADMIXTURE (Alexander et al. 2009), do not consider population information in their analyses, but only inter-individual genetic resemblances. Individuals in these analyses are grouped *a posteriori* into pre-categorized populations each with a given label but this information does not participate in the obtained results themselves.

Altogether, our results based on autosomal data and those of our collaborators as well as those of other independent research groups, indicated since our initial work in 2009 that

1. Pygmy populations are more differentiated from one another than non-Pygmy populations; with Eastern Nsua and Mbuti Pygmies being highly differentiated from Western Baka and Aka, and all other Western Congo Basin Pygmy populations being at intermediate distances between these two groups of populations, on average.
2. Non-Pygmy individuals and populations are highly resembling one another, even at the scale of the entire Congo Basin.
3. Importantly, some Pygmy populations are much more genetically resembling non-Pygmy populations than others, such as the three Bongo populations from Southern Gabon.
4. Linguistics categories are un-informative about genetic differentiation across populations throughout the Congo Basin.
5. Considering inter-individual genetic differentiation, there are signals of possible admixture between Pygmy and non-Pygmy groups of populations, *i)* largely asymmetrical with more non-Pygmy admixture within Pygmy groups than the opposite, and, *ii)*, heterogeneous in intensity across Pygmy groups. Such complex admixture patterns are nevertheless only indicated in our analyses and should be further formally tested as the observed patterns could alternatively stem from other evolutionary mechanisms.

Therefore, as a result of these descriptive analyses, we can conclude that the Pygmy/non-Pygmy binary categorization of Congo Basin individuals based on numerous historical and cultural criteria translates into extensive average genetic differentiation between groups of individuals sampled from the two categories respectively. Nevertheless, this binary categorization system hides tremendous genetic variation as certain Pygmy individuals and populations are resembling non-Pygmy individuals and populations, while others are much more distant genetically. Furthermore, we found tremendous genetic variation across populations within the Pygmy category, and much more genetically homogeneous non-Pygmy populations, even at the scale of the entire Congo Basin. If this result advocates for a common recent origin of all populations labelled as non-Pygmies, the single term Pygmy *de facto* hides tremendous genetic variation further questioning the possible common origins of all populations categorized as such during history. Altogether, genetic differentiation results within the group of Pygmy populations echoes the vast cultural diversity recorded by ethnologists and cultural anthropologists across historical Pygmy groups throughout the Congo Basin as well as at more regional scales (Verdu et al. 2009, 2012; Verdu 2014, 2016).

Finally, note that, as an answer to one of the core questions brought by ethnologists about linguistic diversity throughout the Congo Basin, linguistic resemblance among populations does not predict genetic resemblance among individuals and populations (Verdu and Destro-Bisol 2012). This is due to the fact that Pygmy populations most often speak a language closely related to that of certain non-Pygmy immediate neighboring populations and, since Pygmy populations on average substantially differ genetically from these neighbors, there is no correlation between genetic patterns and linguistic diversity locally in the Congo Basin. This would further suggest that, unlike in numerous other worldwide populations but also far from exceptional among *Homo sapiens* populations, while language differences necessarily participate to identity construction in Central African societies, they do not translate into marital and reproductive preferences or segregation between groups.

2.2. Reconstructing the origins of Central African Pygmy and neighboring non-Pygmy populations using Approximate Bayesian Computations

In this complex anthropo-genetic context, we set out to formally test ***i)*** whether Pygmy populations had a more recent common origin together than their common origin with non-Pygmy populations or, alternatively, if each Pygmy population had an independent genetic origin from one another. Furthermore, ***ii)*** we aimed to test formally whether asymmetrical and heterogeneous admixture occurred between the two groups of populations, or not, in order to explain the observed genetic patterns (see also introduction in **Chapter 1.1**).

2.2.a Briefly introducing Approximate Bayesian Computation statistical inference

Inferring highly complex evolutionary histories from genetic data is not trivial using classical maximum likelihood phylogenetic approaches. Indeed, first, classical such approaches cannot accommodate at the same time numerous population samples with branch-specific changes in effective population sizes and possible gene exchanges, via migration or admixture events, across all pairs of populations throughout history. Such models are often too complex for simply writing the likelihood to be maximized, and even if possible, maximizing this likelihood in a convergent way can be extremely challenging computationally, if not simply out of reach (Foll et al. 2015; Ni et al. 2019). Furthermore, maximum-likelihood approaches' performances to fit a given model to a genetic dataset are, in essence, hard to hierarchize when comparing highly complex models with different numbers and classes of parameters (divergence times, admixture events, effective population size changes), in particular across models performing, *a posteriori*, reasonably similarly to fit the data (Alexander et al. 2009; Gravel 2012). These fundamental limitations of maximum-likelihood approaches were structural when I started the formal reconstruction of the evolutionary history of Central African populations in 2007, and they still remain largely challenging today, even if massive computational and methodological improvements have been achieved since then concerning this specific class of population genetics inference methods.

In this context, a novel statistical inference approach, born in 1997 and rapidly expanding in population genetics at the beginning of the 2000's, seemed promising to investigate the above questions of interest: Approximate Bayesian Computation (Tavaré et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002). As ABC represents a class of inference methods at the core of my research endeavors (see also **Chapters 4** and **5**), I briefly synthetically explain bellow how ABC works (note that I provide a schematic representation in **Chapter 4**):

1. Simulate genetic data under a variety of highly complex genetic scenarios (or models) in competition and assumed a priori to explain the data, by drawing model-parameter values for each simulation in prior distributions set explicitly by the user. This thus result in a vector of parameter value corresponding to a vector of simulated genetic data, for each simulation under each competing scenario.
2. For each simulation separately, calculate a set of summary-statistics thought a priori to be informative about the model-parameters of interest. For instance, population genetics theory classically predicts that genetic differentiation F_{ST} (Weir and Cockerham 1984) between pairs of populations is related to drift parameters (effective population sizes and mutation rates), as well as divergence times and migrations. We can thus use F_{ST} a priori for estimating these parameters with ABC. This step results in a vector of

parameter values for a given simulation associated with a vector of summary statistics calculated on the obtained simulated genetic data, for each simulation under each competing scenario.

3. Calculate the exact same set of summary statistics on the observed real data (thus only once).
4. ABC allows to formally determine which competing scenario produces simulations for which pseudo observed genetic data produced summary statistics closest to the real observed ones. Several such ABC methods for scenario-choice have been produced and rely on simple rejection (*a.* calculate multidimensional Euclidean distances between the vector or observed summary statistics and the vectors of summary statistics calculated separately for each simulated data; *b.* define a threshold value of proximity and retain all simulations within a shorter distance than the chosen threshold, *c.* evaluate the proportion of simulations obtained from each scenario, respectively, for this subset of "closest" simulations); or more complex, accurate, and powerful approaches, such as logistic or local linear regressions (Beaumont et al. 2002), or machine-learning Random Forest or Neural Network more recently (Csilléry et al. 2012; Pudlo et al. 2016).
5. Under the winning scenario identified in procedure 4. ABC allows to estimate the posterior distribution of model-parameters producing simulations for which summary statistics are closest to the observed ones. Based only on simulations from a single scenario, this procedure, again, identifies the closest sets of simulations to the observed data and uses rejection, regression, or machine learning approaches (among others) on the vectors of parameter values used originally for each simulation to estimate posterior distributions for each original or composite parameter in the model.

Therefore, ABC can be used, in principle, to infer the parameters of the models that best mimic the observed data for evolutionary scenarios of arbitrary complexity, provided that one can simulate efficiently enough data to cover the (often large) parameter space and calculate informative summary statistics[38]. In practice, ABC is often limited by the dimensionality of the parameter space that can be hard to explore thoroughly even with very large numbers of simulations, and by the informativeness of statistics that may render certain parameters in the models identifiable while others not (Sisson et al. 2018). More conceptually, ABC has an obvious limitation: one only tests the models that one can think of. Indeed, ABC does not allow to explore unspecified possibilities in the realm of evolutionary scenarios. Rather, ABC allows formally testing hypotheses in the form of clearly defined competing scenarios, with associated prior parameter-distributions set by the user, thought a priori to be underlying the observed data. This nevertheless represents, to my personal views, an ideal framework for most my research endeavors. Indeed, I admit that I am not after the "truth" (which I believe is inaccessible in the case of evolutionary genetic history, at least yet). I am rather interested in comparing the possible influence of different scenarios, all oversimplifying the "real" history, on the observed genetic patterns, within the population genetics paradigm of the synthetic neo-Darwinian theory of evolution (Estoup et al. 2018).

Finally, and importantly, note that while more simple models can also be inferred from genetic data with ABC, if a model can be reasonably treated via likelihood-maximization procedures, there is no reason to conduct instead a full-blown ABC inference. Despite ABC practicality and, I think, intuitive dialectics, it is far less mathematically elegant and philosophically grounded. As its' name reminds the user, ABC relies on a strong approximation, namely approximating the real data by a set of summary statistics calculated on the real data, and cannot therefore be preferred to methods allowing directly to infer the parameters of evolutionary models from the observed genetic data, when possible.

[38] Note that this is why ABC is now often used beyond population genetics, in economy for instance, and why we further developed ABC for linguistic inferences during Valentin Thouzeau's PhD that I co-supervised with Frédéric Austerlitz (see **Chapter 5**).

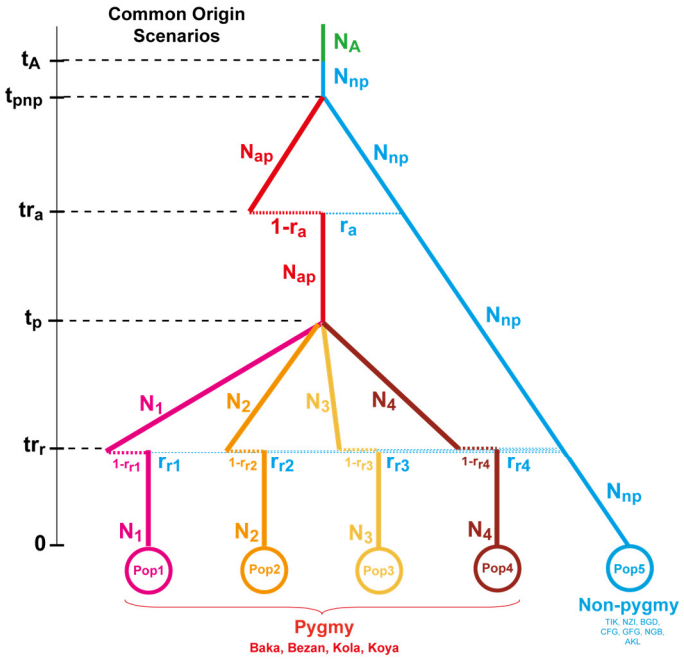
It is important for me to acknowledge here that I initially acquired knowledge and skills in ABC inferences mostly thanks to Dr. Arnaud Estoup from the CBGP lab in Montpellier. He kindly and patiently hosted me for several trips to his lab between 2007 and 2008, in order for me to learn and conduct the ABC scenario-choice and posterior parameter estimation procedures adapted to my Central African project. We did so, using the first beta version, not released at the time, of a massive software package that Dr. Estoup and colleagues had developed and that is still extensively used in the population genetics community, Do It Yourself ABC (Cornuet et al. 2008), and that allowed us to do all of the above with a single integrated computational tool. Our work on Central African human populations published in early 2009 represents one of the very first application of ABC approaches to human genetic data. Since then, we have worked on several occasions with Dr. Estoup who developed specific novel ABC tools with help from me, as well as published a book chapter in the main ABC text-book edited by Beaumont et al. and released in 2018 (Estoup et al. 2019).

2.2.b Approximate Bayesian Computation inferences in practice in Central Africa

Based on these premises, we elaborated (Verdu et al. 2009) eight competing scenarios (different versions of the two most complex scenarios shown in **Figure F2.2.a** below), for explaining the evolutionary history underlying the observed genetic data that we previously described for 28 autosomal tetranucleotide microsatellites genotyped for more than 600 individuals from the Western part of the Congo Basin (see **previous section 2.1** and **Figures 2.1.a-d**).

Four competing scenarios considered a topology where the various populations categorized as Pygmies based on cultural criteria share a common ancestry preceded by the common ancestry between the ancestral Pygmy lineage and the lineage giving birth to the non-Pygmy neighboring populations. Four other scenarios considered that each Pygmy population derive from the gene-pool that give birth to non-Pygmy populations independently, and potentially at different times.

All eight scenarios considered possible admixture events occurring between the Pygmy and non-Pygmy lineages: both after and before the common origin of Pygmy populations in the four "common origin" scenarios, and two separate admixture events occurring at any time along each Pygmy population lineage after it's divergence from the non-Pygmy lineage, for the four "independent origins" scenarios. In four scenarios (two with common origins and two with independent origins of Pygmy populations), admixture intensity could take values, independently for each such event, randomly between 0.1% and 99.9%; meaning that, for 99.9% admixture, the gene-pool of the targeted Pygmy population would be almost completely replaced by that of the non-Pygmy lineage, while 0.1% meant that virtually no admixture occurred at this event. Conversely, in the four remaining scenarios, we set all admixture rates to be equal to zero in all simulations, in order to simulate genetic data that did not experience admixture under these scenarios. Comparing both sets of scenarios would allow us to formally test if admixture was involved during the evolutionary history of our populations or if genetic patterns could be satisfactorily explained without this phenomenon, whichever the common or independent topology.



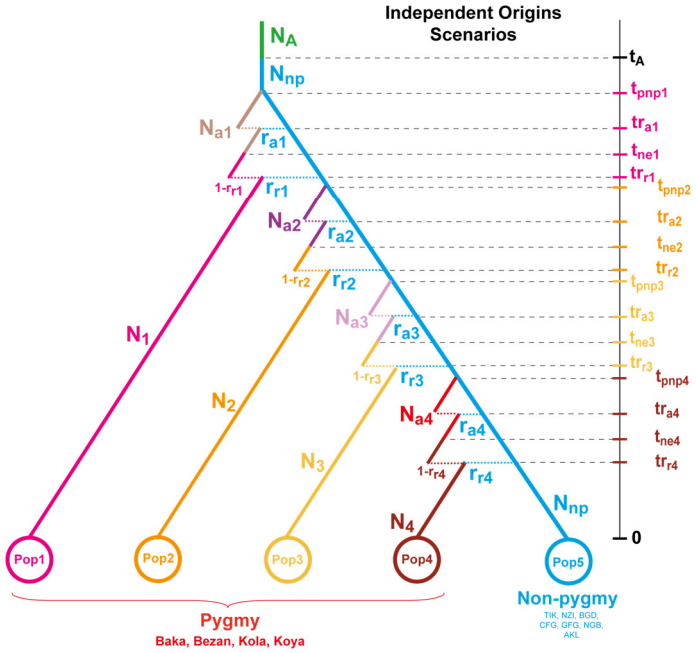


Table T2.2.a.
Model-parameters prior distributions.

Common Origin scenarios	Description	Prior Distribution	Condition
N_1 (Baka)	Effective pop size	Uniform[10-10,000]	
N_2 (Bezan)	Effective pop size	Uniform[10-10,000]	
N_3 (Kola)	Effective pop size	Uniform[10-10,000]	
N_4 (Koya)	Effective pop size	Uniform[10-10,000]	
N_{np} (Non-Pygmies)	Effective pop size	Uniform[10-100,000]	
N_{Ap} (Ancestral Pygmy)	Effective pop size	Uniform[10-10,000]	
N_A (Ancestral population)	Effective pop size	Uniform[10-10,000]	
t_{rr}	Admixture time	LogUniform[1-5,000]	trr<tp
t_p	Divergence time	Uniform[1-5,000]	trr<tp
t_{ra}	Admixture time	Uniform[1-5,000]	tp<tra
t_{pnp}	Divergence time	Uniform[1-10,000]	tra<tpnp
t_A	Time of change in Ne	Uniform[1-10,000]	
r_{r1}	Admixture rate	Uniform[0-1]	
r_{r2}	Admixture rate	Uniform[0-1]	
r_{r3}	Admixture rate	Uniform[0-1]	
r_{r4}	Admixture rate	Uniform[0-1]	
r_{r5}	Admixture rate	Uniform[0-1]	
r_{r6}	Admixture rate	Uniform[0-1]	
r_a	Admixture rate	Uniform[0-1]	
μ	28 Mutation rate param.	Uniform[10^{-4}-10^{-3}]	
p	28 Mutation rate param.	Uniform[0.1-0.3]	

Independent Origins scenarios	Description	Prior Distribution	Condition
N_1 (Baka)	Effective pop size	Uniform[10-10,000]	
Na_1 (Baka)	Effective pop size	Uniform[10-10,000]	
N_2 (Bezan)	Effective pop size	Uniform[10-10,000]	
Na_2 (Baka)	Effective pop size	Uniform[10-10,000]	
N_3 (Kola)	Effective pop size	Uniform[10-10,000]	
Na3 (Baka)	Effective pop size	Uniform[10-10,000]	
N_4 (Koya)	Effective pop size	Uniform[10-10,000]	
Na4 (Baka)	Effective pop size	Uniform[10-10,000]	
N_{np} (Non-Pygmies)	Effective pop size	Uniform[10-100,000]	
N_A (Ancestral population)	Effective pop size	Uniform[10-10,000]	
$t_{rr(1-4)}$	4 Admixture times	LogUniform[1-5,000]	trr<tra
$t_{ra(1-4)}$	4 Admixture times	Uniform[1-5,000]	trr<tra
$t_{pnp(1-4)}$	4 Divergence times	Uniform[1-10,000]	tra<tpnp
t_A	Time of change in Ne	Uniform[1-10,000]	
$r_{r(1-4)}$	4 Admixture rates	Uniform[0-1]	
$r_{a(1-4)}$	4 Admixture rates	Uniform[0-1]	
μ	28 Mutation rate param.	Uniform[10^{-4}-10^{-3}]	
p	28 Mutation rate param.	Uniform[0.1-0.3]	

Figure F2.2.a.
Two competing scenarios for ABC scenario-choice and posterior parameter inferences. Model parameters' prior distributions are indicated in the adjacent **Table T2.2.a**. Model descriptions can be found in text.
Alternative scenarios are described above and can be found in detail and schematically represented in Verdu et al. *Current Biology* 2009 Supplementary Materials.
Figure originally published in Verdu et al. *Current Biology* 2009.

Finally, four scenarios comprised the possibility of an instantaneous expansion of effective population size at any time in the history of the non-Pygmy lineage, while the four other scenarios did not comprise such possibility and the effective population size of the non-Pygmy population remained constant throughout history. This was introduced in our models as numerous previous archaeological, linguistic and genetic works investigated the demographic and genetic expansion of non-Pygmy Bantu-speaking populations. The ill called "Bantu-expansion"[39] was reported to have started at least 5000 years ago with the emergence of agriculture in the region and the concomitant spread of Bantu-speaking populations from Western Cameroon into the rest of Central, Eastern and Southern Africa (e.g. Vansina 1995; Phillipson 2005).

We thus wanted to formally test this hypothesis to find if such scenario *a posterior* best explained our data, or if such scenarios were un-identifiable with our data. Note that for all eight scenarios, Pygmy lineages can each independently punctually change effective population size when diverging from the non-Pygmy lineage or among themselves.

Note that in this initial work, we did not consider the evolutionary history of non-Pygmy neighbors and pooled altogether our samples in a single "population". This simplification was rendered possible as our descriptive analyses showed that, with our data, these populations were barely distinguishable in a first place. As ABC relies on summary statistics informativeness, we already knew that our data was insufficient to capture the evolutionary history of the diverse non-Pygmy populations; a question nevertheless of major interest and which we addressed with vastly more extensive data in other, later, papers produced by Pr. Lluis Quintana-Murci and Dr. Etienne Patin (Pasteur Institute), and their colleagues, to which we had the chance to contribute (see a summary in **section 2.5** below). Similarly, the four Baka Pygmy populations highly resembled one-another and were pooled in a single "Baka" population in these analyses for very similar reasons. Finally, note that the two Bongo Pygmy populations highly resembled non-Pygmy neighbors in our descriptive analyses. We thus designed specific treatments for these populations and whose technics will not be detailed here, but such specificities can be found in extenso in the original 2009 *Current Biology* article.

For each eight scenarios, we performed 500,000 separate simulations drawing parameter values in prior distributions (see **Figure F2.2.a** and associated **Table T2.2.a**), using the classical coalescent backward-in-time simulator implemented in DIY ABC. For each 4 million simulations separately, we calculated, using DIY ABC, 35 different summary statistics, considering within-population statistics such as average Heterozygosities (Nei 1978) or average number of different microsatellite alleles and their variances across loci, as well as statistics calculated between pairs of populations such as F_{ST} (Weir and Cockerham 1984) and Goldstein's pairwise molecular distance $(\delta\mu)^2$ (Goldstein 1995).

We used regression ABC scenario-choice procedures and found that a scenario of common origins of all Western Central African Pygmy populations presented in the top panel of **Figure F2.2.a** largely best explained our data with ABC posterior-probability superior to 95%. Importantly, note that all other scenarios largely failed to best explain the observed summary statistics, with a low type 2 error of wrongly choosing a given scenario *a priori* using in turn 1000 random simulations under each scenario as pseudo-observed data for ABC scenario-choice (0.042 on average ranging from 0.000 to 0.090 across scenarios). Moreover, we found that to best explain the data, the winning scenario had to incorporate a possible increase

[39] Another illustration of how shortening concepts to be used as labels may introduce categorization issues, the linguistic category here transforming surreptitiously into a population name or identifier and subsequent demographic processes attached to its' prejudiced history, although no Bantu population exist as this linguistic category spans local populations speaking more than 300 languages throughout half of one of the largest continents worldwide.

in effective size occurring at some point in the non-Pygmy lineage. Finally, scenarios where admixture events were set to intensities of 0 throughout the history of Central African populations largely lost ABC scenario-choice with respect to the observed data. In other words, ancient and more recent admixture need to occur in our scenarios in order to best mimic the observed data. Finally, note that we obtained highly similar results for our treatments involving the two Bongo Pygmy populations (see Verdu et al. 2009 Supplementary materials), hence pointing to their common origin with all other Central African Pygmy populations despite their high levels of genetic resemblance with non-Pygmy neighbors.

Based on this winning scenario, we then used regression ABC to estimate, *a posteriori*, the distributions of each model-parameters used for the 1% simulations that produced vectors of summary-statistics closest to our observed data (**Figure F2.2.b**). Altogether, numerous parameters could be inferred satisfactorily as they substantially departed *a posteriori* from the prior distribution of parameter values used for simulating data as their 95% Credibility Intervals were relatively narrow.

This was the case for effective population size parameters that indicated that all Pygmy populations did not substantially changed sizes since their common ancestral population. Conversely, our results advocated for a strong increase in effective size in the non-Pygmy lineage.

Furthermore, not only our results qualitatively indicated the need for admixture to have occurred in the evolutionary history of Pygmy and neighboring non-Pygmy populations, but we estimated that admixture was heterogeneous with Bongo populations being much more admixed than Baka populations for instance.

Concerning divergence times, our results strongly supported a very recent common origin for all Western Central African Pygmy populations between 2,600 years before present (95% CI: 725–34,275) and 3,500 YBP (95% CI: 870–41,139) considering a generation time of 25 years or 30 years and based on the modal posterior point estimate for this parameter[40].

Such recent and apparently strong isolation across populations sometimes not very distant geographically was somewhat unexpected in particular for hunter-gatherer populations often reported to have vast exploration ranges and extensive mobility. Thus, we investigated further at an extremely local scale for the three Cameroonian Baka populations in our sample sets, patterns of Isolation-By-Distance following Rousset et al. (1997) methods. In brief, Pr. Rousset and colleagues demonstrated that, at very reduced geographical scale, the slope of the linear correlation between interindividual genetic differentiation and logarithmic geographic distances among individuals (birthplaces in our case) was inversely proportional to effective dispersal in the population (namely the relative distance between an individual birthplace and that of both his/her parents).

[40] Note that the prior distribution for this parameter has a mode (due to constraints on time parameters it is not Uniform albeit drawn from a Uniform distribution) of 9,775 YBP and a 95% CI between 2,050 and 90,875 YBP considering a generation time of 25 years.

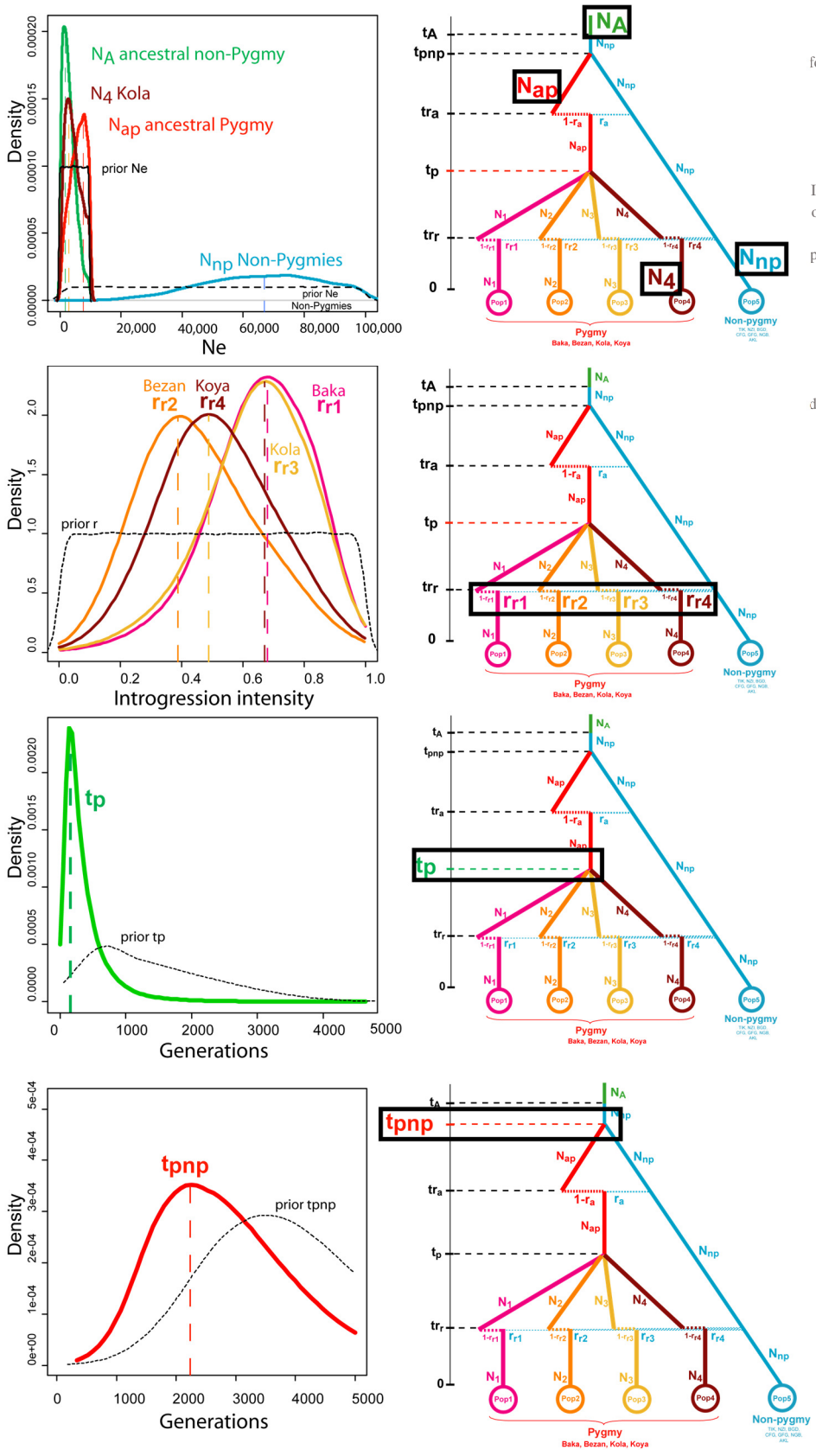


Figure F2.2.b.

ABC estimations of the posterior distribution of four model-parameters under the winning scenario of common origin of Western Central African Pygmy populations, with and expansion of non-Pygmy effective population sizes and possible ancient and recent admixture events across lineages.

In all posterior distribution panels, the distribution of parameters used a priori to perform simulations is given in dashed black line while the parameters' posterior distributions are colored and correspond to parameters shown in the schematic scenario figure (on the left) by a black box.

Intuitively in ABC, we consider that our data and summary statistics withhold identifiable information about model parameters when the estimated posterior distribution substantially departs from their respective priors and when 95% Credibility Intervals are relatively small.

In this context, note here that, as an example, while our data seem to withhold substantial information for posterior estimation of the tp parameter, the posterior estimate of tpnp is only indicative of this parameter as 95% CI is wide.

Figure derived from results presented in a table format in Verdu et al. *Current Biology* 2009.

As our anthropological field-work yielded such detailed birthplace information at a very local scale, we applied this method and found, surprisingly, that Baka Pygmies had very reduced effective dispersal ranging between roughly 12 and 65 km² (Verdu et al. 2010; Verdu 2012). In other words, we found that Baka Pygmies, despite extensive exploration ranges and mobility in their environment for social, ritual, or economic reasons, nevertheless were born very close to where their biological parents were born. Furthermore, note that this result is in fact highly conservative, as any change in reproductive behavior in space or data collection mistakes would instantaneously erase this IBD pattern here observed. Note (*unpublished results*) that we obtained very similar results in the Bezan Pygmies, but not in the Kola albeit the latter were known to have experienced population displacements in the 1980s due to the construction of a pipeline likely having resulted in breaking the IBD signal at the following generation that we sampled.

Altogether such reduced effective dispersal could explain the reproductive isolation among Pygmy populations even at reduced geographical scale and despite large exploration range. Nevertheless, we still do not know since when such dispersal behavior exists in this population, which would need to be assessed in order to propose this mechanism as having contributed to the ABC inferred divergence history here identified.

Finally, despite relatively low information in our observed data, our results provided indications that the more ancient divergence between the ancestral Pygmy population and the lineage that gave birth to non-Pygmy neighbors occurred between 60,000 and 90,000 years ago, albeit, again, with large 95% CI.

Despite these positive results, several parameters could not be well estimated although they contributed to scenario-choice and thus behaved similarly as nuisance-parameters. This was the case for admixture times that were not satisfactorily inferred *a posteriori*, either the most recent time or the more ancient admixture time. Finally, although admixture intensities could be relatively well estimated in most cases, the posterior distribution for the ancient admixture estimate was seldom departing from its' prior. These results show either that we lacked information in our data to satisfactorily identify these important parameters, and/or that scenarios oversimplifications here considered (e.g. a single admixture event in the past over a very large period of evolution), probably mis-fitted the data, although not considering any such events would result in even poorer fit to the data. This may also lead to relatively discrepant results between our inferences and our initial descriptive analyses, for certain aspects. For instance, if STRUCTURE results can be interpreted as due to admixture events as suggested by our ABC results, then we obtain certain ABC posterior estimates, for instance for the Baka, that are higher than what can be inferred from STRUCTURE results. This could be due to constraints on the parameter space for oversimplified scenarios with a single divergence time for all Pygmy populations that could possibly need to be compensated for some populations by a trade-off on posterior estimates of other parameters such as drift and/or admixture intensities. Ultimately, this advocates for further complexifying models which may then hit the dimensionality wall as we will have a hard time thoroughly and evenly explore very large spaces of parameters, as well as be strongly limited in all cases by the lack of information inherent to our very modest marker set used at the time.

Finally, we synthesized our findings from this work and beyond about the neutral demographic history of Central African peopling in a schematic way in **section 2.4** below.

2.3. Complex sex-biased admixture processes between Central African Pygmy and neighboring non-Pygmy populations inferred with Approximate Bayesian Computations

We deployed the same dialectics and inference methodology described above more specifically to investigate admixture processes across Central African populations. Indeed, our previous results established that heterogeneous and largely asymmetrical admixture was at play between Central African populations categorized as Pygmies based on numerous historical and cultural criteria, and the neighboring non-Pygmy populations with whom they share complex socio-economic relationships locally. Importantly, these findings were at odds with the common Western mis-representation of Hunter-Gatherer Pygmy populations living hidden remotely in the heart of the equatorial rainforest, and thus thought to be genetically isolated from their neighbors. Again, as detailed in introduction in **Chapter 1.1**, ethnologists had repeatedly highlighted that this prejudice was indeed a prejudice extremely far from the reality of numerous and complex socio-cultural and economic interactions between Pygmies and neighboring non-Pygmies, as well as the diversity of such relationships across pairs of populations throughout the Congo Basin.

In this context, a review of the ethnographic literature, exchanges with ethnologists including shared experiences on the field in Central Africa, and relatively intense training in cultural anthropology methods and approaches allowed me to further disentangle the socio-cultural behaviors at play between these communities, which allowed us to propose population genetics expectations based on different sociological scenarios regarding intermarriages and reproduction.

As a summary (Kazadi 1981; Bahuchet 1992a; Bahuchet and Guillaume 1992; Joiris 2003; Hewlett 2014; Verdu 2014), complex socio-cultural relationships between Pygmies and neighboring non-Pygmies locally include complex socio-marital and reproductive behavior relationships. Pygmy communities are very often strongly discriminated against by their socio-economically dominant non-Pygmy neighbors, both in reality and in self-reported endogenous discourses and representations. It is very common that non-Pygmy communities represent Pygmy communities as barely human in a classical racist way, where the other-than-self shows cultural differences which are then rendered innate and further hierarchized as lower than self. As a consequence, for non-Pygmies, marriage with a Pygmy individual is often socially detrimental and reported to be largely avoided. However, both endogenous representations and their prevalence in realized behaviors are more complex (**Figure F2.3.a**).

In fact, Central African populations mostly report patrilocal philopatry practices, which is to say that brides move, after marriage, to live close to their husbands' family, against a bride-compensation paid by the husband's family to the bride's family. In addition to socio-economic discriminations against Pygmies, this result in marriages between a Pygmy male and a neighboring non-Pygmy female being highly improbable and strongly detrimental for the non-Pygmy bride and her family. Beyond this endogenous representation, very few such marriages have been reported in the literature and I only witnessed it in a single instance across all my fieldworks. In other words, this endogenous representation and socio-cultural norm seems to be indeed normative and highly prevalent.

However, in this context, marriages between a Pygmy female and a non-Pygmy male are more often tolerated, as I also witnessed. Indeed, several contexts may favor such marriages as, for instance, the frequent case of an economically poor non-Pygmy male unable to pay bride- compensations within his community, who will possibly have more opportunities to find a Pygmy female for marriage as bride-compensations to be paid would likely be much less expensive. Another example emerges when a non-

Pygmy male wishes to establish a beneficiary alliance with a specific Pygmy family renowned for their hunting skills and/or medical and magical knowledge of the forest[41].

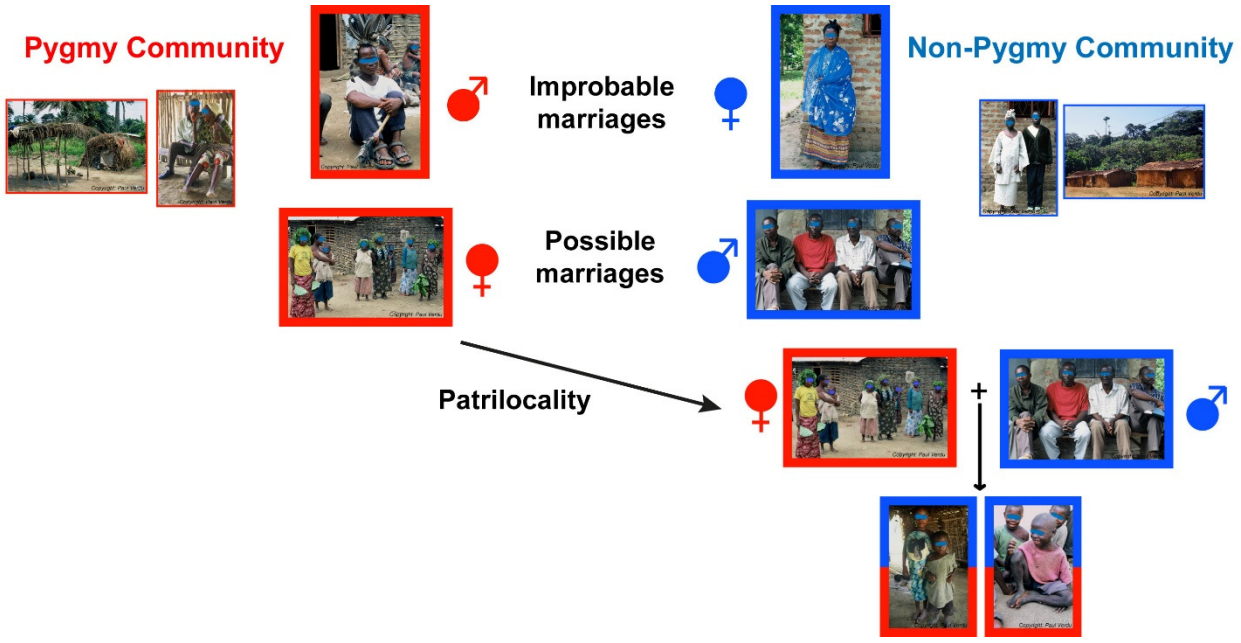


Figure F2.3.a.
Part 1 of the schematic representation of spouse and offspring mobility in intermarriages between Pygmies and non-Pygmies base on ethnographic investigations. Photographs ©*Paul Verdu*, Gabon and Uganda, 2006-2007.

Therefore, in the patrilocal context, the Pygmy bride moves to her non-Pygmy in-laws and give birth to admixed children within the non-Pygmy community. As illustrated above, this should result in genetic patterns opposite to what we observed and inferred in our data, with substantial amounts of "red" into the "blue", and much lesser "blue" into the "red" … Further investigating ethnographically, we noticed that these marriages almost systematically end up in a divorce within five years, mostly due to the strong discriminations experienced by the Pygmy female in the non-Pygmy community. Beware, divorces occur frequently in either community separately, but these specific inter-community marriages seem to always rapidly fail. As a consequence of the divorce, or in the case of the non-Pygmy husband's death, the Pygmy female returns to her community of origin together with her children also discriminated against in the non-Pygmy community (**Figure F2.3.b**).

Altogether, such purely socio-cultural process would result in substantial asymmetric genetic admixture from the non-Pygmy gene-pool into the Pygmy gene-pool, with a relatively much lower gene-flow from Pygmies into non-Pygmies that would correspond to exception to this reported socio-cultural behavior. Finally, note that illegitimate children between Pygmy females and non-Pygmy males, whether due to dramatic abuses or simply extra-marital reproductive events, would result in the same genetic patterns. Such scenario, albeit complex, would thus satisfactorily explain the observed genetic patterns and previous ABC inference results based on autosomes only.

[41] Note that Pygmy females are often attributed a fantastic fertility in the equatorial forest region where secondary infertilities are frequent. This myth, never formally demonstrated, may result in increased sexual abuses against Pygmy females, but was rarely presented as a reason for marriage in my investigations. I am still not sure how this representation about fertility is prevalent and normatively influencing reproductive behaviors across communities.

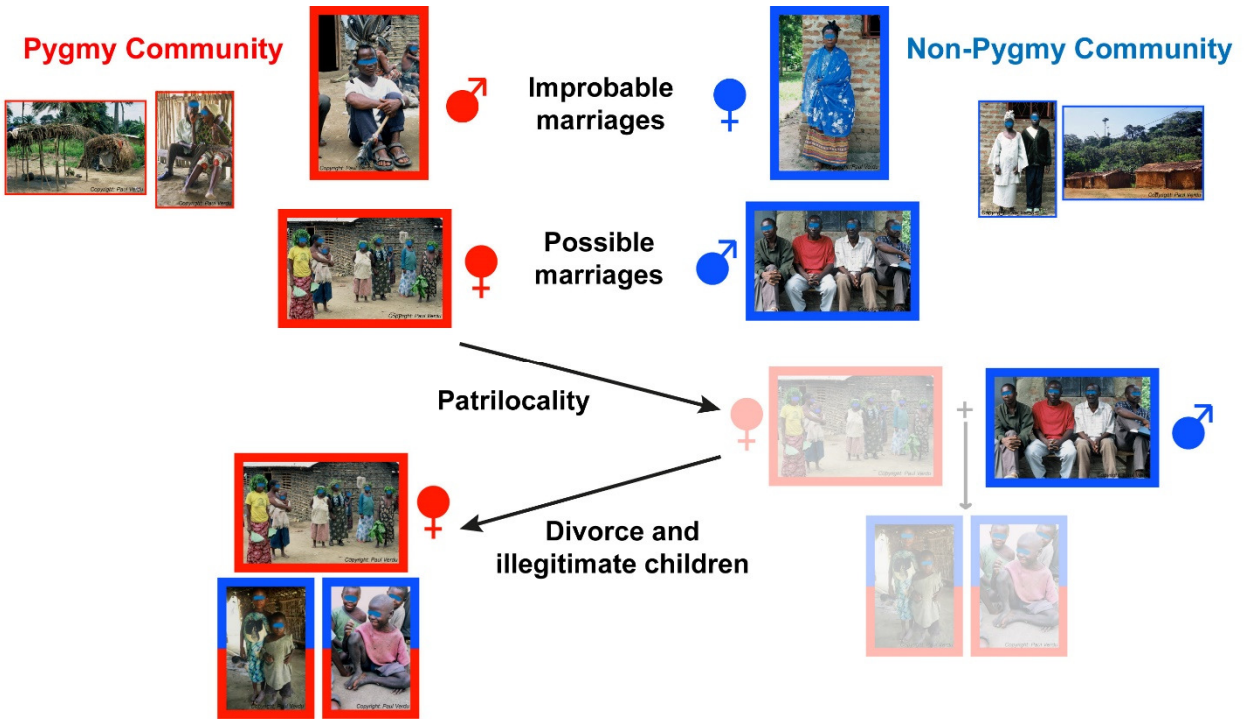


Figure F2.3.b.

Part 2 of the schematic representation of spouse and offspring mobility in intermarriages between Pygmies and non-Pygmies base on ethnographic investigations. Photographs ©*Paul Verdu*, Gabon and Uganda, 2006-2007.

Finally, how do we explain, under this framework, variation in introgression intensities across the various Pygmy groups throughout the Congo Basin? We observed qualitative variation across pairs of community in levels of discriminations against Pygmies reported by either the Pygmy victims and their non-Pygmy abusers. As a consequence, certain Pygmy communities, such as some Bongo groups, seem to be relatively more socially integrated with neighboring non-Pygmies than others, such as the Koya or the Baka for instances who are strongly discriminated against. Ultimately, the apparent strength in discrimination against Pygmies locally seem to qualitatively predict relatively well the observed intensity of genetic introgressions.

Based on these expectations, we set out to formally investigate patterns of sex-specific admixture and formally test different scenarios, using the same ABC dialectics and methodology as explained in the previous section (and which we will not further detail here).

Together with Master students Viola Grugni and Noémie Becker which I supervised, and again with the crucial technical support from Myriam Georges, we generated microsatellite data on our samples for the X-chromosome, for the non-recombining portion of the Y chromosome, as well as HVR-1 mitochondrial sequences (Verdu et al. 2013). This provided us with sex-specific genetic data which could be incorporated explicitly in the novel version of DIY ABC that Dr. Estoup and colleagues released in 2011. We therefore set out to evaluate, for each chromosomal compartment separately, ABC posterior parameter distributions specifically concerning male and female effective population sizes as well as male and female-mediated gene-flows in each Pygmy population, by considering simplified versions of the winning model previously identified, restricted to only three populations: two Pygmy populations and a single non-Pygmy sample merging the two neighboring populations. After extensive simulations specific to each population-samples trio for each chromosomal compartment separately, we conducted ABC posterior parameter

inferences for all pairs of Pygmy populations in our dataset (and associated neighboring non-Pygmy populations), and synthesized results obtained for each Pygmy population separately (Verdu et al. 2013).

Our ABC posterior parameter inferences showed that the complex sex-specific scenarios proposed by ethnographic work was indeed compatible with the observed genetic data. For compelling instances in addition to other results not discussed here, **Figure F2.3.c** shows that levels of female-mediated mitochondrial DNA admixture in each Pygmy population separately were always extremely low. In contrast, levels of male-mediated introgression via the Y-chromosome were always much higher. Finally, and most interestingly, while our hypothesis about the prevalence of discriminations influencing the absolute levels of observed genetic admixture still held, we found here that, in fact, the Pygmy communities most discriminated against also exhibited the strongest sex-biased admixture patterns. In interaction with reduced effective population sizes in Pygmy populations, this resulted in the extreme case of the Koya Pygmies, strongly discriminated against by their non-Pygmy immediate neighbors, where virtually non-Pygmy mitochondrial DNA are found, and where, conversely, Y-chromosome variation seem to have been replaced almost completely by that of the non-Pygmy neighbors. Finally, while we were unable at the time to investigate admixture patterns in non-Pygmy populations with our modest microsatellite dataset, further work conducted by Dr. Patin and Quintana-Murci and their team to which we could collaborate, identified, using genome-wide data for hundreds of thousands of SNP markers, that the low levels of Pygmy introgression into the non-Pygmy gene-pool was, conversely, largely female mediated rather than male mediated, as expected with the above complex socio-cultural scenario.

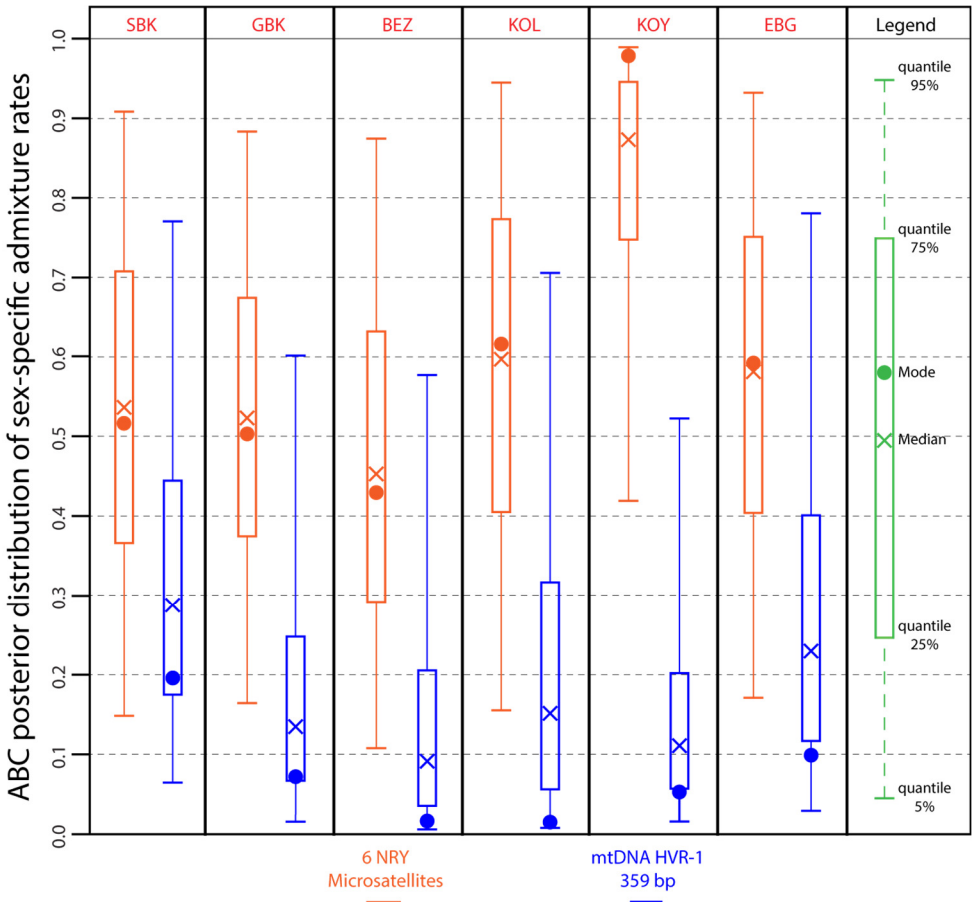


Figure F2.3.c.

ABC estimations of the posterior distribution of admixture introgression rates from non-Pygmy neighboring populations into each Pygmy population, based on Y-chromosome microsatellites and mitochondrial DNA HVR-1 sequences. Figure is simplified from Verdu et al. *Molecular Biology and Evolution* 2013.

2.4. A synthesis of anthropological genetics perspectives on the neutral demographic history of Central African peopling in the framework of the Pygmy/non-Pygmy complex categorization

In schematic **Figure 2.4.a** below, we synthesized the descriptive and inference results described above (**Sections 2.1 – 2.3**), and further explored in other publications following the seminal works of Cavalli-Sforza et al. (1986) and Destro-Bisol et al. (2004), and our first formal testing for the origins and admixture history of Central African populations in the complex historical and cultural binary categorization of populations into Pygmies and non-Pygmies as detailed in **Chapter 1.1**. Furthermore, we reproduced below the schematic figure (**Figure F2.4.b**) previously published in Verdu et al. (2013) and specifically focusing on the influence of sex-specific socio-cultural behavior on admixture and effective population sizes patterns in Central African populations.

Altogether, interdisciplinary data collections between population genetics and cultural anthropology allowed obtaining diverse types of genetic data from numerous populations extensively described anthropologically and categorized as Pygmies and neighboring non-Pygmies based on numerous cultural criteria (and no phenotypic criteria). Using these data and population genetics hypotheses carefully translated from ethnological and linguistic previous questioning, anthropological geneticists could show that populations gathered under the blanket term "Pygmy" today in Central Africa show genetic patterns consistent with a common, shared, genetic origin more recent than their common origin with the ancestral non-Pygmy lineage. More specifically, we found an ancient divergence between ancestral Pygmy and non-Pygmy lineages between 70,000 and 130,000 years ago[42], for reasons that remain, to my knowledge, unelucidated and rarely hypothesized.

Then, more recently some 25,000 years ago, Eastern and Western Congo Basin Pygmy populations seem to have diverged. It has been hypothesized that such reproductive isolation might have been due to a fragmentation of the rainforest environment at this time in this region. However, we do not know where the ancestral Pygmy populations lived, nor whether such forest fragmentation indeed resulted in population fragmentations, in particular since Pygmy populations live in different environments today.

Finally, we unveiled the very recent divergence across the various Western Central African Pygmy populations during the last 5000 to 3000 years, concomitant with the last climatic crisis and forest areas reductions in the region, and also contemporary with the expansion of agriculture accompanied by demic migrations of non-Pygmy populations throughout the Congo Basin. Furthermore, we found that Pygmy and non-Pygmy neighboring populations are admixed as a result of complex sex-specific asymmetric and heterogeneous admixture process in the context of patrilocality and social discriminations against Pygmies. Given the importance of these complex sociocultural behaviors to explain admixture patterns today, it may be plausible that the expansion of agricultural non-Pygmy populations was accompanied by profound socio-cultural changes for Western Pygmy populations possibly triggering the fragmentation of the ancestral genetic population. Then, we showed that differential admixture strongly drove the genetic differentiation across Pygmies, in addition to different effects of genetic drift.

[42] Remember that population genetics do not estimate times in years, but in coalescent events, mutation rates, and generation times. Translating this evolutionary-genetics time into years necessarily suffers several profound approximations that can be discussed albeit some of these key approximations do not stem from population genetics but from other disciplines such as functional genetics or demography.

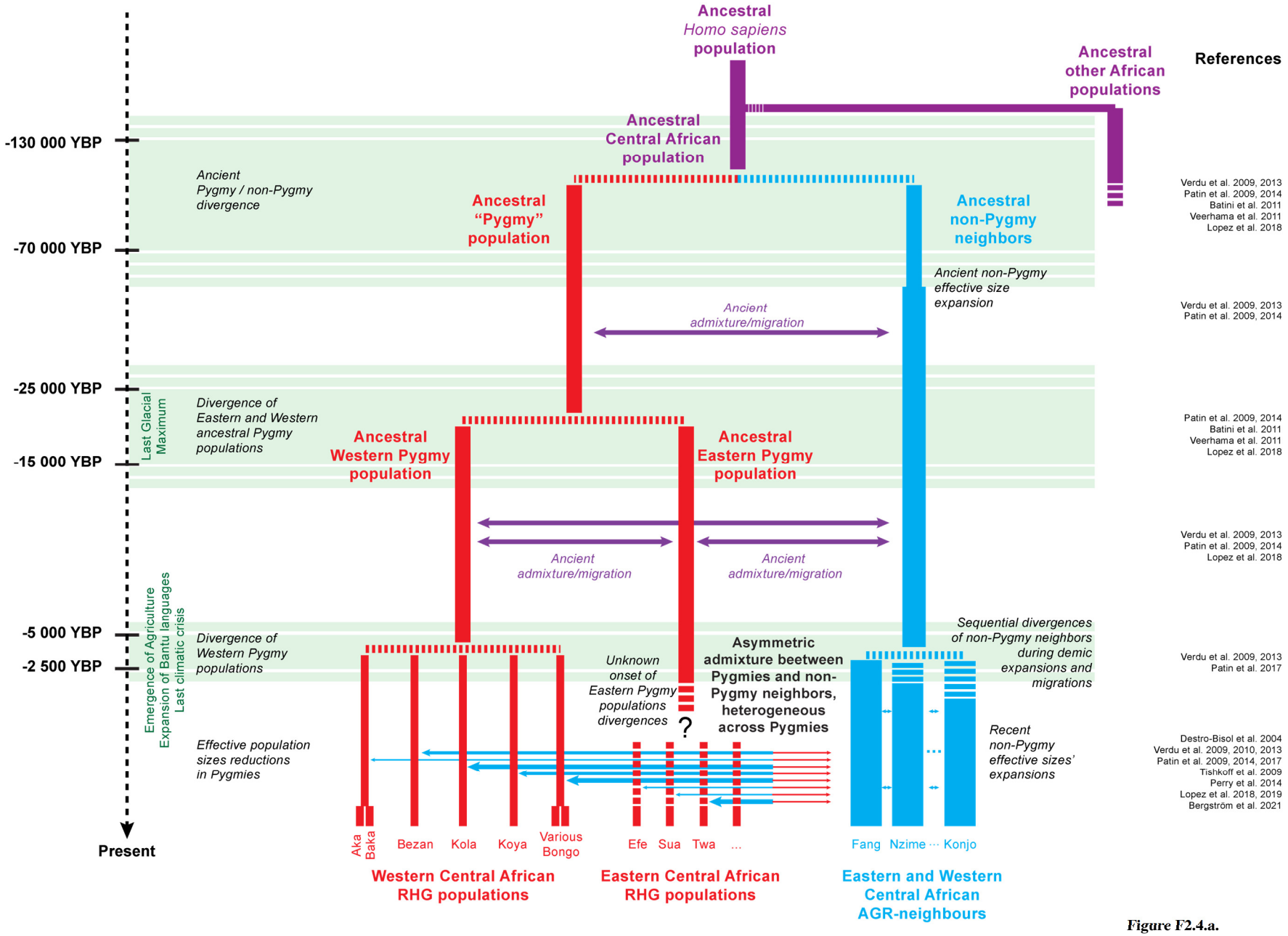


Figure F2.4.a.

Synthetic schematic representation of the genetic peopling history of Central African populations as inferred by anthropological genetics and human population genetics methods from observed, extant, genetic and anthropological data. Note that the "Pygmy" and "non-Pygmy" populations' categorization labels rely on numerous historical and cultural categorization criteria characterized and discussed in **Chapter 1.1**. This figure is largely updated from a previous similar attempt published in Verdu P (2014*) in "Hunter-Gatherers of the Congo Basin : Culture, History and Biology of African Pygmies"*, Ed. Barry S. Hewlett, Transaction Pub.

Moreover, our inferences of sex-specific effective population sizes variation across Central African populations echoed a very long-standing debate in the cultural and biological anthropology community. Indeed (**Figure F2.4.b**), we found that populations categorized as Pygmies throughout the Congo Basin often had sex-specific effective population sizes (estimated taking into account admixture with neighboring non-Pygmies) that fitted expected patterns under matrilocal philopatric marriage behaviors. Indeed, in matrilocal populations, where males move to live close to their in-laws after marriage, we expect high genetic variation across male genetic lineages within and across populations, and relatively lower genetic variation across female genetic lineages; which is what we find for Central African Pygmy populations. The opposite genetic patterns are expected in patrilocal populations, this latter pattern being observed within and across non-Pygmy populations. However, today, both Pygmy and non-Pygmy populations report exclusively patrilocal philopatric practices in marriages within populations (Verdu et al. 2013).

In this context, it was hypothesized by cultural and biological anthropologists in a seminal work *"Man the Hunter"* (DeVore and Lee 1968), that patrilocal behaviors emerged with agriculture and that hunter-gatherer human populations were likely matrilocal before their transition to agriculture. While our results may support such hypothesis, where hunter-gatherer Pygmies still retained genetic signatures of past matrilocal practices, I strongly believe that our population genetics inferences are inconclusive about this particular question, largely out of reach, in fact, of our paradigms and methods. Indeed, other alternative scenarios are equally probable and other mechanisms are likely to be at play to explain the observed patterns. For instance, the complex sex-specific admixture processes here observed could be sufficient, if they have been ongoing for reasonably long, to explain the matrilocal genetic signature here observed despite Pygmy populations practicing patrilocality. In this case, genetic patterns today reflect this former history, and genetic signatures from philopatric practices in a more remote past are irremediably lost for genetics inferences. Furthermore, from a cultural anthropology perspective, while Pygmy populations report practicing patrilocality, numerous other socio-cultural rules may apply to make this simplistic endogenous speech not as prevalent in Pygmy populations than in non-Pygmies. Finally, note that all three possibilities are not mutually exclusive… In all cases, future work will be needed, but it is plausible that it will never be possible to address this question with extant genetic data alone, and without the insights of ancient DNA data from the region, which are crucially lacking, still (Verdu et al. 2013; Verdu and Austerlitz 2015).

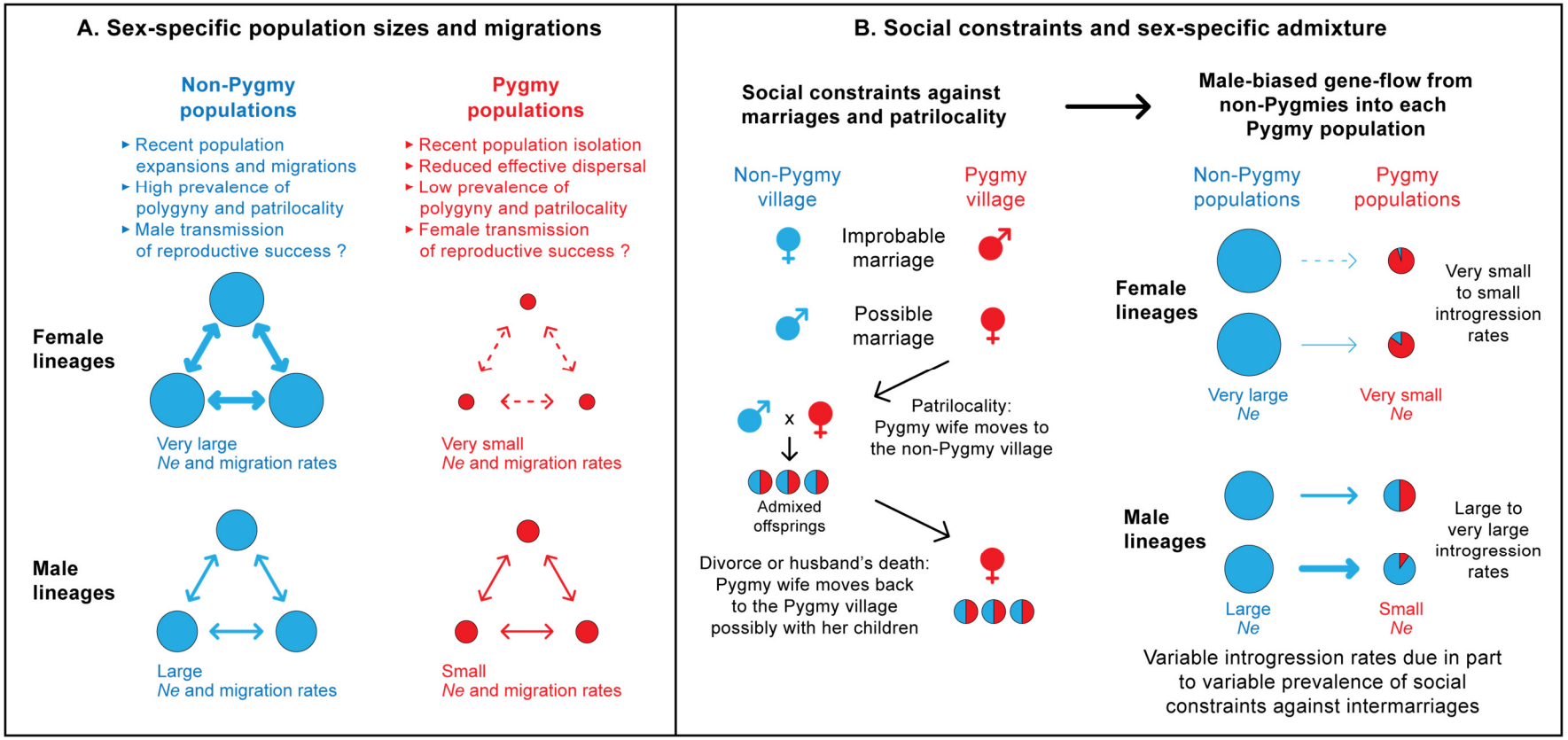


Figure F2.4.b.

Synthetic schematic representation of sex-specific effective population sizes, migrations and admixture within and across groups of Pygmy and non-Pygmy populations in Central Africa as inferred by anthropological genetics and human population genetics methods based on observed extant genetic data.

Note that the "Pygmy" and "non-Pygmy" populations' categorization labels rely on numerous historical and cultural categorization criteria characterized and discussed in **Chapter 1.1**.

A black and white version of this exact figure was previously published in Verdu et al. *Molecular Biology and Evolution* 2013.

In this context, numerous questions remain to be assessed for further elucidating the genetic history of Central African peopling. For instance[43], the admixture process itself remains to be dated and better understood, despite one of our early attempts to do so (Patin et al. 2014). Indeed, it is still debated in the anthropology community whether the complex socio-cultural relationships observed today between communities existed before the European colonization of the region, or not. Furthermore, recent seminal work about the demographic expansion of Bantu-speaking populations throughout the Congo Basin further question whether admixture processes emerged synchronically with the first encounters or whether they occurred only later after substantial societal changes for both groups of ancestral populations. More technically, we proposed in 2013 that sex-specific *cultural* transmission of the reproductive success varied across Pygmy and non-Pygmy populations throughout the Congo Basin, echoing previous propositions across human populations with different lifestyles (Blum et al. 2006). This has not been tested yet, and most, if not all, previous inference approaches were conducted under random matting assumptions; and, therefore, may be strongly biased. However, in general, a major difficulty for further investigating the evolutionary history of Central African populations remain the difficult access to extended datasets from the region. For instance, little is known today about the demographic history of Eastern Central African Pygmy populations, largely under-represented in studies and only recently investigated by ourselves, Pr. Luis Barbosa Barreiro (University of Chicago), and Pr. George H. Perry (Pennsylvania State University), and their respective teams (Perry et al. 2014). Most crucially, detailed anthropological information about the same individuals sampled for DNA is imperative, as I hope to have demonstrated here, and further renders the challenge… challenging. Furthermore, remember that population demographic census sizes are small throughout the Congo Basin and in particular in Pygmy populations. I thus fear that we may never reach sufficient sample sizes to achieve the statistical power needed to disentangle complex evolutionary forces having shaped the genetic diversity patterns observed today.

Finally, I hope to have illustrated here how interdisciplinary work between cultural anthropology and population genetics, rooted in joint field-work and extensive exchanges and discussions, can be fruitful for anthropological genetics and human population genetics. In return, I often published cross-disciplinary articles reporting these findings and discussions directly to the cultural anthropology community (Verdu 2012, 2014, 2016, 2019; Verdu and Austerlitz 2015; Perry and Verdu 2017). I would like to stress here that this is a substantial work in itself. Diffusing population genetics methods and paradigms is, I believe, essential for other non-experimental and non-biological disciplines to better understand what can and what cannot be said based on genetics results, as well as which novel questions population geneticists indeed bring to cultural anthropologists, and which questions are beyond the reach of population genetics. This has massively helped me to foster novel research projects and collaborations over the years, thus, in practice, illustrating the richness of such interdisciplinary approach. Again, the genetic history of a population is only one of its many histories, and while it is tempting to draw analogies across disciplines, they all have to be considered as hypothetical and further tested with the specific scientific methods deployed by either discipline, even in none experimental sciences.

Beyond its self-interest, reconstructing the neutral demographic history of Central African populations was absolutely crucial for better understanding the evolutionary history of these populations and identifying genomic signature of differential natural selection processes having shaped, or not, observed genetic variation at the root of phenotypic differences, as we will briefly overview in the following section.

[43] See also **Section 2.6** for ongoing perspectives on other key questions.

2.5. Admixture as a major force driving phenotypic diversity and evolutionary trajectories of Central African populations: what about height?

Identifying the genetic factors determining non-pathologic phenotypic variation across individuals and populations rely classically in comparing allelic frequencies within and across groups of individuals categorized based on the phenotype of interest, similarly to the case-control approach extensively deployed in medical genetics. The goal is thus to identify correlations between given mutations and phenotypic variation providing candidate mutations further explored by functional genetics (real biology) to establish, or reject, a causality linking genotypes to the observed traits, and decipher the biological and bio-chemistry mechanisms from genotype to phenotype. In the statistical framework of association studies, the evolutionary history of mutations is not strictly necessary to identify them. However, the statistical testing itself relies on comparing allelic frequencies across groups of individuals, and sample stratification and cryptic genetic structure may strongly reduce the power needed to identify candidate mutations, and may dramatically increase rates of false-positive discovery, therefore rendering the whole investigation futile. It is thus crucial for bio-statisticians involved in association studies, and aiming at identifying the genetic determination of phenotypic traits, to consider subtle genetic patterns within and across studied groups, themselves the product of the evolutionary and demographic histories of the individuals.

Furthermore, identifying mutations whose distribution within and across populations and associated frequencies have been targeted by natural selection processes also heavily relies on knowledge of the neutral evolutionary history of humans. Indeed, human population geneticists looking for signatures of differential natural selection processes across human populations formally need to demonstrate that observed allelic frequencies cannot be obtained by just flipping a coin, a.k.a. by "sole" neutral demographic and drift processes. In this context, our work about the demographic reconstruction of neutral genetic variation patterns across Central African populations was also essential to further investigate possible different evolution across populations. Indeed, what about non-pathologic adult standing height genetic determination and evolution across shorter Pygmies and taller non-Pygmies?

I hereby honestly reckon that I personally lack interest in working myself to identify the genetic determination of phenotypic variation in humans and its' evolutionary relevance… I am, unsurprisingly at that point of this document, vastly more interested in deciphering neutral evolutionary mechanisms. As a consequence, most of my work revolves around demographic inference from genetic data rather than exploring natural selection processes in humans. Nevertheless, I also reckon, and know from extensive experience, that "determining the genetic basis and evolutionary relevance of the short adult height of Central African Pygmies" is perhaps the essential question that drew the interest of Western evolutionary biologists to investigate these populations for more than a hundred years (see **Chapter 1.1**); which subsequently and incidentally allowed me to work…

Indeed, numerous anthropological biology hypotheses have been proposed to explain the observed shorter stature of Pygmy populations in Central Africa (Froment 1993; Perry and Dominy 2009; Becker et al. 2010, 2011). Evolutionary hypotheses proposed that it was:

i) a thermoregulatory adaptation to the hot and humid environment of the equatorial rainforest, or

ii) a morpho-mechanical adaptation to hunting and gathering efficiently in a dense environment, or

iii) a physiological adaptation to irregular and low availability of hunting and gathering nutritional produces in a diluted environment, or

iv) the by-product of an immunity adaptation to a highly challenging epidemiological environment, or

v) the by-product of an earlier onset of reproduction as an adaptation to high mortality rates triggered by both the hunting and gathering lifestyle and the challenging infectious environment, or, finally,
vi) a combination of these hypotheses.

Conversely, numerous other anthropologists simply refused the evolutionary explanation to the reduced Central African Pygmy height, and rather considered this observed phenotypic trait as due to phenotypic plasticity resulting from high infection rates and insufficient nutritional intakes experienced by individuals during growth, but not adaptive on evolutionary scales in any ways.

Altogether, the debate has been raging since the 1950s at least, as the overall lack of demographic, epidemiological, and anthropological data from Central Africa never allowed to formally test these varied hypotheses satisfactorily. In fact, the heritability of the "pygmy phenotype" was not even assessed due to lack of data. Beyond the lack of data for formal testing, anthropologists had nevertheless gathered sufficient data to show the tremendous variation in adult height across the various Pygmy populations throughout the Congo Basin, a variation that was never explicitly explained in the previously proposed evolutionary hypotheses, and thus impaired their plausibility in a first place. Nevertheless, these hypotheses slowly transformed, over the years, into scientific known facts in the eyes of the non-specialized audience as well as in the eyes of numerous scientists.

In this context, and despite my relative lack of enthusiasm to dig into these very difficult questions, Pr. Evelyne Heyer and my student at the time Dr. Noémie Becker incited me to investigate the genetic determination of Pygmies' height. Our initial Central African dataset comprised extensive non-pathological adult height data, that I participated to collect on the field, and often associated with DNA samples from several Pygmy and some non-Pygmy populations categorized as such based only on cultural criteria. We initially only had microsatellite data for these individuals which rendered association studies impossible. However, Dr. Becker had the intuition that it might be worth to compare our phenotypic data with one major aspect of genetic variation across individuals and populations identified in our previous work: admixture levels (Verdu et al. 2009). I reckon I really did not think this could work, but I was wrong… Indeed, we found that individual standing and sitting height variation in Western Central African individuals categorized as Pygmies based on cultural criteria only, significantly strongly positively correlated with individual levels of admixture from the non-Pygmy gene-pool (**Figure F2.5**). In other words, the more a Western Central African Pygmy individual was genetically resembling the non-Pygmy gene-pool, the taller he or she was, on average (Becker, Verdu et al. 2011). To our knowledge, this was the first indirect proof that height differences between Pygmies and non-Pygmies were reflected in neutral genetic differentiation patterns, a single step beyond the possibility of a genetic determination of height differences evolutionarily relevant. Indeed, several other research teams replicated the result since, and we extended it at the genome-wide scale using tens of thousands of independent SNPs genotyped throughout the genome of Pygmy and non-Pygmy neighbors throughout the entire Congo Basin (Pemberton, Verdu et al. 2018).

Most importantly to my views, identifying that phenotypic differences significantly correlated with neutral admixture patterns had one essential dialectical consequence. It empirically showed that, beyond the elusive genetic determination of the "pygmy phenotype" chased by biological anthropologists, it might be much more productive for population geneticists to investigate the genetic determination of height differences between Pygmies and non-Pygmies based on cultural criteria. Indeed, the focus had been almost exclusively set on explaining why Pygmies were shorter, and virtually never on why non-Pygmies were

taller. Instead, our results showed the link between height differences and genetic admixture patterns, rather than either phenotype taken separately[44].

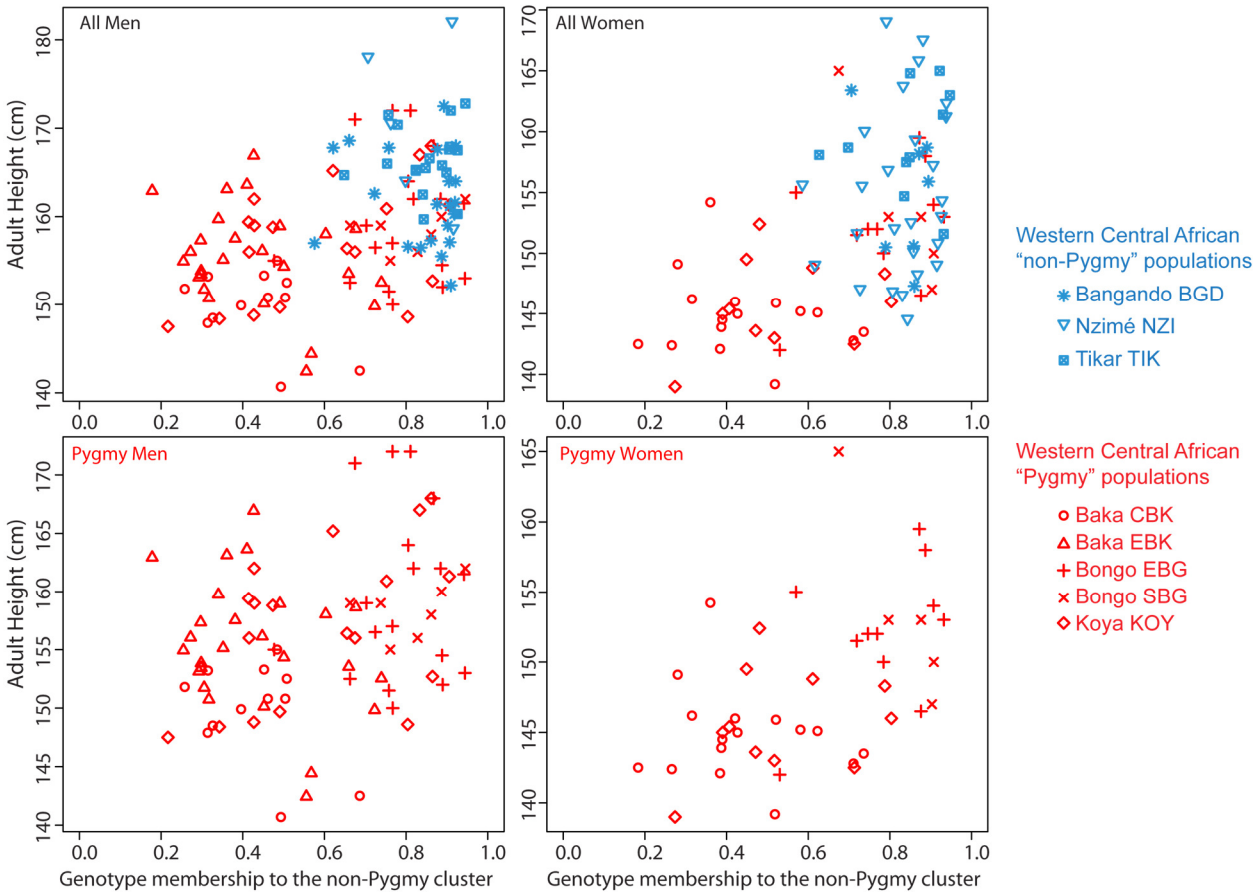


Figure F2.5.

Positive correlation between individual height and individual levels of genetic resemblance with the non-Pygmy gene-pool.
Figure previously published in Becker, Verdu et al. *American Journal of Physical Anthropology* 2011
Note that the "Pygmy" and "non-Pygmy" populations' categorization labels rely on numerous historical and cultural categorization criteria, without considering height, characterized and discussed in **Chapter 1.1**.

Based on these results, Pr Barreiro and Pr Perry conducted the first full-blown association study on novel anthropometrical and genetic data they had gathered in the Ba.Twa Pygmies from Southwest Uganda and in their non-Pygmy neighbors, the Ba.Konjo. Most importantly, they used an admixture-mapping approach, taking-into-account the genomic admixture structure of individuals and increased their power to identify mutations potentially involved in height-differences determination between groups of populations (Perry et al. 2014). Importantly, they did not identify the same genomic regions associated with height differences between the few pairs of Pygmy and non-Pygmy populations they investigated in the West and, respectively, in the East of the Congo Basin. Furthermore, they showed that either candidate mutations had changed in frequency more rapidly than expected under neutrality. They thus concluded that natural selection was at play for determining height differences across Pygmy and non-Pygmy populations, but that

[44] As a side note, this conceptual angle further echoes the ethnological approach of Frederik Barth that I favored in my anthropological genetics work; and approach interested in studying ethnic groups from the perspective of the dynamic boundaries they share rather than on a putative cultural nucleus defining them separately (see **Chapter 1**).

different regions were the target of such evolution in different Congo Basin population. Hence, they advocated for convergent evolution of this trait after the divergence of Eastern and Western Pygmy populations, rather than a common origin of the genetic determination of height differences throughout the Congo Basin acquired in the ancestral Pygmy population.

Following this major work, we generated genome-wide SNP data in the much larger sample set that we had, but with much lower numbers of SNPs due to financial constraints (genotyping chips were really more expensive ten years ago). Our association study incorporating interindividual genetic differentiation and admixture patterns identified possible genomic regions and groups of mutations involved in height differences determination, further expanding the already vast list of genetic mutations possibly involved in non-pathologic height variation determination in humans in general. However, we could not produce convincing results as to natural selection signatures at these loci, in the face of the fundamental lack of power due to the high levels of genetic stratification and admixture across Central African populations investigated with too small a number of genetic markers (Pemberton, Verdu et al. 2018).

In this context, the team of Pr. Quintana-Murci and Dr. Patin set out, implicating us in their projects, to generate more data on our samples, and scanned differential natural selection signatures throughout the genomes of Central African Pygmy and non-Pygmy neighbors, without aiming at a particular phenotype, and, most importantly, taking-into-account explicitly the neutral demographic history of the investigated populations. This approach yielded tremendous novel insights into the evolutionary history of Central African populations in general, and further fed the debate about height differences' evolutionary relevance in particular.

In very brief, in a work led by Dr. Marie Lopez, PhD student in the Quintana-Murci lab at the time, we showed that the frequency of deleterious mutations and the efficiency of negative natural selection to sweep them was rather equivalent in Pygmies and non-Pygmy neighbors, based on whole-exome sequences. This was rather surprising: we expected genetic-load to be increased in Pygmy populations compared to their neighbors, as the formers did not experience effective size growth and even underwent recent bottlenecks (see previous sections in this chapter and **Figure 2.4.a** above). However, our results comparing Pygmies and non-Pygmies, taking-into-account explicit demographic "neutral" models of evolution for these populations via extensive simulations, showed that admixture from non-Pygmy neighbors effectively counter-balanced the accumulation of deleterious mutations in Pygmy populations (Lopez et al. 2018).

Furthermore, Dr. Lopez further showed in another important work that natural selection occurring after admixture from non-Pygmies into Pygmy populations maintained genetic variation acquired by this mechanism, further contributing to the genetic adaptation of Pygmies to their local environment, in particular with respect to immunity related pathways (Lopez et al. 2017). Importantly, in this paper, we further found strong signals of polygenic adaptation directly targeting genes and mutations previously shown to be involved in morphological traits and height. In turn, the large pleiotropy of these genomic variants resulted in a signal of adaptation for other pathways including life-history traits and the reduction of reproductive age. This result thus rejected the previous anthropological biology hypothesis stating that height phenotypes would be, instead of what we found, the indirect by-product of adaptation to high mortality rates assumed to be favoring an earlier onset of reproduction.

Finally, we showed that non-Pygmy populations had also benefited from the low levels of admixture from their Pygmy neighbors (Patin et al. 2017). Indeed, in another seminal article led by Dr. Patin and Pr. Quintana-Murci, we identified strong signatures of positive natural selection targeting immunity-related mutations in regions of the genomes of non-Pygmies particularly enriched in admixed segments from Pygmies compared to the rest of the genome and to expectations under neutral demographic models

incorporating admixture events as previously reconstructed. This result further showed the importance of admixture in natural selection processes, positive or negative, as a determinant force driving human evolution and possible adaptation to local environments.

Altogether, while the exact biological and bio-chemical mechanisms underlying non-pathologic phenotypic differences, including adult height differences, across Pygmy and non-Pygmy populations are still largely un-elucidated, anthropological genetics and population genetics' work demonstrated over the past 15 years the evolutionary relevance of these phenotypic differences in general and of height-differences in particular, and showed the importance of admixture processes in having shaped them throughout the Congo Basin.

2.6. Ongoing and future perspectives about the evolutionary history of Central African populations

After more than 15 years of intense population genetics work on the evolutionary history of Central African populations, numerous fascinating questions remain to be addressed, some of which I investigate myself. As illustrated throughout this Chapter, ancient and recent admixture processes have played a key role in the genetic evolution of all human populations settling the Congo Basin today. However, while we acquired detailed knowledge about recent admixture processes, ancient admixture histories are still largely mis-understood (**Figure F2.4.a**). Indeed, while we know that we need to consider possible gene-flow exchanges in the remote past across ancestral lineages to explain the genetic diversity patterns observed today, we do not know when did they happen, between which ancestral populations, and how much gene-flow did they involve.

As a matter of fact, the question of ancient population structure and migration in the African continent as a whole is a very "hot" topic today in the human population genetics community (Schlebusch and Jakobsson 2018; Ragsdale and Gravel 2019; Bergström et al. 2021). The first essential question about ancient admixture in Africa stems from the past ten years' discoveries of admixture events between the *Homo sapiens* lineage and other, now extinct, Neanderthal or Denisovan "archaic" hominid species, dated several tens of thousands of years ago in different regions of the world. Indeed, by comparing modern *Homo sapiens* genomes with newly generated ancient genomes from various Neanderthal samples in Europe, or Denisovan remains in Central Asia, population geneticists identified portions of non-African populations' genomes originated from these other species, thus settling the fundamental debate of the existence of such events among paleo-anthropologists. In the lack of ancient remains clearly attributed to another species than *Homo sapiens* since its' emergence some 300,000 years ago on the African continent, and in the lack of ancient genomes older than several thousands of years in Sub-Saharan Africa in general, it is extremely difficult to rule out that certain highly diverging portions of the genomes of extant African populations may derive from analogous ancient admixture events; a challenge that several research labs have taken.

Furthermore, the classical paleo-anthropological view of a single ancestral *Homo sapiens* population emerged in East Africa has long-lived. Paleo-anthropologists continuously revisit their models in the light of new discoveries and currently form a view of *Homo sapiens* emergence from numerous proto *Homo sapiens* from different regions of the continent in interaction with one another over long periods of time having eventually given birth to *Homo sapiens*. Population geneticists have also considered, for a long time now, that the speciation event for our species was likely rather a long, complex, and regionally sub-structured process, rather than a unique "Adam and Eve" event. Thus, two competing scenarios are explored today by populations geneticists at the root of genetic diversity patterns observed today in Africa (Ragsdale et al. 2022). The first model considers that Homo sapiens evolution in Africa in a remote past occurred in a tree-like manner, with sub-populations isolated for long periods of time after their respective divergences, with low levels of genetic migrations among them until much more recently. The second model considers rather a complex network of geographically separated populations nevertheless regularly exchanging genes over long periods of time followed by more recent geographic and cultural isolation until further gene exchanges in a much more recent past. These two views are, in fact, philosophically divergent as the former considers a somewhat classical isolation with possible migration Wright-Fisher island-model (Fisher 1922; Wright 1931) for the evolution of *Homo sapiens*, while the latter considers instead a continuous network of genetic movements in a "population" only defined by its contours rather than demes. It is from this ancestral

"meta-population" that the lineages ancestral to the extant populations would have emerged, those we investigate throughout the continent and at the root of all human genetic diversity worldwide.

I do not yet explore the fundamental questions detailed above first about ancient "archaic" admixture in African populations. Indeed, I think that without any reference of very ancient (> 10,000 years old) African genomes from unquestionably non-*Homo sapiens* remains as defined by paleo-anthropologists, such endeavors, while extremely interesting conceptually and methodologically, will be unable to convincingly bring new knowledge about the subject.

However, I currently extensively work with Pr. Mattias Jakobsson and Pr. Carina Schlebusch from Uppsala University, to reconstruct ancient admixture histories among *Homo sapiens* populations in Africa since the last 300,000 years. We conduct model-based inferences across numerous Central African and Austral African novel high-quality whole-genome sequences to address several nested questions. Namely, we investigate several possible tree-like topologies having given birth to observed genomic diversities and further try to distinguish whether ancient migration models, with recurring constant gene-flow over large periods of time, better fit the observed data than admixture models, where ancient admixture events occurred more punctually, all models taking-into-account the known much more recent gene-flow events described in our previous studies. This ongoing work was conducted mainly by Dr. Gwenna Breton during her PhD co-mentored by us three, and is currently being prepared for publication (Breton et al. 2020).

Beyond these fundamental questions about deciphering the influence of ancient admixture and migration events across ancestral African populations, as stated in **section 2.4** above, the recent history of admixture across Central African populations remains to be better understood. Indeed, our initial attempt at reconstructing the admixture history between Pygmies and non-Pygmies showed that these events occurred no earlier than 1000 years before present. However, the methods we used at the time could only consider over-simplistic models of admixture, and our results would be more accurately described as: "if admixture occurred at a single point in time in the history of Pygmy and non-Pygmy populations, then it would take at most 1000 years since then to obtain the genetic patterns that we observed today"; which is unsatisfactory when the question to address is rather "when and how did admixture events occurred in Central Africa"… In this context, the following chapters **3** and **4** recapitulate most of the theoretical and methodological research I have developed until today to build the population genetics tools which would allow us to reconstruct complex admixture histories from genetic data; which we will apply to the Central African context in a proximal future.

Finally, investigating the evolutionary history of natural selection's influence on genetic diversity patterns in Central Africa still represents a substantial amount of my research endeavors in collaboration mainly with Pr. Quintana-Murci and Dr. Etienne Patin's team, as well as with Pr. Zachary Szpiech (Pennsylvania State University). For this purpose, we have collected anew more than 400 DNA samples from four Cameroonian populations (2 "Pygmies" and 2 neighboring "non-Pygmies"), together with more than 20 anthropometric measurements and several basic physiological data for each volunteer participant. This project started in 2015 with Dr. Pemberton and initially funded by a national Canadian grant that we had obtained as co-PIs. Since then, my colleague completely retired academic research, and I am now in the process of rebuilding the project and securing funding to thoroughly investigate this unique novel data.

Chapitre 3

A general theoretical framework for investigating complex admixture histories

Ribeira Grande de Santo Antão at dusk.
Santo Antão, Cabo Verde, 2016
©Paul Verdu

Chapter 3. A general theoretical framework for investigating complex admixture histories

During my PhD, I showed how complex socio-cultural behaviors shaped genetic diversity patterns in Central Africa, and, in particular, how admixture processes were deeply determined by these cultural forces. It was then clear to me that I wanted to further investigate complex admixture processes in human populations. My goal was to understand how the history of complex sociocultural relationships within and between communities triggered, or not, genetic admixture events; leaving a signature, or not, in genetic diversity patterns across human populations. This would also allow me to understand how complex admixture processes themselves acted to shape human genetic evolution.

In this context, I was very interested in studying the influence of recent socio-cultural changes on admixture processes in-turn having influenced genetic diversity patterns during the history of European colonization since the 15th century and the Trans-Atlantic Slave Trade (TAST). My initial, spontaneous, questions were: When did admixture between enslaved and non-enslaved communities started in the various colonial empires? Did the expansion of the Plantation Economy in the 17th century and the concomitant establishment of strong socio-marital segregation between communities throughout colonial empires, as enacted in the *Code Noir* or the *Sistema de Castas* for instances, modify admixture processes? Did the abolition of the TAST, during the first half of the 19th Century, change admixture processes? Did the abolition of slavery, during the second half of the 19th century, change admixture processes? Did the end of racial segregation and extensive voluntary migrations, in the second half of the 20th century, modify admixture processes?

In other words, I aimed at using observed genetic data in admixed populations descending from enslaved-Africans since the 15th century to reconstruct when did admixture events occur and how did they occur (intensity and duration), compare the obtained genetic results with major known historical processes, and determine which such processes had possibly influenced admixture patterns, and which did not translate in the book of our genetic evolution.

On these premises, I investigated the methodological tools available at the time in 2008 to the population genetics community for inferring complex admixture histories from genetic data. As explained briefly in **Chapter 1.2.a** above, admixture in general, and admixture among enslaved-African descendants in the Americas in particular, had been the focus of countless theoretical, methodological, and data-analyses investigations in population genetics since at least the 1930's, when the first statistical formalization of expected allelic frequencies in admixed populations was proposed (Bernstein 1931).

Section 3.1 presents the classical population-genetics statistical framework for investigating admixture and migration processes from genetic data. **Section 3.2** then presents our general mechanistic admixture model. In **section 3.3**, we show how the distribution of admixture fractions in the admixed population may be informative about the underlying parameters of historical admixture models. **Section 3.4** then briefly presents our latest developments about the sex-specific generalization of our mechanistic admixture model. Finally, in **Section 3.5**, we present ongoing perspectives about theoretical investigations of complex admixture models.

3.1. Classical mechanistic models of admixture

Classically in population genetics, admixture is defined as a punctual exchange of genes between previously genetically isolated populations resulting in a novel "admixed" population. In this admixed population, allelic frequencies are expected to be a simple linear combination of allelic frequencies in the source populations, respectively, proportional to the relative contribution of each source to the gene-pool of the novel admixed population. Alternatively, classical migration models between two populations, or demes, as in Wright and Fisher's island-models (Fisher 1922; Wright 1931), involved recurrent and constant exchanges of genes between otherwise genetically isolated populations over entire periods of time, resulting also in allelic frequencies in the receiver population being linear combinations of allelic frequencies in the receiver and source populations (**Figure F3.1** below). Classical models of admixture thus considered a single founding admixture event (or pulse) over the entire history of the admixed populations, and migration models considered recurrent constant gene-flows over entire evolutionary periods (Long 1991).

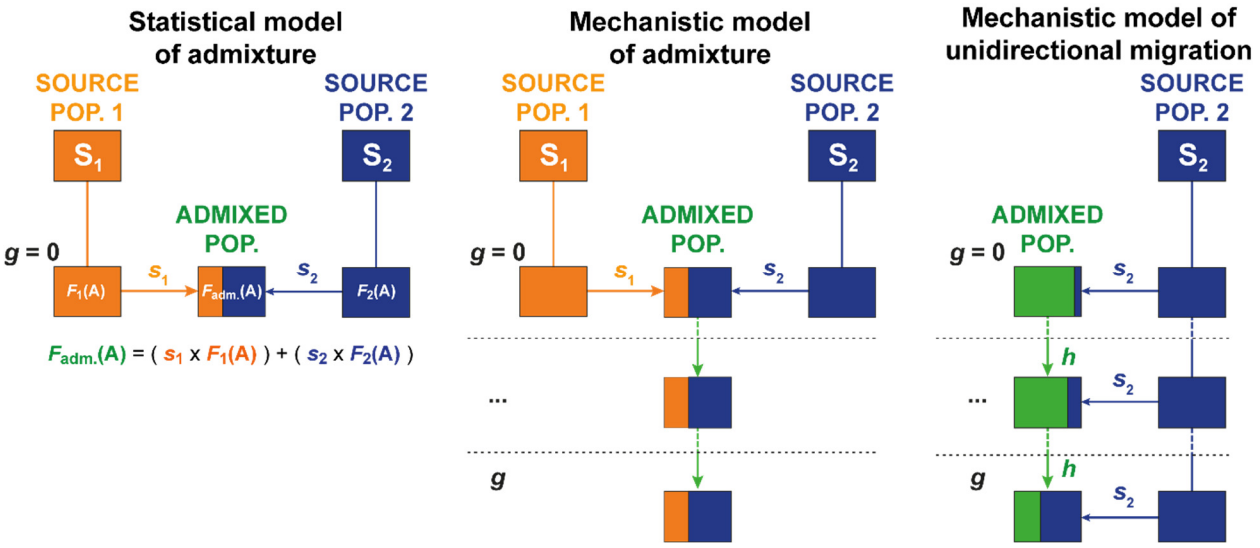


Figure F3.1.
Classical models of admixture and migration. Adapted from Bernstein (1931), and Long (1991). $F_1(A)$, $F_2(A)$, and $F_{adm.}(A)$ represent, the frequency of allele A in Source population 1, Source population 2, and the Admixed population after the founding admixture event, respectively. s_1, s_2, and h are in [0,1] with $s_1 + s_2 = 1$ or $s_2 + h = 1$ respectively.

I was surprised to discover that virtually all the literature about admixture and genetic migration since then considered simple mechanistic models very marginally differing[45] from the original models of the founders of the discipline. Until the 2005's, it was as if the entire population genetics community fundamentally agreed that admixture and genetic migration were major mechanisms of biological evolution of all species, including humans, but were satisfied by the performances of over-simplistic such models for their daily research, and despite perfectly admitting that these models were often not realistic at all. Indeed, we have countless examples, in both humans and non-human species, for which we expect gene-flows to have occurred at several points in time, or even during relatively short periods of time, and varying in

[45] For instance, asymmetric admixture and migration models where rapidly proposed as in **Figure F3.1**, but migration itself remained a constant recurrent process, and admixture remained a punctual event.

intensity from either source populations and in time. Such likely more realistic models were thus impossible to approach with classical models considering a single event of admixture or constant migrations over time. I was under the impression, then, that population geneticists mostly pragmatically preferred to primarily investigate genetic evolution within trees of isolated lineages, each under drift and mutation and natural selection, and admixture and migration were "just" a supplementary layer of complexity, often only considered as a necessary nuisance to the tree-like evolution, and only sometimes considered as the process of interest. From my very reduced experience at the time, it rather intuitively seemed to me that all populations may be practically seen by population geneticists as non-admixed, but just take a look and you will inevitably always actually find complex admixture histories…

Numerous, elegant, efficient, and sometimes sophisticated statistical methods had been developed, and are still developed, to identify admixture and estimate it in observed genetic data (Cavalli-Sforza and Bodmer 1971; Chakraborty and Weiss 1988; Long 1991; Bertorelle and Excoffier 1998; Pritchard et al. 2000; Tang et al. 2005). To do so, one does not, strictly speaking, need to explicit the underlying mechanistic model of admixture, whether complex or simple. Indeed, these statistical methods ultimately aim at estimating the parameters of the linear combination of allelic frequencies from the sources in the observed admixed population (**Figure F3.1**), rather than at disentangling when and how admixture and migration occurred to produce, in the observed data, said admixture patterns.

Therefore, researchers over the years had sometimes punctually implemented more complex mechanistic models for their specific statistical needs without generalizing nor thoroughly investigating them, but most of the community interested in admixture inference to determine when and how did admixture occur still relied on simplistic mechanistic approaches to admixture.

In this context, it became clear to me that I would need to build some complex mechanistic models from scratch in order to be able to apply them to observed genetic data to address my specific questions of interest. However, despite me not being afraid of mathematics and probabilities, at least as long as they were not too complex, I had no empirical experience in mathematical formalization nor theoretical population genetics. I therefore set out to look for a post-doctoral contract in order to fill this gap with this project, and had the chance and luxury to interest Pr. Noah A. Rosenberg at the University of Michigan (at the time) who enthusiastically proposed to me a position to pursue this endeavor with him (Verdu and Rosenberg 2011).

3.2. A general mechanistic model for admixture histories

Despite my detailed specific questions about the TAST, I wanted to investigate a sufficiently flexible general model possibly of interest to other problematics and other diploid and sexually-reproducing species. Also, for my first theoretical work, I was afraid of having to treat the mathematics of continuous-in-time genetic models and therefore, right from the beginning, decided to opt for discrete-in-time models without possible overlapping generations. While this initial constraint was strongly reducing the array of possible applications of my model, it was nevertheless of interest to numerous cases and, in any case, represented a reasonable starting point. I therefore proposed the general model described schematically in **Figure F3.2.a**.

The first founding element of the model was how to define the admixed population itself (Verdu and Rosenberg 2011). Indeed, should the admixed population be composed only of genetically admixed individuals? Or should it be composed of individuals among which admixture was possible but not necessary? As explained in **Chapter 1.2.a** above, this was a categorization choice fundamentally influencing both the range of applicability of the model and how the model would be treated mathematically and probabilistically. Here, I chose to be comprehensive, which also incidentally made the model more general and easier to treat analytically: I decided to define the admixed population as a population where admixture could potentially occur, but may or may not occur. In other words, imagine an un-inhabited island at time 0, where and when two or more previously isolated populations disembark. Admixture between members of these source populations is then possible and may or may not occur. Under such definition, we expect *a priori* that admixture fractions from either source populations should be between 0 and 1, at each generation after the initial disembarkation.

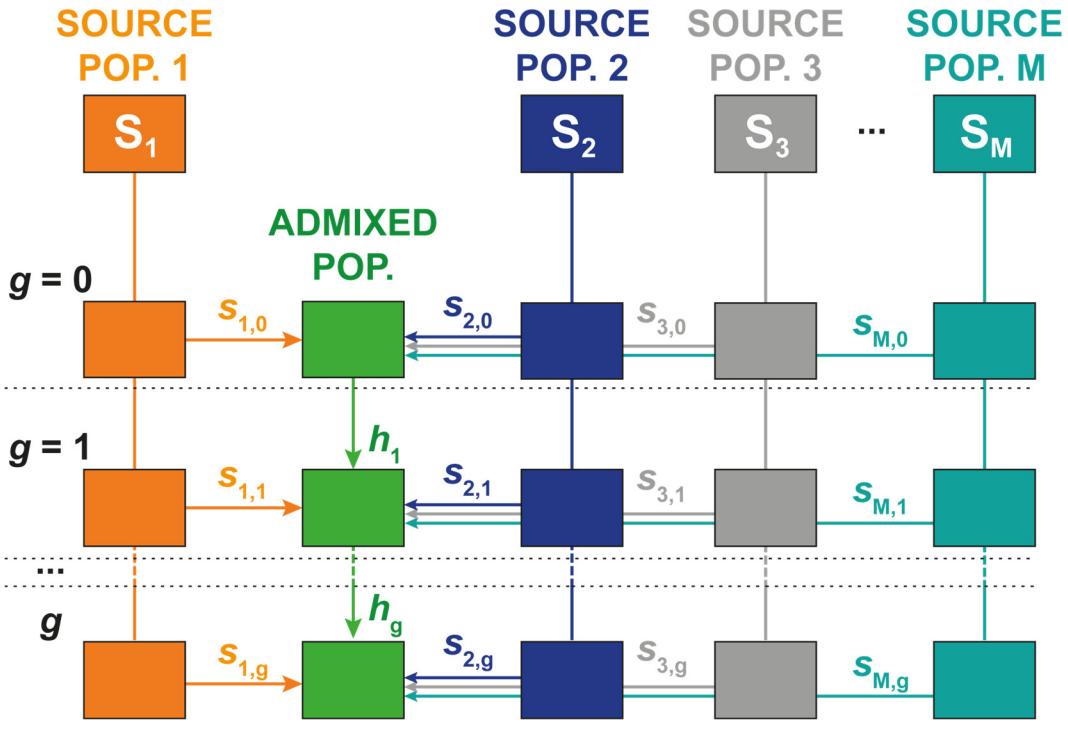


Figure F3.2.a.
General mechanistic model for admixture histories of hybrid populations.
Figure published as is in Fortes-Lima, Laurent et al. *Molecular Ecology Resources* 2021, where it was re-drawn from its initial conception in Verdu and Rosenberg *Genetics* 2011.

We define the parameters of the general model in **Figure3.2.a** as follows:

- g in [0..G], is the discrete number of generations since the founding of the admixed population at "time" 0 and until generation G. The duration of the evolutionary process here considered, whether admixture occurs or not, is thus G generations.
- $s_{i,g}$ in [0,1], is the parental contribution of source population i in [1..M] to the admixed population at generation g.
- h_g in [0,1], is the parental contribution of the admixed (hybrid) population to itself at generation g.
- *Note that*, henceforth, we often call $s_{i,g}$ and h_g the "admixture parameters" at generation g.

By "parental contribution of source population i to the admixed population", or "by the admixed population to itself", at generation g, we mean the proportion of biological parents coming from either source population, or the admixed population itself, having given birth to all individuals in the admixed population at generation g. In other words, pick a random individual in the admixed population at generation g, he/she has a probability $s_{i,g}$ (between 0 and 1) of having exactly one of his/her parent originating from source population i at the previous generation, and a probability h_g (between 0 and 1) of having exactly one parent from the admixed population itself at the previous generation. With such definitions, at each generation, all $s_{i,g}$ and h_g sum to one.

This general model heavily differed from classical mechanistic models in that it could consider varying gene-flow at each generation from the same source as well as possibly different sources contributing to the gene-pool of the admixed population over time. Furthermore, defined as such, our model encompassed all classical admixture and migration models for particular values of model parameters. For instance, for constant values over time of the genetic contribution from a given source (for all values of g in [0..G] and for a given i in [1..M], $s_{i,g} = s_{i,g+1}$), and all other parameters set to 0 (for all values of g in [0..G] and whichever j in [1..M] different from i, $s_{j,g} = 0$), this general model formally becomes the classical unidirectional migration model from Long (1991) described in **Figure F3.1** above. Similarly, if only two source populations contribute to the gene pool of the admixed population synchronically at generation g summing source contributions to 1, and all other admixture parameters from all other sources at all subsequent generations are set to 0, this comes down to the classical admixture model with two-source populations founding the admixed population at generation g.

Most importantly, I was obviously not the first one to have had thought off such intuitive, but complex, model for mechanistic histories of admixture. I much later found out, in fact while writing my paper, that this model had been proposed and investigated by Ewens and Spielman in 1995. However, these authors had not investigated this model for historical inference from genetic data, but rather to benchmark the influence of population subdivision and gene-flow on transmission disequilibrium tests (TDT). Similarly, note that it had been used once since, to my knowledge, by Guo et al. (2005) to examine linkage-disequilibrium statistics, and again not for developing demographic inferences nor to investigate the influence of the admixture process itself on the genetic evolution of admixed populations[46].

Ideally, I would use genetic data to infer all admixture parameters at each generation; a somewhat utopic goal intuitively. However, I thought that genetic information might help identify at least some of

[46] Note that I was relieved after feverishly reading and re-reading these papers, thinking that all my work was in fact redundant before even submitted to a journal… I was then told by Pr Rosenberg and other colleagues versed in theoretical population genetics, that this very often happened in math: someone had sometimes thought about it before you. Fortunately, they also taught me that it was also rarely an issue as different points of views and formulations by different researchers on a given problem were also of major interest to researchers in theoretical mathematicised disciplines, even if the model was no longer as novel as I initially arrogantly thought…

these parameters therefore allowing me to, *a posteriori*, compare results with expectations from historical data.

As a perfect illustration of my proposed research on genetic admixture in societies from the TAST-era, and further echoing **Chapter 1**'s thesis about categories and categorization processes, the *Sistema de Castas* in Hispanic Americas (Ramos Pérez and Díaz-Trechuelo López Spínola 1992) in the 17th and 18th century defined social categories and accompanying segregation systems following a genealogical model recapitulated incompletely in the general admixture model here proposed.

As exemplified in the painting from the 18th century Mexico reproduced below (**Figure F3.2.b**): individuals were assigned to a named category of the society based on the origins of their parents (from either source or admixed populations in our model above), over no less than the *sixteen* previous generations, and with different possible source populations across continents: Europe/Spain (*"Español/a",* beware, see next), Sub-Saharan Africa (*"Negro/a"*), North Africa (*"Moro/a"*), or the Americas (*"Indio/a"*). However, compared to our model, far from all reproductive possibilities are evenly considered (not to say the least in this racist system…), as, for instances:

i) no categories are reported in this painting for admixed individuals between "Sub-Saharan Africa" and "America",

ii) some categories confusingly use the same name over generations across different origins of the parents, as for instance certain types of marriages over several generations produce offspring with a name similar to that of one of the source population: *"Español"* designated either a Spanish person from Spain, the offspring of Spanish parents in the Americas, or the offspring of a *"Castizo"* with an *"Española"*, knowing that the *"Castizo"* is himself the offspring of a *"Meztiso"* with an *"Española"*, and the *"Meztiso"* himself the offspring of an *"Español"* with an "*India"* from the Americas...

iii) note that reproduction events considered in this system are heavily sex-biased. For instance, while marriages with females from Sub-Saharan Africa (*"Negra"*) or North Africa (*"Mora"*) exist in this painting, there are no reports of marriages between males from either region…

This perfectly illustrated the type of questions I was interested in and why I needed a complete general mechanistic model of admixture to formally test them: how effectively did this highly complex social categorization and segregation system (which designated categories of admixed offspring with specific names over at least 16 generations!), translated, or not, in genetic admixture patterns observed in descendant populations today? Where the "missing categories" also missing in observed genetic admixture patterns? If yes, this would have meant that the system was highly prevalent and normative for all reproductive events which strictly conformed to it over many generations. If not, that would have highlighted a discrepancy between endogenous representations, social norms, and the realized behaviors of individuals, which would in-turn need to be explained by other social mechanisms allowing for the transgression of the norm.

Figure F3.2.b.
Painting of the Sistema de Castas "a cuadretes", 18th century.
Museo Nacional del Virreinato, Tepotzotlán, Mexico.
https://commons.wikimedia.org/w/index.php?curid=4642698

3.3. Distribution of admixture fractions across individuals in the admixed population

Now that we had the general mechanistic model, which quantity of interest did we want to investigate under it? My previous investigations of admixture patterns in Central Africa (**Chapter 2**), had shown that variation of genetic admixture patterns across populations reflected socio-cultural behaviors regarding inter-marriages. I therefore thought to investigate how the parameters of our general admixture model influenced the distribution of admixture fractions across individuals within the admixed population. As numerous population genetics methods already existed to estimate individual admixture fractions from observed allelic frequencies in the admixed and the source populations, we could then, in principle, use this observed distribution to infer the parameters of the complex history of admixture that had given birth to our observations.

We therefore decided to define $H_{i,g}$, the probability for a random locus in a random individual in the admixed population at generation g to have ultimately originated from source population i (Verdu and Rosenberg 2011).

Thus, for a random individual in the admixed population at generation g, if we consider an infinite mutation model, then $H_{i,g}$ is the proportion of the genome of this individual ultimately deriving from source population i. In this framework, the density of probabilities of $H_{i,g}$ would represent the distribution of admixture fractions from population i at generation g in the admixed population; and all the moments of $H_{i,g}$ would fully describe this distribution; with the expectation, $E[H_{i,g}]$ being the mean admixture from source population i in the admixed population at generation g, the variance $V[H_{i,g}]=E[H_{i,g}^2]-(E[H_{i,g}])^2$ would be the inter-individual variance of admixture fractions from source i at generation g, and so on for all the higher moments of the distribution of $H_{i,g}$.

With the definitions of our general model, we could easily write a recursion relation to calculate $H_{i,g}$, for any one source population i, as a function of all model parameters based on all possible pairs of parents for a random individual in the admixed population, denoted Y, at the previous generation (see Verdu and Rosenberg 2011 Supplementary materials p.2):

"For the first generation ($g = 1$), for any mutually distinct values of i, j, and l between 1 and m, we have

$$H_{i,1} = \begin{cases} 1 & \text{if } Y = S_i S_i, \text{ with } P[Y = S_i S_i] = s_{i,0}^2 \\ \frac{1}{2} & \text{if } Y = S_i S_j, \text{ with } P[Y = S_i S_j] = 2 s_{i,0} s_{j,0} \\ 0 & \text{if } Y = S_j S_j, \text{ with } P[Y = S_j S_j] = s_{j,0}^2 \\ 0 & \text{if } Y = S_j S_l, \text{ with } P[Y = S_j S_l] = 2 s_{j,0} s_{l,0}. \end{cases} \quad \text{(S1)}$$

For all subsequent generations ($g \geq 2$), we have

$$H_{\mathrm{i,g}} = \begin{cases} 1 & \text{if } Y = \mathrm{S_iS_i}, \text{with } \mathrm{P}[Y = \mathrm{S_iS_i}] = s_{\mathrm{i,g-1}}^2 \\ \dfrac{H_{\mathrm{i,g-1}} + 1}{2} & \text{if } Y = \mathrm{S_iH}, \text{with } \mathrm{P}[Y = \mathrm{S_iH}] = 2s_{\mathrm{i,g-1}}h_{\mathrm{g-1}} \\ \dfrac{1}{2} & \text{if } Y = \mathrm{S_iS_j}, \text{with } \mathrm{P}[Y = \mathrm{S_iS_j}] = 2s_{\mathrm{i,g-1}}s_{\mathrm{j,g-1}} \\ \dfrac{H_{\mathrm{i,g-1}}^{(1)} + H_{\mathrm{i,g-1}}^{(2)}}{2} & \text{if } Y = \mathrm{HH}, \text{with } \mathrm{P}[Y = \mathrm{HH}] = h_{\mathrm{g-1}}^2 \\ \dfrac{H_{\mathrm{i,g-1}}}{2} & \text{if } Y = \mathrm{S_jH}, \text{with } \mathrm{P}[Y = \mathrm{S_jH}] = 2s_{\mathrm{j,g-1}}h_{\mathrm{g-1}} \\ 0 & \text{if } Y = \mathrm{S_jS_j}, \text{with } \mathrm{P}[Y = \mathrm{S_jS_j}] = s_{\mathrm{j,g-1}}^2 \\ 0 & \text{if } Y = \mathrm{S_jS_l}, \text{with } \mathrm{P}[Y = \mathrm{S_jS_l}] = 2s_{\mathrm{j,g-1}}s_{\mathrm{l,g-1}}. \end{cases} \quad \text{(S2)}$$

Here, $H_{\mathrm{i,g-1}}^{(1)}$ and $H_{\mathrm{i,g-1}}^{(2)}$ are fractions of ancestry from source population $\mathrm{S_i}$ for the two parents of a hybrid individual at generation g with $Y = \mathrm{HH}$. We use the superscripts (1) and (2) only to indicate that $H_{\mathrm{i,g-1}}^{(1)}$ and $H_{\mathrm{i,g-1}}^{(2)}$ are independent and identically distributed (IID) random variables, so that if an individual in population H at generation g has two parents from H, the admixture fraction is distributed as the mean of the admixture fractions for two IID random individuals from H in the previous generation."

From there, using the law of total expectations and the binomial theorem, we derived recursion relations for all k-moments of the distribution of admixture fractions from one source population i at generation g (see Verdu and Rosenberg 2011 Supplementary materials p.5):

"For $g = 1$, we have for $k \geq 1$ and any i from 1 to m,

$$E\left[H_{\mathrm{i,1}}^{\mathrm{k}}\right] = s_{\mathrm{i,0}}^2 + \frac{s_{\mathrm{i,0}}}{2^{\mathrm{k-1}}} \sum_{\substack{\mathrm{j=1} \\ \mathrm{j \neq i}}}^{\mathrm{m}} s_{\mathrm{j,0}}. \quad \text{(S9)}$$

For $g \geq 2$, we have

$$\begin{aligned} E\left[H_{\mathrm{i,g}}^{\mathrm{k}}\right] = s_{\mathrm{i,g-1}}^2 + \frac{s_{\mathrm{i,g-1}}h_{\mathrm{g-1}}}{2^{\mathrm{k-1}}} \sum_{\mathrm{r=0}}^{\mathrm{k}} \binom{\mathrm{k}}{\mathrm{r}} E\left[H_{\mathrm{i,g-1}}^{\mathrm{r}}\right] + \frac{s_{\mathrm{i,g-1}}}{2^{\mathrm{k-1}}} \sum_{\substack{\mathrm{j=1} \\ \mathrm{j \neq i}}}^{\mathrm{m}} s_{\mathrm{j,g-1}} \\ + \frac{h_{\mathrm{g-1}}^2}{2^{\mathrm{k}}} \left(\sum_{\mathrm{r=0}}^{\mathrm{k}} \binom{\mathrm{k}}{\mathrm{r}} E\left[H_{\mathrm{i,g-1}}^{\mathrm{r}}\right] E\left[H_{\mathrm{i,g-1}}^{\mathrm{k-r}}\right] \right) + \left(\frac{h_{\mathrm{g-1}}}{2^{\mathrm{k-1}}} \sum_{\substack{\mathrm{j=1} \\ \mathrm{j \neq i}}}^{\mathrm{m}} s_{\mathrm{j,g-1}} \right) E\left[H_{\mathrm{i,g-1}}^{\mathrm{k}}\right]. \end{aligned} \quad \text{(S10)"}$$

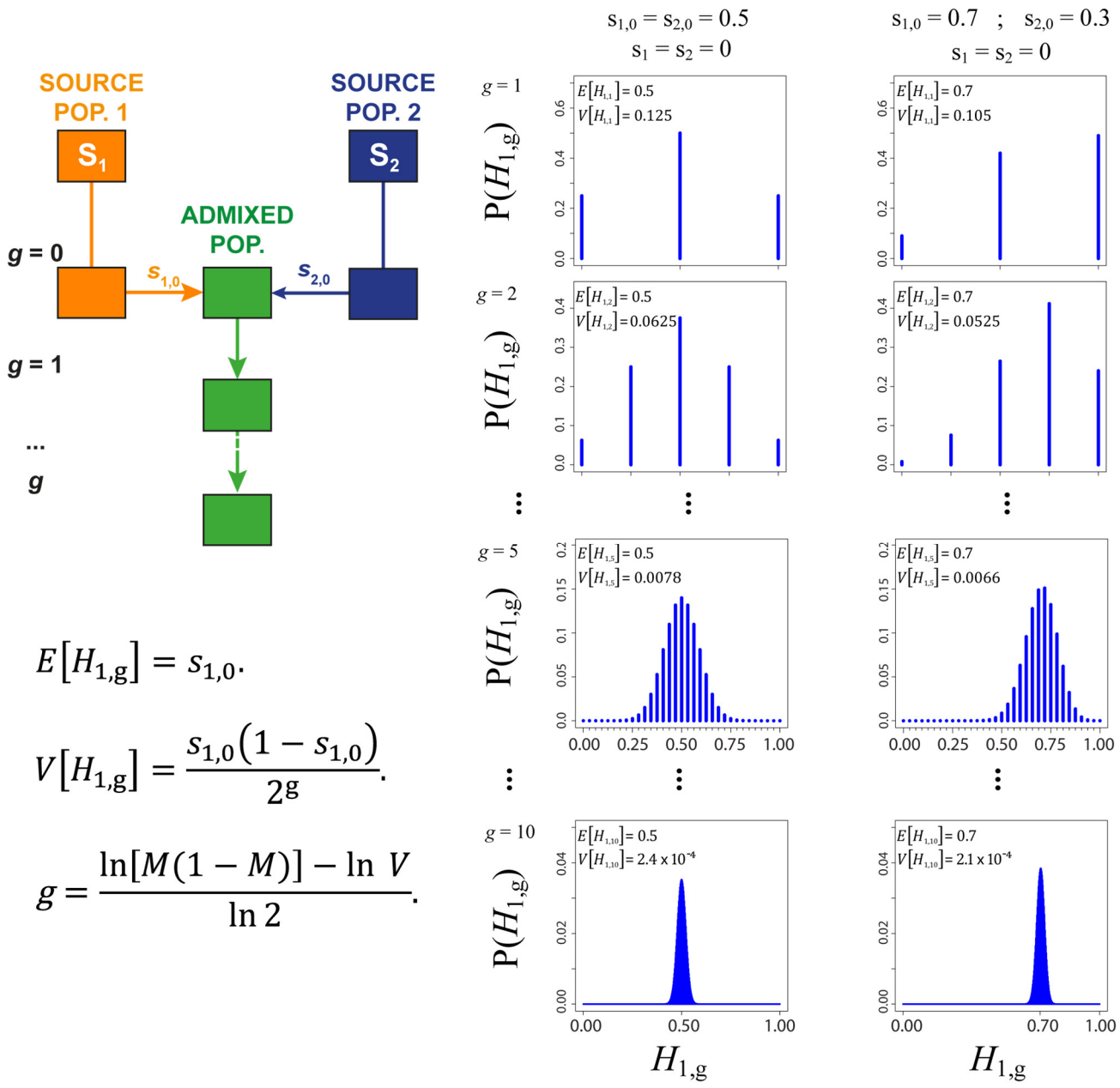
From here, and without further re-doing here the whole previous publication, we could analytically solve, or numerically compute, these recursions to show numerous things detailed in our initial 2011 publication and following published developments.

For the simplest example, under the classical admixture scenario with two source populations founding the admixed population at time 0, and then no longer further contributing to the admixed population (**Figure F3.3.a**), the above equations simplify to give us, for all $g > 0$, $E[H_{1,g}] = s_{1,0}$; meaning that the average admixture fraction across individuals in the admixed population from source population 1, should reflect the initial contribution of source population 1 to the gene-pool of the admixed population under this simple model.

Furthermore, under the same model, we could solve the second moment to then obtain, for all $g > 0$, the variance $V[H_{1,g}] = \frac{s_{1,0}(1-s_{1,0})}{2^g}$, therefore a function of the funding admixture parameter and time since that event.

Therefore, if admixture occurred in such a way in an observed population today (a single event back $g+1$ generations since the DNA sampling), by estimating the distribution of individual admixture fraction from one of the source population from genetic data, calculating simply it's mean (M) and variance (V) would directly give us (within all the fundamental assumptions, and thus limitations, of our general model described above), its initial intensity ($s_{1,0} = M$) and when did the admixture event occurred with $g = (\ln[M(1-M)] - \ln V)/\ln 2$ (see **Figure F3.3.a** below).

More generally, we hereby showed that, theoretically, the shape of the distribution of admixture fractions across individuals within the hybrid population, its observed range and estimated mean, variance, kurtosis, skewness, etc, carried substantial information about all the parameters of the underlying highly complex admixture process that had given birth to our observations. Furthermore, different distributions of admixture proportions from a given source population between different target admixed population would reflect different admixture histories experienced respectively.



$$E[H_{1,g}] = s_{1,0}.$$

$$V[H_{1,g}] = \frac{s_{1,0}(1 - s_{1,0})}{2^g}.$$

$$g = \frac{\ln[M(1-M)] - \ln V}{\ln 2}.$$

Figure F3.3.a.

Examples of the density distribution of the admixture fractions $H_{i,g}$ from one of the source populations (i = 1) in the simplest version of Verdu and Rosenberg (2011) general mechanistic model of admixture presented in the top-left panel, over 10 generations after the founding admixture event. One can see that, over time, the variance of admixture fractions in the admixed population diminishes while the mean remains constant in such simple classical admixture models.

See Verdu and Rosenberg (2011) equations 3-5 for the analytical expression of P($H_{i,g}$) under <u>any</u> two-population version of the general mechanistic model presented in Figure F3.2.a.

All this theoretical ambition was challenged, in practice, by a simple fact, beyond the inherent simplifying assumptions of the model for representing "reality". Can we estimate the distribution of admixture fractions, even in an ideal pseudo-observed population, with sufficient accuracy to identify the parameters of the underlying complex admixture history model in a first place?

With Pr. Erkan O. Buzbas (University of Idaho), at the time a post-doctorate fellow in Pr. Rosenberg lab, we set out, schematically, to evaluate when past admixture parameters of a complex admixture model would cease to be identifiable from the distribution of admixture fractions in the admixed population (Buzbas and Verdu 2018). We considered two two-source-populations versions of our general model: one with a single period of recurring constant admixture, and the other with two consecutive periods of recurring constant admixture, with different constant admixture parameters, respectively. First, we found analytically that, for any model with constant recurring admixture, admixture fractions estimated from diagnostic bi-allelic markers (whose alleles discriminated between the two source populations), inevitably would reach a stationary distribution, and would do so rather rapidly as a function of the precision of our estimations of admixture fractions, within 50 generations at most. This fundamental result is numerically illustrated in **Figure F3.3.b** below, reproduced and adapted from Figure 2 in Buzbas and Verdu (2018).

Second, we found that, when this stationarity would be achieved, model parameters from the preceding period of admixture as well as the duration of the second period of admixture would irremediably be un-identifiable only from the admixture patterns observed in the admixed population (see Figure 5 in Buzbas and Verdu 2018).

These results demonstrated what could be intuitively appreciated from the results of our analytical developments of the general model of complex admixture histories (e.g. **Figure F3.3a**). First, admixture fractions in the admixed population rapidly reach stationarity when the admixture process is recurrent and constant over time. Importantly, note that, a model with a single founding pulse of admixture without subsequent admixture falls into this definition: after the pulse of admixture, the model is a recurrent constant admixture model of intensity 0 from either source. Second, once this stationarity is achieved, observed admixture fractions are only informative about the intensity of this recurring admixture process, and no longer informative about its duration. Finally, when stationarity is achieved, all other older admixture processes having preceded the last period of admixture experienced by the population, are irremediably impossible to infer from admixture fractions only. Thus, while admixture fractions inherently carry extensive information about the underlying complex admixture process, model parameters may still be un-identifiable.

In practice, this meant that a relatively recent admixture process could completely obliterate the signatures in the distribution of admixture fractions inherited from the process having occurred before, and it could do so as a function of the duration of the last process ongoing until sampling in the admixed population. The limit would be 50 generations of no admixture until sampling which would only leave the footprint of the intensity of the last admixture pulse in the admixture fraction and not even how old was this last event, but could also be much more rapidly un-identifiable if the distribution of admixture fractions were not accurately estimated.

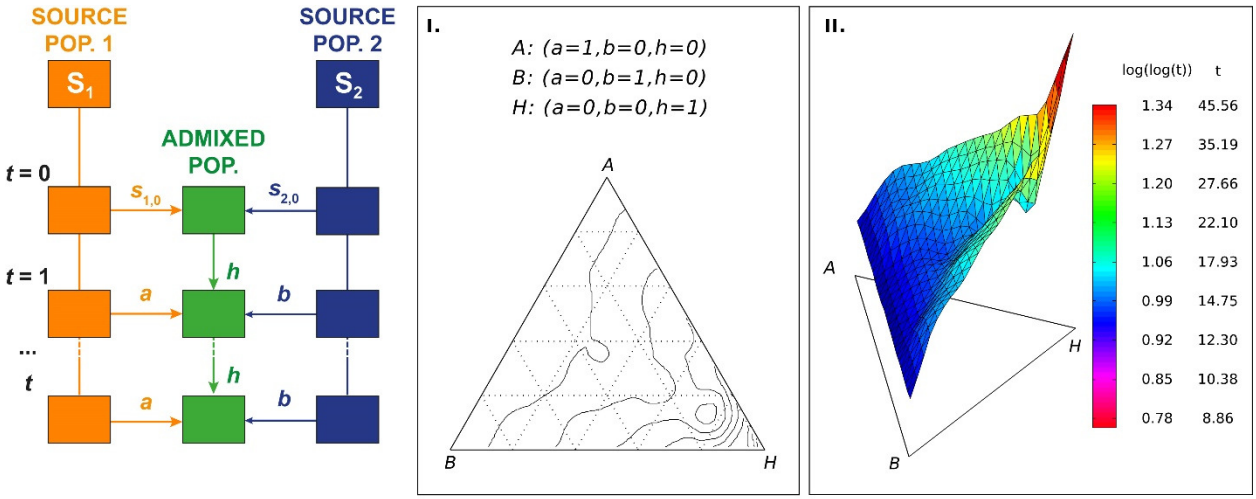


Figure F3.3.b.
Time to convergence to the stationary distribution of admixture fractions under the model described in the top-left panel: after a founding pulse of admixture, the admixed population experiences recurring constant admixture from both source populations until sampling in generation *t*.
Time to convergence increases with the much larger contribution of the admixed population to itself over time (h~1), and with estimation precision of admixture fractions. Note that for low precision estimates (10^{-1}), convergence to stationarity is achieved within a few generations for relatively low values of h (~10%), thus for historical models where the admixed population is largely founded anew at each generation.
Reproduced and adapted from Figure 2 in Buzbas and Verdu *Theoretical Population Biology* 2018.

III.

		Precision							
		10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-10}	10^{-12}	10^{-16}
h	0.09	2	3	3	4	6	9	11	13
	0.19	2	3	4	5	7	12	14	16
	0.29	2	4	5	6	13	14	17	19
	0.39	3	4	5	7	15	17	19	22
	0.49	3	4	6	8	17	19	22	27
	0.59	2	4	6	8	19	21	25	29
	0.69	2	4	7	9	22	24	28	34
	0.79	2	4	7	9	24	27	32	39
	0.89	2	4	7	10	27	30	35	42
	0.99	1	2	6	9	28	32	38	46

3.4. A sex-specific version of the general mechanistic admixture model

With Dr Amy Goldberg (Duke University), a master student under my supervision for one year at Pr. Rosenberg lab, we developed our general admixture model to encompass explicitly sex-specific contributions from either source populations to the gene-pool of the admixed population, as well as that of the admixed population to itself, over time (**Figure F3.4.a**). Indeed, admixture process often occur in a sex-biased manner in humans and non-human species (e.g. Verdu et al. 2013, see **Chapter 2**), and we wanted to investigate how such processes may influence the distribution of admixture fractions over time. Furthermore, this development would allow us to investigate expected admixture fractions from different types of chromosomal compartments, namely autosomes, X-chromosome, Y-chromosome, and mitochondrial DNA, separately, to further gain information on the underlying complex admixture process from comparisons among autosomal and sex-specific genetic data (Goldberg et al. 2014).

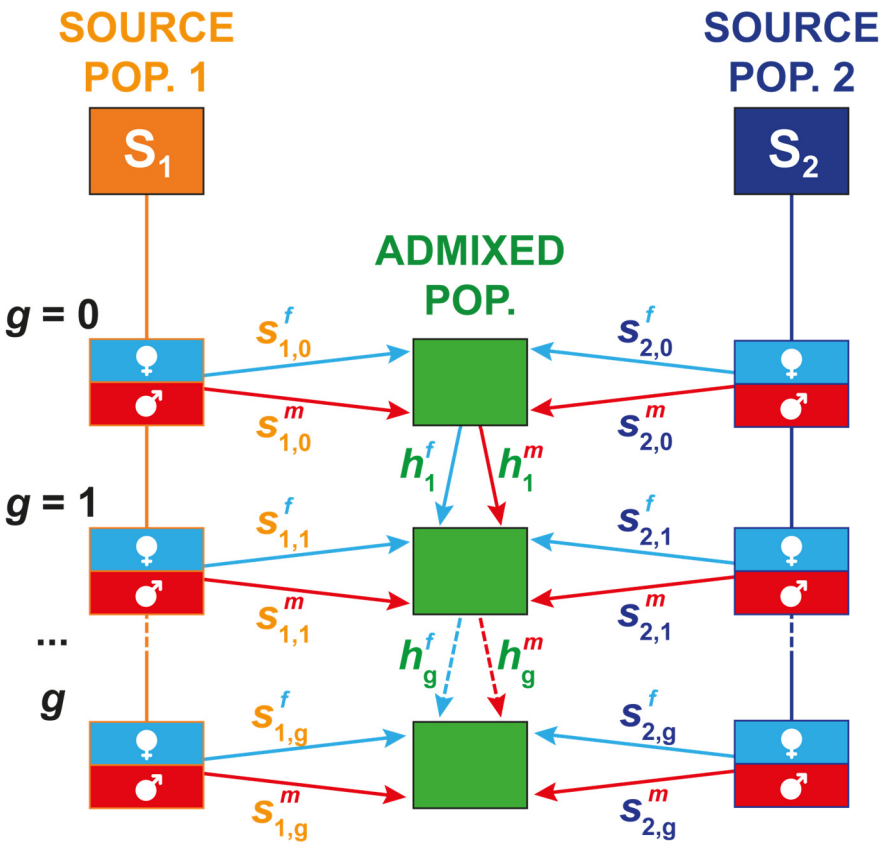


Figure F3.4.a.
Sex-specific two-population version of the general mechanistic model of Verdu and Rosenberg 2011.
Figure reshaped from Figure 1 in Goldberg, Verdu and Rosenberg *Genetics* 2014

We investigated analytically and numerically the distribution and moments of sex-specific admixture fractions in the admixed population as a function of all model parameters, with the exact same dialectic as in Verdu and Rosenberg 2011. We did so "simply" by explicitly decomposing in our recursions the two parents of a random individual in the admixed population as a male and a female parent, for each source and admixed population at each generation. In brief, we found out that, even for classical simple or recurring

constant admixture processes, the distribution of admixture fractions in the admixed population estimated from autosomal data was influenced by sex-biased admixture processes. Indeed, we found that a strongly sex-biased admixture process, with all males from one source population and all females from the other (for an extreme instance), diminished the variance of the distribution of admixture fractions over time more rapidly than a more equilibrated sex-specific admixture process, while the mean admixture fraction would not be affected. For one possible intuitive explanation of the phenomenon: under such strongly sex-biased admixture process and in particular when the total contributions from the source populations are uneven (one contributes more than the other), the effective sizes in the source populations from which parents are drawn are *de facto* reduced, compared to non-sex-biased processes, hence producing admixture patterns less variable in the admixed population (Goldberg et al. 2014).

This theoretical result was both good and bad news. Inferring complex admixture parameters from autosomal data only could result in biased estimates, if a sex-biased admixture process in fact occurred. In other words, in certain parts of the parameter space, a given distribution of admixture fractions estimated from autosomal data could be obtained with two very different complex admixture models with different parameters (duration and admixture rates), one sex-biased and the other not. The good news was that the parameters of complex sex-specific admixture models could be identified by contrasting the distribution of admixture fractions obtained from the admixed population using, separately, autosomal, X-chromosome, Y-chromosome, and mitochondrial data; each chromosomal compartment carrying specific information about the different parameters of the model.

3.5. Perspectives for theoretical developments of complex admixture histories models investigated with genetic data.

Numerous developments of this general theoretical framework are of interest for future analytical work investigating the behavior of complex admixture processes influencing genetic diversity patterns. Among others, a major primary interest to me in my future endeavors is to introduce parameters specifically controlling non-random matting processes in the admixed population. Indeed, we considered here only random matting in the admixed population, whether sex-biased or not. However, admixture processes in numerous populations are likely not random.

For instance, hybrid depression in plants can often be due to individual admixture fractions determining different flowering periods across individuals differently admixed, thus preventing matting among individuals with different genomic admixture landscapes. This classical phenetic consequence of admixture in plants would thus be much better modeled by a general model with an assortative matting parameter itself a function of individual admixture fractions (e.g. Escobar et al. 2008).

For another example of primary interest to my research: admixture processes during the TAST are known to have been strongly sex-biased, in particular during the Plantation Economy era (see **Chapter 5**), but were also highly stratified across socio-economic classes during the colonial history of European monarchic empires. Indeed, social norms would also influence (prevent) marriage opportunities among aristocrats, bourgeois, and lower classes, bourgeois and lower classes being more likely to over-represent genetically admixed individuals compared to aristocrats (Berlin 1998; Carreira 2000; Eltis 2002; Eltis and Richardson 2015). Therefore, complex socio-cultural behavior regarding marriages in socially stratified populations may result in non-random matting processes further related to the genetic admixture histories of the admixed population. I am therefore in the process of explicitly parameterizing non-random processes in admixed populations under our general model, and similarly investigate its effect analytically on the distribution of admixture fractions over time; a work still in progress today (see **Chapter 5**).

Finally, an intuitive and very promising development for the novel theoretical framework here developed would be to investigate, under the same general mechanistic model, another quantity of major interest to the investigation of admixture processes: the distribution of length of admixed-chunks from each source populations along the genomes of individuals in the admixed population over time. Indeed, since at least the formalized "junction" theory of Fisher (Baird 2006), introducing the notion of admixed-chunks of chromosomes due to recombination in the genome of admixed individuals, we expect the distribution of length of admixture segments in high-LD along the genome of admixed individuals to bear extensive information about the admixture process and in particular about the time since admixture events. This was thus a very promising development for our theoretical framework… which was brilliantly developed independently from us and published the year after our initial paper by Pr. Simon Gravel (McGill University), a post-doctorate student in a neighboring Stanford lab at the time (Gravel 2012). The fantastic paper from Pr. Gravel deployed the exact same mechanistic model inspired from Ewens and Spielman (1995) as the one we proposed, and derived under it the expected distribution of admixed-chunks lengths from each source populations. Most importantly, Pr. Gravel did not "only" conduct analytical investigations as we did in our paper, but further provided a readily usable algorithm implemented in a software, TRACTS, allowing to infer model parameters deterministically from observed distribution of admixture LD. This, very legitimately, brought extensive attention from the human population genetics community that, since then, has extensively used TRACTS to infer the complex admixture models underlying observed genetic patterns. Despite this immense success, however, similarly to us, Pr. Gravel could not derive closed forms

of the expected distribution of admixed fragments lengths for admixture models encompassing more than two punctual pulses of admixture. As a consequence, observed data could only be fitted to exponential curves expected under either single or two-pulse scenarios over the entire admixture history of the admixed population. As acknowledge and extensively discussed by the authors in methodological papers developing on Pr. Gravel's seminal approach, this was a computational limit of these approaches forbidding empirically to investigate more complex admixture processes possibly underlying the observed data (Gravel 2012; Hellenthal et al. 2014).

We thus proposed to overcome at least in part this limitation with the novel ABC framework presented in the following Chapter.

Chapitre 4

MetHis: a novel Approximate Bayesian Computation framework for reconstructing complex admixture histories from genetic data

Alto Mira.
Santo Antão, Cabo Verde, 2016
©Paul Verdu

Chapter 4. *MetHis*: a novel Approximate Bayesian Computation framework for reconstructing complex admixture histories from genetic data

I introduced Approximate Bayesian Computation methods in **Chapter 2.2** above (Tavaré 1997; Pritchard et al. 1999; Beaumont et al. 2002), emphasizing on how flexible and promising these inferences approaches were. Indeed, in principle, models of an arbitrary complexity could be inferred from genetic data with ABC, provided that simulations under said models could be efficiently performed, and provided that summary-statistics would carry identifiable information about the underlying model-parameters. Therefore, as other maximum-likelihood approaches were inherently unable to consider admixture models more complex than two-admixture pulses' models (see **Chapter 3.5**), I decided to opt for ABC inference using the general model presented in **Chapter 3**.

Indeed, we had shown theoretically that the distribution of admixture fractions in the admixed population carried substantial information about the model-parameters of the underlying complex admixture process (Verdu and Rosenberg 2011). Therefore, as numerous population genetics statistical methods already existed to infer individual admixture fractions from varied source populations in the observed admixed population, we could readily use them to obtain the distribution of admixture fractions observed from genetic data, and use this distribution, or rather the statistics describing it such as its' mean, quantiles, variance, kurtosis, and skewness, as summary-statistics for ABC inference with our general admixture model. Therefore, I "just" needed to be able to simulate large amounts of realistic genetic data under the general model, calculate summary statistics on each simulation including statistics describing the distribution of admixture fraction in the sampled admixed population, and then use the obtained reference tables of vectors of model parameters used for each simulation associated with the corresponding vector of summary-statistics, to apply existing ABC algorithms to perform ABC scenario-choice and posterior-parameter inferences using observed genetic data (**Figure F4**).

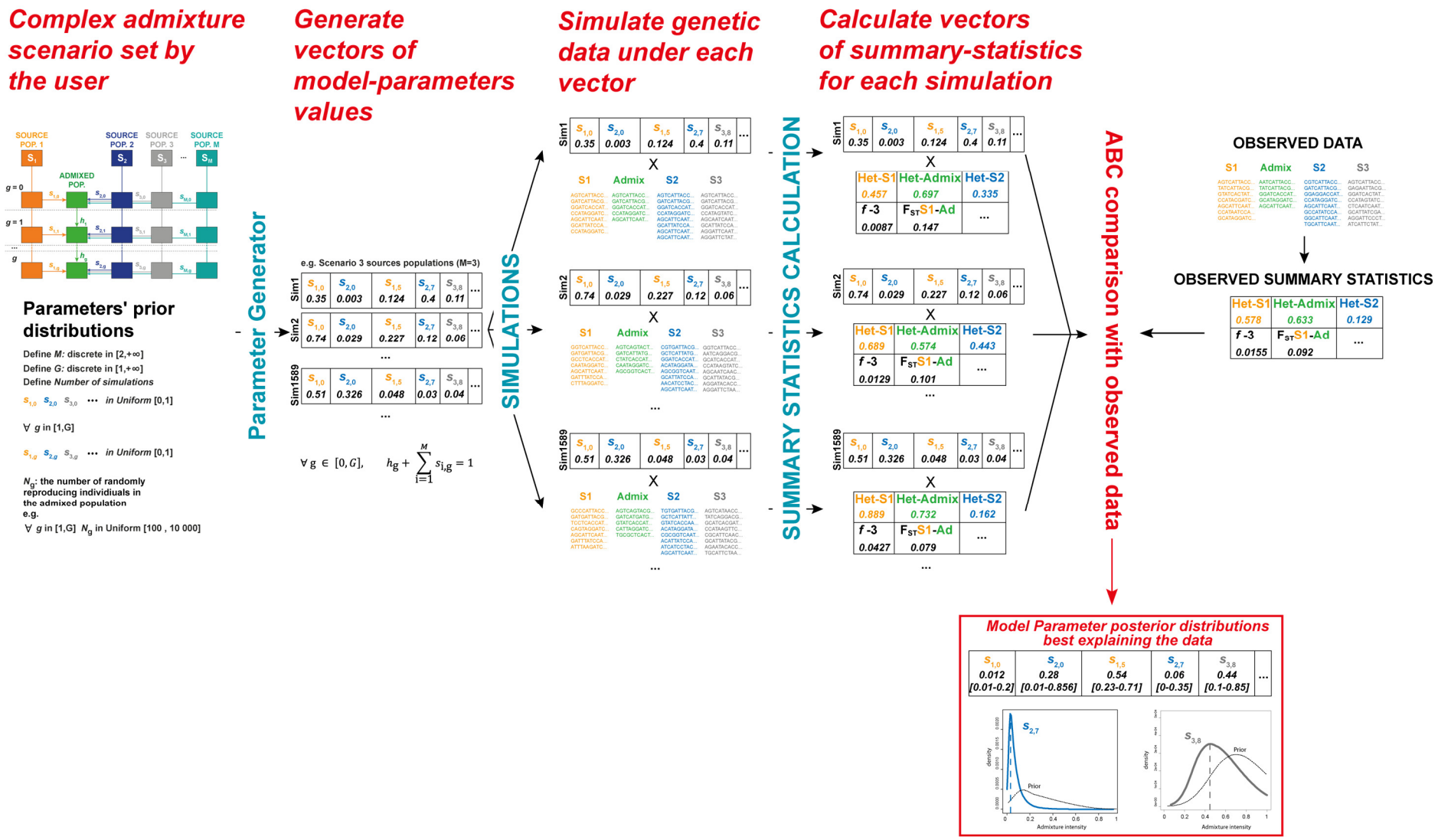


Figure F4.

Schematic general workflow of complex admixture history inference from genetic data using Approximate Bayesian Computation

Freshly recruited in the CNRS in 2012, and appointed to my former lab UMR7206 Eco-anthropology under the direction of Pr. Bahuchet and Pr. Heyer, I set out to fulfill this apparently sound endeavor (**Figure F4**), but rapidly hit a completely foreseeable wall I had nonetheless not expected: simulating data under the general admixture model we had proposed was not trivial. Indeed, I naïvely thought that I could readily use existing fantastic coalescent-based simulators such as *MS-MS* (Ewing and Hermisson 2010) or *FastSimcoal* (Excoffier and Foll 2011), and simply implement the versions of admixture or migration fitting my complex admixture models. It turned out that, while indeed these simulators can be efficiently adapted to simulate data under the general model, the nature of classical coalescent simulations where not fitting the fundamental definition of our models.

Indeed, classical coalescent approaches reconstruct different genetic pedigrees underlying each observed genetic marker while, in our models, each individual's genome stem from a single genealogical pedigree. As a result, when conducting classical coalescent simulations under the simplest admixture scenario from two source populations, at the first generation after the admixture event, all simulated individuals show the same admixture fraction from one of the source populations, proportional to the intensity of the admixture parameter set for the simulation. We thus do not have individuals with 0%, 100%, or 50% admixture from this source at the first generation, with a mean of the distribution of individual admixture fractions in the admixed population close to $s_{1,0}$, the intensity of the founding admixture event from one of the source populations, as expected in Verdu and Rosenberg 2011 (see **Chapter 3**). Instead, we obtain all individuals with a proportion of admixture close to $s_{1,0}$ at the generation after the admixture event under the coalescent, as each independent markers are coalescing within this source population with probability $s_{1,0}$, per definition. To convince one-self of this possible caveats, I had the chance to supervise Dr. Marguerite Lapierre, master student at the time, to conduct series of coalescent simulations showing exactly the patterns expected above at odds with the theoretical expectations from our general admixture model.

Therefore, to mimic our general admixture model using existing backward-in-time simulators under the coalescent, we first needed to define pedigrees for each sampled individual in the admixed population, and then conduct coalescent simulations of genetic markers within these pedigrees. Such approach first simulating a full pedigree, and then simulating under the coalescent within this pedigree, had been brilliantly proposed in 2012 by Wakeley et al., to further investigate natural selection processes under the coalescent, for models possibly comprising admixture across populations. However interesting, I was left with the need to first simulate pedigrees, a complex task in itself as I needed to set priors and processes that I barely knew anything about in human populations in general, nor in admixed populations related to the TAST.

In this context, I decided to build my own genetic data simulator under our general mechanistic model of complex admixture histories. I decided to opt for a forward-in-time simulator as only such approach would allow the flexibility that I needed to consider any version of this general model fully parameterizable by the user, and as I thought this would yield the most potential for future developments implementing non-random matting or admixture-related natural selection processes (see **section 4.3** below). Furthermore, I would need to design this simulator for performing efficiently under ABC; which meant with parameter-priors easily set by the user, drawing large numbers of vectors of parameter values from these priors readily usable to perform numerous simulations as needed for ABC. In turn, we would need to efficiently calculate summary statistics for each simulation in order to obtain the two reference tables, one for the vectors of parameter values and the other for the corresponding vectors of summary statistics.

4.1. *MetHis* software for simulating data and calculating summary-statistics under the Verdu and Rosenberg 2011 general mechanistic model of complex historical admixture

After obtaining funding from the French ANR in 2015-2016 for five years, I could hire a post-doctoral student, Dr. Cesar A. Fortes-Lima (University of Uppsala), to start implementing such simulator and conduct an initial proof of concept. Indeed, ABC is an *a posteriori* class of inference method. In practice, this means that, as explained above, once it is determined that ABC can work in principle, then one has to perform the ABC analysis until the end to actually determine whether it has worked in the specific study-case, or not. The good news is that, even if it does not work *a posteriori*, one will be able to determine in details which step of the process failed, and why; whether simulations were unable to mimic the observed data, whether summary-statistics were overall uninformative about model parameters, whether the parameter space was too large producing high-dimensional summary-statistics spaces in which different ABC inference approaches get lost and fail to produce posterior parameter distributions departing from their priors, or else.

Dr. Fortes-Lima managed to build a first pre-alpha version of the software following the algorithm I had designed (below), in Python language. After we had shown that indeed, this could produce genetic data whose behavior were very closely predicted by our analytical expectations under the model, we decided that a full-blown C version of our preliminary software was worth producing for publication to the community. However, after 18 months of post-doctoral contract with me, Dr. Fortes-Lima had another opportunity at Dr. Schlebusch's lab in Uppsala, which I supported him for as it provided him with data analyses opportunities more aligned with his ambitions than the methodological work conducted with me.

As I did not know how to code in C, I started looking for private services possibly interested in translating our initial software into C, and failed to find any for such specific fundamental research endeavor. Fortunately for me, I managed to get Dr. Romain Laurent, Research Engineer in the lab, interested in the project, thanks to our initial proof of concept. I had long known Dr. Laurent, and his skill, enthusiasm and scientific rigor ultimately allowed this project to be completed with success. We have worked extensively together on this project on a daily basis for more than eight months, but I need to emphasize here that he is the one who really built the software's code, and I have only assisted him for general algorithmic and detailed scientific strategic choices and posterior benchmarking and checks, but I did not write the code myself for this software, even if I conducted myself most subsequent ABC inferences.

We thus produced and benchmarked a novel software, called *MetHis*, coupled with existing ABC tools for scenario-choice and posterior-parameter inferences using genetic data, and specifically designed for investigating complex admixture histories (Fortes-Lima, Laurent et al. 2021). While this method and results are extensively described in the resulting publication in *Molecular Ecology Resources* (2021), where interested-readers shall find in a synthetic scientific way all the needed information to re-use our methods and tools, I decided to explain below how and why several choices were made during the process, to better explain its dialectics. I will then very briefly summarize results from the publication. Finally, I will further expand on ongoing methodological implementations, as well as future perspectives for this new set of population genetics tools.

4.1.a. Simulating genetic data under the Verdu and Rosenberg 2011 model

As previously said in **Chapter 1** and in **Chapter 3**, I was ultimately interested in reconstructing the complex admixture histories of enslaved-African descendants admixed populations since the 15th century on either side of the Atlantic. Previous extensive human population genetics investigations had shown that, in numerous such populations in the Americas, including the Afro-American communities in the USA, admixture patterns had occurred mainly between European and African continental source populations (e.g. Baharian et al. 2016; Mathias et al. 2016; Micheletti et al. 2020; Fortes-Lima and Verdu 2021).

How many source populations did we want to consider in our simulations?

I therefore decided to opt for building our genetic data simulator for the two-source population version of our software. This was a pragmatic choice substantiated by the very definition of our general mechanistic model. Indeed, while M possible source populations for the admixture of an admixed population would have been, in many study-cases, more suitable, it was far easier to start with a two-source population model. Furthermore, in **Chapter 3** we investigated the influence of complex M population admixture models on the distribution of admixture fraction across individuals within the admixed population, but calculated for a single given source population at the time. In other words, while admixture could stem from many sources, we would investigate the history of admixture specific to each source population, separately and in turn. In this framework, a two-source population version of our model could be used sequentially, one source population being the targeted population from which we would reconstruct introgression into the admixed population, and the other source population could, conceptually, embed all other source populations involved in the admixture process ("one minus the targeted source population"). Then another of these populations could be the focus, while the second source population would embed all other remaining sources, including, this time, the first source that we had previously investigated. Therefore, while conceptually not as satisfactory as readily building a simulator for M source populations, the simple two source population models could, in principle, be hijacked for investigating indirectly M source population models, an interesting trade-off knowing the complexity of building, from scratch, a genetic data simulator with numerous possible source populations.

Which populations in our models did we want to simulate?

We then decided that our simulator would focus on simulating data in the admixed population over the mechanistic history of admixture from $g = 0$ until the present, when a sample mimicking the observed one would be drawn from the admixed population and used for summary statistics calculation. We thus decided not to simulate the genetic evolution of the source populations during the admixture process, but instead obtain the source population data, either from observed data for recent admixture histories (such as the TAST histories of admixture which only occurred within the roughly last 20 generations until today), or using simulations under a classical coalescent to generate data from isolated source populations that our simulator would then take as input for simulating the admixed population.

Which type of genetic data did we want to simulate?

Previous complex admixture histories reconstructed with maximum-likelihood approaches, such as TRACTS (Gravel 2012) or GLOBETROTTER (Hellenthal et al. 2014), required massive amounts of data to operate satisfactorily, as patterns of admixture LD decays had to be estimated along the genomes of admixed individuals. This often required several hundreds of thousands of SNPs, which represent still a

substantial financial investment not easily obtainable even for human genetics investigations, and clearly out of reach of most non-human species.

We did not need a priori such extensive data to estimate inter-individual distribution of admixture fractions in the admixed population, from which we hoped to gain sufficient information for parameter estimation, theoretically. We therefore opted to simulate any number of autosomal independent SNPs chosen by the user, and then evaluate how our method performed for different absolute amounts of data. From previous experiences with investigating publicly available SNPs sets in Afro-Americans and putative source populations from continental Europe and Africa, we evaluated that 100,000 autosomal independent SNPs were sufficient to obtain estimates of individual admixture fractions from either Africa or Europe that would not massively change when considering more SNPs (Mathias et al. 2016; Micheletti et al. 2020).

Ultimately, the question "how many markers do you need?" is a non-trivial and circular question a priori. This all depends on how reproductively isolated the populations at the source of admixture have been prior to the admixture event. The closer genetically the two source populations are, the more SNPs one will need to estimate accurately the admixture patterns in the admixed population. Alternatively, the more genetically distant the source populations are, the less markers will be needed to do so.

In this framework, investigating human populations is often much more challenging than other non-human admixture processes. Indeed, *Homo sapiens* is a very recent species on evolutionary time scales, and so are divergences across populations, leaving little amounts of time for mutation and drift to differentiate isolated sub-populations. Therefore, when considering admixture histories between continental Europe and Sub-Saharan Africa at the time of the TAST, we are working with F_{ST} values between the source populations in the order of 0.07-0.10 at the genome-wide scale, a much lower value than when investigating hybridization histories between populations genetically isolated for tens of thousands of generations. As a consequence, while investigating human genetic history may sound "easy" thanks to the wealth of data already accessible, or possibly generated, it is in practice often very hard, as even such massive amounts of data may be insufficient to distinguish source populations with a very recent evolutionary history in a first place.

Nevertheless, we did not know a priori how our ABC inferences would perform as a function of the number of genetic markers, and in all cases, numerous non-model species do not have access to genome-wide SNP data[47] but rather investigate microsatellite markers, as we had investigated in Central Africa (see **Chapter 2**). We therefore decided to build the simulator to allow different types of data to be used. We readily implemented independent autosomal SNPs, and any type of microsatellite data specified by the user and with a General Mutation Model allowing for insertion and deletion, fully parameterized by the user and whose parameter-values could also be drawn from priors for ABC inferences (Fortes-Lima, Laurent et al. 2021).

How did we simulate individual genetic data under the two-source population version of Verdu and Rosenberg 2011?

We therefore aimed at simulating independent genetic markers in the admixed population as a function of model parameters and in a two-source population model. We simply applied a random matting process, creating random gametes from randomly drawn parents (without selfing) in each source and admixed

[47] Note that is less and less the case with the advances of RAD sequencing and Next Generation Sequencing, now allowing to collect numerous makers from non-model species even lacking well known reference genomes. It is still a financially costly endeavor for the investigation of non-model species requiring extensive amount of highly skilled work in molecular genetics and bio-informatics.

population at each generation, following model parameters values, and pairing them to produce the offspring in the admixed population at the following generation. At the end of the admixture process at time G, we sample individuals with sample-sizes matching observed sample-sizes, either randomly or setting a threshold of parental relatedness as is commonly done in human genetics by flagging individuals' genealogies over the last 2 generations. Note that while we implemented a mutation model for microsatellite data, we did not implement a mutation model for genotyped independent SNPs. Indeed, the values of a mutation rate for such data is not consensual, and in any case low and slowly affecting population allelic frequencies. While it is reasonable to neglect such mutation for such type of markers for recent evolutionary processes, such as those having occurred during and after the TAST, this might be an issue when considering, in the future, much older admixture histories. Finally, note that while only these types of markers were readily implemented in *MetHis*, Dr. Laurent coded them as "boxes" that could in principle easily be filled with independent genetic sequences rather than SNPs. Nevertheless, developing *MetHis* for sequences rather than SNPs or microsatellites, will need to elaborate on sequence mutation and recombination models which is not trivial, albeit completely feasible without changing the core architecture of the software (Fortes-Lima, Laurent et al. 2021).

4.1.b. Calculating summary-statistics for ABC inferences

For our proof of concept with Dr. Fortes-Lima, we simply had resorted to a constellation of existing software to calculate summary statistics on each simulation. This was computationally tedious and costly as simulated data had to be read by multiple different software. We thus obviously opted for embedding summary statistics calculation within a "Summary-Statistics calculation" tool within *MetHis*. This allowed to substantially diminish memory costs and calculation efficiency, a code-writing task nevertheless non-trivial and tedious that, again Dr. Laurent supervised, with the help of Antoine Cools, Ferdinand Petit, and Maël Pretet, Licence and Master students in bio-informatics at the time, that I had hired and supervised for these tasks over the years.

We thus chose to implement a number of classical population genetics ABC summary statistics, likely informative about our model parameters. As detailed in Supplementary note of Fortes-Lima, Laurent et al. 2021, we computed a series of "within-population" statistics for the source populations and the admixed population, separately, including, for SNP data, SNP-by-SNP expected heterozygosities (Nei 1978), their means and variances, the means and variances of individual pairwise ASD within each population, and an inbreeding coefficient F (Danecek et al. 2011). For microsatellite data, we computed the mean number of alleles per markers within each population, the mean expected heterozygosity, the mean allele size variance across markers, the mean and variance of individual pairwise ASD (Bowcock et al. 1994) within populations. As per population genetics theory, and empirically shown in ABC classical approaches since Beaumont et al. (2002), these statistics were often informative about effective sizes and drift within populations, and also indirectly reflecting admixture across populations (detected possibly from an increase in heterozygosities due to gene-flow for instance). We also computed several classical "between-populations" statistics such as, for both SNP and microsatellite data, multilocus pairwise F_{ST} (Weir and Cockerham 1984), f3-admixture statistics (Patterson et al. 2012), as well as mean pairwise population ASD.

In addition to these classical ABC summary-statistics, and as extensively explained above and in the previous **Chapter 3**, we developed sets of summary-statistics specifically describing the distribution of admixture fractions from one of the two source populations (the other one being one minus this one in such model) across individuals sampled in the admixed population. As said, while numerous methods exist for

estimating such values from genetic data, most are tedious or very tedious computationally and sometimes conceptually, rendering most methods out of reach for application in an ABC framework where they would need to be computed for tens or hundreds of thousands of simulations. We therefore implemented a new a particular estimate of individual admixture fractions based on pairwise ASD that we already calculated for other summary-statistics. Based on the statistical description of allelic frequencies in an admixed population since Bernstein (see **Chapter 3**), we expected admixed individuals to be at intermediate genetic distances from either source populations, proportionally to their average admixture fraction. Thus, projecting the ASD matrix using Multi-Dimensional Scaling could provide us with an estimate of such distance proportional to admixture fractions (**Figure F4.1.b**), very efficiently computationally. We therefore implemented such calculation to obtain the estimated distribution of admixture fractions from one source population within the admixed population. From this distribution, we used minimum, maximum, all 10% quantiles, mean, variance, kurtosis, and skewness, as 16 separate summary statistics for ABC scenario-choice and posterior parameter inferences.

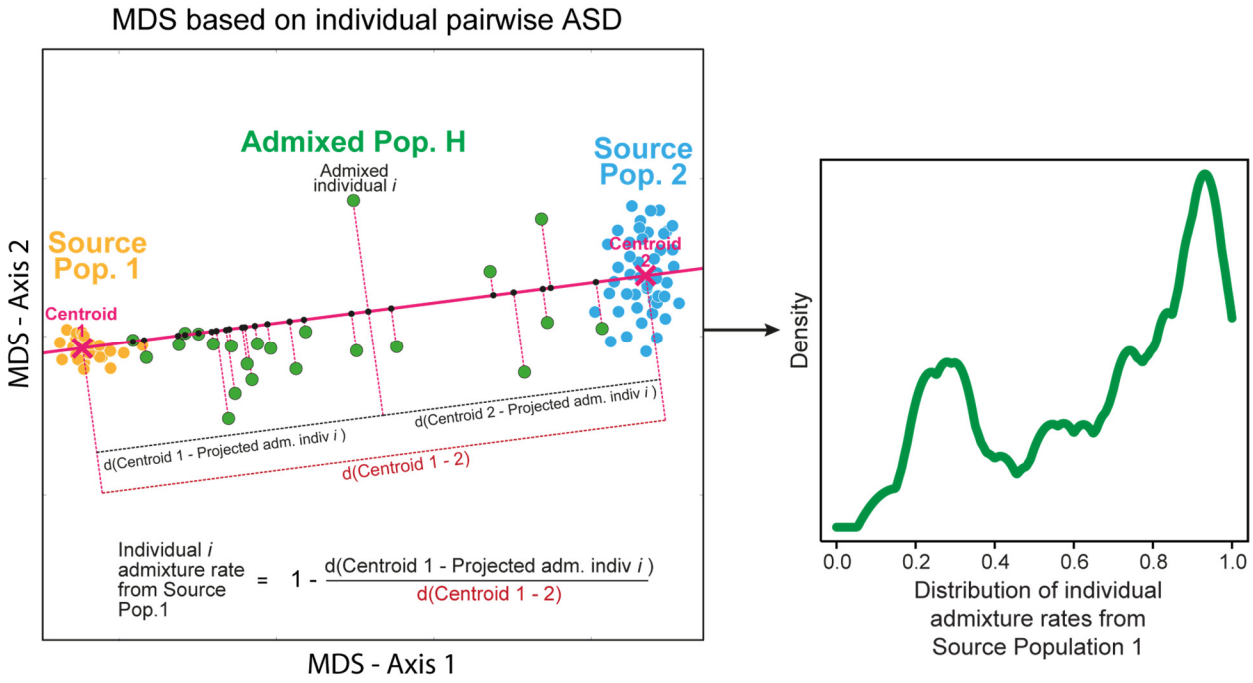


Figure F4.1.b.

Schematic representation of ASD-MDS statistical estimation of average individual admixture fraction from one of the source populations in a two-source population admixture model implemented in *MetHis*. Schematics represent estimation in a 2-Dimension MDS projection for simplicity, but note that estimates obtained with *MetHis* consider, instead, a 3D MDS projection for such calculations.
Figure originally presented in Supplementary Figure SF12 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

Note that, by doing the calculation presented in the schematic **Figure F4.1.b**, we could obtain a priori "negative" admixture fractions for admixed individuals projected in the MDS on the other side of the centroid of a given source population. This would ultimately reflect a high resemblance of this individual with the other individuals in the source population. This was not a problem per se in ABC, as this could happen with real observed data, and our goal in ABC is to best mimic summary statistics obtained from real data, whatever their meaning.

4.2. *MetHis* coupled with machine-learning ABC in practice

4.2.a. MetHis-Random Forest ABC scenario-choice; MetHis-Neural Network ABC posterior parameters joint inferences

MetHis produces reference tables ready to be used for ABC scenario-choice and parameter inference procedures using existing software implemented in R (Csilléry et al. 2012; Pudlo et al. 2016; Raynal et al. 2019). In 2016, during the initial development of *MetHis*, a novel ABC scenario-choice approach relying on Random Forest machine learning (Breiman 2001), was proposed and rapidly adopted by the community as a massive improvement from previous scenario-choice approaches relying on regression among other methods (Beaumont et al. 2002). Indeed, in Random Forest ABC (Pudlo et al. 2016), authors proposed to build a random forest of decision trees based on the summary statistics calculated from simulations, and then used this random forest for predicting which class of competing scenarios produced vectors of summary statistics closest to the observed ones, comparing, to do so, each scenario performances.

ABC scenario-choice had always been, to me, of major interest for formally testing competing scenarios from the data, compared to the conceptually and statistical difficult comparison of maximum-likelihood values obtained by fitting real data separately to scenarios comprising different sets of parameters. Random Forest ABC in this context represented a major practical, and even philosophical, improvement compared to previous approaches (**Figure F4.2.a**). Indeed, it is a categorization algorithm rather than a regression or MCMC algorithm, thus to my views much better suited for the categorization question at stake for scenario-choice procedures. And indeed, it proved to perform much more efficiently with equal accuracy compared to previous methods: one needs ten to a hundred folds less simulated data to perform robust scenario-choice, and, furthermore, RF-ABC is not affected by correlations among summary-statistics, a very concrete frequent issue of previous methods (see Sisson et al. 2019). Indeed, one can even formally use RF-ABC to classify the importance of each statistic on the shape of the random forest and on the prediction, to illustrate how each summary-statistics relatively contributes to the finale decision (Pudlo et al. 2016).

While Random-Forest ABC has also been developed for posterior-parameter estimation (Raynal et al. 2019), it still faces major inherent difficulties. Indeed, as we just said, RF is a classification algorithm, and therefore only provides posterior estimates of each parameter separately, rather than jointly, and for quantiles of the posterior distribution rather than the full distribution itself. While one can, *a posteriori*, reconstruct the distribution from each estimated quantile, the risk of over-fitting or, alternatively, of increased noise is high. In all cases, we prefer to estimate all parameters jointly rather than independently, as our evolutionary models incorporate possible trade-offs across the various demographic parameters which would be better accounted for with a joint estimation of vectors of parameters, rather than when considering original parameters separately (Fortes-Lima, Laurent et al. 2021). We therefore opted for machine-learning Neural Network ABC posterior estimation (Csilléry et al. 2012), as our primary ABC posterior parameter inference method, albeit note that the reference tables produced by *MetHis* are readily usable with other classical ABC software implementing other such methods.

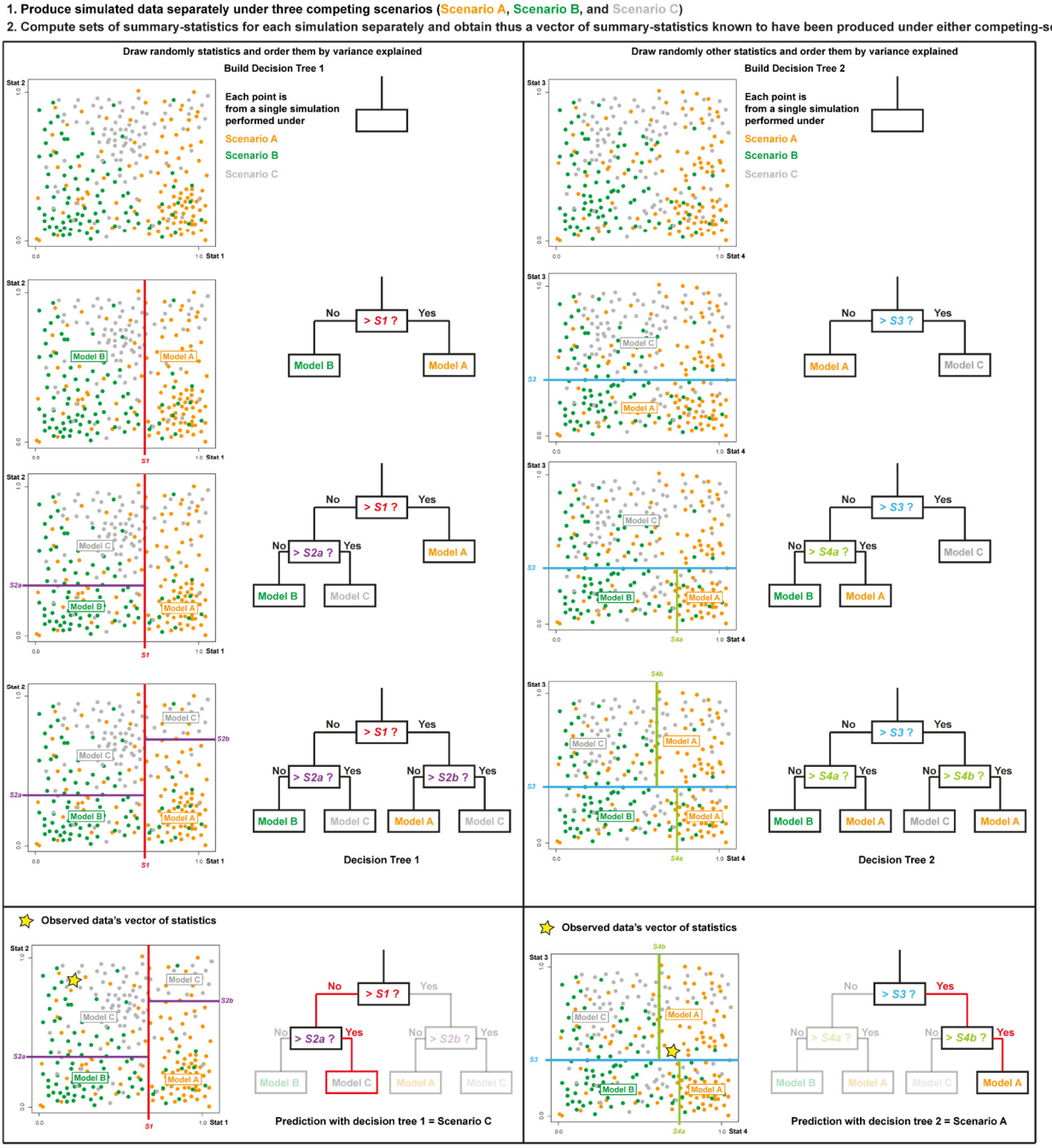


Figure F4.2.a.

Schematic representation of Random Forest ABC scenario choice for simulations performed under three competing scenarios

4.2.b. A case study for evaluating MetHis-ABC performances in practice

As I was interested, initially and primarily, in investigating complex admixture processes having occurred during the TAST, we evaluated the "real-life" performances of our novel *MetHis*-ABC approach for a case study of two separate enslaved-African descendant admixed populations in the Americas, for which extensive data were readily available in public database: a sample from an Afro-American community from South-West USA (Afro-American ASW), and a sample from Barbados (Barbadian ACB). Previous studies had already shown that both populations were admixed mainly between North-Western European populations and West Central African populations for which, in existing databases, the British GBR and Nigerian Yoruba YRI represented good proxy source populations in the 1000 Genomes Project (2015) database[48] . Beyond my personal interest in the outcome of such case-study, this was a challenging task that could challenge the limits of our approach while explicitly showing its' novelty compared to previous approaches.

First, as said before, African and European populations are, in absolute, not that genetically different from one another, a thus challenging problem for reconstructing admixture histories that rely primarily on the level of genetic differentiation of putative source populations. Second, this was a recent (roughly 20 generations ~ 500 years) admixture history, which further challenged parameter estimation in such a small time-frame. Finally, we were also interested in evaluating whether recent voluntary migrations since the 1960's influenced admixture patterns, thus allowing us to evaluate empirically our power to identify such recent events from genetic data. This latter option is not trivial with previous maximum-likelihood methods relying on admixture LD decay (Gravel 2012; Hellenthal 2014). Indeed, very recent admixture events result in long stretches of admixture from either source populations, associated with high LD within stretches, in the genomes of first-generation admixed individuals. These very recent stretches prove hard to scale and fit to the expected curves of LD decay, and they further may be confused with long stretches of admixture in lower LD due to shorter (older) admixture stretches having recently recombined. They are thus often excluded from analyses in practice, which *de facto* prevents the researcher to evaluate very recent admixture events with these methods.

We considered nine competing scenarios of highly complex admixture histories, combinations of three classes of models from either source population (**Figure F4.2.b1**). We considered scenarios where either African or European source populations (or both) contributed to the gene-pool of the admixed population in two separate independent pulses drawn randomly in the 20 generations after the original founding admixture pulse. Alternatively, we considered scenarios where either source population (or both) contributed in a recurring monotonically decreasing way to the gene-pool of the admixed population after the independent initial founding admixture pulse at generation 0. Finally, we considered scenarios where either source population (or both) contributed in a recurring monotonically increasing way to the gene-pool of the admixed population after founding. This latter class of scenario was put into competition to explicitly evaluate whether recent admixture processes may overwhelm previous signals of admixture, in particular as historical demographic data demonstrated, at least for the USA, that more than ten times more people voluntarily emigrated to the USA from Africa since the 1960's than the total amount of enslaved-Africans forcibly deported from Africa to the USA over the last, most intense, two hundred years of the TAST until the 1810's (Berlin 1998, 2010; Eltis and Richardson 2015).

[48] Which was also consistent with expectations from historical records as both regions (the USA and Barbados), had been British peopling colony during the TAST, and the Bight of Biafra being a major zone of embarkation of enslaved-Africans within the British Empire TAST commercial network.

Note that recurring admixture scenarios were implemented minimizing the number of parameters thanks to Dr. Bruno Toupance (Université Paris Cité), who adapted a classical rectangular hyperbola function to simply control the shape of the recurring admixture process between the onset time and associated intensity and the offset time and intensity, for a period of time set with priors by the user. In between, admixture parameter values at each generation are the numerical solutions of this scaled rectangular hyperbola function for which the steepness parameter "u" is drawn between 0 and 0.5. In practice, for decreasing admixture scenarios, a value of $u = 0$ corresponds to a very steep recurring admixture, virtually constant until the end of the admixture period after an initial pulse at the beginning of the period; $u = 0.5$ instead corresponds to a linear process between the onset and offset intensities. Note that for equivalent onset and offset admixture intensities and a value of u equal to 0.5, the period of recurring admixture corresponds to a constant introgression from the source population of that intensity, thus corresponding to a classical unidirectional migration model during that period (see **Chapter 3**).

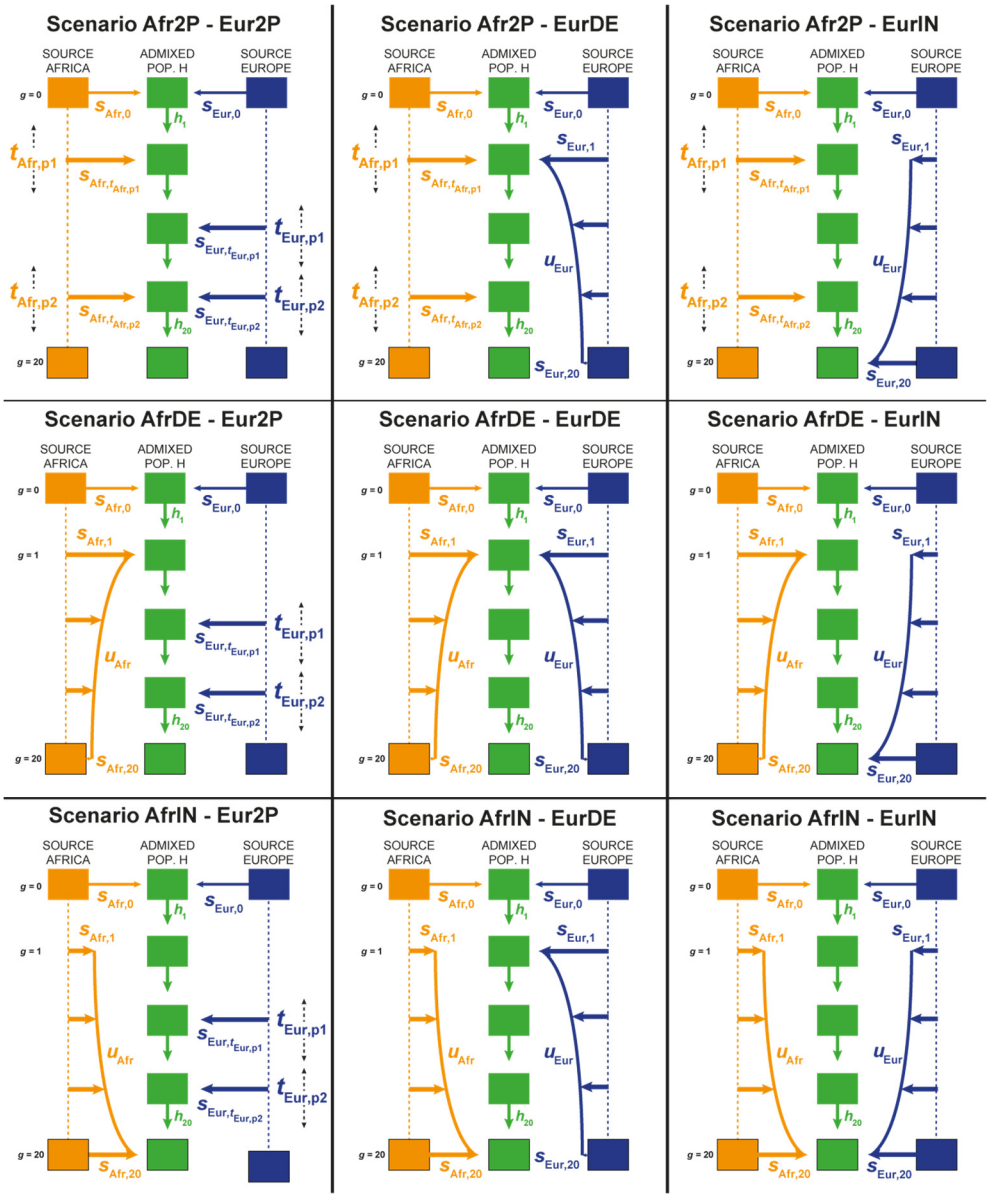


Figure F4.2.b1. Nine competing scenarios of complex admixture histories for the genetic patterns of the Afro-American (ASW) and Barbadian (ACB) population samples from the 1000 Genomes Project, investigated with *MetHis*-ABC. **We conducted 10,000 separate simulations under each competing scenario for ABC scenario-choice with Random Forest, and 100,000 simulations under the winning models for further ABC posterior parameter estimations with Neural Network ABC.** Detailed description of the models and parameter priors are presented in Fortes-Lima, Laurent et al. *Molecular Ecology Resources* (2021). Figure originally published as Figure 1 from this publication.

This competing scenario design was thus aimed at disentangling whether genetic and admixture diversity patterns observed in the Afro-American ASW and the Barbadian ACB were better explained by pulse scenarios (three possible pulses per source population), compared to recurring decreasing and increasing admixture scenarios. Indeed, demographic migrations from Europe and Africa to the Americas, whether forced or voluntary, occurred at least on a yearly basis since the 16th century, with periods during the Plantation Economy era, between the 17th and 19th century, when population movements occurred much more frequently (Eltis and Richardson 2015; Fortes-Lima and Verdu 2021). It could thus, a priori, be expected that admixture between European and African source populations also occurred in a recurring way over time during the TAST, instead of more punctual, pulse-like, events; pulses-models that had been, until now, the only investigated models from a genetics perspective.

Importantly, note that I firmly believe that *MetHis*-ABC ought to be used only if more complex models are investigated than the ones approached by other maximum likelihood methods. Otherwise, I incite the user to instead use these ML methods, more elegant mathematically and often much more powerful statistically and computationally than ABC.

MetHis-ABC prior checks

Before launching any ABC inference, it is necessary to evaluate whether simulated data under the proposed models could produce sets of summary statistics encompassing the observed values calculated from the real data; otherwise, ABC inference is meaning-less. We thus conducted classical goodness-of-fit tests, visually inspected that each summary statistics observed in the real data fell within the spaces of values obtained from simulations, and verified that observed vectors of summary statistics also fell within the multidimensional space of summary-statistics calculated from simulations using PCA. All was in order, our design for *MetHis* simulations were indeed able to produce simulations whose summary-statistics values mimicked observed values separately obtained from the Afro-American (ASW) and the Barbadian (ACB) real genetic samples (Fortes-Lima, Laurent et al. 2021).

MetHis-ABC Random Forest scenario-choice

With our *MetHis*-ABC RF procedure, we found that the nine competing scenarios could be distinguished from one-another accurately in a vast majority of cases: without using observed data but instead considering in turn each 90,000 simulations as pseudo-observed data for RF scenario-choice. This was rather surprising to me given the high level of scenario-nestedness in certain parts of the parameter space (Robert et al. 2010). Indeed, for instance, parameter values corresponding to a recurring admixture model with a very sharp reduction of introgression rates the generation immediately after the onset, perfectly mimic a pulse-like model of admixture. Analogously, values of the model-parameters corresponding to a very sharp increase in recurring admixture at the last generation will a priori be confused with a single recent pulse of admixture occurring close to the present.

ABC cross validation approaches are of particular interest as they easily allow to partition *a posteriori* the space of parameter values to evaluate empirically the influence of such nestedness on RF scenario-choice confusions (Fortes-Lima, Laurent et al. 2021). Indeed, we found that the scenario-choice confusions made with our approach between different scenarios were most often occurring in spaces of the parameter-values where models were highly nested and thus biologically un-discriminated. Finally, note that despite these promising scenario-choice results, in real-data inferences with ABC, the question is not so much to perform well in distinguishing scenarios in general, i.e. over the entire space of parameter-values. Instead,

we want good performances for distinguishing among scenarios in the vicinity of our observed data, which can be the case even if scenarios are hard to distinguish overall.

Furthermore, we found that, as predicted by our analytical investigation of the general admixture model, RF-ABC scenario-choice was massively driven by extensive information from the summary-statistics describing the distribution of admixture fractions in the admixed population, largely before the amount of information brought by other more classical statistics. In particular, note that higher moments of the distribution, namely its shape as determined by variance, kurtosis and skewness, as well as minimal/maximal admixture values and the first and last 10% quantiles, were the most informative statistics.

Note that our cross-validation performances (before using real data as above), remained very satisfactory when considering, respectively, 50,000 or even 10,000 SNPs instead of 100,000 (**Figure F4.2.b2**). Furthermore, considering five times fewer individual samples from either source and admixed populations lowered our power to accurately distinguish among scenarios, but only very reasonably so and in any case, did not change patterns of correct and incorrect scenario prediction.

Altogether, this showed that our *MetHis*-ABC approach performed very satisfactorily, even with reduced data. This was to some extent expected by the nature of ABC inferences. Indeed, they rely on information carried by summary-statistics, rather than the absolute amount of data. Thus, as long as data is sufficient to estimate reasonably well the statistics, and if they are informative about model parameters, ABC will perform well, whichever the absolute amounts of data considered to calculate them.

Finally, using observed data, we found that for both the Afro-American ASW and the Barbadian ACB, the best scenario was the "AfrDE-EurDE" scenario, which encompassed monotonically decreasing recurring admixture from both the African and European source populations; a result that we will not detail more extensively here and encourage interested readers to refer to Fortes-Lima, Laurent et al. 2021 for further information and discussion.

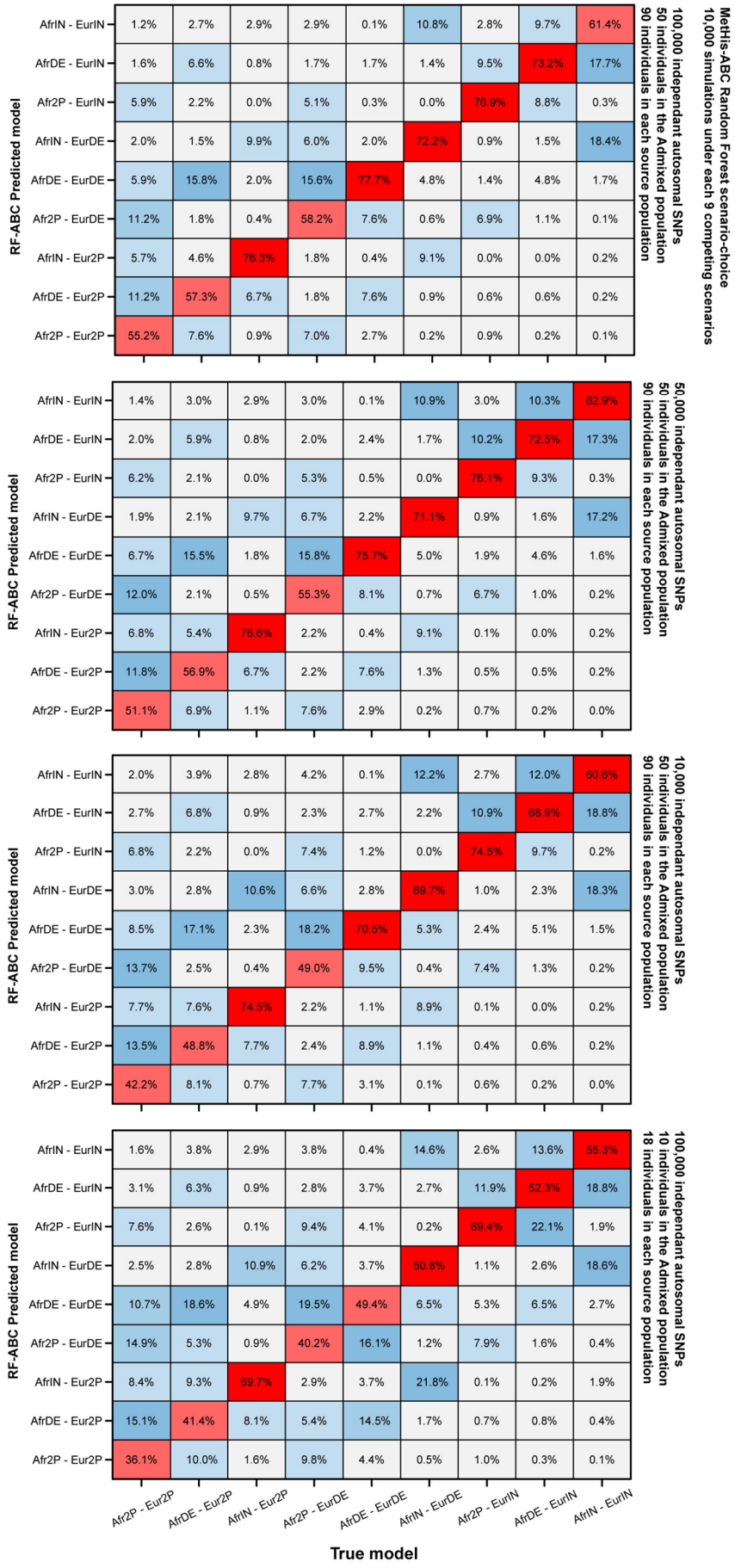


Figure F4.2.b2.
Cross-validation results for the discriminatory power of *MetHis*-ABC using Random Forest scenario-choice with 1,000 decision trees in the random forest, as a function of the number of SNP markers or of the numbers of individuals sampled in the admixed and source populations. Figure built from Figure 2 and Supplementary Figure SF5 from Fortes-Lima, Laurent et al. *Molecular Ecology Resources* 2021.

MetHis-ABC Neural Network posterior parameters joint estimations

Under the winning model, separately for each target population, we then performed additional simulations to feed the Neural Network learning for further ABC posterior parameter inferences (Csilléry et al. 2012; Jay et al. 2019). Note that the very tedious parts of Neural Network ABC procedures are not the inferences using real data themselves, but rather determining the optimal basic parameters of the NN, namely the tolerance level, which is the number of closest simulations to be used for NN inference, and the numbers of neurons in the hidden layers of the network, using empirically simulations as pseudo-observed data. Indeed, there are no theoretical or empirical ways of determining these parameters for specific case-studies *a priori* (Jay et al. 2019). They thus need to be tested in turn, in order to determine *a posteriori* which couple of NN-parameter values minimize posterior model-parameters' error rates. Once these are determined, a "final" Neural Network can be trained based on the observed data using the chosen tolerance level and number of neurons in the hidden layer, to obtain posterior parameter distributions for all model-parameters estimated jointly. Note that we empirically showed that this particular ABC posterior parameter estimation procedure outperformed other methods such as rejection, random-forest, or NN for each parameter separately, at least in our study-case.

We reproduced below our obtained results for each target admixed population for exemplification (**Figure F4.2.b3**). We can see here, in general, that older model-parameters are less distinguished from their respective priors than more recent events, as expected: more recent events erase the identifiability of previous parameters, albeit not completely in this recent admixture process, also as expected from our previous theoretical work (see **Chapter 3** about Buzbas and Verdu 2018). Furthermore, we can see that certain parameters are extremely precisely inferred *a posteriori*, which we confirmed by extensive ABC posterior-parameter errors and Credibility-Intervals accuracies estimations, another interesting possibility provided by any ABC approach. Indeed, note that previous ML methods for dating admixture events do not readily provide confidence intervals or even standard errors around the point estimations. Here, not only the inference provides posterior-parameter distributions, but, in addition, simulations in the vicinity of the observed data can be re-used as pseudo-observed to further evaluate the statistical power of the inference, or the lack of satisfactory power.

Finally, we will not recapitulate here the discussions of the obtained results from an anthropological genetics perspective. Indeed, our goal was here to propose a novel method and evaluate how well it performed using real, complex, genetic data. The scenarios here explored are somewhat un-satisfying as they still lack complexity if we were to reconstruct the detailed admixture histories of these populations as the primary focus of interest; instead, they served as a proof of concept for our methods. For instance, among many others, we considered in these analyses, for simplicity, constant reproductive sizes over generations in the admixed population; a likely problematic oversimplification as we expected changes in effective sizes in the admixed population to interact with admixture and genetic diversity patterns. Finally, we did not dwell on anthropological genetics interpretations of our results, as both Afro-American (ASW) and Barbadian (ACB) sample sets from the 1000 Genome Project were not accurately informed and categorized, hence lacking crucial information for better detailing models to be tested, and, most importantly, for interpreting our inference results in the light of historical information; a crucial categorization problem detailed in **Chapter 1**.

In the following **Chapter 5**, we will use this *MetHis*-ABC framework with the explicit objective of investigating the anthropological genetics history of admixture experienced in each Cabo Verdean island separately, as we initially intended as the root for the whole project (**Chapters 3** and **4**) started in 2009.

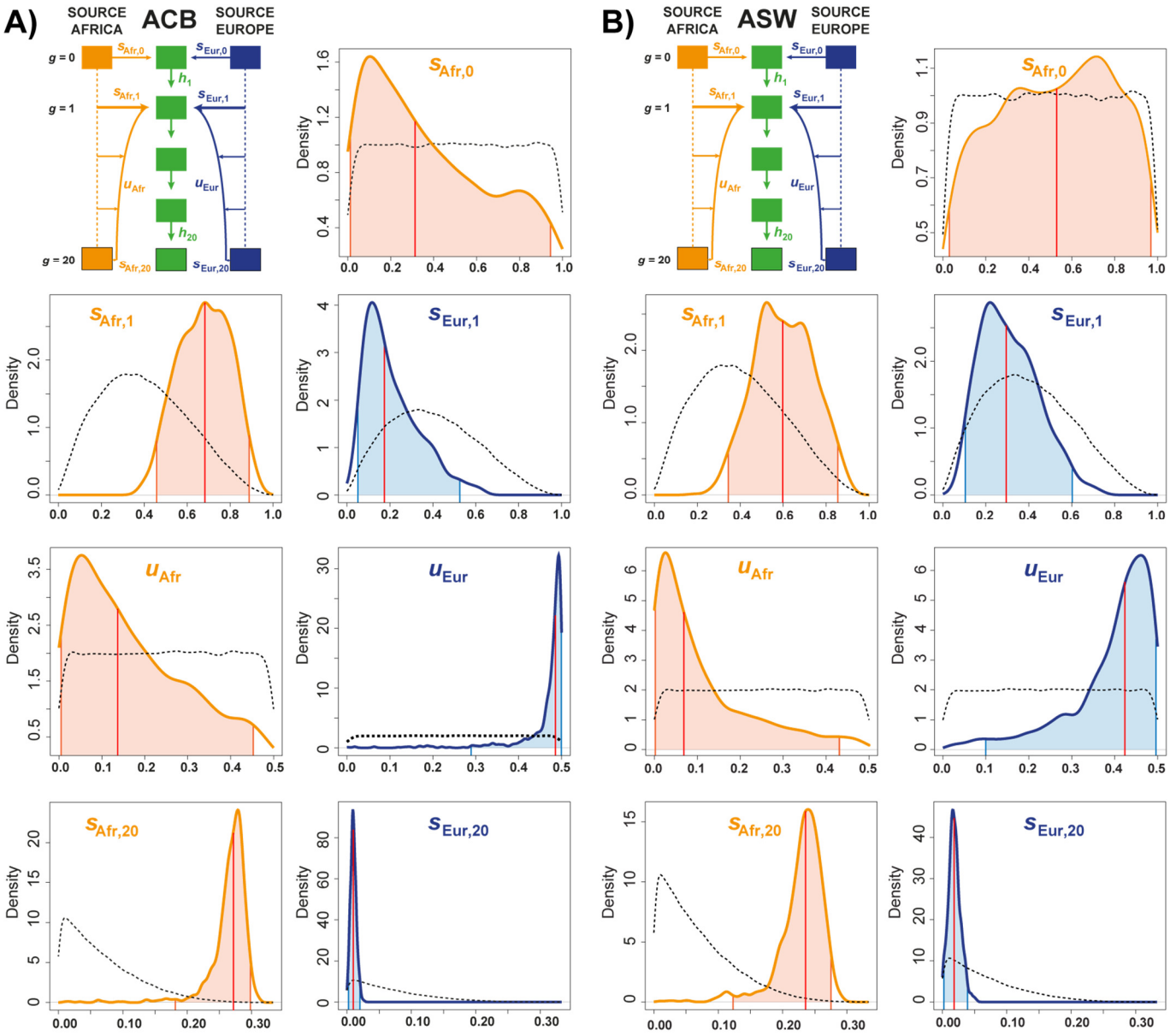


Figure F4.2.b3.

MetHis-ABC posterior parameter estimations under the winning model using Neural Network approaches, for the Afro-American ASW and the Barbadian ACB respectively.

Figure originally published as Figure 4 in Fortes-Lima, Laurent et al. *Molecular Ecology Resources* (2021).

4.3. Ongoing developments and future perspectives for the inference of complex admixture histories from genetic data using Approximate Bayesian Computation

Our *MetHis*-ABC framework released in 2021 has been used, with success, for disentangling migrations from admixture processes at play for the human peopling of Oceania in a recently published article in *Nature* (Choin et al. 2021). Beyond the anthropological genetics question of interest in this paper, we could empirically confirm here that *MetHis* simulations of autosomal independent genotyped SNP data over much older admixture histories (300 generations) than the ones previously explored, were still in good agreement with the observations, even without explicitly implementing a mutation model. Furthermore, ABC scenario-choice and posterior parameter inferences still performed similarly well as in our proof of concept. Finally, we deployed this version of our *MetHis*-ABC framework for studying Cabo Verde in the following **Chapter 5**.

4.3.a. Novel summary-statistics for MetHis-ABC

Since the original developments presented above, several novel features have been implemented and released publicly in *MetHis*. First, we built a novel summary-statistics for ABC inference in the hopes to better capture genetic variation patterns in the admixed populations. Indeed, the user can now consider, in addition to the distribution of admixture fractions in the admixed population, the ASD-MDS two-dimensional distribution of angles between admixed individuals and either centroid of the source populations (**Figure F4.3.a**). We noticed from previous experience, and further deterministic use of *MetHis* (see next), that the individuals in the admixed population departed, over time, from the trajectory between the centroids joining source populations projected on a 2D MDS based on individual pairwise ASD. Without formal proof, we think that this behavior may be due to drift and mutation in the admixed population since the last admixture event, until the admixed population forms a specific ASD-MDS cluster when sufficiently drifted from the sources after the last admixture event. We thus try to capture this information with angles as in **Figure F4.3.a**, and use the distribution of these angles similarly as previously (min, max, 10%-quantiles, mean, variance, kurtosis, skewness), as a set of possible summary statistics for ABC inference, in the hopes that it would provide further information for scenario-choice and posterior-parameter estimations. For instance, we expect the variance of this distribution to be very low in ASD-MDS projections, where admixed individuals cluster tightly, whichever their relative distance to either source populations' clusters on the 2D projection. We expect to find such pattern when the last admixture event has occurred sufficiently long ago (see **Chapter 3**). Note, however, that we do not have, to our knowledge, a formal proof of the behavior we think we may capture, at least in part, with this set of summary-statistics. Future analytical work on MDS dimensionality reduction from ASD matrices will need to be endeavored to rigorously justify the interest in this novel summary-statistics, that we can nevertheless readily play-with empirically with *MetHis*.

Finally, on the ABC summary-statistics front, as said before, admixture linkage-disequilibrium statistics are extremely informative about complex admixture histories' parameters. Using them in an ABC inference framework is nevertheless highly challenging, as such statistics have to be compiled for numerous independent simulations similarly as they are computed on real data; a task out of reach computationally as these complex statistics already require substantial time and power to be calculated only once on the real data. However, with Pr. Zachary A. Szpiech (Pennsylvania State University) and Dr. Romain Laurent, we are currently trying to develop efficient ways to calculate Runs-of-Homozygosity patterns (ROHs) and

ancestry-specific ROHs (see **Chapter 5**), for further use as summary-statistics in an ABC framework. While ROHs are not exactly admixture-LD patterns, they may capture parts of the same information for the benefit of ABC inferences performances. In case of success, we will thus further need to implement recombination and linkage disequilibrium in *MetHis* simulations rather than only considering biologically independent SNP markers as done currently (see previous sections in this chapter and also **section 4.3.b** next).

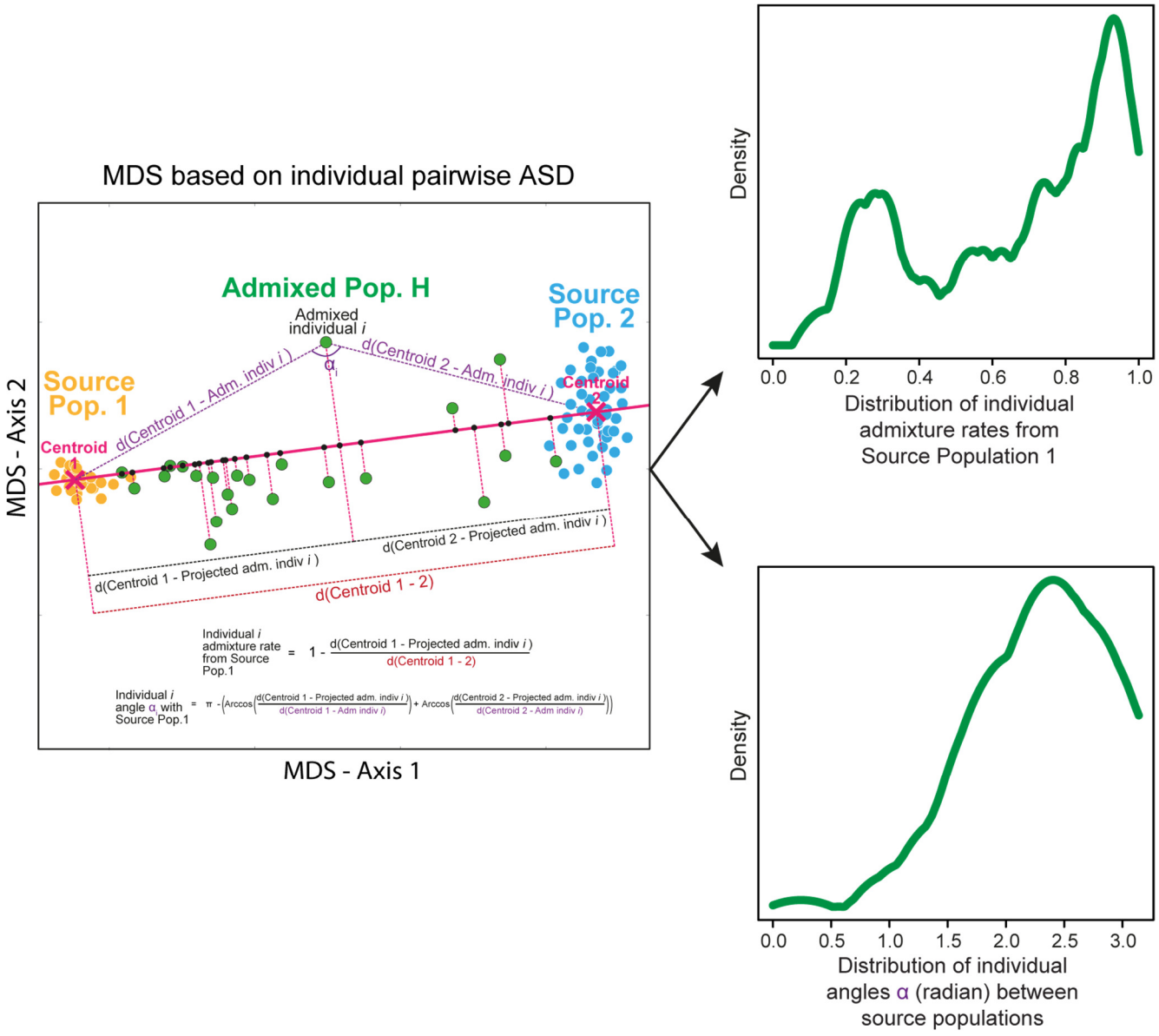


Figure F4.3.a.

Schematic representation of ASD-MDS statistical estimation of average individual admixture fraction from one of the source populations, as well as that of the distribution of angles formed by admixed individuals from the line-joining the centroids of the source populations, in a two-source population admixture model and implemented in *MetHis*.

Figure originally presented in Supplementary Figure SF12 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

4.3.b. Other types of genetic data and models simulated with MetHis

Most importantly, with a PhD student, Marta Ciccarella, co-supervised by Jorge M. Rocha (CIBIO-University of Porto) and myself, we currently aim at developing a sex-specific version of *MetHis*, as presented in our theoretical analyses in **Chapter 3**. Further implementing sex-specific summary-statistics will allow to expand our approach for investigating complex sex-biased admixture histories from genetic data. Ms. Ciccarella is also interested in developing the parameterization of non-random matting processes in the admixed population over time, in the current version of *MetHis*. This will allow us, as explained in the **Chapter 3**'s theoretical perspectives, to empirically estimate deviations from random matting in the admixed population and possibly better understand the role of complex socio-marital stratification possibly at play in the genetic evolution of human admixed populations, for instance (Zaitlen et al. 2017; Versluys et al. 2021).

MetHis was designed forward-in-time in part to ease future implementation of natural selection processes in the admixed population, in particular to investigate post-admixture selection processes exemplified in our empirical findings in Central African populations (Patin et al. 2017). In principle, *MetHis* could relatively easily be tuned for such purpose by controlling the reproductive success and survival of individuals in the admixed population with a novel selection parameter, which remains yet to be fully defined and implemented. Thus, in collaboration with Dr; Patin's team, we hope to be able to further explore such possibilities provided by future significant developments of our inference framework and novel computational tools. In particular, we will need to develop *MetHis* for simulating sequences rather than SNPs or microsatellites, which, as said in previous sections of this chapter, can relatively easily be done based on the existing design of our simulator.

Finally, I am currently co-supervising the PhD project of Ms. Juliette Sauvage with Dr. Céline Bon (MNHN) and Dr. Marie-Claude Marsollier-Kergoat (CEA), which aims at reconstructing the history of the Paris Basin peopling from ancient DNA data. For this project, we are generating genomic sequences of human remains un-earthed by archaeologists in the region of Paris from different archaeological sites since the Mesolithic period, more than 5000 years ago, up-until the middle-ages. Specifically concerning *MetHis*-ABC, we aim at reconstructing the admixture history of individuals from each archaeological site over time. To do so, we will need to implement, in the *MetHis* simulator, the type of data that we expect to obtain from ancient DNA, and associated summary-statistics for ABC inferences. Rather than trying to simulate such data under the complex admixture-history models themselves, we aim at conducting the same types of simulations performed for modern DNA, and extracting then haploid partial genomic markers mimicking the observed ancient DNA data. Using then haplotypic summary-statistics classically used in paleo-genetics, this will allow us to perform ABC inferences projecting ourselves in past populations, at each point in time and for each archaeological site separately.

4.3.c. Using MetHis deterministically

I supervised, recently in 2021, a Master student in bio-informatics, Maël Pretet, for investigating, deterministically, how changes in reproductive sizes in the admixed population interacted with admixture and genetic diversity patterns, as captured by summary statistics calculated in *MetHis*, over time and as a function of admixture parameters. To do so, Mr. Pretet implemented as an option for *MetHis* users, with Dr. Laurent's help, the possibility to follow over time a given simulation, by stopping the simulation at each generation and calculating summary-statistics, before starting the simulation again. Furthermore, in

order to empirically evaluate the behavior of computationally intensive Bayesian clustering algorithms, such as STRUCTURE or ADMIXTURE, and beyond that, for any other possible software of interest, Mr. Pretet also implemented the possibility for the simulated genetic data to be written in full at any generation of *MetHis* simulations (the user should beware of usable storage space on his/her disks when choosing this option).

Beyond the results that Mr. Pretet produced and that we are currently in the process of further exploring to envision a possible future publication, these sets of tools and options are readily implemented in *MetHis*. They will be, I believe, of interest for exploring and benchmarking the behavior of numerous population genetics statistics under highly complex admixture scenarios. These tools will be of particular interest for several widely-used population genetics statistics that are lacking analytical expectations, or those statistics which have not been benchmarked thoroughly, under such complex models, such as Paterson's f_3-statistics (2012) for an instance. Furthermore, based on these developments of *MetHis*, we hope to be able to use our software for future predictions of genetic and admixture patterns expected in populations having undergone or still undergoing admixture, for instance for conservation biology purposes.

4.3.d. Ongoing projects using MetHis-ABC for non-human species

We sincerely hope that our *MetHis*-ABC framework will be interesting beyond the study of human populations. Therefore, we started to initiate two projects investigating the admixture history of diploid plants populations, from diverse types of genetic data including microsatellites and RAD sequencing SNPs. The first project, initiated in collaboration with Pr. Hardy (Université Libre de Bruxelles), is about the ancient admixture history of *Terminalia superba* in the Dahomey Gap in Western Africa. Pr. Hardy had previously shown that this tree species was born from hybridization between previously isolated sub-populations from West Western Africa and from Central Africa respectively, when the two forest massifs overlapped 190,000 years ago, before being re-separated 20,000 years ago and since (Demenou et al. 2018). We are thus in the process of testing different possible past admixture processes having given birth to observed genetic patterns.

The other plant-project for *MetHis*-ABC, that we have initiated with Pr. Nathalie Machon (MNHN) and a previous Master Student, Claire Stragier, that I supervised, is interested in admixture patterns in four local populations of *Arenaria grandiflora* in the Fontainebleau Forest. These populations have been the focus of a hybrid reintroduction program since 1999 from plant clones salvaged in a botanical conservatory before their extinction locally in the 1990's, and from other clones from a separate population from the region of Chinon in North-Central France. Since then, populations are followed and genotyped for a variety of markers every generation (roughly 4 years for this plant) in a genetic-rescue framework. We ought to use *MetHis* here to better understand the admixture process ongoing under our eyes, and in particular possible departures from random matting, in turn possibly indicative of hybrid depression or heterosis due, for instance, to phenetic differences as a function of individual admixture levels. In addition, *MetHis* could possibly be used to explore, predictively, the expected genetic and admixture patterns of these populations in the future, which may help managing the populations in the future. This project has been funded by the ANR in 2022, as part of a larger project (ANR FloRes) piloted by Dr. Ophélie Ronce (CNRS-University of Montpellier) aiming at evaluating plants' conservation and rescue programs in the face of climatic changes in a variety of case-studies in France.

Chapitre 5

The genetic and linguistic admixture histories of Cabo Verde

Alto Mira.
Santo Antão, Cabo Verde, 2016
©Paul Verdu

Chapter 5. The genetic and linguistic admixture histories of Cabo Verde

In **Chapter 1.2** above, I introduced, in details, the genesis and categorization and sampling protocol at the root of our interdisciplinary project aiming at reconstructing the genetic and linguistic histories of Cabo Verde, in the context of European colonization since the 15th century and the Trans-Atlantic Slave Trade (TAST). We further generalized the fundamental population genetics questions of interest about Cabo Verde, to the investigation of the admixture history of enslaved-African descendant admixed population during the TAST in the same **Chapter 1.2**, and in the population genetics' theoretical **Chapter 3** and methodological **Chapter 4**.

As a recapitulation from the introductions in **Chapter 1.2**, **Chapter 3**, and **Chapter 4**, the key questions of interest to the genetics side of the Cabo Verde project were:

i) Which continental European and African populations contributed to the gene-pool of Cabo Verde today?

ii) What are the patterns of genetic diversity and admixture within Cabo Verde at reduced geographical scale? Are their islands or group of islands relatively more or less reproductively isolated than others due to serial founding events and isolation triggered by political changes over the 400 years of the peopling of the archipelago? Or, instead, are genetic diversity and admixture patterns evenly distributed throughout the archipelago as suggested by numerous population movements across islands throughout history? (**Figure F1.2.b**)

iii) When and where did the admixture events that built genetic diversity and admixture patterns today occur? Were there more admixture events identifiable during the massive population movements triggered by the expansion of Plantation Economy between the mid-17th and the early 19th century? Did the abolition of the TAST reduced the opportunity for admixture events to shape extant genetic patterns? Did the abolition of slavery and the major sociological changes that followed increased the opportunity for admixture events to occur? Did recent 20th century work-related migrations resulted in novel, recent, admixture events?

Furthermore, remember that the origins of the project stemmed from a linguistic question brought to us by Pr. Marlyse Baptista: could population genetics help, indirectly, informing which languages may have contributed to the lexical, grammatical, and/or syntactic diversity of Cabo Verdean Kriolu? Comparative and historical linguistics already identified numerous African and European languages having contributed to the lexical, grammatical, and syntactic diversity of Cabo Verdean Kriolu. Intuitively, identifying previously unsuspected populations at the source of Cabo Verde genetic diversity may, in turn, propose novel hypotheses for the languages that may have contributed to the linguistic diversity of Kriolu.

We decided to collect linguistic utterances, i.e. individual realizations of Kriolu in a given context, that of academic semi-spontaneous interviews conducted individually (see the end of **Chapter 1.2**), rather than "Languages" as per classical linguistic definitions. We collected such data among the same individuals we sampled genetically to address the above, genetics, disciplinary questions. Indeed, while several linguists, including Pr. Baptista, had described linguistic variation across Cabo Verdean islands, we did not have systematic data representative of each Kriolu linguistic variants, such as classical word-lists, based on which we could conduct classical computational linguistics investigations whose results could then be

compared with the outcome of classical population genetics descriptive analyses (**Figure F1.2.c**)[49]. However, by collecting anew individual Kriolu utterances as we did, we could compare at the same scale centered on individuals, a specific type of linguistic variation, individual's "manners of speaking Kriolu", with inter-individual genetic variation. Several pluri-disciplinary questions spontaneously emerged from this interdisciplinary fieldwork setting:

i) What is the linguistic diversity of Kriolu within Cabo Verde at reduced geographical scale, estimated as inter-individual utterance variation? Are there correlations between the geographic distribution of genetic and linguistic diversity within Cabo Verde, possibly reflecting parallel historical trajectories or phenomenon having given birth to either type of biological and cultural variation?

ii) What is the linguistic admixture diversity of individual manners of speaking Kriolu throughout Cabo Verde? Are specific utterances of a particular African or European origin more often employed in individual discourses in certain parts of Cabo Verde than in others? Are linguistic and genetic admixture histories correlated or disjointed in Cabo Verde?

Altogether, why answering all these questions could be of importance?

First, Cabo Verde is the first European peopling colony in Sub-Saharan Africa of the TAST era. Furthermore, Cabo Verde was a major slave-trade platform during most of the TAST. Probably as a result, Cabo Verdean Kriolu is also the first Atlantic Creole language from the era, born from linguistic contacts between African and European languages. As such, investigating genetic and linguistic admixture histories in Cabo Verde allows to investigate the influence of the TAST on cultural and biological diversity from the very beginning of this historical era, more than 30 years before Columbus (**Figure F1.2.a**). Finally, while numerous American and Caribbean population have been the focus of anthropological genetics studies in the context of the TAST, the influence of this era on genetic diversity patterns in Africa have been seldom explored.

Second, numerous populations of European and African origins were forcibly deported or voluntarily migrated to Cabo Verde since the 15th century, as attested by historical records, and which are known by linguists to have left an identifiable trace in Cabo Verdean Kriolu. Which ones contributed or not to the genetic diversity of Cabo Verde is still debated among historians and geneticists (**Figure F1.2.a**), similarly as for all enslaved-African descendant populations having emerged during the TAST.

Third, the islands of the Cabo Verdean archipelago were peopled successively over the course of more than 300 years during the TAST (**Figure F1.2.b**). This represents an ideal geographic and historic framework to investigate, at a micro-geographical scale, the histories of genetic and linguistic founding events, demographic migrations, and admixture. Furthermore, Cabo Verde peopling history could reveal possible such mechanisms also at play in other enslaved-African descendant populations born from the TAST, and possibly reflects the histories and socio-demographic processes having influence the genetic and linguistic diversity of numerous populations at a macro-geographical scale across the three continents.

Finally, classical genetic-linguistic comparisons are often conducted between isolated languages and genetic populations, albeit their respective descriptions of variation stem from observations collected at very different scales and with different methods. Indeed, genetic variation stem from observed genetic data

[49] a type of genetic-linguistic comparison that we had endeavored, for instance, in the Central African project (**Chapter 2.1**).

compared across individuals grouped into populations, while linguistic variation stem from comparisons across complex meta-constructed languages obtained from specific (few) informants (see **Figure F1.2.c**). Studying jointly genetic and linguistic variation collected in the same individuals would thus allow us, not only to investigate genetic-linguistic correlations from a contact-language, but also to anchor genetic-linguistic comparisons on different objects (genes and utterances), nevertheless collected at the same scale centered on individuals.

Therefore, answering these questions could both provide invaluable knowledge about the possible mechanisms having influenced genetic and linguistic variation in numerous other populations descending from enslaved-Africans deported throughout the TAST since before Columbus, as well as, more directly, about the construction of genetic and linguistic diversity in Cabo Verde at a micro-geographical scale.

Since 2009 and the beginning of this particular project, we have developed the theoretical and methodological tools to address the population-genetics aspects of these questions using the genetic and anthropological data collected during six interdisciplinary fieldworks between 2010 and 2018. During these fieldworks, we also collected systematically for the same individuals, semi-spontaneous Kriolu utterances, fully transcribed with the morpho-phonetic orthographic norm of Kriolu (the ALUPEC), which we could describe with classical statistical tools. Finally, we further collected word-list utterances from a substantial subset of our dataset (see **section 5.3** below). Based on both types of data, we developed a novel "population linguistics" general framework, and further developed a novel joint genetic and linguistic inference methodology and associated computation tools for reconstructing, from word-list utterance data and genetic data collected in the same individuals, the genetic and linguistic histories of populations.

In this chapter, I will first (**section 5.1**) report on the results we obtained about the genetic admixture history of Cabo Verde, finishing with ongoing work furthering these results (**section 5.2**). Then, I will focus on the pluri-disciplinary statistical comparison of genetic and linguistic variation within the archipelago (**section 5.3**). Finally, I will present the inter-disciplinary aspects of our ongoing project about joint genetics-linguistics historical inferences (**section 5.4**). Note that, after 13 years piloting this project, some of this work has already been published, some other parts are currently (while writing these lines in mid-2022), under review and released as pre-prints in open access archives, and some are in preparation or still on the research workbench.

5.1. The genetic admixture histories of Cabo Verde

First, we set out to infer the European and African populations having contributed to the gene-pool of extant Cabo Verdeans. Second, we investigated in details the distribution of genetic and admixture diversity at micro-geographical scale, within Cabo Verde. Finally, we reconstructed, using *MetHis*-ABC (see **Chapter 4**), the complex admixture histories of each Cabo Verdean island separately.

Note: The results described and summarized in this section 5.1 are currently in pre-print format by Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833. Since April 2022, this article is under review for publication in eLife, and still in progress at the time of writing this chapter in June 2022.

5.1.a. European and African populations at the source of Cabo Verdean genetic diversity today.

As detailed in **Chapter 3**, population genetics methods to infer the possible sources of genetic admixture in a target population do not necessarily rely on explicit mechanistic models of complex admixture histories. Instead, they rely on statistical methods whose goal is to estimate the proportion of the genetic material observed in the admixed population ultimately originating from putative source populations.

Intuitively, these approaches rely on the systematic comparison of allelic or haplotypic frequencies between the target admixed population and, preferably, numerous other worldwide populations. These comparisons aim at identifying, among these extant populations, which ones share most recent common ancestries with each part of admixed individuals' genomes, respectively. These extant populations are then considered as descendants from the proxy source populations for the past admixture processes, which eventually gave birth to the observed proportions of genetic admixture in the target population.

We thus had to merge the 2.5 million SNPs genotyped from the 261 Cabo Verdean individual DNA samples collected during our fieldwork, with previously published data. We decided to merge our novel data with genetic data from 2504 individuals from the 1000 Genomes Project in order to ubicate the genetic diversity patterns from Cabo Verde in the worldwide human genetic diversity. Furthermore, we wanted to compare Cabo Verdean data with extensive population samples from Africa, where enslaved-Africans had been forcibly deported in part to Cabo Verde during the TAST. Thus, we obtained access to the African Genome Variation Project data comprising genome-wide autosomal genotyping data for 1307 individuals from more than 20 Western, Eastern and South African populations (Gurdasani et al. 2015; EGA accession number EGAD00001000959); and further merged the data from Patin et al. (2017), a similar genome-wide genotyping dataset, comprising 1235 individuals mainly from Cameroon, Gabon, and Angola (EGA accession number EGAS00001002078), a later study to which we had contributed extensively in our Central African projects (see **Chapter 2**). We then conducted the quality control and merging procedures, among Cabo Verdean genotyping batches over the six fieldworks first, and then with these previously published datasets, and summarized the pipeline in the **Figure F5.1.a1** below.

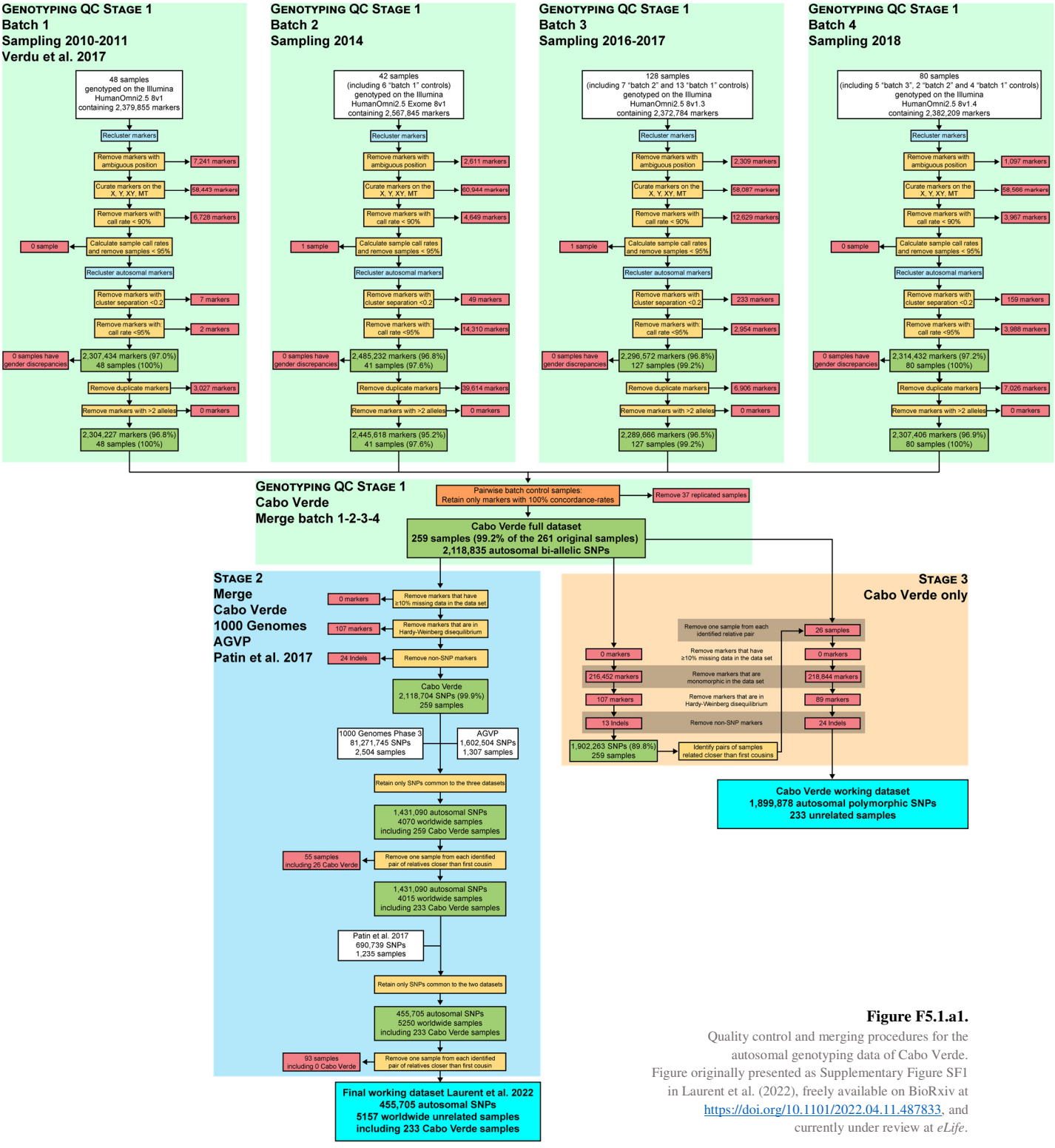


Figure F5.1.a1.
Quality control and merging procedures for the autosomal genotyping data of Cabo Verde. Figure originally presented as Supplementary Figure SF1 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

I will not detail here all the procedures described in this **Figure 5.1.a1**, but it is important for me to remind here that those are extremely complex and tedious procedures. I believe that, unfortunately, they do not obtain the recognition they deserve either in publications nor in the human population genetics community in general. Indeed, these procedures require excellent knowledge of databases produced by other independent researchers, handling efficiently massive amounts of complex data, and performing, in principle, simple population genetics checks and calculations rendered non-trivial by nestedness among the various steps and by the very large amounts of data. I feel that some researchers often under-estimate the challenges of these quality-control procedures and the numerous bio-informatics skills needed to achieve them. In turn, this results in often under-estimating the time needed for students to acquire the skills to perform these procedures, nevertheless at the root of all the data analyses one might endeavor. To my views, these bio-informatics quality-controls and merging procedures largely replaced, in the human population genetics training, the place occupied previously by tedious point-by-point DNA extraction, PCR, and old-generation sequencing molecular genetics procedures needed to generate working data. Indeed, while I spent roughly 8 months, during my PhD, generating microsatellite variation by hand, and only a few weeks subsequently thoroughly curating the data, it is not uncommon, today, that massive genome-wide data can be generated (when DNA samples are available) in a few weeks' time, while curating them often require numerous months of intense work for untrained researchers.

In this case, I conducted myself the curation of all genotyping data from the different batches, and Dr. Laurent brilliantly conducted all subsequent data merging procedures (**Figure F5.1.a1**), resulting in a merged working dataset for identifying sources of admixture to the Cabo Verdean gene-pool comprising more than 450,000 autosomal SNPs genotyped in more than 5,157 worldwide samples including 233 family unrelated Cabo Verdeans (see **Figure F5.1.a2**).

Based on this data, we first explored individual pairwise ASD patterns using MDS 3D projections (see **Chapter 4**), sub-sampling populations, in turn, and recomputing the MDS on the subsampled population set, in order to empirically explore the sets of individuals apparently more closely resembling one another. Ultimately, we found that (**Figure F5.1.a3**), Cabo Verdean individuals were on a different ASD-MDS trajectory between continental European and African individuals than other enslaved-African descendant populations from the TAST in the Americas, represented in our dataset by a sample from USA Afro-Americans (ASW), a sample of Barbadians (ACB), and a sample of Puerto Ricans (PUR). Indeed, we found Cabo Verdeans aligned on a trajectory between individuals from West Western African Senegambia and individuals from South-Western Europe. Instead, we found both Afro-American and Barbadian samples aligned on a trajectory between individuals from West Central African Nigeria and individuals from North-Western Europe. Finally, and again in contrast, we found Puerto-Rican individuals on a trajectory between individuals from Central Western Africa and individuals from South-Western Europe.

These patterns would be expected if these four populations had experienced admixture between different European and African populations (see **Chapter 3** and **4**). However, it is important to emphasize here that these real data analyses do not represent a proof of such admixture process. Indeed, dimensionality reduction in an MDS projection can bias our observations by hiding in higher dimensions preferred genetic relationships among other groups of individuals. In any case, we describe genetic dissimilarities among real data in these procedures without formally testing the origins of the observed patterns, which could be produced by other mechanisms than admixture.

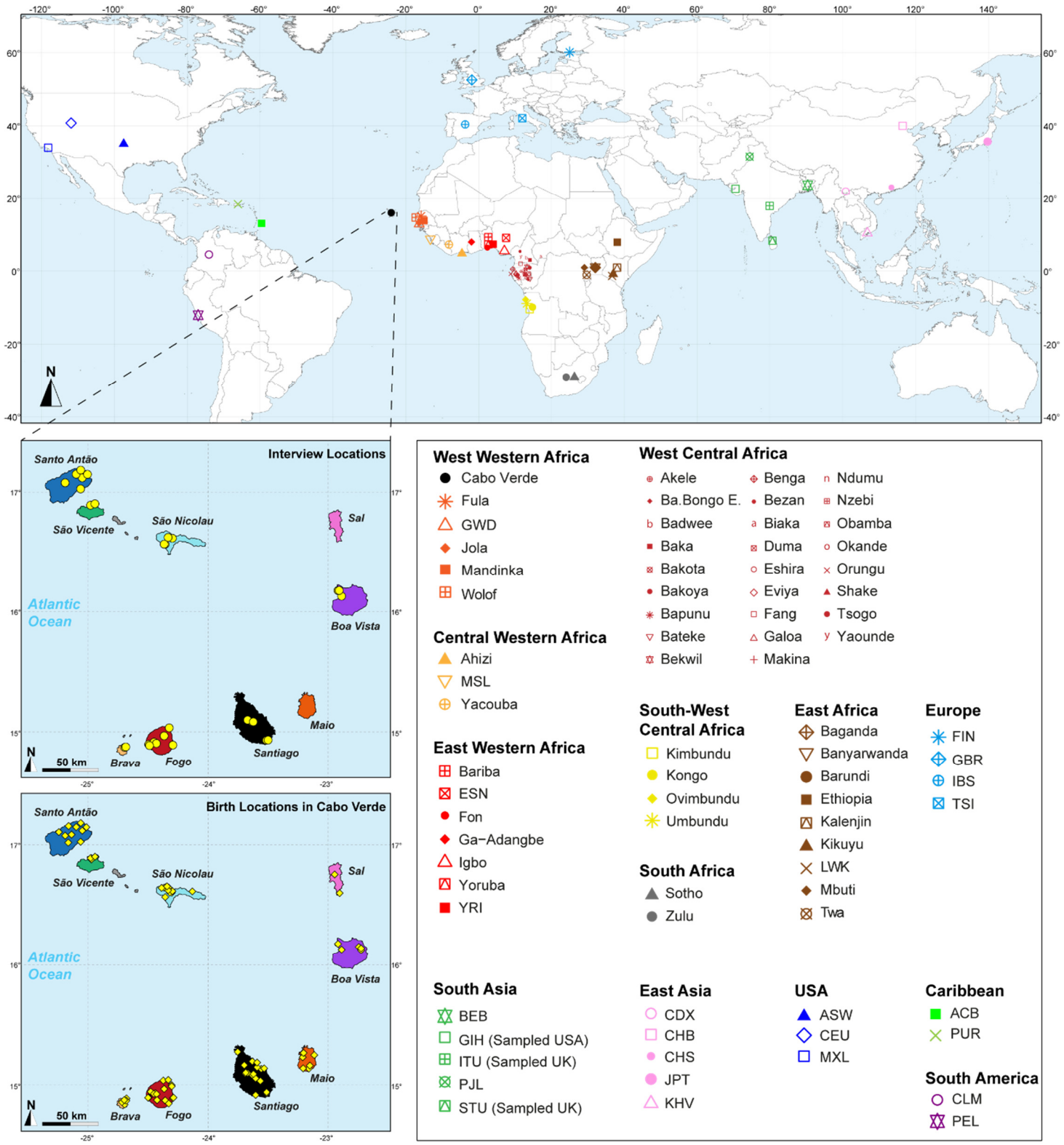


Figure F5.1.a2.

Population sample map and interview and birth-place location in Cabo Verde.

Figure originally presented as Figure F1 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

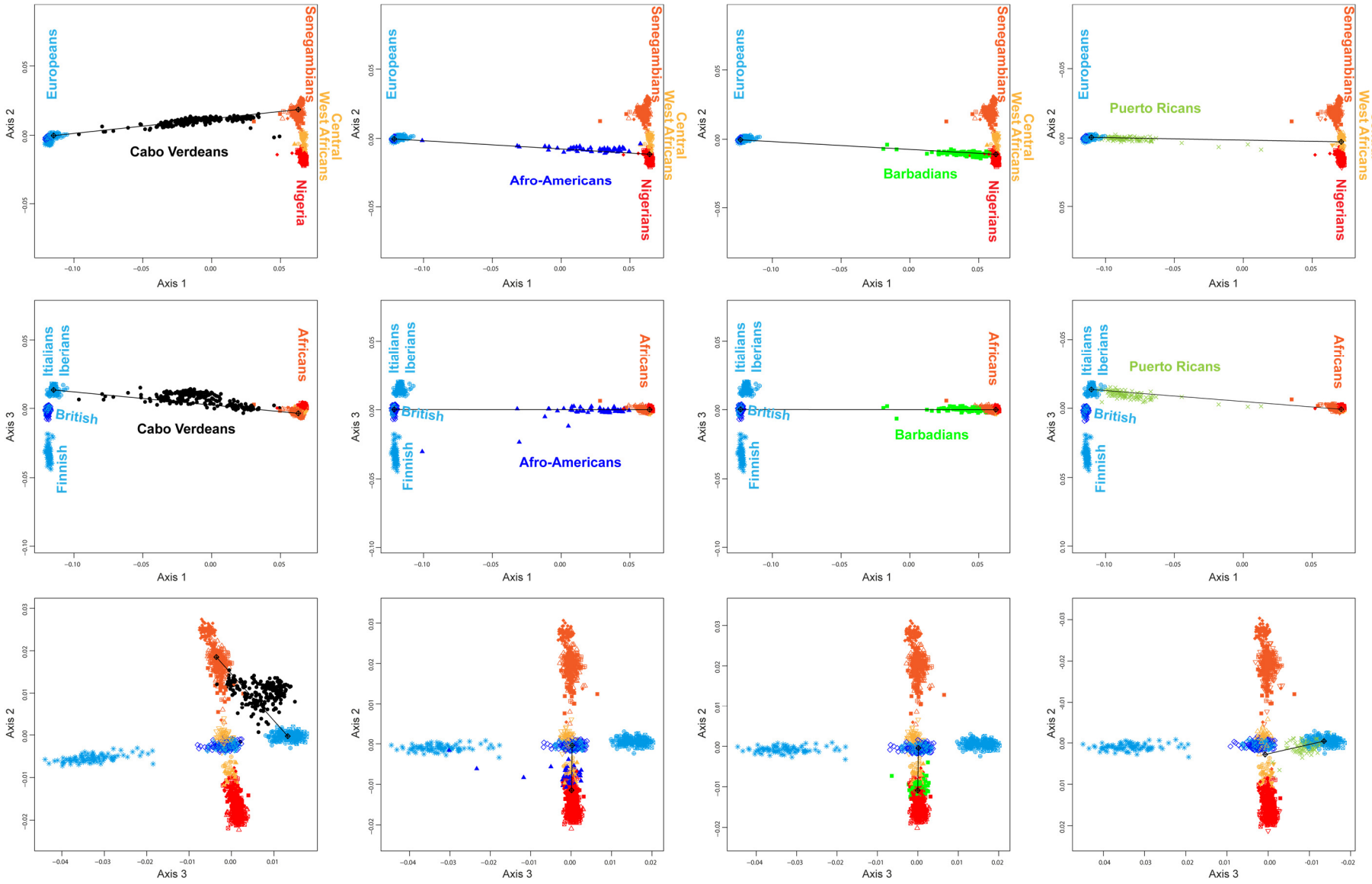


Figure F5.1.a3.

Three-dimensional ASD-MDS projections considering the same European and African populations and, separately, Cabo Verdeans, Afro-Americans, Barbadians, and Puerto-Ricans. Symbols can be found in Figure F5.1.b. Figure reshaped from Figure F2 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

We obtained very similar results, as expected, using the maximum-likelihood clustering algorithm implemented in ADMIXTURE (Alexander et al. 2009), for values of K-clusters ranging from 2 to 10 (**Figure F5.1.a4**).

Indeed, while ADMIXTURE results are more than often over-interpreted in the human population genetics literature (Lawson et al. 2018), this clustering algorithm in fact captures very similar information as the ASD-MDS decomposition and should be interpreted similarly. In fact, ADMIXTURE genotype membership clustering patterns for individuals should be interpreted as levels of genetic similarity among individuals: two individuals showing similar membership proportions patterns are more resembling one another than two individuals showing different patterns. Interpreting membership proportions from a single cluster (color) across individuals exhibiting substantial levels of several colors as "ancestries" or "admixture" is dangerous: "German flags" patterns often indicate unresolved clustering of the individual at this value of K, similarly as when short Euclidean distances in 2D ASD-MDS projections might hide substantial differentiation in higher-order dimensions. In other words, similarly to ASD-MDS, ADMIXTURE recapitulates inter-individual levels of genomic resemblance and dissimilarities, which can be interpreted as admixture levels but which does not formally test for such underlying model (despite the name of the software), and could very well be due to either unresolved clustering or, alternatively, other demographic processes than admixture which may result in similar patterns. This very common misconception was nevertheless extremely explicitly warned against by the original conceivers of these methods since Pritchard et al. (2000) with STRUCTURE, both in their papers and in the software manuals; and equations are un-ambiguous. However, these authors used the word "ancestry" to translate the statistical model they used for inferring levels of genotype membership in a given individual from, in fact, the K virtual genetic clusters built by the method and maximizing allelic frequencies differentiations among all individuals included in the analysis. It seems that numerous anthropological geneticists and human population geneticists misconceived this "virtual ancestry" for a modelized explicit "genealogical ancestry", which is what everyone is ultimately after including me, when the methods actually do not model such process in their computations.

So, why use STRUCTURE/ADMIXTURE on top of ASD-MDS if they are redundant? In fact, ADMIXTURE allows to visually investigate numerous axes of variation (K) at the same time, which is extremely tedious and confusing when exploring combinations of 2D MDS for numerous axes. So, while MDS provides a better visual intuition of levels of inter-individual genetic differentiation, which STRUCTURE/ADMIXTURE do not provide easily as one does not readily see the "distance" between colors, this latter analysis allows to explore efficiently levels of resemblances from higher-order dimensions and is thus highly complementary for the initial statistical description of genetic diversity patterns in any population genetics data analysis approach.

In our case and in summary (Laurent et al. 2022), indeed, while ADMIXTURE results recapitulated our ASD-MDS decompositions, we can see from K=6, and on, that genetic patterns across several groups of Cabo Verdean islands are in fact substantially differentiated from one another, a pattern barely previously seen in the previous ASD-MDS analyses. At K=10, we can identify three different clusters differentiating individual genetic patterns born respectively in Santiago-Fogo-Brava, Maio-Boa Vista-Sal-São Nicolau, and Santo Antão-São Vicente patterns, thus showing high levels of genetic differentiation at reduced scale within the Cabo Verde archipelago (see **section 5.1.b**). These results and the previous ASD-MDS ones thus significantly expanded previous results obtained at much larger continental scale (Belezza et al. 2012; Verdu et al. 2017).

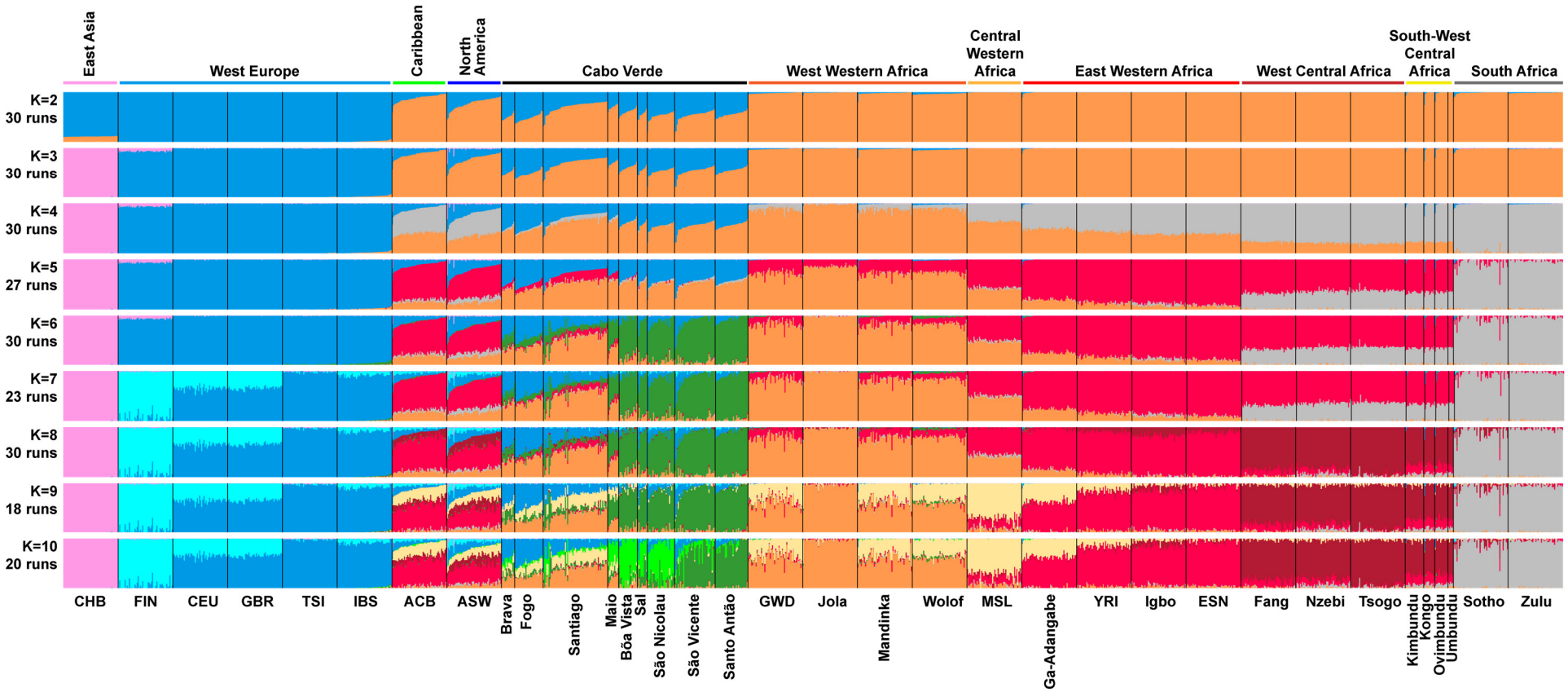


Figure F5.1.a4.

ADMIXTURE analysis for K between 2 and 10, based on 1,369 individuals genotyped at 102,543 independent low-LD SNPs.
Figure originally from Figure F3 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

So, are these observed patterns indeed stemming from admixture between the specific continental African and European populations? To answer this question, we conducted local-ancestry explicit inferences implemented in the ShapeIT2-Chromopainter-Globbetrotter-SourceFind pipeline (Lawson et al. 2012; Delaneau et al. 2012; Hellenthal et al. 2014; Chacón-Duque et al. 2018). This analysis relies on inferring by a formal testing, the most recent shared haplotypic ancestries within a targeted individual phased genome, with specific populations among a set of putative sources chosen by the user. Here, we decided to consider all populations from the ADMIXTURE analysis (**Figure F5.1.a4**) as possible sources for the haplotypic patterns of each Cabo Verdean individual genomes, considering either 4 putative sources or six putative sources, and then summarized results for individuals within islands of birth in Cabo Verde (Laurent et al. 2022; **Figure F5.1a5**).

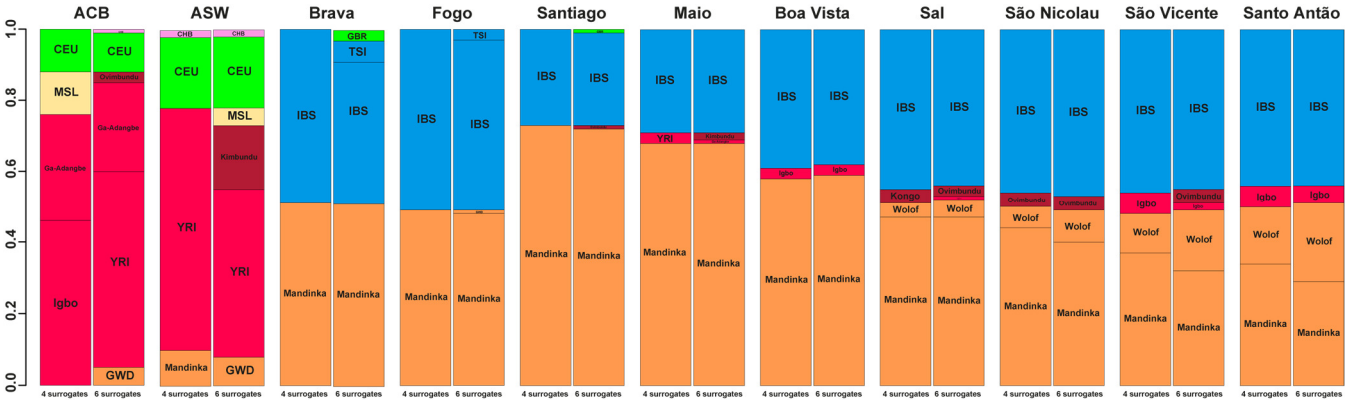


Figure F5.1.a5.

Shared haplotypic inference of putative source populations for the genomic patterns of Afro-American, Barbadian, and Cabo Verdean individuals born in different islands of the archipelago. Figure originally from Figure F3 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

We thus found somewhat surprising and unexpected results to our first question: where did the genetic ancestors of Cabo Verdeans come from in continental Europe and Africa?

First (Laurent et al. 2022), despite numerous voluntary and forced migrations from Europe and Africa to the Cabo Verdean archipelago since the establishment of the colony in the 1460's, throughout the TAST era and the Portuguese colonization, and until today, only a very reduced number of these populations contributed to the genetic diversity of Cabo Verdeans today. Indeed, Cabo Verdeans from all islands almost exclusively share ancestry with extant Iberian populations from South Western Europe, with some very limited contributions from Italians and British populations in a couple islands. On the African side, Cabo Verdeans almost exclusively share recent haplotypic ancestries with certain Senegambian populations, Mandinka and Wolof-speakers, the latter of the two proxy source populations being only found substantially in São Vicente and Santo Antão and to some extent in São Nicolau and Sal. Finally, albeit Western Central and Angolan populations are known to have been forcibly displaced to Cabo Verde during the TAST, even leaving unquestionable linguistic contributions to Caob Verdean Kriolu, these populations only left a very limited genetic signatures only in certain Cabo Verdean islands. Altogether, our results show a fascinating discrepancy between historically known extensive demographic movements to Cabo Verde from all over the Atlantic coast of Africa and from several European regions, and genetic contributions from only certain populations from the, small and nearby Cabo Verde, Senegambian region.

In contrast, our analyses for the Afro-American and Barbadian samples from the 1000 Genomes project showed, similarly to previous studies, much more substantial contributions from various populations from West Western Africa until Angola, hereby qualitatively reflecting, in the case of these two population samples, what could have been expected from extensive historical demographic migrations during the TAST (Laurent et al. 2022).

Finally, we further identified in this analysis (and other in Laurent et al. 2022), significant variation in relative admixture proportions from either source across individuals born on different Cabo Verdean islands, further echoing the different clustering patterns across islands found with the ADMIXTURE analysis.

5.1.b. Isolation-By-Distance genetic patterns at reduced geographical scale within the archipelago.

Together with Pr. Szpiech, we thus set out to investigate in more details genetic diversity patterns within individuals born in Cabo Verde only, using patterns of Runs-Of-Homozygosities (Szpiech et al. 2019), ASD-MDS (**Figure 5.1.b**), and Mantel and partial-Mantel testing. In brief (Laurent et al. 2022), we found substantial genetic differentiation across individuals from different islands, a priori irrespective of inter-island distances. Furthermore, we found ROH patterns also substantially differing across islands but all indicating substantial reproductive isolation within islands. Finally, together with Mantel testing, we found significant substantial genetic Isolation-By-Distance at very reduced geographical scale (Rousset 1997, and see the end of **Chapter 2.2.b**), among Cabo Verdean-born individuals across, and even within, islands of birth. The strength of this signal was also of some surprise to us, as sampled individuals reported, in our anthropological interviews, extensive exploratory mobility across Cabo Verde, mainly for familial, professional, religious, or other social reasons. Furthermore, historians had also reported extensive population movements across islands throughout the history of Cabo Verde. However, again, these demographic movements do not seem to have entirely resulted in the expected homogenization of genetic diversity across Cabo Verdean islands, as we found, instead, substantial differentiation resulting from long-term reproductive isolation strongly anchored in birth places. Finally, these patterns were also consistent, in particular ROH patterns, with possible deviations from random-matting locally, albeit we did not formally test this hypothesis, yet (see **section 5.2** below).

We also investigated specifically the distribution of admixture patterns across individuals' islands of birth. While we did not find a specific geographic distribution of admixture patterns correlated with island proximity throughout Cabo Verde, we nevertheless found that, overall, two individuals born far away from one another, had higher chances of exhibiting more differentiated admixture patterns than two individuals born nearby. Furthermore, investigating ancestry-specific ROH patterns (ROH specifically within African or European admixed genomic segments respectively, or ROH spanning ancestry break-points), we found some differences across islands, indicating that, for some islands, admixture likely predated the subsequent reproductive isolation across islands.

Altogether these results were strongly indicative of a serial founder model followed by relative reproductive isolation across islands in the Cabo Verdean archipelago. Furthermore, our results were consistent with certain islands being initially peopled by individuals already admixed between African and European source populations (Laurent et al. 2022).

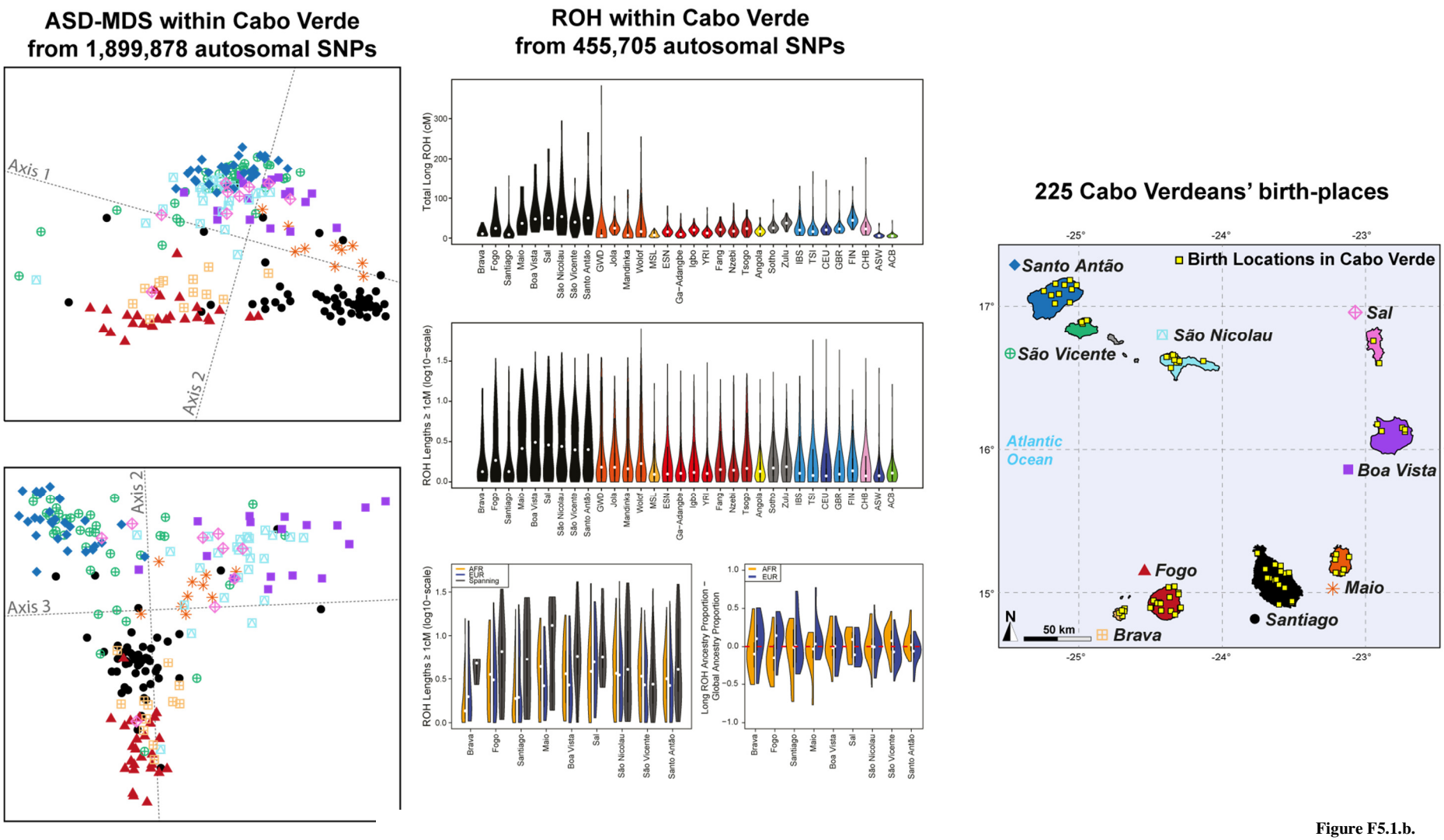


Figure F5.1.b.

3D ASD-MDS projection within Cabo Verdean born individuals and ROH patterns. Note that ASD-MDS projections have been procrusted to the actual geographical coordinates of individuals' birth-places.

Figure adapted from Figure F4 and F5 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

5.1.c. Complex admixture histories for each Cabo Verdean islands reconstructed with MetHis-ABC

On these bases, we set out to reconstruct the detailed admixture histories of each Cabo Verdean island separately from independent autosomal SNPs and using the *MetHis*-ABC framework described in the previous **Chapter 4**. We found out that the European and African admixture patterns in Cabo Verde stemmed from very few populations from either continent, and that the African contributions mostly stemmed from genetically very close populations only differentiated using extensive haplotypic information rather than a more limited SNP set. Therefore, we deemed reasonable in our case with *MetHis*-ABC to simplify possible admixture histories by considering only a two-source population model with the South-Western European Iberian IBS and the Senegambian African Mandinka at the root of the admixture process for each island.

We thus conducted 9 independent *MetHis*-ABC full-blown inference procedure, each corresponding to genetic data from individuals' born on the nine Cabo Verdean islands represented in our dataset. We considered four competing scenarios represented in **Figure 5.1.c1** below, for which detailed descriptions can be found in Laurent et al. 2022 at https://doi.org/10.1101/2022.04.11.487833, following the same parameter descriptions previously detailed in **Chapter 4**.

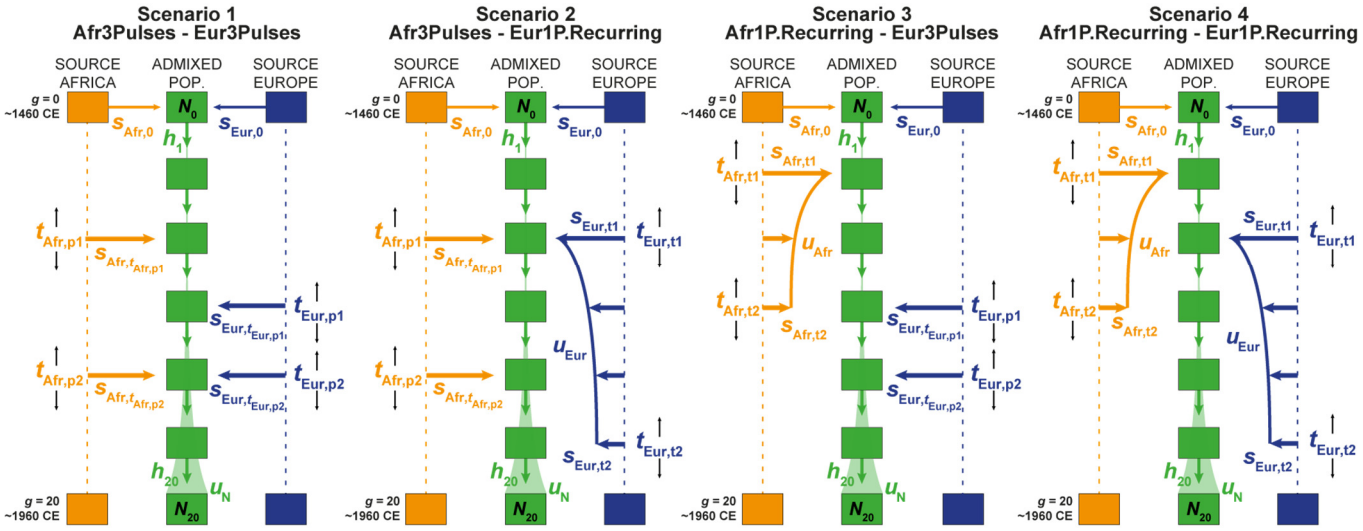


Figure F5.1.c1.

Four competing complex admixture models inferred with *MetHis*-ABC for each 9 Cabo Verdean island, separately.
Figure adapted from Figure F6 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

Scenario 1 considered two possible independent admixture pulses from both European and African source populations, after the founding pulse at generation 0. Scenario considered the same scenario of 3 admixture pulses from the African source, and a scenario with a period of recurring admixture of any short or long duration from the European source. Indeed, we here draw the onset and offset time for this period randomly within the 20 generations of the admixture process after the founding admixture pulse. Note that, also differing from models in Chapter 4, we here considered possible constant recurring admixture processes rather than strongly constraining a decrease in recurring admixture as we did in **Chapter 4**. Scenario 3 considered the same scenarios as the Scenario 2, but exchanging source populations of origins.

Finally, Scenario 4 considered a possible period of recurring admixture independently from either source, after the founding admixture pulse. Finally, also complexifying from previous competing scenarios in Chapter 4, we wanted here to evaluate the influence of changes in the reproductive sizes within the admixed population on our inferences. Therefore, we implemented a possibly increasing reproductive size within the admixed population with founding parameter values drawn in a Uniform [10 , 1000] distribution, and in Uniform [100 , 100,000] in the present, reproductive sizes in between being the numerical solution of a rectangular hyperbola function scaled between beginning and end sizes of "steepness" parameter u drawn in [0, 0.5].

Therefore, note that scenario nestedness was high in substantial spaces of parameter-values, more nested a priori than in our initial proof of concept (Fortes-Lima, Laurent et al. 2021). However, as we demonstrated then and explained in Chapter 4, our goal was not to test whether we were able to always discriminate among these subtle different complex models, but rather focus on the vicinity of our observed data. Furthermore, as we also demonstrated, nestedness and scenario-choice confusion with *MetHis*-ABC occurs in spaces of parameter values where scenarios are, in fact, biologically equivalent concerning the admixture mechanistic process.

Therefore, our competing scenario design aimed at identifying qualitatively whether highly complex models, with three pulses of admixture from either source, or if periods of recurring admixture, or combinations of both scenarios, best explained our observations on each Cabo Verdean islands separately. Under the winning model for each island, we would then estimate posterior parameter distributions, and contrasting the obtained genetic inference with historical data on the peopling history of each island that we compiled based on numerous historical previous publications, to propose possible explanations for the inferred admixture mechanisms reconstructed from genetic data.

Based on these scenario definitions, we conducted *MetHis* simulations and subsequent ABC Random Forest scenario-choice and ABC Neural Network posterior parameter estimations similarly as presented in **Chapter 4**, for each island of birth of Cabo Verdean individuals separately.

Our Random Forest Scenario-choice procedures favored scenarios comprising three pulses of admixture from both European and African sources for explaining genetic diversity patterns in 6 Cabo Verdean islands out of 9. Furthermore, we found that a scenario encompassing pulse-like models from the African source and, instead, a period of recurring admixture from the European source were favored in two islands out of nine. Finally, our inferences favored a scenario with pulses of admixture from Europe and a period of recurring admixture from Africa in a single island out of nine. Finally, note that Scenario 4, comprising a period of recurring admixture from both Africa and Europe, was the least favored model by Random Forest ABC inferences for all nine Cabo Verdean islands (**Figure F5.1.c2**).

Based on the winning scenario for each island, we then performed *MetHis* 100,000 simulations and conducted nine separate ABC Neural Network posterior parameter inferences (**Figure F5.1.c3**), following the same methodology as in Fortes-Lima, Laurent et al. 2021, described in **Chapter 4**.

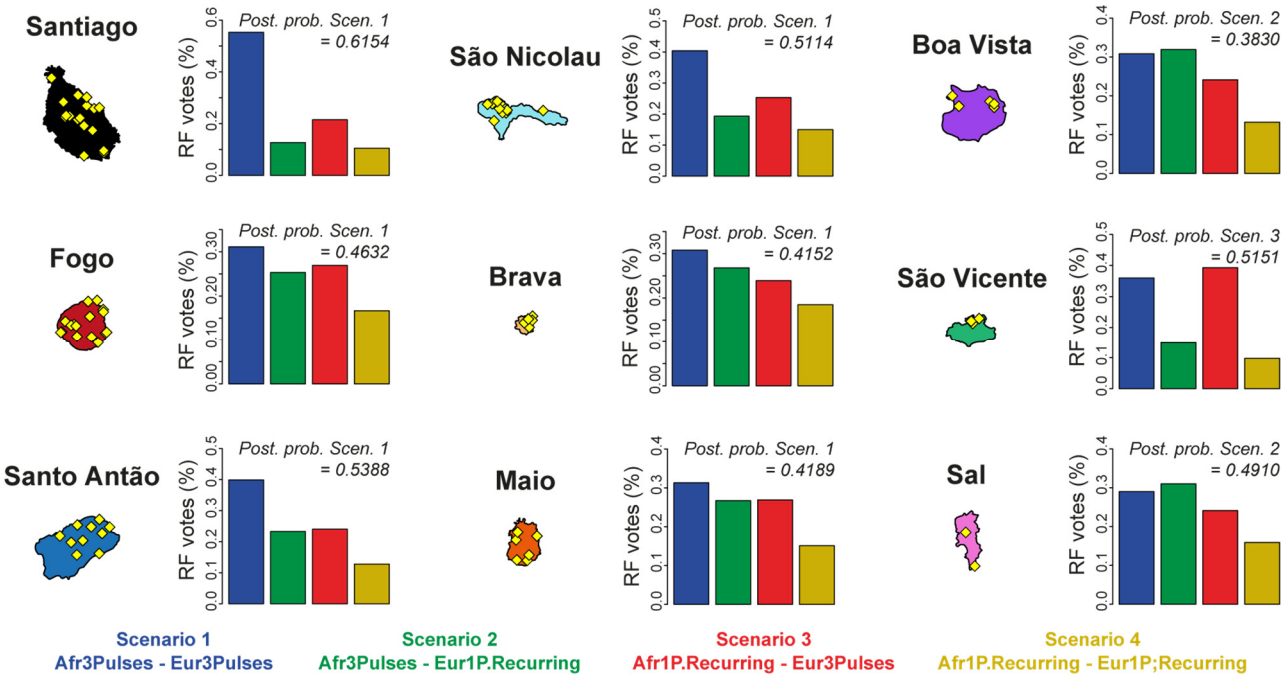


Figure F5.1.c2.

Random Forest ABC scenario-choice for the four competing complex admixture models inferred with *MetHis*-ABC for each 9 Cabo Verdean island, separately. Figure adapted from Figure F7 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

In parallel, we explored the historical literature about the peopling history of Cabo Verde, with a specific focus on census data on each island gathered by historians based, among others, on Royal decrees, real estate contracts and notary legacy deeds, parish establishments and records, tax and commercial records, and later, official demographic census. We reported our detailed findings associated with precise references from this literature outside of my disciplinary fields, in an extensive table in Supplementary Table ST2 from Laurent et al. 2022.

Based on these previous studies, we thus determined the historical date for the first administrative official settlement of each island separately, and also the first historical date when records unquestionably showed a perennial number of residents above 100 individuals until today, i.e. the first time in Cabo Verde history when 100 individuals reside on a given island and henceforth never got below this number until today. Finally, also based on historical records, we determined that the height of TAST, corresponding to a sharp demographic increase of enslaved-Africans deportations started in the mid-17th century until the abolition of the TAST in the early 19th century (see Fortes-Lima and Verdu 2021 for a review of anthropological genetics perspectives on the TAST, and see below). This period corresponds to the expansion of Plantation Economy, and the concomitant shift from a Society with Slaves to a Slave-Society (Berlin 1998; Chaudenson and Mufwene 2001), in the Caribbean and the Americas as well as in Cabo Verde and São Tomé e Principe, this latter archipelago of the Bight of Biafra being the original testing-grounds of this economic system. Finally, we defined a transition period during the 19th century corresponding to the period between the abolition of the TAST and that of slavery throughout the entire Portuguese empire in 1878. Note that the independence of Cabo Verde from Portugal was enacted in 1975 (Albuquerque et al. 1991; Carrera 2000).

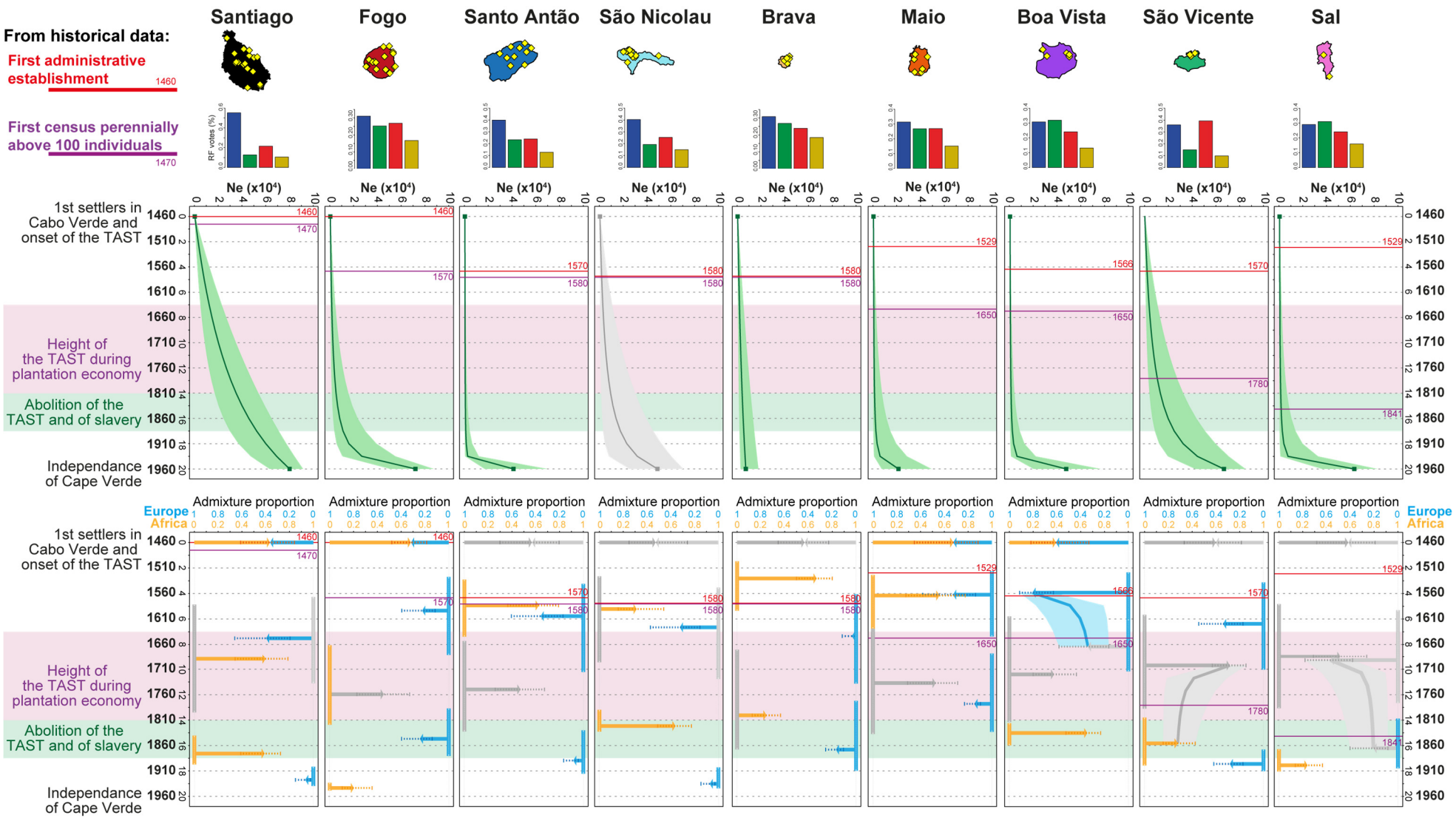


Figure F5.1.c3.

Neural Network ABC posterior parameter estimates for the complex admixture history of each Cabo Verdean island inferred, separately, with *MetHis*-ABC.
Median posterior point-estimates are provided with corresponding 50% Credibility Intervals. Parameters posterior distributions departing little from their priors and that our inference thus failed to identify are reported in gray.
Historical periods stem from historical records as described in text.
Top panels correspond to Reproductive Size parameters only and bottom panels to the admixture process itself, albeit posterior estimations were conducted jointly for all model-parameters.
Figure adapted from Figure F7 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.
Further refer to this publication for full-blown posterior distributions for each parameter and each island.

The interested reader should refer to Laurent et al. 2022 and Supplementary Text herein, for extended discussions of these results for each Cabo Verdean island separately in the light of historical knowledge about the peopling of each island. Here, we briefly synthesize and discuss these results at the scale of the Cabo Verde archipelago.

First (**Figure F5.1.c3**, top panels), our *MetHis*-ABC Neural Network posterior parameter estimations showed that Cabo Verdean reproductive sizes experienced a sharp increase very recently in the 20th century for almost every island, except Santiago, the historical economic and administrative center of Cabo Verde since the establishment of the colony in the 1460's, where reproductive size increase is estimated to have been more linear (and with the exception of São Nicolau where we largely failed to identify these parameters). These result echo historical reports of the difficult perennial peopling of Cabo Verde until as late as the 19th century on Sal, and the often large time-lapse between the first administrative establishment of an island and the first census perennially above 100 settlers until today in certain islands. Such difficult settlement and low attractiveness of the islands were due mainly to the dry Sahelian climate with overall difficult access to fresh-water surface resources[50], to chronic droughts triggering massive famine outbreaks accompanied by deadly epidemic outbreaks, to shifting political dominion and changing commercial opportunities, to piracy disrupting the fragile settlements during the 17th and 18th centuries, and to the distance to continental Europe. Interestingly, historical records show the repeated strong incentives to settle Cabo Verde issued regularly by the Portuguese crown between the 15th and early 17th centuries, delivering extremely liberal commercial opportunities and freedom to emigrants to the new colony, in the hope to maintain the Portuguese dominion on the archipelago which occupied a primary geostrategic position in the commercial routes between the three continents (Albuquerque et al. 1991; Carreira 2000).

Concerning the admixture histories themselves (**Figure F5.1.c3**, bottom panels), note first that only Santiago, Santo Antão, and to a less clear extent, São Nicolau, show the first identifiable admixture event synchronically to the first peopling of an island. All other islands show founding admixture events occurring before, sometime long before, the historical dates indicating perennial peopling. Thus, together with ROH results presented above (**section 5.1.b**), this indicated that Cabo Verdean islands, apart from the three islands cited above, were initially settled by already admixed individuals coming from other Cabo Verdean islands. Second while we identify numerous admixture events between the 1460's and the 1650's from either Europe or Africa separately, or more synchronically from both continents, we massively fail, for all islands, to infer significant admixture events during the height of the TAST between the 1650's and the early 19th century. This indicated very tenuous information in genetic data for identifiable admixture events from this entire period, when in fact far most enslaved-Africans were brought to the archipelago, and when European emigration waves were also substantial. Furthermore, note that during the period between the abolition of the TAST and that of slavery, we precisely estimate (with posterior estimations sharply departing from their priors, reduced 95% Credibility Intervals, and low error rates), once again, numerous admixture events from either Africa or Europe depending on the island. Finally, we identify several precise events of admixture during the 20th century, with low intensities, albeit very accurately inferred with our procedures.

Altogether, these results, together with local ancestry inferences' results in **section 5.1.a** and ROH results in **section 5.1.b**, are consistent with a very limited contribution of the numerous deportations from all over Africa to Cabo Verde, during the height of the TAST (**Figure F5.1.c4**), to the genomic diversity and genetic admixture histories of the archipelago. This could be due to two, non-mutually exclusive

[50] Cabo Verde is named in reference to the proximity of "Cap Vert" in Senegal, rather than because it is "green", which is in fact very far from reality during most of the year on the archipelago

phenomenon. First, it is plausible that during this era, most enslaved-Africans deported to Cabo Verde only transited via the archipelago before being re-deported to the Americas and the Caribbean mainly, and to Europe to a much lesser extent. It has been long-debated by historians whether Cabo Verde during this era was, indeed, only a transit platform, or whether enslaved-Africans stayed for much longer amounts of time or indefinitely. Our results cannot conclude to either hypothesis (see below), but we can confidently assess that, whichever the duration of the stay in Cabo Verde of these individuals, they probably did not substantially contribute to the genetic diversity of the archipelago until today. Note however, that the contribution of these migration waves to the cultural diversity of Cabo Verde is unquestionable as Cabo Verde Kriolu still bears traces of the influence of numerous languages from continental Africa from Senegambia to Angola (Quint 2000; and see also **section 5.3** below).

We cannot literally conclude to settle this demographic and historical debate, as we, in population genetics, only witness genetic (and not genealogical) histories of extant populations, and not their life-history of mobility. As such, another, non-mutually exclusive, socio-cultural process may very well explain the observed genetic and admixture patterns, irrespective of the duration of the stay of enslaved-Africans on the archipelago.

Historians often describe the shift from a "Society with Slaves" to a "Slave-Society" with the expansion of the Plantation Economic system in the mid-1650's (Berlin 1998; Chaudenson and Mufwene 2001; Eltis and Richardson 2015). This period would be characterized by major socio-cultural changes, and, in particular, concerning marital and reproductive relationships between enslaved and non-enslaved communities, as well as within enslaved-communities between newcomers and previously established enslaved-African descendant communities. Indeed, this historical era was characterized by the emergence, everywhere throughout colonial empires, of legally enforced socio-marital segregations strongly forbidding marriages between enslaved and non-enslaved communities, as well as providing legal support to slave-owners for controlling unions among enslaved individuals. This is best exemplified by the enactment of the Code Noir in French colonies in the Americas (which represented at the time vast regions of the Caribbean and most of the North American continent) in 1685 and its strict enforcement from the early 18th century and on, whether under the French rule or not. Similar legally enforced changes also occurred in Cabo Verde, as exemplarily demonstrated in the seminal work of Carreira (2000).

This period also saw the normalization of racist discriminations and increase in brutal dehumanization and violence exerted in particular towards enslaved-Africans, notwithstanding the undeniable horrors committed during slavery prior to this era. Importantly, this Plantation Economy era also corresponds to a demographic shift with massive deportations of enslaved-Africans (**Figure F5.1.c4**), accompanied punctually and locally, by immigration waves (forced or voluntary) from Europe, resulting in global increases of non-native population censuses in European colonies on both sides of the Atlantic (Eltis 2002; Fortres-Lima and Verdu 2021).

Therefore, it is plausible that these major socio-cultural shifts regarding relationships between enslaved and non-enslaved communities triggered with success marital and reproductive isolation among communities and therefore, prevented substantial admixture to occur; despite massive demographic movements putting into contact previously genetically isolated continental European and African populations, and notwithstanding known dramatic sexual abuses exerted by non-enslaved individuals on enslaved individuals. Furthermore, the increased control of slave-owners on marital and reproductive relationships within enslaved communities might have further prevented admixture between newcomers and previously established enslaved-communities, thus also further limiting genetic admixture events. Note,

finally, that such process could have effectively limited genetic admixture, irrespective of the duration of the stay of enslaved individuals, locally.

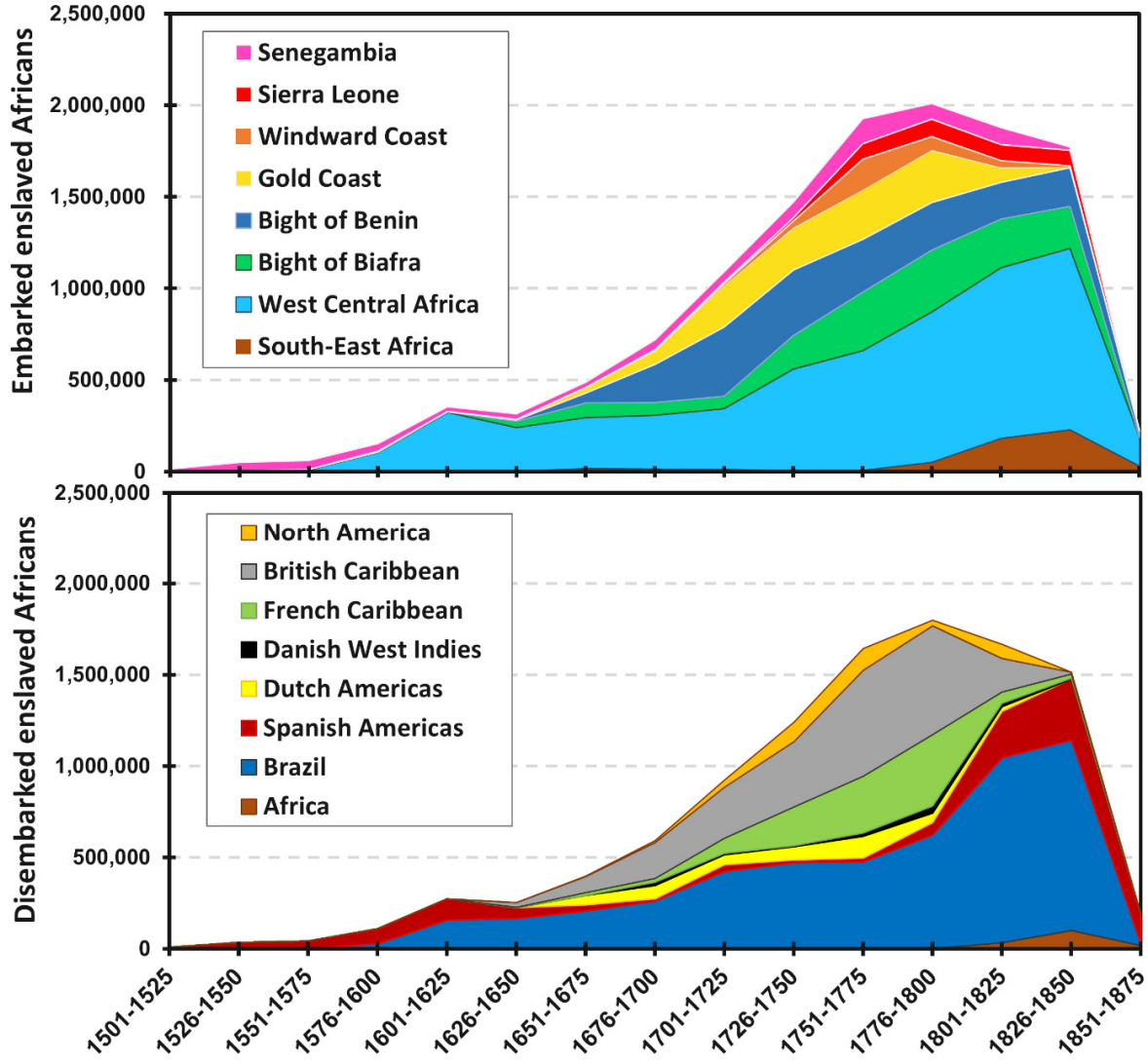


Figure F5.1.c4.

Historical records for cumulated estimates of enslaved-Africans demographic forced deportation from Africa and disembarkations in the Americas and the Caribbean, based in part on Eltis and Richardson 2002. Source tables openly available at https://www.slavevoyages.org/estimates/L8KDKZCH.

The massive expansion of the Plantation Economic system, tested first in São Tome e Principé, and rapidly implanted in the Americas and the Caribbean as well as European colonies in Africa in the first half of the 17th century, triggered the massive expansion of enslaved Africans deportations until the abolition of the TAST starting in 1807. The 19th century still witnessed extensive enslaved-African deportations, often illegal, until the effective abolition of Slavery in European colonial empires in the second half of the 19th century.

Figure previously published in Fortes-Lima and Verdu *Human Molecular Genetics* 2021.

In this context, the observed genetic patterns today and detailed genetic inference results would be highly consistent with the following schematic historical scenario:

1460's-1630's: Initial settlement of European migrants in the island of Santiago, low in numbers and strongly sex-biased towards males. Deportations of enslaved-Africans to Cabo Verde, including women, forced to participate in the building of the colony. Rapidly, as witnessed by historical records (see **Chapter 1.2**), admixture between European males and enslaved-African females occurred and significantly contributed to the genetic make-up of the new colony. Further enslaved-African deportations, massively from the nearby Senegambia, contributed, via admixture with pre-existing populations, to the genetic diversity of the archipelago. During this period, successive settlements of other islands either readily involved admixed individuals or rapidly funded anew admixed populations via the same process as described in Santiago.

1630's-1800's: Normative and prevalent socio-cultural barriers strongly segregating enslaved and non-enslaved communities deployed with the expansion of Plantation Economy, and intensified enslaved-African deportations, would have successfully prevented admixture between communities as well as between the varied newly arrives continental African populations and pre-existing slave-communities mainly of Senegambian origins.

1800's-1875's: the abolition of the TAST and that of slavery triggered numerous illegal enslaved-African deportations to Cabo Verde (and elsewhere). In parallel, socio-cultural changes in European colonies led to these abolitions and further echo profound societal changes in the relationships between enslaved and non-enslaved communities, despite segregation and racism being durably socio-legally established in most colonies. Nevertheless, these changes could have triggered, once again, substantial admixture events as identified in our analyses.

1875's-present day: Strong demographic population expansion of Cabo Verde limited the influence of novel admixture events to strongly modify populations' genetic diversity patterns in Cabo Verde. Nevertheless, we identified significant and very accurate; albeit reduced in intensity, admixture events from either Europe and Africa, possibly related to known historical workforce and economy-related migrations among the former Lusophone empire colonies in Africa and beyond, in particular in the 20th century.

5.2 Ongoing work and perspectives for section 5.1: the genetic admixture histories of Cabo Verde.

Future work will need to evaluate the plausibility of this detailed socio-historical causes for explaining the observed genetic admixture processes here described. Indeed, population genetics can only infer the "how", but not directly explain the "why". However, such hypothetical, but likely, scenario would imply strongly sex-biased admixture processes in a strongly stratified society, possibly both mechanisms being dynamic over time. We did not evaluate such possibilities, as our approaches only consider non-sex-specific admixture processes and random matting in the admixed population (see **Chapter 4**).

We are currently, with my PhD student Marta Ciccarella, developing a sex-specific version of *MetHis* and further parameterizing non-random matting. This will soon allow us to formally test the above hypotheses, and evaluate whether the genetic expectations of this history of socio-cultural changes may match, or not, the genetic patterns observed today. Note, however, that investigating non-random matting processes in a very small census population is not conceptually trivial, as mate choices are already limited.

Importantly in this dialectical approach, finding a match would not validate that the socio-historical scenario here proposed is indeed causative of the observed genetic and admixture patterns. However, finding a mis-match between genetic expectations under such historical causative model and observed genetic patterns would allow us to formally reject this particular historical process as causative; a result which would further require building novel hypotheses for future testing.

Furthermore, Ms Ciccarella is also in the process of reconstructing the admixture histories for São Tomé e Principe, using the same type of genome-wide genotyping data that I generated anew thanks to a collaboration with Pr. Jorge M. Rocha (CIBIO-University of Porto). Indeed, Pr. Rocha sampled São Tomé e Principe populations more than twenty years ago for reconstructing the genetic peopling history of this Central African archipelago. We decided that I would genotype remaining DNA extracts from these collections and that Pr. Rocha would co-supervise Ms Ciccarella's PhD, in order to widen our questions of interest to compare the detailed admixture histories at play separately in the two former Portuguese peopling colonies in Africa. Furthermore, this ongoing project will allow us to investigate migrations and diasporas between the two archipelagos. Indeed, there is still a substantial community of Cabo Verdean descendants in São Tomé e Principé, due to known ancient and recent workforce and economic migrations, as also witnessed in my anthropological questionnaires within Cabo Verde about individual mobility histories. This project will likely soon further our understanding of the influence of historical socio-cultural dynamics on genetic diversity and admixture patterns in human populations, in the context of the TAST.

Finally, in addition to the methodological perspectives for *MetHis*-ABC developments presented previously in **Chapter 4**, this Cabo Verdean study highlights the empirical need to evaluate the sensitivity of our inferences to the use of extant populations as proxy source populations. Indeed, a long-standing question in admixture statistical or mechanistic inferences regards the influence of how far genetically the proxy population may be from the real population at the source of admixture in the past, before strongly biasing the results of our analyses. To formally test these issues empirically, we are currently testing several cases with *MetHis* by, *i)*, comparing inference results obtained using different source populations, or, *ii)*, combining samples from different populations, including the Cabo Verde admixed one, in a single source; or, *iii)*, by using, instead of observed genetic data, explicit sequential-coalescent simulations for producing source populations genetic data. We started to conduct this methodological-empirical project with Dr. Laurent and Dr. Toupance in the lab, and Dr. Laval and Dr Patin at the Pasteur Institute.

5.3. Parallel trajectories of genetic and linguistic admixture histories of Cabo Verde

Our initial project was not only to reconstruct the genetic admixture of Cabo Verde, but investigate also a certain type of linguistic diversity: individual utterances in the context of semi-spontaneous discourses conducted by academic researchers, across individuals for which we had sampled linguistic, genetic, and anthropological data (see **Chapter 1.2**, and introduction to **Chapter 5**). As said, this was thus a pluri-disciplinary project in nature as, despite having built the project and conducted sampling in a largely interdisciplinary way, subsequent analyses and problematics were disciplinarily and conducted separately between genetics and linguistics, and only then results would be compared between disciplines.

In this context, we had gathered extensive semi-spontaneous discourses, fully transcribed in the morpho-phonetic orthographic norm for writing Cabo Verdean Kriolu (ALUPEC), a tedious task requiring extensive linguistic skills, mostly conducted by Sergio S da Costa over the years. Based on this data, I aimed at investigating linguistic utterance diversity patterns across individuals, as a function of the anthropological co-variables such as residence location, birth-places and parental and grand-parental birth-places, academic education duration, age, and sex.

Furthermore, I initially thought that linguistic expertise could establish a form of qualitative distance between individual utterances and African and, respectively, European source languages. Indeed, as creolists already referred to complex concepts such as "basilectal" or "acrolectal" Kriolu language variants, I candidly thought that these concepts designated a form of relative distance to, respectively, African or European languages. With such index, I could compare linguistic variation across individuals with genetic admixture patterns, and evaluate whether linguistic and genetic patterns matched or mismatched (as, at the population scale, in the Central African example in **Chapter 2**), and further elaborate on the different cultural and biological histories of admixture in Cabo Verde.

In 2011, drowning in post-doctorate life juggling with multiple projects and job-applications, I was immersed in the genetic analyses of the project and could not devote all the time required to explore the possibilities of the linguistic data we had gathered on the field. Luckily for me, a theoretical population genetics PhD student supervised by Pr. Rosenberg, Dr. Ethan M. Jewett, was working on a side project about N-grams and linguistic diversity within the Anglo-American language. Discussing my Cabo Verde project with him, he proposed to help us with parsing and describing the linguistic data we had already gathered. His help was seminal to our project, and we published together, in 2017, the first article, to our knowledge, about genetic-linguistic descriptive comparisons among the same individuals at the within-population and language scale, for a preliminary sample of 44 Cabo Verdean individuals collected in the 2010 and 2011 fieldworks. To further help Dr. Jewett to better represent the pluri-disciplinary challenges of our project and data, he came to observe our fieldwork sampling with Pr. Baptista, Sergio da Costa, my students, and myself, for the 2014 and 2016 fieldtrips. Dr. Jewett soon-after left academic research for private research in bio-informatics, and, several years after the initial publication, we could replicate and expand our initial methods to the full Cabo Verdean dataset, confirming that our results were not due to small preliminary sample sizes, and furthering them and their interpretations. I will, henceforth, synthesize these results published preliminarily in Verdu, Jewett et al. (2017), and in full in Laurent et al. (2022), pluri-disciplinarily alongside the genetic history of admixture detailed above in **section 5.1**, freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

We first parsed all individuals' semi-spontaneous discourses of Kriolu "spoken every day" (see **Chapter 1.2**), by considering various types of morpho-phonetic or syntactic variation of the same lexical root transcribed explicitly in ALUPEC as different itemized "utterances"; and excluded onomatopoeia,

names and interjections from parsing. Note that we kept as different uttered items several, very rare, English or French words that fitted our definition of utterances (**Chapter 1.2**).

This parsing resulted in a total of 92,432 uttered items across the 225 genetically unrelated individuals born on Cabo Verde in our dataset, out of which 4,831 were unique. We then simply counted the occurrence of each unique uttered item for each individual discourse, and then calculated the relative frequency of usage of each unique item in an individual's discourse: his/her manners of uttering Kriolu in this context. Finally, we computed from there the individual pairwise matrix of Euclidean distances between unique utterances' relative frequencies, a measure of inter-individual linguistic distances between individuals' manners of speaking Kriolu (**Figure F5.3.a**). Note that, again (**Chapter 1.2**), we did not decide whether individual's spoke expert Kriolu or no Kriolu at all: individuals just self-reported speaking Kriolu and we asked them to speak in their own Kriolu from every day, for instance, that they spoke at home. Furthermore, we did not interrupt people during the narration, and fully transcribed them even in the cases (rare, but not exceptional either) when participants drifted to speak about something completely different.

Individual X's uttered discourse:

N ta lenbra sin. Na kel filmi, kel rapas staba xintadu si.. kel rapas panha kel masan…kel masan e ruma si… e ruma-l asi… e dura ku el ta ruma asi… dipos di kel, e staba xintadu pa si… e tene si kel otu kolega ki staba di la… aian, N odja si…
[etc.]

Individual X's corresponding unique utterance count:

N	ta	lenbra	sin	na	kel	filmi	rapas	staba	xintadu	si	panha	masan	e	ruma	ruma-l	[etc.]
2	2	1	1	1	7	1	2	3	2	4	1	2	5	2	1	[etc.]

Individual unique utterances' relative frequencies:

	a	á	abakate	abakáti	abaná	ábitu	ábitus	abizu	abô	abo	abon	aboreside	abra	abrí	abri	[etc.]
CV2018_BOR_04	0.00214	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
CV2018_BOR_05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
CV2018_BOR_06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
CV2018_BOR_07	0	0	0	0	0	0	0	0	0.00146	0	0	0	0	0	0	
CV2018_BOR_08	0.00203	0	0	0.00203	0	0	0	0	0	0	0	0	0	0	0	
[etc.]																

Pairwise matrix of inter-individual Euclidean distances between vectors of uttered relative-frequencies:

For each pair of individuals i and j, and for $L = 4831$ unique uttered items:

$$d(f_i, f_j) = \sqrt{\sum_{l=1}^{L} (f_{i,l} - f_{j,l})^2}$$

Figure F5.3.a.
Schematic representation of the parsing of semi-spontaneous discourses collected for each sampled individual in Cabo Verde, further counting unique utterances and individual unique utterance relative frequencies from which we compute a pairwise Euclidean distance matrix of linguistic variation. Method originally presented in Verdu, Jewett et al. *Current Biology* 2017, and further described in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

We described the obtained matrix of pairwise utterance distances using similar statistical-approach and tools as previously for genetics descriptions at the within Cabo Verde level, and could thus juxtapose results alongside the obtained genetics patterns for comparison.

First, we conducted MDS projections of this utterance Euclidean distance matrix (**Figure F5.3.b**), which showed that the first axis of variation interestingly showed a gradient of inter-individual distances separating, apparently, different manners of speaking Kriolu across individuals born on different islands. The second axis distinguished mostly across certain individuals within Cabo Verde. Furthermore, considering instead the first and third axes, further revealed clearly differentiated manners of speaking Kriolu among individuals born on different islands, or groups of islands, with some striking similarities with what we had identified in the first three axes of the genetic ASD-MDS for the same individuals. Indeed, we found substantial differentiation between Santiago-born individuals' manners of speaking Kriolu from that of Fogo-, Maio-, and Brava-born individuals, differing from Santo Antão- and São Vicente-born individuals, and with São Nicolau-, Boa Vista-, and Sal-born individuals being at intermediate distances between the three groups. To some extent, such patterns of linguistic differentiation was expected by linguists having investigated the different Kriolu linguistic variants, although they did not formally evaluate inter-individual distances in manners of realizing in utterances these different language variants, as this was not their object of study (**Chapter 1.2**).

Investigating in details Mantel correlations between utterance distances and the corresponding distances between individuals' anthropological co-variables, we found that differences in manners of speaking Kriolu were significantly positively correlated with age differences: as classically expected in socio-linguistics, tow individuals with major age differences speak, on average, in a more differentiated way than two individuals closer in age. Beyond this result, we found, surprisingly that differences in utterance frequencies significantly correlated with geographic distances among individual birth-places within Cabo Verde (Spearman rho = 0.2855, two-sided Partial-Mantel $p<2.10^{-4}$, correcting for age differences). Furthermore, we found significant positive correlations also with paternal and maternal birth-places' distances respectively. Finally, we did not find correlations wit other co-variables such as residence locations (people do not speak Cabo Verdean variants according specifically to where they live), nor with education duration (people do not speak Cabo Verdean variants according to whether they receive longer or shorter academic education).

Altogether (Verdu, Jewett et al. 2017; Laurent et al. 2002), these results indicated, first, that, similarly to genetic differences, individual manners of speaking Kriolu as captured by differences in utterance frequencies, are for a significant part anchored in individual birth-places: two individuals born far apart in Cabo Verde are more likely to speak Kriolu in different ways than two individuals born close-by. Furthermore, correlations with parental birth-places, as well as in our 2017 publication, with grand parental birth places, demonstrated that at least a part of inter-individual variation in manners of speaking Kriolu is transmitted from one generation to the next, via anchorage in birth-places.

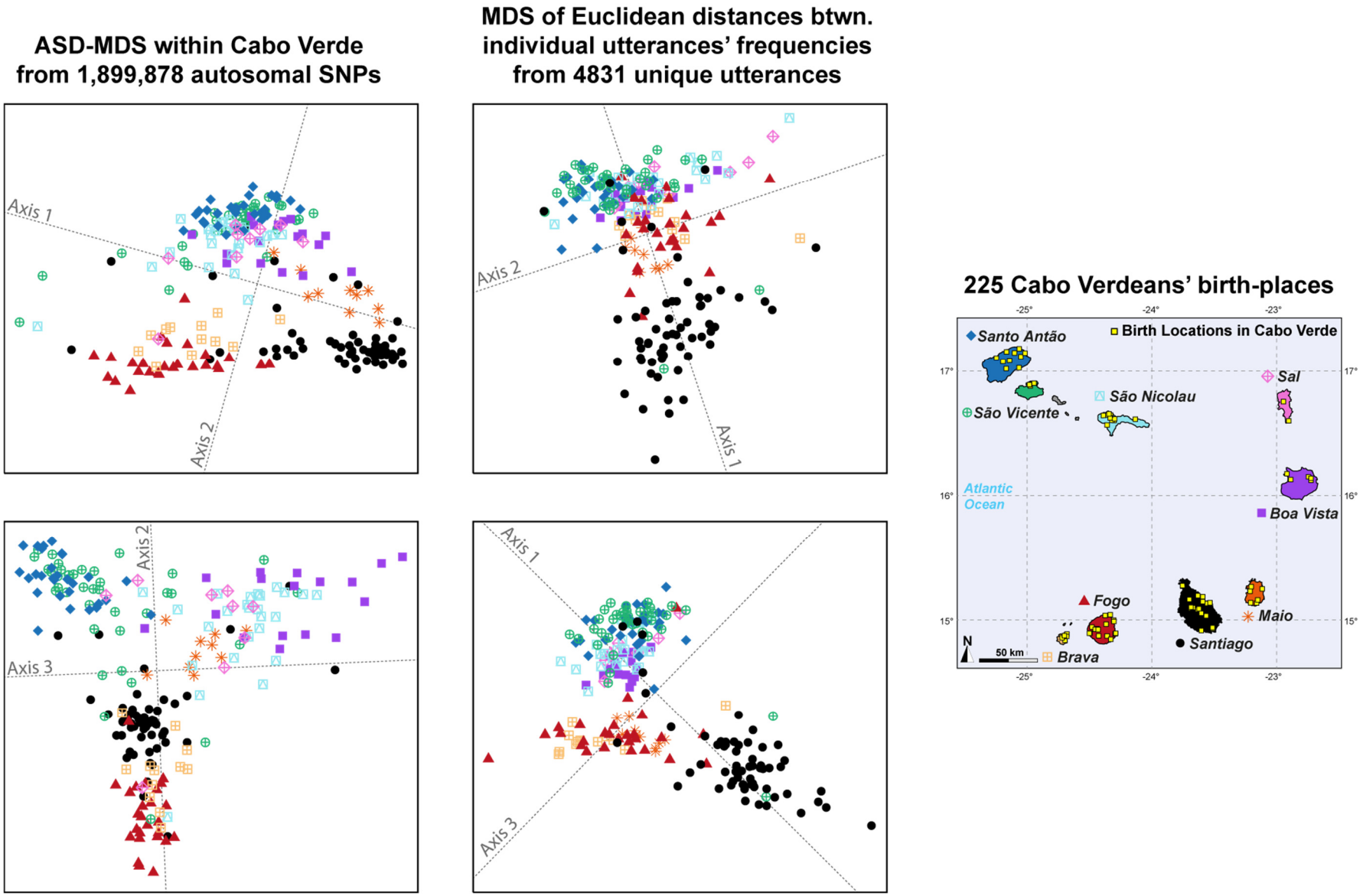


Figure F5.3.b.

3D ASD-MDS projection within Cabo Verdean born individuals as in Figure 5.1.b above. 3D MDS of the matrix of pairwise Euclidean distances between individual utterance frequencies as in Figure F5.3.a
Note that all MDS projections have been procrusted to the actual geographical coordinates of individuals' birth-places.
Figure adapted from Figure F5 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

Based on these results, we further aimed at comparing genetic and linguistic admixture patterns. However, my initial thoughts of establishing an individual "basilectal" or "acrolectal" score that could be translated into levels of "African or European linguistic admixture levels" were long-lived… Indeed, as per linguistic paradigms (see **Chapter 1.2**), the "basilectal" or "acrolectal" qualifications apply to language variants, not to individual realizations of these linguistic variants in a given context, as what we had with our utterances. Nevertheless, we could do something else (Verdu, Jewett et al. 2017; Laurent et al. 2022).

We decided to classify each 4831 unique utterances in one of four different categories. Category A comprised uttered items identified by linguists as directly coming from a known African language. Category B comprised uttered items identified by linguists as having a dual African-European etymology. Category C comprised items identified by linguists as of a strictly European (mainly Portuguese) origin, not bearing any traces of "Africanization", whether phonetically, morphologically, or semantically. Finally, Category D comprised words of a recognized Portuguese origin but showing possible traces of influences from African languages, whether phonetically, morphologically, syntactically, or semantically. We thus set out, Pr. Baptista, Mr. Costa, and myself, to evaluate in turn each one of 4831 the unique uttered items and classify them into each one of these classes. Note that we found 88 unique items in category A, but only uttered in total 3803 times over the 92,432 utterances, thus relatively rarely. Category B comprised 256 other items uttered 6960 times in total. Category D comprised only 26 items uttered 6762 times in total. Category C comprised all other items.

We then decided to count specifically the relative frequency of usage of items in each category specifically, and compute thus an "African linguistic score" as: for individual *i* and the set of utterances in each category denoted *Cat* (in [A; B; A&B]), $Z_{i,Cat} = \sum_{l=1}^{L_{Cat}} f_{i,l}$, with L_{Cat} the number of uttered items in the corresponding category, and $f_{i,l}$ the frequency of utterance of item *l* by individual *i*, defined as previously.

Results showed that the inter-individual differences in relative frequency of Kriolu uttered items specifically marking intense past contacts with African languages, were positively significantly correlated with individual birth-places distances (for Category A&B, Spearman rho =0.1297, two-sided Mantel $p<2x10^{-4}$) and even marginally significant at very close geographical scale within each island and without considering inter-birth-island distances. In other words, two individuals born far apart in Cabo Verde were more likely to use different frequencies of African-origin utterances than two individuals born nearby.

Ultimately (**Figure F5.3.c**), we found a strong positive correlation (Spearman rho =0.2070, p = 0.0018) between individuals' linguistic African scores, measured as the relative frequency of usage of uttered items with African origins or strong African influences, and individual levels of genetic admixture from African origin estimated as previously (see **section 5.1**).

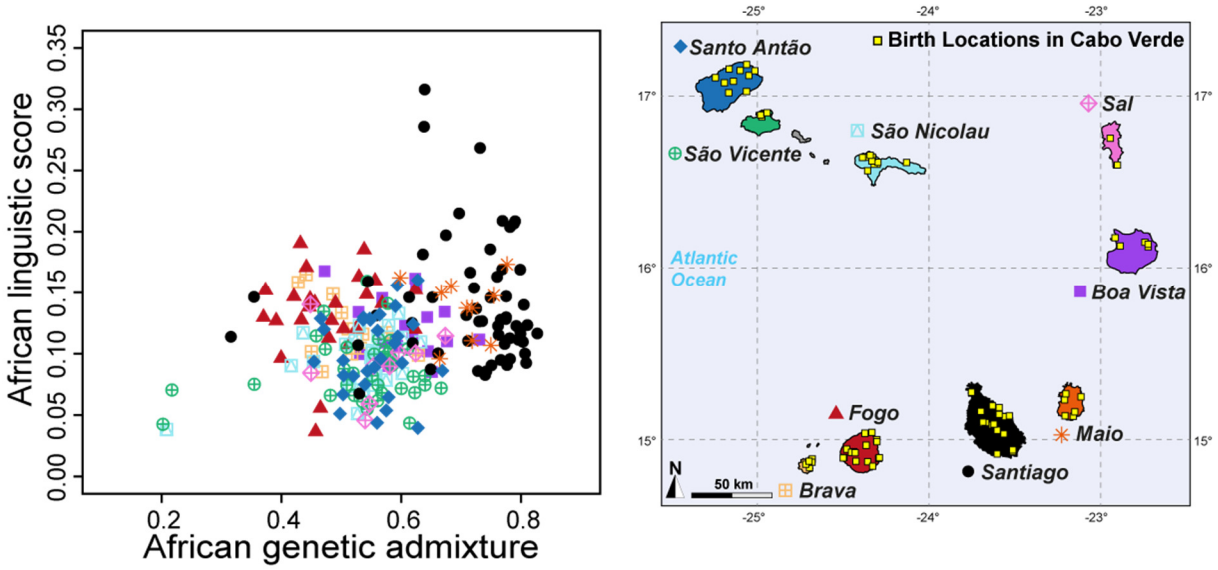


Figure F5.3.c.

Positive correlation between individual genetic admixture levels from Africa, and individual relative frequency of using uttered items marked by a strong influence of African languages.

Figure adapted from Figure F5 in Laurent et al. (2022), freely available on BioRxiv at https://doi.org/10.1101/2022.04.11.487833, and currently under review at *eLife*.

Altogether (Verdu, Jewett et al. 2017; Laurent et al. 2022), we found that a significant part of individuals' respective manners of uttering Kriolu were inherited vertically from one generation to the next, and anchored in birth-places. Furthermore, we found a positive correlation between genetic admixture levels and linguistic African admixture levels, also anchored in birth-places from one generation to the next. We therefore proposed two non-mutually exclusive scenarios for explaining the observed patterns, presented in **Figure 5.3.d** below.

First, genetic and linguistic admixture histories seem to have occurred in parallel, anchored in individual birth-places under a mechanism involving genetic and linguistic Isolation-By-Distance, including, or not, some form of non-random matting possibly based on manners of speaking of individuals.

Second, indiscriminately whether the linguistic manner of speaking Kriolu is transmitted vertically over generations, horizontally among peers within generations, and/or obliquely via other sources such as medias, individuals could self-construct their linguistic identity to use more or less frequently Africanisms in their Kriolu utterances, reflecting their genealogically known or putative ties to African or European sources, or doing so based on pre-assumed levels of African or European genetic ancestry anchored in imperfect phenotypic markers of such origins such as skin color, or else.

In all cases, our results showed that genetic and linguistic admixture histories occurred in parallel in Cabo Verde, whether mechanistically and/or whether the social construction of individual utterance of the Kriolu language followed indirectly the genetic admixture histories of the archipelago.

Importantly, we have no formal way of testing these hypotheses, as such inferences would require knowledge of an explicit linguistic model of transmission over generations with possible variation (mutations) and borrowings (migration) which, to our knowledge and that of our linguists' colleagues, simply does not exist for such complex linguistic traits as utterances…

Which brings us, finally, to the last section of this dissertation.

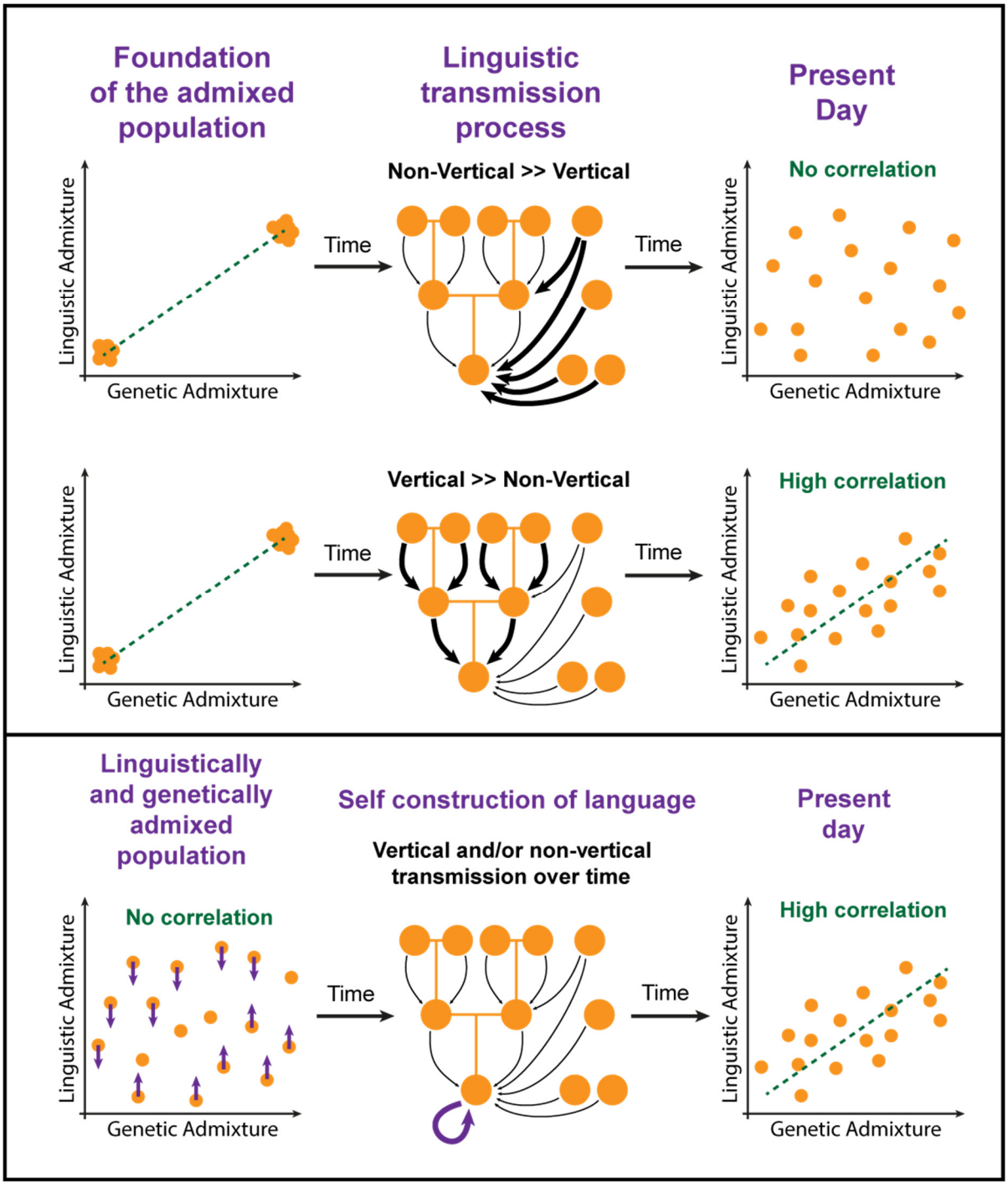


Figure F5.3.d.

Schematic representation of possible mechanisms to explain the observed correlations between genetic and linguistic admixture within Cabo Verde. Note that the first panel represents what would be expected if vertical transmission of linguistic features among individuals from one generation to the next was marginal, as opposed to what we observed in our data.

Figure adapted from Figure 1 and 6 in Verdu, Jewett et al. *Current Biology* (2017).

5.4. Ongoing perspectives: A novel "population linguistic" framework, and novel joint inferences of genetic and linguistic histories from observed data

5.4.a. ABC inference for reconstructing the linguistic history of populations

Between 2015 and 2018, Dr. Valentin Thouzeau's PhD, co-supervised by Dr. Frédéric Austerlitz and myself, aimed initially at comparing genetic and linguistic diversities and reconstructing the respective phylogenetic evolutionary trees using newly developed maximum likelihood approaches, focusing on extant genetic and linguistic data gathered in the lab over the years from Central Asia by Pr. Evelyne Heyer and Dr. Philippe Mennecier (see Thouzeau et al. 2017). However, these powerful methods were nevertheless limited due to difficult explicit implementation of possible migrations or admixtures among branches of the genetic or linguistic trees, as the likelihood itself of such highly complex models when considering numerous population samples is often impossible to write or, in all cases, computationally very hard to maximize in a convergent way (see **Chapter 4**).

Facing such classical issues, we decided to try to see what could be done with an ABC approach. For the genetics part, numerous methodological tools already existed and we were already versed in such analyses (see **Chapter 2** and **4**). But for linguistics data, the fundamental requirements of ABC, namely the ability to simulate efficiently data mimicking the observed data and the ability to calculate summary-statistics a priori informative about the underlying parameters of the simulated models, were virtually inexistent. As we just said above, this was mainly due to the lack of a formal synthetic model of linguistic evolution. Indeed, while certain fields of linguistics, such as glottochronology, phonetics, historical linguistics, or syntactics, indeed sometimes formalized models of linguistic transmission with variation over time, such theoretical models differed across linguistic traits and items, and, in any case, the vast majority of linguistics highly complex objects lacked explicit consensual such models. As a result, while we had extensive classical lexical linguistics data characterizing linguistic diversity across numerous Central Asian populations, we had no mechanistic models for simulating such data, and thus no possibilities of conducting ABC inferences for reconstructing the complex history of linguistic isolation, borrowing, and variation having given birth to the observed data.

In this context, Dr. Thouzeau set out to build, not a "synthetic theory of linguistic evolution", but an empirical mechanistic model of linguistic transmission with changes and possible borrowings over time, in order to be able to simulate, not words, but linguistic variation across languages observed from a specific type of linguistic construct: lexical cognates. From lexical comparisons across languages, historical and comparative linguists classically define "lexical cognates" as sets of lexemes, or words, from different languages with a common meaning[51] and a common etymological root in a putative common language of origin (Pagel 2009). For instance, the English "brother", the German "Bruder", and the French "frère" form a lexical cognate as they have a common etymological root, but the Spanish "hermano" forms another cognate as it has the same meaning but derives from a different etymological root.

Based on these definitions, Dr. Mennecier determined the cognates for all items in the word-lists he had established to characterize the different languages from which Pr. Heyer obtained genetic samples in joint fieldworks over the years. We thus had population genetics data from grouped individual samples and corresponding cognate diversities across languages spoken by these populations. In this context, and further

[51] Same meaning is not always used by linguists for defining cognates, but was considered in the particular definition of cognates by Dr. Mennecier in the study-case of Central Asian languages.

expanding the "common origin" definition of cognates, Dr. Thouzeau built two mechanistic models, one for classical admixture and one for classical migration (see **Figure F3.1** and **Chapter 3**), able to produce cognate diversity in a set of populations. He implemented a software package, *PopLingSim* (Thouzeau et al. 2017), to simulate efficiently lexical cognate diversity as a function of parameters drawn from prior distributions set by the user (**Figure F5.4.a1**). Under these models, a set of *N* lexemes with different meanings evolved as a vector of independent items in three-populations topologies, each lexeme possibly changing punctually with a probability "μ" (the cognate "mutation" rate) at each "linguistic generation".

Then, in unidirectional migration models, each lexeme could be borrowed from one branch to another one with a probability "δ" (the cognate "migration" rate); while in the linguistic two-population admixture model, at a given point in time, a source population would contribute a proportion of *r* lexemes' cognates to the lexical list of the new admixed population (and 1 – *r* lexemes contributed by the other source population). Note that these models of linguistic evolution only generate diversity among the objects defined by linguists following these rules. The objects in question are not directly observed in language practices (i.e. one does not speak in cognates, but in utterances), but are, as classically in linguistics, objects built buy linguists and relying on complex scientific knowledge and elaborations.

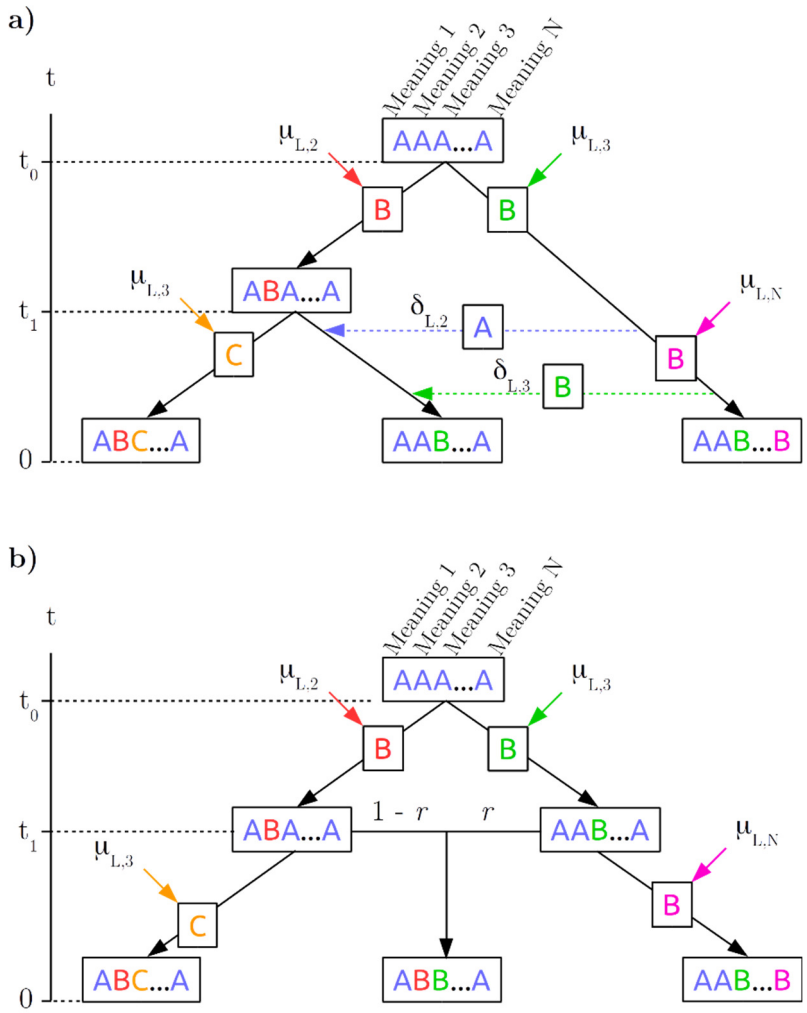


Figure F5.4.a1.
Schematic representations of mechanistic models of transmission with changes over (linguistic) time lexical cognate diversity among three language varieties. Such models are possibly implemented in the *PopLingSim* software published in Thouzeau et al. *Proceedings of the Royal Society B* (2017). Figure originally published in Supplementary Material of Thouzeau et al. (2017).

I would like to elaborate a bit further on the linguistic generation time implemented in these models. Indeed, we implicitly define a "linguistic time" that is scaled in units of lexical changes, analogously to the coalescent time in genetics which is scaled in units of genetic mutations. Thus, both "times" are self-sufficient in their circular disciplinary definition, and are never equivalent to absolute chronological time. This is an important fundamental and axiomatic issue: the genetic "time", which is the only one of explicit interest to population genetics methods, can be translated to a chronological time by a number of assumptions and approximations on the biological-generation chronological time of the species of interest, and on the mutation rate, rendered possible by a synthetic theory of molecular evolution anchored in extensive experimental observations. Here, for the "linguistic time", it might be, in principle, possible to also translate it into an absolute chronological time, but it would require to have assumptions on the linguistic generation time and on the "real" linguistic mutation rate, in a first place. To our knowledge, we do not have such assumptions yet in the lack of a synthetic theory of linguistic evolution of cognates, which probably may not even exist in reality, but this is another debate for linguists...

Nevertheless, as in genetics, it is already of major interest to compare linguistic time ratios, thus freeing ourselves from the needed assumptions to translate linguistic times in absolute chronological times. For an intuitive schematic example, it is of interest to know that, in the unspecified realm of linguistic changes over time, it took twice more time to obtain a given observed cognate diversity in two given language varieties, than the time needed to obtain another given observed cognate diversity between two other language varieties.

Based on this modelling framework, Dr. Thouzeau thus set out to conduct simulations by drawing the parameters of the models (split and admixture times, mutation and migration rates, etc.) in prior distributions. He then built a number of basic summary-statistics to describe cognate diversity such as the number of cognates and the variance of the number of cognates etc., mostly stemming from the realm of multivariate analyses in statistics, and thus not necessarily specifically built by linguists for linguistic purposes. He also considered a number of classical computational linguistic statistics specifically designed for cognates and equivalent to dissimilarity statistics in genetics.

Briefly, he ultimately designed five competing scenarios for the topological and admixture or migration history of language varieties, completely analogous to such scenarios in population genetics with the corresponding definitions of the parameters drawn, obviously, from different priors (**Figure F5.4.a2**). Based on these five scenarios and associated genetic and linguistic simulation tools, he conducted ABC inferences to determine which scenarios best explained observed data, separately for genetic and cognate diversity data observed in Central Asia, but, again, concerning the same groups of individuals from which genetic and linguistic data were both sampled originally. Without entering the details of the obtained results specific to the peopling history of Central Asia described in this article (Thouzeau et al. 2017), my bottom line here is that it worked substantially well! Indeed, Dr. Thouzeau could identify different scenarios underlying the genetic history of certain populations while, for the same populations, he found alternative scenarios best explaining the lexical cognate diversity data. In other populations, both the genetic and linguistic scenarios best explaining both data, matched. Furthermore, he could, for each data separately, formally test competing scenarios both in terms of the topological histories and/or of whether linguistic borrowings occurred in an admixture-like mechanism or in a migration-like mechanism (Thouzeau et al. 2017).

These promising results, the first to our knowledge to have formally tested alternative complex scenarios to reconstruct the genetic and linguistic histories of the same set of populations, paved the way

for our on-going project specifically focusing on Cabo Verde linguistic and genetic history, at the within-population and within-language levels.

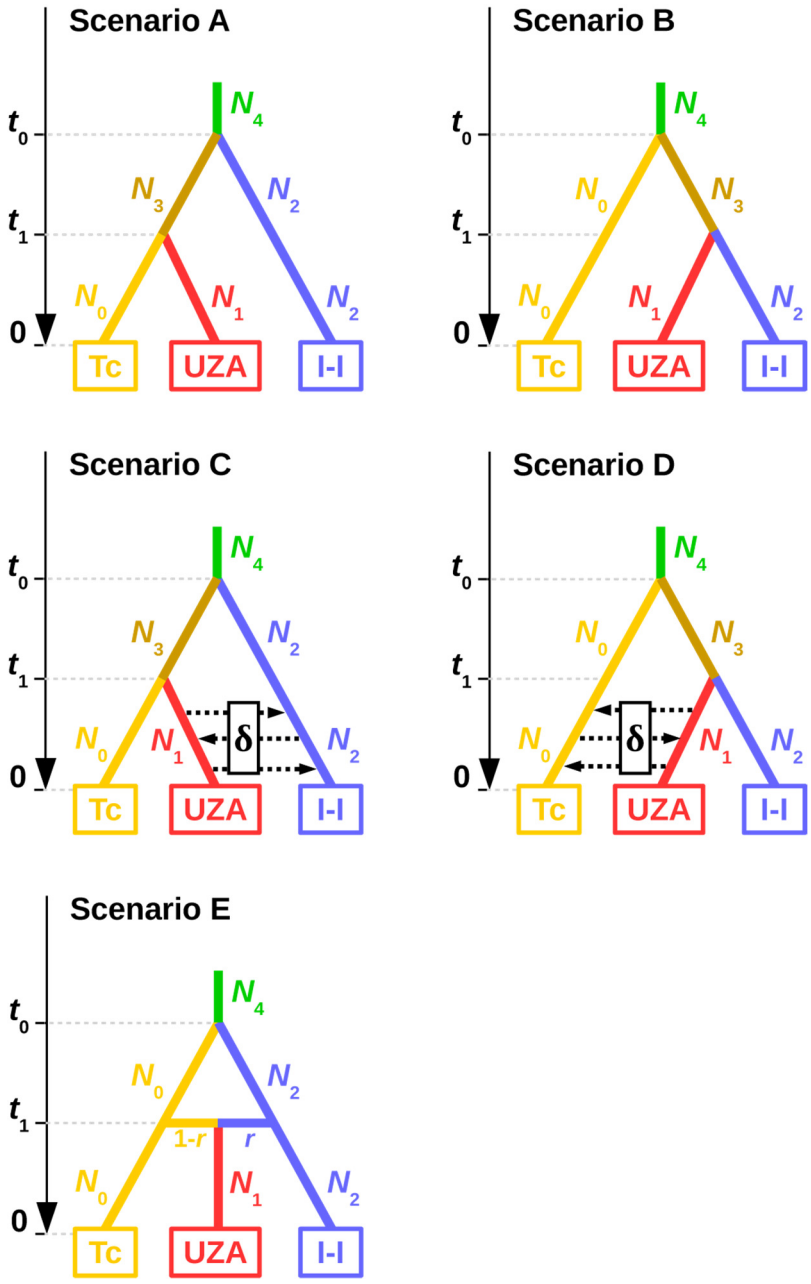


Figure F5.4.a2.

Five competing scenarios for the genetic or, respectively, linguistic history of three Central Asian populations or, respectively, language varieties, inferred using ABC using observed genetic data or, respectively, lexical cognate diversity.

Figure originally published in Figure 1 of Thouzeau et al. *Proceedings of the Royal Society B* (2017).

5.4.b. ABC inference for reconstructing jointly the genetic and linguistic history of groups of individuals

Based on this initial seminal work, we worked with Dr. Thouzeau to confront the asymmetry of classical genetic-linguistic comparisons (**Figure F1.2.c** and **Chapter 1.2**), by building a novel population linguistics paradigm that would allow to conduct historical inference jointly for genetic and linguistic data gathered at the same scale and centered on individuals. We had already built the framework allowing statistical descriptions and comparisons of genetic and linguistic data at the same scale, which we presented above in **section 5.3**, but, as mentioned, while we had built an ABC framework to conduct genetic inferences from observed data, we did not have equivalent tools for conducting ABC linguistic inferences at the scale of individuals within languages.

Involving Dr. Thouzeau in the Cabo Verde project, we could not find nor elaborate a mechanistic model describing inter-individual changes over time of utterances from semi-spontaneous discourses… We have not completely dropped the idea, but it still seems out of reach today and we are not actively pursuing this research direction. Instead, in 2016, Dr. Thouzeau proposed to me to hijack another classical linguistic method and object that I had initially rejected with Pr. Baptista (see **Chapter 1.2.c**): Swadesh word-lists (Swadesh 1971).

Indeed, we could not use the same ABC methods based on simulation of cognate diversity between language varieties that we had developed in *PopLingSim*, at the individual level within language for semi-spontaneous discourses. However, we could perhaps modify our simulation approach to consider individual word lists within language.

A first problem was that cognates are uninformative for individual utterances within language, as they stem from comparing lexemes across language varieties. Individual utterances are just realization of said varieties, and while two different utterances for a given lexeme may differ between individuals, they are most often classified as the same cognate as they extremely often stem from the same lexical item (and thus meaning and etymological root) within language. Therefore, Dr. Thouzeau decided to use, instead of cognate variation, utterance variation in the same way that we had previously investigated utterance variation across individuals' semi-spontaneous discourses (Verdu, Jewett et al. 2017; Laurent et al. 2022).

The goal was to present individuals with a list of pictures representing, hopefully as unambiguously as possible, a single meaning, and record the "word" individuals would use as a given utterance of this meaning. We would then simply consider each utterance as a fixed item, and compare "brutally" utterances for each meaning, across individuals and within the language, whichever was the source of the difference between uttered items; that is, we would count similarly any difference between utterances for a given meaning, whether the difference was morphological, phonetical, or, more drastically, lexical.

Dwelling on my prior experiences conducting word-lists linguistic collections in Central Africa, and after discussing the many known issues of such sampling with Dr. Mennecier, we built a Swadesh-like list of pictures representing colors, numbers, body parts, basic natural and anthropic objects, animals, or certain basic verbs. We then set out with Dr. Thouzeau for our 2016 fieldwork in Cabo Verde and tested this Swadesh-like protocol, in addition to the genetic anthropological and linguistic protocols we continued, with Pr. Baptista, from previous fieldwork. Remember that (**Chapter 1.2.c**), with Pr. Baptista, we had initially rejected such protocol and instead went for semi-spontaneous discourses, because, by definition of the Swadesh word-list, these words were likely very little variable across individuals within a language.

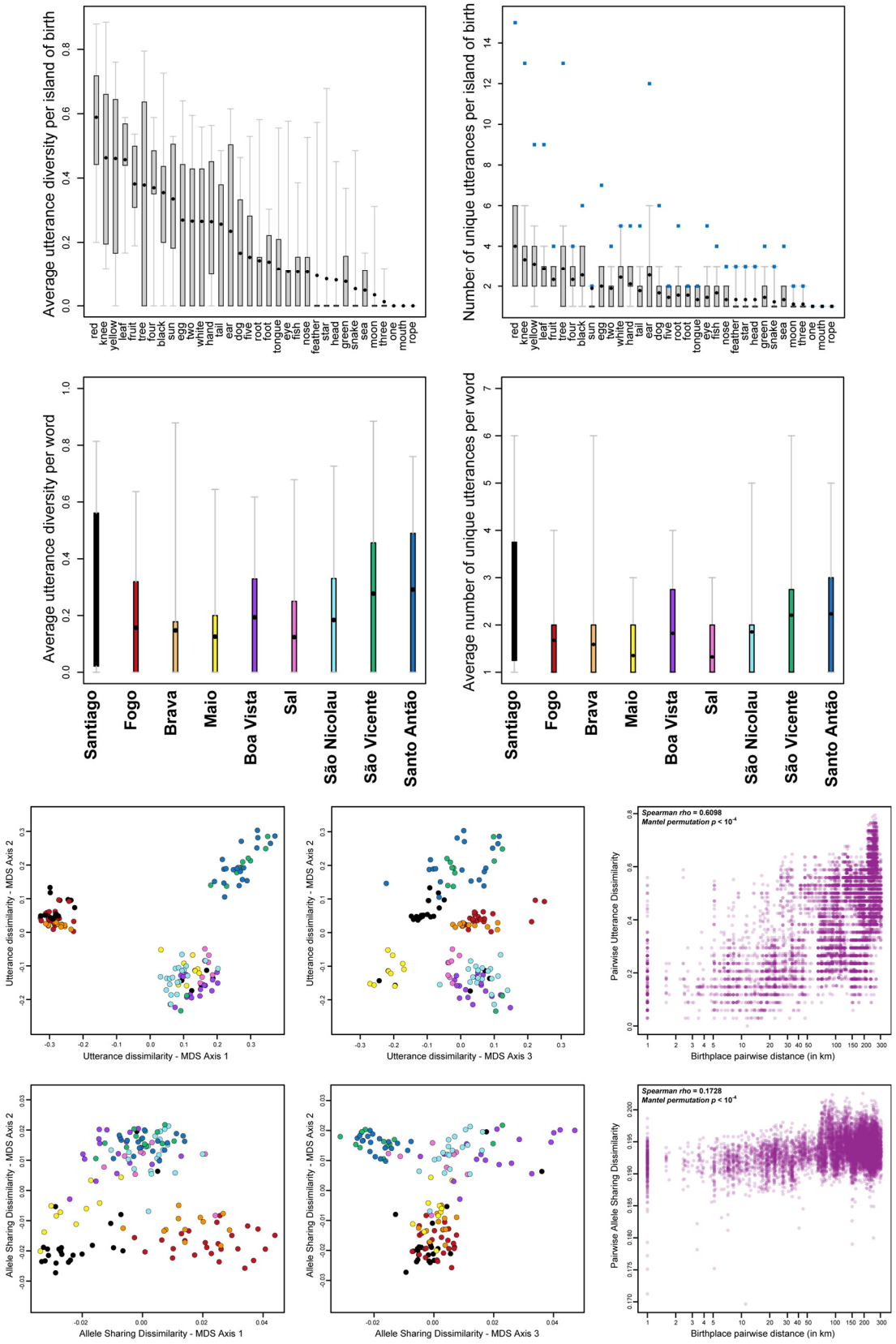


Figure F5.4.b1.

Utterance variation from word-lists in Cabo Verde (top 2 panels). Linguistic utterance variation correlates with individuals' island of birth (third horizontal panel). Genetic variation correlates with individuals' islands of birth (fourth horizontal panel), as previously seen for more SNP markers in Chapter 5.1. Thouzeau et al. *in Prep*

However, we discovered during our first recordings on the field, that there were really many ways of saying (phonetically and morphologically) the word "tree" in Kriolu, a word that came up systematically in the discourses we had recorded. Some individuals would say "arvi", "arvori", "arv'", "spinu", etc. Thus, while I did not have much hopes about Dr. Thouzeau's idea to capture a form of systematic inter-individual linguistic diversity with such word-list protocol, I reckoned that there might also be unsuspected utterance variation in other very basic and usually invariable words. The fact that the goal was, ultimately, to be able to conduct explicit simulations for linguistic ABC inference among individuals within languages, further made me believe that it was worth a try. Even if it failed, Dr. Thouzeau would have acquired field experience which would unquestionably enrich his perspectives for his PhD dissertation…

But it did not fail, and thus became an additional part of our protocol for the following fieldwork. Indeed, we had identified a set of 34 meanings that performed well with individuals, in that they would very rapidly utter the corresponding item, as little ambiguous in meaning as possible. Analyzing this data for the 147 genetically unrelated individuals born in Cabo Verde, we found tremendous amounts of highly structured variation across Cabo Verde with respect to individual birth-places, even more geographically structured than the variation observed previously with semi-spontaneous uttered discourses (**Figure F5.4.b1**).

We thus had built a novel linguistic object: inter-individual utterance variation for a given list of meanings, collected from the very same individual for which we had collected genetic and anthropological data. Dr. Thouzeau then set out to build a mechanistic model for simulating inter-individual variation of such objects within a language, implementing, similarly as he did with *PopLingSim*, specific models of changes (mutation) of utterances over time drawing parameter priors from large prior distributions (as we had no idea *a priori* what such mutation rate might be). Furthermore, he implemented such simulation software to be centered on individuals forward-in-time, continuous-in-time following a Moran (1958) model, and allowing for groups of individuals to diverge from one-another. Finally, as if it was not enough, he built a software so as to simulate jointly genetic (autosomal independent SNPs) and linguistic such data for the same individuals, where genetic population divergences would be distinct from linguistic varieties divergences, thus essentially allowing the genetic and linguistic histories to diverge, or not, for the group of individuals investigated. Using this simulator, *GeLiS*, coupled with Random Forest ABC scenario-choice and Neural Network ABC posterior parameter estimation procedures (see **Chapter 4**), we could reconstruct jointly the genetic and linguistic histories of Cabo Verde, from observed genetic and linguistic data gathered in the same individuals.

Therefore, based on the historical data compiled for the census and settlement history of Cabo Verde (**section 5.1**), we set out to reconstruct the history of founding events, associated with bottleneck changes of population sizes, and successive settlement of each island in the archipelago; the history from a genetics and linguistic point of view, separately or jointly (**Figure F5.4.b2**).

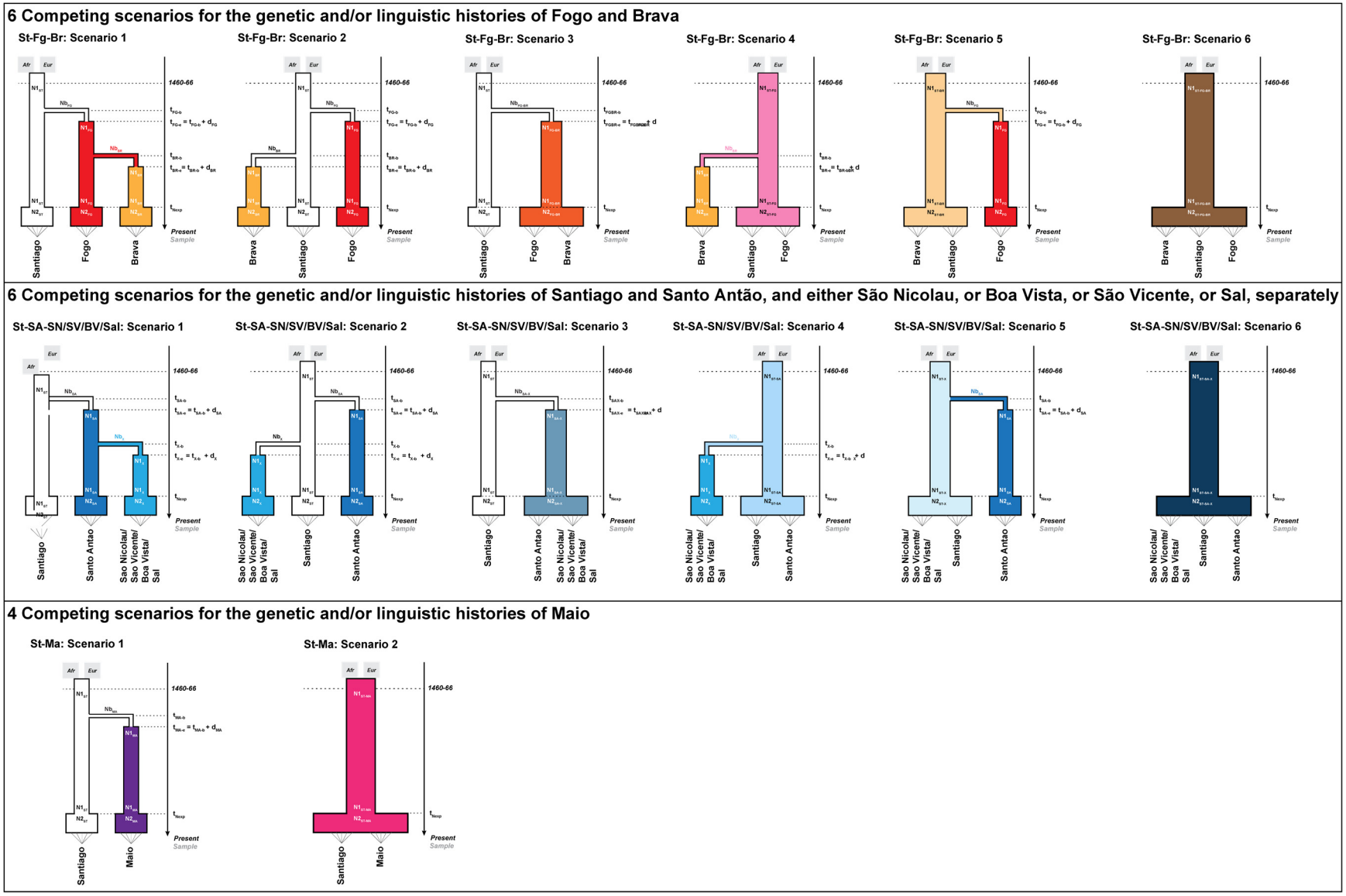


Figure F5.4.b2.

Competing scenarios for the genetic and linguistic peopling history of each Cabo Verdean island, based on prior historical information, simulated with GeLiS and inferred using Random Forest ABC scenario-choice and Neural Network ABC posterior parameter inference, from observed individual-centered genetic and linguistic data.

Thouzeau et al. *in Prep*

We are in the process of writing the article corresponding to this extensive and exhaustive work. As this dissertation is already long, I will thus only rapidly tease some of our not-yet peer-reviewed results, and leave the reader to follow up on this work, hopefully published in the next few months, or year.

In practice, we found out that it worked extremely satisfactorily. We could disentangle the different possible successive settlements of each island, and found, in all cases, but that of Boa Vista, congruencies between the linguistic and genetic histories of the founding peopling of each Cabo Verde islands: the people who historically gave birth to the gene-pool of an island also gave birth to its linguistic utterance diversity. The exception of Boa Vista showed, instead, that it's genetic peopling came from a different island than its linguistic peopling. Furthermore, we found that linguistic utterances' diversity derived more rapidly than genetic diversity everywhere in Cabo Verde: while manners of speaking Kriolu differentiated rapidly after the establishment of a novel settlement on an island, reproductive isolation among islands took more time. In other words, linguistic differentiation predated genetic differentiation everywhere in Cabo Verde

Therefore, and to conclude here, this work sets the bases for a novel population linguistics paradigm analogous to the classical population genetics paradigm, anchored on inter-individual genetic and linguistic variation from the same sets of individuals (**Figure F5.4.b3**). This framework allows us, for the first time to our knowledge, to formally infer the genetic and linguistic joint or separated genetic and linguistic admixture histories, and will allow, in the future, to further disentangle the fundamental influence of socio-cultural behaviors on the genetic and linguistic histories of complex admixture processes having given birth to human biological and cultural variation observed today.

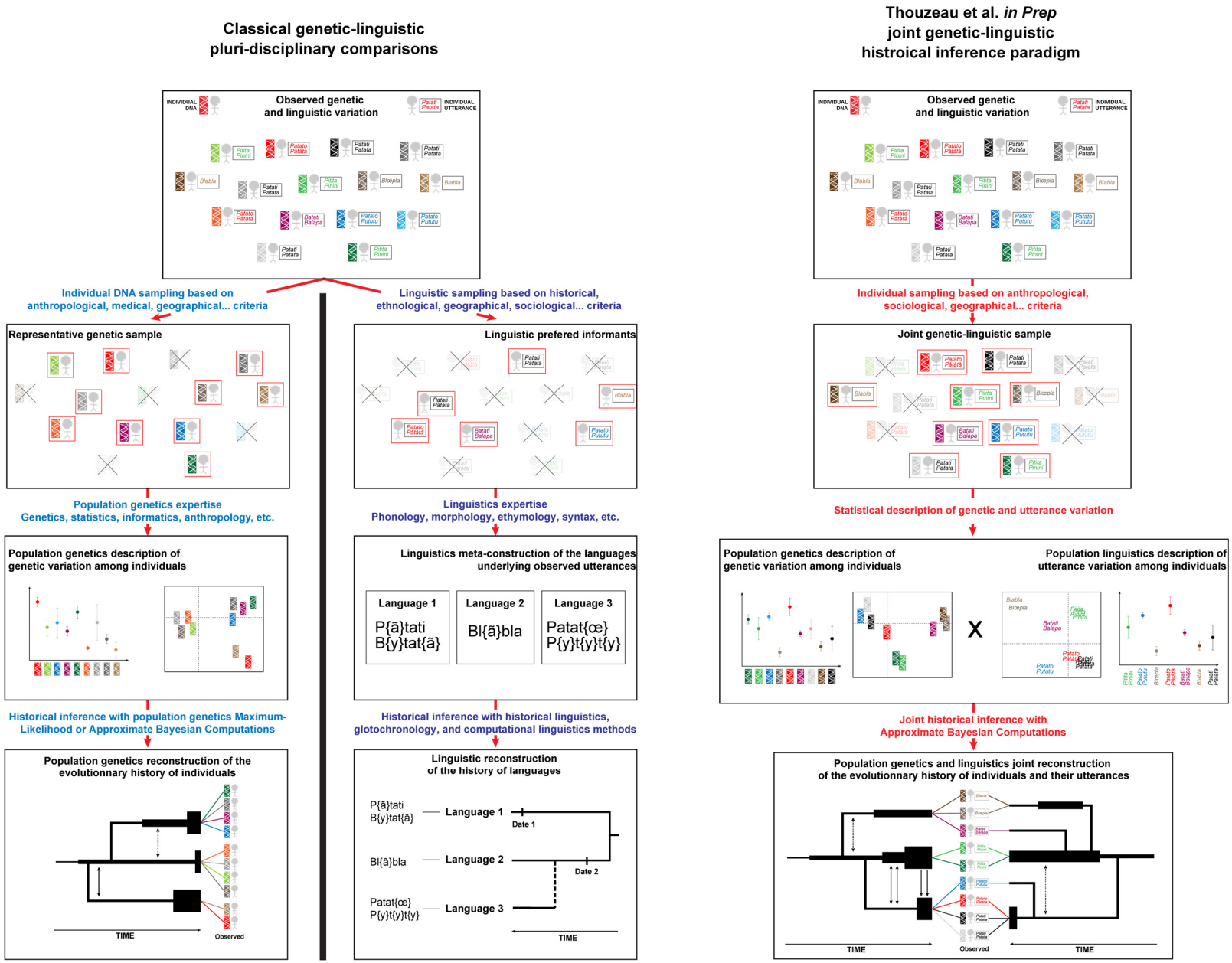


Figure F5.4.b3.

A novel paradigm for joint genetic-linguistic historical inference from inter-individual genetic and linguistic data.
Thouzeau et al. *in Prep*

From left to right: Sergio da Costa, Valentin Thouzeau, and Ethan Jewett (back) during word-list utterance sampling in the village of Paul Santo Antão, Cabo Verde, 2016
©Paul Verdu

Conclusion

Histories of Admixture

Conclusion. Histories of Admixture

I hope to have illustrated here how the past 20 years of research in anthropological genetics and human population genetics, to which I have tried to contribute, significantly elaborated on the classical synthetic theory of evolutionary genetics presented schematically in introduction (see updated figure below).

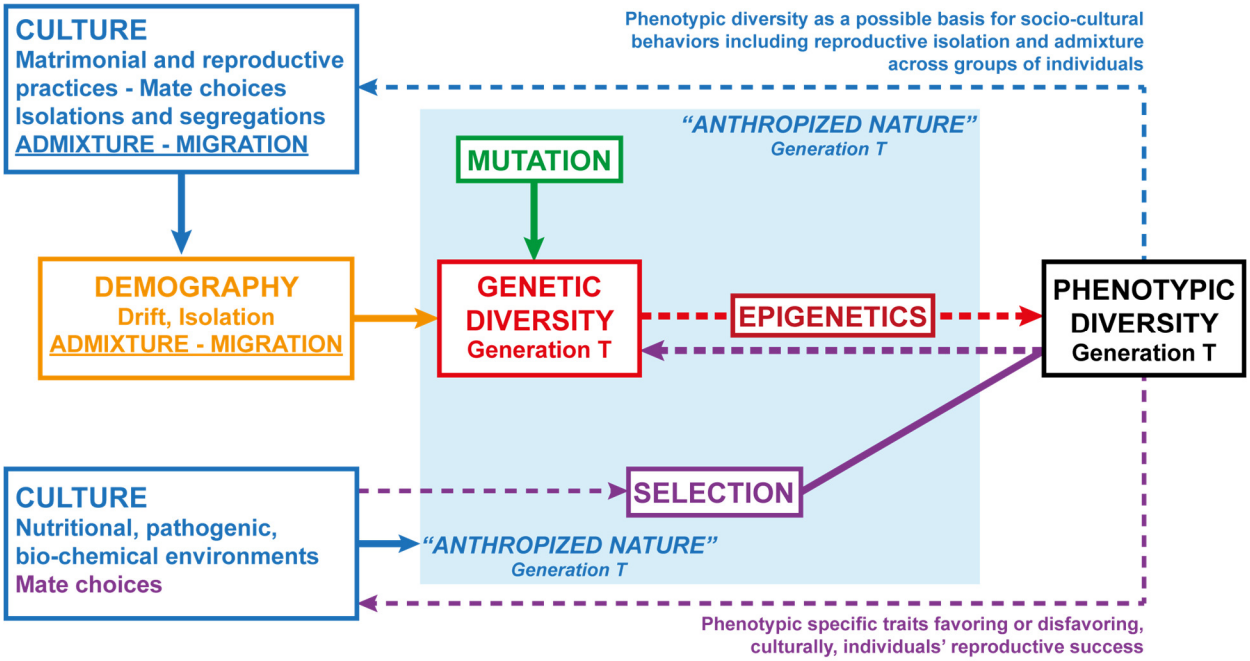


Figure Conclusion

Schematic representation of the influence of evolutionary forces (mutation, demography, selection) on genetic diversity and partially associated phenotypic diversity, in the updated neutral, neo-Darwinian, synthetic theory of evolution. While formally productive in Human Sciences disciplines, the Nature-Nurture, or Nature-Culture, separation is largely un-operational from a population genetics evolutionary perspective: "Nature" and "Culture" interact constantly in a dynamic way to influence genetic diversity patterns observed within and among populations.

In particular, I hope that I have shown how complex admixture processes have shaped and continue to shape rapidly the genetic diversity and evolution of our species, heavily determined by complex socio-cultural behaviors. In that, our work and that of numerous other groups, in my eyes, highlight that Culture can be a massive force influencing rapidly the biological diversity and evolution of our species, through genetic admixture and migration processes in particular. Based on the examples provided above and numerous others in the anthropological genetics' literature, I strongly think that the Nature-Culture schism in anthropology, mainly stemming from cultural anthropology disciplines but also often fully adopted by numerous human biologists, is profoundly un-operational for investigating human genetic diversity and evolution. Indeed, we showed how genetic diversity and evolution was profoundly shaped by cultural processes, themselves often intertwined with complex relationships with heavily anthropized ecological environments and, sometimes, rooted in genetically-determined, in part, phenotypic variation.

Furthermore, as numerous studies start to investigate ancient admixture events across *Homo sapiens* populations and with other, now extinct, hominid species, it is clear now that admixture has been pervasive

throughout *Homo sapiens* history. While we cannot directly access the socio-cultural mechanisms at play in a remote past, it is nevertheless consistent to consider, as a starting point, that socio-cultural mechanisms were indeed at play in the past to determine admixture and migration processes. Indeed, considering that our species is a social species, and that it has always been a social species since its emergence, there is no reason, in the light of the fundamental, recent or more ancient, mechanisms highlighted in this dissertation, to think that equivalent mechanisms were not also at play in very ancient times.

More generally, I thus hope to have participated in developing novel ways to slightly shift the classical paradigm for admixture and migration investigation in population genetics; from a history of isolation where admixture and migrations are secondary, often nuisance, mechanisms, to a framework where admixture and migration events can be constitutive of our diversity and evolution, separated by periods of reproductive isolation. In other words, instead of imagining human populations as being isolated and sometimes admixed, considering them as related by dynamic admixture and migration processes and, sometimes, isolated, can produce novel insights in both the cultural and biological evolution of our species. Future work investigating more or less ancient human DNA will allow us, for the first time, to literally travel back in our evolutionary history, and further decipher the evolutionary forces, including possible cultural mechanisms, having led to the genetic landscape observed in human populations today.

To further our understanding of the influence of cultural behavior in shaping our evolution, I think we need, collectively, to overcome a fundamental challenge. Population geneticists can investigate mechanistically the evolution of genetic diversity, thanks to a unified robust model of inter-generational transmission of genetic material with mutation. Of course, numerous fundamental genetics disciplines still extensively work on deciphering how mutations arise, with interesting advances from quantum genetics lately, as well as on how DNA is indeed transmitted to offspring. These may or may not lead to fundamentally revising the synthetic theory of evolution in the future. For now, this theory remains our best paradigm and hypotheses-testing dialectics to build novel knowledge and understanding of our biological evolution. In this context, it is crucial to understand that there is no equivalent of such synthetic theory for exploring the transmission of culture with possible changes over time. This is practically due to the fact that mendelian inheritance is, in the end, far much simpler than cultural transmission for virtually any cultural trait of interest.

More conceptually, genetic diversity stands only on observed inter-individual genetic variation; the object of population genetics is thus solely the marginal product of what is observed at the inter-individual level, and mutation, demographic, and selection forces act on the living individuals and the genetic diversity they carry by determining its transmission, or not, to the next generation. Conversely, whether cultural variation is born from the marginal product of individual behaviors in different groups in interactions or, alternatively, whether cultural variation answers to its own dynamics from which individuals and groups of individuals may realize certain differentiated behaviors, remains largely unclear. Therefore, while I use cultural variation as facts observed at a given point in time and influencing genetic diversity patterns, I cannot predict its mechanistic influence on biological evolution over time, as I have no formal conception of how it may change over time. One ultimate goal would thus be to determine the mechanisms, and their parameters, underlying cultural transmission, changes, and innovations, to better integrate cultural and genetic anthropology, at least to the benefit of the latter discipline.

Personally, without proof and based on my direct experiences investigating cultural variation, I have the intuition that cultural variation answers to its own mechanisms of transmission with or without changes over time and space, and that individual cultural behaviors stem from possibilities available within the

frames they have been taught by social interactions through life (and which include the transgression of the frame). I thus do not think, but again may be wrong, that cultural variation is the marginal product of individual behavioral variation. From there, it would be necessary to build anew a complete mechanistic theory of cultural evolution answering to its own forces and mechanisms, without ever even trying to hijack some aspects of the synthetic theory of genetic evolution. This may seem obvious: we all intuitively know that utterances, for instance, are not transmitted in a diploid mendelian way. Nevertheless, the task I propose is immense, in a large part due to the fact that such questions are of interest to me, but are not always the core focus of interest of the vast majority of researchers in numerous cultural anthropology fields. Advancing in this direction would thus require, from my part, a renewed intense inter-disciplinary effort, as it is not honest to simply ask cultural anthropologist to handle my problems. In all cases, this would likely produce novel objects of study, similarly to our novel population linguistics objects for historical linguistic inference, which may or may not be of interest to other researchers from diverse fields.

Altogether, as illustrated by the numerous experienced and proposed pluri-disciplinary and inter-disciplinary challenges detailed throughout this dissertation, I believe that there is a vast need, today, to take the time for mediating the scientific paradigms, methods, and dialectics from one discipline to the other. It is to some extent obvious for vastly differing disciplines, such as cultural anthropology and population genetics here illustrated, but it is also the case, I think, for more closely related disciplines. Indeed, the "phylogenetic" proximity between evolutionary anthropologists and anthropological geneticists often leads to an apparently common specialized lexicon, which, in fact, fundamentally differs. For instance, drift or population sizes are vastly differing concepts between computational linguistics, archaeology, ecology, population dynamics, and population genetics. Thinking they are actually "roughly similar" leads, to my views, to major confusions, misunderstandings, and sterile disagreements. Another example is the importance of explicating the scientific categories used from different disciplinary perspectives, as detailed in the first chapter of this dissertation: "***what is a population?***" is probably the major unifying question that needs to be at least thought about prior to starting any pluri-disciplinary research across evolutionary anthropology disciplines. I reckon that taking the time for such essential, but often not efficient, dialogue is costly. I nevertheless think that it is a necessity, unless one wants to take a tremendous risk to mutilate others' disciplines leading to false interpretations of one's disciplinary results.

Finally, the diffusion and mediation of the scientific methods to the broader, non-specialized and non-academic, audience is equally crucially necessary to my views. Since its discovery and the rapid advances to our understanding of biology that it brought, DNA and genetics has gained incredible traction in our societies and in our daily life. Today, molecular genetics and population genetics tools are extensively deployed in practice for judicial and medical purposes, and even more recently for "recreational" purposes (as explained in **Chapter 1.4**), *de facto* hiding still tremendous amounts of scientific disagreement and debate. DNA is often seen, in my numerous repeated experiences of interactions with journalists, tribunals, policy makers, and broad-audience public debates and conferences, as the ultimate grail, endorsed with mathematical certainty and objectiveness, while I have shown that, in reality, neither qualificatives truly apply in the probabilistic realm of genetics. I think that scientists in general, and geneticists in particular, should spend more time explaining how they obtain their results rather than spending most of their "on air" time highlighting the results themselves, without ever trying to detail the how? and the from where? I have very often been opposed by scientists that "no one cares about these details; people are interested in the results". I think that this is purely and simply wrong: it is not because mediating the scientific method is

much more difficult than simply communicating the obtained, grandiose, results, that people are interested in the latter and not the former. I think that people are interested in stories, and the process underlying scientific research is a story, a fascinating and exciting one, both to tell and to listen to.

Remerciements

Je tiens à remercier ici, vivement et très sincèrement, tous les participants à mes projets de recherche au Cameroun, au Gabon, en Ouganda, et au Cap Vert. Ils m'ont accordé patiemment et avec bienveillance leur temps, afin de partager avec moi des bouts de leur vie, de leur langue, et me fournir des petits échantillons de leur salive et autres mesures anthropométriques. Sans eux, je n'aurais pas pu mener les différents projets de recherche présentés ici. Merci aussi à tous ceux qui m'ont aidé sur le terrain, comme assistants, traducteurs, et/ou comme collaborateurs, et notamment Alain Fezeu, Omarou Mendjok, Nyantzi Yosam, Mugisa Ezekiel, Walina Zephanus, et Angelo « Djiño » Barbosa.

Je tiens également à remercier les étudiants de tous les niveaux avec qui j'ai eu la chance de pouvoir travailler, même brièvement, et celles et ceux que j'ai pu encadrer, ou co-encadrer, au cours de mes recherches. Vous m'avez toutes et tous très souvent bien plus appris que ce que j'ai pu vous enseigner. Merci donc (dans l'ordre chronologique de nos premières rencontres) à Héloïse Bastide, Viola Grugni, Noémie Becker, Amy Goldberg, Margeritte Lapierre, Mirian Barbosa, Claire Stragier, Antoine Cools, Alexandra Blant, Ferdinand Petit, François Mallordy, Marie Lopez, Cesar Fortes-Lima et Kazunari Matsudaira. Une mention toute spéciale pour les thésards et thésardes (du passé et du présent), Valentin Thouzeau, Gwenna Breton, Marta Ciccarella, et Juliette Sauvage : merci pour votre patience, votre enthousiasme communicatif, et la richesse de vos regards, tout comme pour m'avoir fait l'honneur, parfois, de partager avec moi vos doutes, scientifiques ou non, vos états d'âmes et vos interrogation sur l'avenir académique ou non ; dans l'espoir que nous continuions les collaborations scientifiques tous azimuts !

Un grand, un immense MERCI à mes compagnons de science, et en tout premier lieu à Marie-France Mifune, Romain Laurent, Zachary Szpiech, Bruno Toupance, Samuel Pavard, José Utgé, Erkan Buzbas et Etienne Patin, pour votre indéfectible soutien et intérêt, parfois salutairement critique, pour mes travaux et mes lubies scientifiques et, je l'espère, pour votre amitié. Sans vous, aucun de mes travaux, même ceux en solitaire, n'auraient été menés à bien. Et merci, bien sur à tous les autres compagnons, peut-être moins en première ligne, mais pourtant bien tous présents ; vous qui m'avez forgé tant scientifiquement qu'humainement au fil de ces années : Franz Manni, Laure Ségurel, Ethan Jewett, Guillaume Laval, Raphaël Leblois, Cécile Garcia, Victor Narat, et Céline Bon.

Merci aux rapporteurs de ce documents et aux membres du jury d'avoir accepté ce travail, long (… désolé), de discussion de mes travaux de recherche.

Merci à Philippe Chambon et à Aline Thomas, pour leur convivialité, leurs échanges scientifiques avec moi sur le monde magique de l'archéologie, et aussi un peu pour avoir aimablement lu et corrigé un morceau de ce document, la partie de perspective sur les catégorisations archéologiques.

Merci à Sylvie Le Bomin, Serge Bahuchet, et Marie-France Mifune pour m'avoir patiemment formé sur le terrain à la méthode ethnographique, et avoir fait de mes campagnes d'échantillonnage, des succès.

Merci à Marlyse Baptista, pour m'avoir proposé de travailler au Cap Vert, pour m'avoir fait découvrir ce pays et ses habitants ; et merci à elle et à Sergio Costa, d'avoir partagé leurs savoirs et savoirs faire en linguistique.

Merci à Myriam Georges, Begoña Martinez-Cruz et Hélène Quach, de m'avoir appris la génétique moléculaire et de m'avoir toujours épaulé dans mes projets.

Merci au Plateau de Paleogénomique et Génétique Moléculaire (P2GM) du MNHN sur le site du Musée de l'Homme, et en particulier à Françoise Dessarps-Freichey et José Utgé, pour avoir permis la genèse de très nombreuses données génétiques exploitées dans le Chapitre 5 de ce document.

Merci à Arnaud Estoup, d'avoir pris le temps de m'enseigner l'ABC, et d'avoir pris le temps de continuer !

Merci à Céline Bon, José Utgé, Sophie Lafosse, Françoise Dessarps-Freichey, et Amélie Chimènes pour l'incroyable aventure P2GM que vous m'avez offert et qu'on continue à partager ; et merci d'avoir absorbé toutes les tâches et les soucis quotidiens m'ayant permis de trouver le temps d'écrire ce document ! Et en route pour de nouvelles aventures !

Merci à Taouès Lahrem et à Florence Loiseau, pour avoir encadré administrativement l'immense majorité de mes projets. Sans vous, en pratique, aucun projet de recherche ou d'encadrement n'aurait pu être mis en œuvre. Et surtout, merci de m'avoir tant aidé, avec le sourire !

Merci à Emilie Masson, Estelle Bervas-Clerc, et Iris Boh, pour nos collaborations et vos enseignements et réflexions juridiques et déontologiques sur mes projets et au-delà.

Merci au vaste peuple du Musée de l'Homme, pour m'avoir accueilli et soutenu professionnellement et pour sans cesse avoir partagé vos nombreux savoirs avec moi en cette formidable maison de science, et ce depuis 2002. Je ne peux pas tous vous citer ici, mais vous savez pouvoir vous attribuer une partie, méritée, des travaux rapportés ici.

Et merci à toutes celles et ceux qui m'ont donné à penser, et souvent plus, si généreusement au cours du temps : Frank Alvarez-Pereyre, Alain Froment, Alain Epelboin, Barry Hewlett, Bonnie Hewlett, Fernando Ramirez-Rossi, Renaud Vitalis, Marie-Claude Marsollier-Kergoat, Nathalie Machon, Mattias Jakobsson, Carina Schlebusch, et Jorge Rocha.

Merci à tous mes anciens encadrants ayant accompagnés mes premiers pas en recherche scientifique : Kenneth Kidd, Judith Kidd, et Luis Barreiro, je n'aurais surement pas continué sans votre aide et votre exemplarité.

Et merci, merci, merci à mes mentors, officiels ou non, mais à qui je dois tout scientifiquement et académiquement. Dans l'ordre chronologique donc : Lluis Quintana-Murci, le premier qui a cru que c'était possible, confiance sans cesse renouvelée depuis, et qui continuera à l'être, je l'espère, pour le futur. Frédéric Austerlitz, qui m'a littéralement forgé à la méthode scientifique et à la recherche en génétique des populations, qui m'a soutenu, guidé, et patiemment relu au cours de ces nombreuses années, sans oublier ses enseignements essentiels à la « convivialité » en milieu académique. Serge Bahuchet, qui a eu la patience de m'enseigner l'anthropologie en générale et celle de l'Afrique Centrale en particulier, de la théorie au terrain, pour m'avoir introduit dans le monde étrange des « Pygmologues » et, bien au-delà, pour

toutes les discussions scientifiques pluridisciplinaires que nous avons eues et qui m'ont tant marquée. Noah Rosenberg, pour sa patience infinie, sa science, et sa méthode, qui m'a guidé sur le chemin de la théorie en génétique des populations, qui m'a soutenu et épaulé pour l'obtention de mon poste académique, et qui continue à être un collaborateur et un relecteur essentiel à mes travaux de recherche.

Et, bien évidemment et avant toutes et tous, merci infiniment à Evelyne Heyer, ma directrice de thèse, ma collaboratrice, et ma directrice d'unité. Merci pour la science, pour toutes les opportunités incroyables que tu m'as offertes sur un plateau toutes ses années, pour ton soutien actif et inébranlable, pour ton écoute de mes (trop) nombreux états d'âmes et autres coups de gueule, pour les encouragements, pour l'accueil dans ton équipe puis dans ton laboratoire et pour ton travail immense, souvent invisible, à m'avoir fourni le cadre idéal de réalisation de mes travaux passés, présents et futurs. Merci donc, chère Evelyne.

Merci enfin à toutes celles et ceux que j'aurais bien malencontreusement oublié ici.

Plus personnellement, merci à ma famille : mes parents, Jean-Paul et Geneviève, pour m'avoir toujours soutenu et encouragé dans mon parcours, et ma sœur, Caroline, pour son enthousiasme et son optimisme sans faille. Merci aussi à tous mes amis et amies, vous saurez vous reconnaitre. Vous savez que ce travail n'existerait pas sans vous, vos conseils, vos soutiens, et vos critiques, scientifiques ou autres. Merci à #payetonconfinement, groupe ultra-moderne, réseau-socialisé mais restreint, le seul dont je fais partie en fait… qui se reconnaitra et qui m'a soutenu tout au long de cette écriture ardue.

Et merci à Marie-France, à qui se travail est dédié, et grâce à qui, je le maintiens, tout est possible.

References

References

Albuquerque, Luís de, and Maria Emília Madeira Santos. 1991. *História geral de Cabo Verde*. 3 vols. Lisboa; Praia [Cape Verde]: Lisboa: Centro de Estudos de História e Cartografia Antiga, Instituto de Investigação Científica Tropical.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Allentoft, Morten E., Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, et al. 2015. "Population Genomics of Bronze Age Eurasia." *Nature* 522 (7555): 167–72. https://doi.org/10.1038/nature14507.

Alvarez-Pereyre, Frank. 2003. *L'exigence Interdisciplinaire: Une Pédagogie de l'interdisciplinarité En Linguistique, Ethnologie et Ethnomusicologie*. Paris: Maison des sciences de l'homme.

Baharian, Soheil, Maxime Barakatt, Christopher R. Gignoux, Suyash Shringarpure, Jacob Errington, William J. Blot, Carlos D. Bustamante, et al. 2016. "The Great Migration and African-American Genomic Diversity." Edited by Greg Gibson. *PLOS Genetics* 12 (5): e1006059. https://doi.org/10.1371/journal.pgen.1006059.

Bahuchet, Serge. 1992a. *Dans la forêt d'Afrique Centrale: les pygmées Aka et Baka*. Histoire d'une civilisation forestière 1. Paris: Peeters-Selaf.

———. 1992b. "Spatial Mobility and Access to Resources among the African Pygmies." In *Mobility and Territoriality: Social and Spatial Boundaries among Foragers, Fishers, Pastoralists, and Peripatetics*, 1st ed.

———. 1993. "L'invention Des Pygmées (Inventing Pygmies)." *Cahiers d'Études Africaines* 33 (129): 153–81. http://www.jstor.org/stable/4392434.

———. 2012. "Changing Language, Remaining Pygmy." *Human Biology* 84 (1): 11–43. http://www.jstor.org/stable/41466784.

Bahuchet, Serge, Henri Guillaume, and S.M. Van Wyck. 1982. "Aka-Farmer Relations in the Northwest Congo Basin." In *Politics and History in Band Societies*, 189–211. Cambridge (GB) ; Paris: Cambridge University Press ; Maison des Sciences de l'Homme. https://www.documentation.ird.fr/hor/fdi:16055.

Baird, S J E. 2006. "Fisher's Markers of Admixture." *Heredity* 97 (2): 81–83. https://doi.org/10.1038/sj.hdy.6800850.

Baptista, Marlyse. 2002. *The Syntax of Cape Verdean Creole: The Sotavento Varieties*. Linguistik Aktuell = Linguistics Today, v. 54. Amsterdam ; Philadelphia: John Benjamins Pub.

———. 2015. "Continuum and Variation in Creoles: Out of Many Voices, One Language." *Journal of Pidgin and Creole Languages* 30 (2): 225–64. https://doi.org/10.1075/jpcl.30.2.02bap.

Barth, F. 1969. *Ethnic Groups and Boundaries: The Social Organization of Culture Difference*. Scandinavian University Books. Universitetsforlaget. https://books.google.fr/books?id=Cza0AAAAIAAJ.

Batini, C., J. Lopes, D. M. Behar, F. Calafell, L. B. Jorde, L. van der Veen, L. Quintana-Murci, G. Spedini, G. Destro-Bisol, and D. Comas. 2011. "Insights into the Demographic History of African Pygmies

from Complete Mitochondrial Genomes." *Molecular Biology and Evolution* 28 (2): 1099–1110. https://doi.org/10.1093/molbev/msq294.

Beaumont, Mark A, Wenyang Zhang, and David J Balding. 2002. "Approximate Bayesian Computation in Population Genetics." *Genetics* 162 (4): 2025–35. https://doi.org/10.1093/genetics/162.4.2025.

Becker, Noémie S. A., Paul Verdu, Barry Hewlett, and Samuel Pavard. 2010. "Can Life History Trade-Offs Explain the Evolution of Short Stature in Human Pygmies? A Response to Migliano et al. (2007)." *Human Biology* 82 (1): 17–27. https://doi.org/10.3378/027.082.0101.

Becker, Noémie S.A., Paul Verdu, Alain Froment, Sylvie Le Bomin, Hélène Pagezy, Serge Bahuchet, and Evelyne Heyer. 2011. "Indirect Evidence for the Genetic Determination of Short Stature in African Pygmies." *American Journal of Physical Anthropology* 145 (3): 390–401. https://doi.org/10.1002/ajpa.21512.

Beleza, Sandra, Nicholas A. Johnson, Sophie I. Candille, Devin M. Absher, Marc A. Coram, Jailson Lopes, Joana Campos, et al. 2013. "Genetic Architecture of Skin and Eye Color in an African-European Admixed Population." *PLoS Genetics* 9 (3): e1003372. https://doi.org/10.1371/journal.pgen.1003372.

Bergström, Anders, Chris Stringer, Mateja Hajdinjak, Eleanor M. L. Scerri, and Pontus Skoglund. 2021. "Origins of Modern Human Ancestry." *Nature* 590 (7845): 229–37. https://doi.org/10.1038/s41586-021-03244-5.

Berlin, Ira. 1998. *Many Thousands Gone: The First Two Centuries of Slavery in North America*. Cambridge, Mass: Belknap Press of Harvard University Press.

Berlin, Ira. 2010. *The making of African America : the four great migrations*. New York: Penguin Books.

Bernstein, Felix, , Comitato italiano per lo studio dei problemi della popolazione.,. 1931. *Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung*. Roma: Istituto poligrafico dello stato.

Bertorelle, G., and L. Excoffier. 1998. "Inferring Admixture Proportions from Molecular Data." *Molecular Biology and Evolution* 15 (10): 1298–1311. https://doi.org/10.1093/oxfordjournals.molbev.a025858.

Blum, Michael G B, Evelyne Heyer, Olivier François, and Fréderic Austerlitz. 2006. "Matrilineal Fertility Inheritance Detected in Hunter–Gatherer Populations Using the Imbalance of Gene Genealogies." *PLoS Genetics* 2 (8): e122.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd, and L. L. Cavalli-Sforza. 1994. "High Resolution of Human Evolutionary Trees with Polymorphic Microsatellites." *Nature* 368 (6470): 455–57. https://doi.org/10.1038/368455a0.

Breiman, Leo. 2001. "Random Forest." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Breton, Gwenna, Per Sjödin, Panagiotis Zervakis, Romain Laurent, Alain Froment, Agnès E. Sjöstrand, Barry S. Hewlett, et al. 2020. "Deciphering Early Human History Using Approximate Bayesian Computation and 74 Whole Genomes from Central and Southern Africa." University of Uppsala: Palaeo-Research Institute, University of Johannesburg, P.O. Box 524, Auckland Park, 2006, South Africa. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1455514&dswid=-2911.

Buzbas, Erkan Ozge, and Paul Verdu. 2018. "Inference on Admixture Fractions in a Mechanistic Model of

Recurrent Admixture." *Theoretical Population Biology* 122 (July): 149–57. https://doi.org/10.1016/j.tpb.2018.03.006.

Callaway, Ewen. 2018. "Divided by DNA: The Uneasy Relationship between Archaeology and Ancient Genomics." *Nature* 555 (7698): 573–76. https://doi.org/10.1038/d41586-018-03773-6.

Carreira, António. 2000. *Cabo Verde : formação e extinção de uma sociedade escravocrata (1460-1878).* Praia: IPC.

Cavalli-Sforza, Luigi Luca. 1986. *African Pygmies*. Academic Press.

———. 1997. "Genes, Peoples, and Languages." *Proceedings of the National Academy of Sciences* 94 (15): 7719–24. https://doi.org/10.1073/pnas.94.15.7719.

———. 2001. *Genes, Peoples and Languages*. London: Penguin.

Cavalli-Sforza, L L, E Minch, and J L Mountain. 1992. "Coevolution of Genes and Languages Revisited." *Proceedings of the National Academy of Sciences* 89 (12): 5620–24. https://doi.org/10.1073/pnas.89.12.5620.

Cavalli-Sforza, L L, and Marcus W. Feldman. 2003. "The Application of Molecular Genetic Approaches to the Study of Human Evolution." *Nature Genetics* 33 (S3): 266–75. https://doi.org/10.1038/ng1113.

Cavalli-Sforza, L L, and Walter Fred Bodmer. 1971. *The Genetics of Human Populations*. Courier Corporation.

Chacón-Duque, Juan-Camilo, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuña-Alonzo, Rodrigo Barquera, Mirsha Quinto-Sánchez, et al. 2018. "Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance." *Nature Communications* 9 (1): 5388. https://doi.org/10.1038/s41467-018-07748-z.

Chafe, Wallace L., ed. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Advances in Discourse Processes, v. 3. Norwood, N.J: Ablex Pub. Corp.

Chakraborty, Ranajit, and Kenneth M Weiss. 1988. "Admixture as a Tool for Finding Linked Genes and Detecting That Difference from Allelic Association between Loci." *Proceedings of the National Academy of Sciences* 85 (23): 9119–23.

Chaudenson, Robert, and Salikoko S. Mufwene. 2001. *Creolization of Language and Culture*. London ; New York: Routledge.

Choin, Jeremy, Javier Mendoza-Revilla, Lara R. Arauna, Sebastian Cuadros-Espinoza, Olivier Cassar, Maximilian Larena, Albert Min-Shan Ko, et al. 2021. "Genomic Insights into Population History and Biological Adaptation in Oceania." *Nature* 592 (7855): 583–89. https://doi.org/10.1038/s41586-021-03236-5.

Cornuet, Jean-Marie, Filipe Santos, Mark A. Beaumont, Christian P. Robert, Jean-Michel Marin, David J. Balding, Thomas Guillemaud, and Arnaud Estoup. 2008. "Inferring Population History with DIY ABC: A User-Friendly Approach to Approximate Bayesian Computation." *Bioinformatics* 24 (23): 2713–19. https://doi.org/10.1093/bioinformatics/btn514.

Creanza, Nicole, Merritt Ruhlen, Trevor J. Pemberton, Noah A. Rosenberg, Marcus W. Feldman, and Sohini Ramachandran. 2015. "A Comparison of Worldwide Phonemic and Genetic Variation in Human Populations." *Proceedings of the National Academy of Sciences* 112 (5): 1265–72. https://doi.org/10.1073/pnas.1424033112.

Croft, William. 1991. “Linguistic Selection: An Utterance-Based Evolutionary Theory of Language Change.” *Nordic Journal of Linguistics* 19 (2): 99–139. https://doi.org/10.1017/S0332586500003358.

Csilléry, Katalin, Olivier François, and Michael G. B. Blum. 2012. “Abc: An R Package for Approximate Bayesian Computation (ABC): *R Package: Abc*.” *Methods in Ecology and Evolution* 3 (3): 475–79. https://doi.org/10.1111/j.2041-210X.2011.00179.x.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Darwin, Charles, 1809-1882. n.d. *The Descent of Man, and Selection in Relation to Sex. By Charles Darwin ... In Two Volumes...With Illustrations.* London: London: John Murray, 1871.

Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. 2012. “A Linear Complexity Phasing Method for Thousands of Genomes.” *Nature Methods* 9 (2): 179–81. https://doi.org/10.1038/nmeth.1785.

Demenou, Boris B., Jean-Louis Doucet, and Olivier J. Hardy. 2018. “History of the Fragmentation of the African Rain Forest in the Dahomey Gap: Insight from the Demographic History of Terminalia Superba.” *Heredity* 120 (6): 547–61. https://doi.org/10.1038/s41437-017-0035-0.

Destro-Bisol, Giovanni, Francesco Donati, Valentina Coia, Ilaria Boschi, Fabio Verginelli, Alessandra Caglià, Sergio Tofanelli, Gabriella Spedini, and Cristian Capelli. 2004. “Variation of Female and Male Lineages in Sub-Saharan Populations: The Importance of Sociocultural Factors.” *Molecular Biology and Evolution* 21 (9): 1673–82. https://doi.org/10.1093/molbev/msh186.

Devezer, Berna, Danielle J. Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. 2021. “The Case for Formal Methodology in Scientific Reform.” *Royal Society Open Science* 8 (3): rsos.200805, 200805. https://doi.org/10.1098/rsos.200805.

DeVore, Irven, and Richard B Lee. 1968. *Man the Hunter*. Aldine Publishing Company.

Du Chaillu, P.B. 1892. *Adventures in the Great Forest of Equatorial Africa and the Country of the Dwarfs*. John Murray. https://books.google.fr/books?id=nEM6AQAAMAAJ.

Eltis, David, ,. 2002. *Coerced and Free Migration : Global Perspectives*. Stanford, Calif.: Stanford University Press.

Eltis, David, and David Richardson. 2015. “Atlas of the Transatlantic Slave Trade.”

Escobar, Juan Sebastián, Antoine Nicot, and Patrice David. 2008. “The Different Sources of Variation in Inbreeding Depression, Heterosis and Outbreeding Depression in a Metapopulation of *Physa Acuta*.” *Genetics* 180 (3): 1593–1608. https://doi.org/10.1534/genetics.108.092718.

Estoup, Arnaud A, Paul Verdu, Jean-Michel Marin, Christian Robert, Alexandre Dehne Garcia, Jean-Marie Cornuet, and Pierre Pudlo. 2019. “Application of ABC to Infer the Genetic History of Pygmy Hunter-Gatherer Populations from Western Central Africa.” In *Handbook of Approximate Bayesian Computation*, edited by Scott A. Sisson, Yanan Fan, and Mark A. Beaumont, Chapter 18. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall. https://hal.inrae.fr/hal-02787321.

Estoup, Arnaud, Louis Raynal, Paul Verdu, and Jean-Michel Marin. 2018. “Model Choice Using Approximate Bayesian Computation and Random Forests: Analyses Based on Model Grouping to Make Inferences about the Genetic History of Pygmy Human Populations.” *Journal de La*

Société Française de Statistique 159 (3): 167–90. http://www.numdam.org/item/JSFS_2018__159_3_167_0/.

Ewens, W. J., and R. S. Spielman. 1995. "The Transmission/Disequilibrium Test: History, Subdivision, and Admixture." *American Journal of Human Genetics* 57 (2): 455–64.

Ewing, Gregory, and Joachim Hermisson. 2010. "MSMS: A Coalescent Simulation Program Including Recombination, Demographic Structure and Selection at a Single Locus." *Bioinformatics* 26 (16): 2064–65. https://doi.org/10.1093/bioinformatics/btq322.

Excoffier, L, P E Smouse, and J M Quattro. 1992. "Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data." *Genetics* 131 (2): 479–91. https://doi.org/10.1093/genetics/131.2.479.

Excoffier, Laurent, and Matthieu Foll. 2011. "Fastsimcoal: A Continuous-Time Coalescent Simulator of Genomic Diversity under Arbitrarily Complex Evolutionary Scenarios." *Bioinformatics* 27 (9): 1332–34. https://doi.org/10.1093/bioinformatics/btr124.

Falush, Daniel, Matthew Stephens, and Jonathan K Pritchard. 2003. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies." *Genetics* 164 (4): 1567–87. https://doi.org/10.1093/genetics/164.4.1567.

Fisher, R. A. 1922. "Darwinian Evolution of Mutations." *The Eugenics Review* 14 (1): 31–34.

Foll, Matthieu, Hyunjin Shim, and Jeffrey D. Jensen. 2015. "WFABC: A Wright-Fisher ABC-Based Approach for Inferring Effective Population Sizes and Selection Coefficients from Time-Sampled Data." *Molecular Ecology Resources* 15 (1): 87–98. https://doi.org/10.1111/1755-0998.12280.

Fortes-Lima, Cesar A., Romain Laurent, Valentin Thouzeau, Bruno Toupance, and Paul Verdu. 2021. "Complex Genetic Admixture Histories Reconstructed with Approximate Bayesian Computation." *Molecular Ecology Resources* 21 (4): 1098–1117. https://doi.org/10.1111/1755-0998.13325.

Fortes-Lima, Cesar, and Paul Verdu. 2021. "Anthropological Genetics Perspectives on the Transatlantic Slave Trade." *Human Molecular Genetics* 30 (R1): R79–87. https://doi.org/10.1093/hmg/ddaa271.

Froment, Alain. 1993. "Adaptation biologique et variation dans l'espèce humaine : le cas des Pygmées d'Afrique." *Bulletins et Mémoires de la Société d'anthropologie de Paris* 5 (3): 417–48. https://doi.org/10.3406/bmsap.1993.2371.

Fürniss, Susanne. 2011. "Partages et emprunts de musiques rituelles au Sud-Est-Cameroun." In *Territoires musicaux mis en scène*, edited by Monique Desroches, Marie-Hélène Pichette, Claude Dauphin, and Gordon E. Smith, 263–77. Presses de l'Université de Montréal. https://doi.org/10.4000/books.pum.9077.

Goldberg, Amy, Paul Verdu, and Noah A Rosenberg. 2014. "Autosomal Admixture Levels Are Informative About Sex Bias in Admixed Populations." *Genetics* 198 (3): 1209–29. https://doi.org/10.1534/genetics.114.166793.

Goldstein, D B, A Ruiz Linares, L L Cavalli-Sforza, and M W Feldman. 1995. "Genetic Absolute Dating Based on Microsatellites and the Origin of Modern Humans." *Proceedings of the National Academy of Sciences* 92 (15): 6723–27. https://doi.org/10.1073/pnas.92.15.6723.

Gravel, Simon. 2012. "Population Genetics Models of Local Ancestry." *Genetics* 191 (2): 607–19. https://doi.org/10.1534/genetics.112.139808.

Gray, Russell D., and Quentin D. Atkinson. 2003. “Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin.” *Nature* 426 (6965): 435–39. https://doi.org/10.1038/nature02029.

Gray, Russell D., and Fiona M. Jordan. 2000. “Language Trees Support the Express-Train Sequence of Austronesian Expansion.” *Nature* 405 (6790): 1052–55. https://doi.org/10.1038/35016575.

Guo, Wei, Wing K. Fung, Ningzhong Shi, and Jianhua Guo. 2005. “On the Formula for Admixture Linkage Disequilibrium.” *Human Heredity* 60 (3): 177–80. https://doi.org/10.1159/000090119.

Gurdasani, Deepti, Tommy Carstensen, Fasil Tekola-Ayele, Luca Pagani, Ioanna Tachmazidou, Konstantinos Hatzikotoulas, Savita Karthikeyan, et al. 2015. “The African Genome Variation Project Shapes Medical Genetics in Africa.” *Nature* 517 (7534): 327–32. https://doi.org/10.1038/nature13997.

Guthrie, Malcolm, ,. 1967. *Comparative Bantu; an Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages.* Farnborough: Gregg.

Hellenthal, G., G. B. J. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. 2014. “A Genetic Atlas of Human Admixture History.” *Science* 343 (6172): 747–51. https://doi.org/10.1126/science.1243518.

Hewlett, Barry S. 1996. “Cultural Diversity among African Pygmies.” *Cultural Diversity among Twentieth-Century Foragers: An African Perspective* 1: 215–44.

Hewlett, Barry S., ed. 2014. *Hunter-Gathers of the Congo Basin: Cultures, Histories, and Biology of African Pygmics*. 1st ed. Routledge. https://doi.org/10.4324/9780203789438.

Hewlett, Bonnie L., ed. 2019. *The Secret Lives of Anthropologists: Lessons from the Field*. Abingdon, Oxon ; New York, NY: Routledge.

Jay, Flora, Simon Boitard, and Frédéric Austerlitz. 2019. “An ABC Method for Whole-Genome Sequence Data: Inferring Paleolithic and Neolithic Human Expansions.” Edited by Ryan Hernandez. *Molecular Biology and Evolution* 36 (7): 1565–79. https://doi.org/10.1093/molbev/msz038.

Joiris, Daou V. 2003. “The Framework of Central African Hunter-Gatherers and Neighbouring Societies.” *African Study Monographs: Supplementary Issue*, no. 28: 57–79. http://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/68426/1/ASM_S_28_57.pdf.

Kazadi, Ntole. 1981. “Meprises et Admires: L’ambivalence Des Relations Entre Les Bacwa (Pygm�es) et Les Bahemba (Bantu).” *Africa: Journal of the International African Institute* 51 (4): 836–47. https://www.jstor.org/stable/1159357.

Kimura, Motoo. 1968. “Evolutionary Rate at the Molecular Level.” *Nature* 217 (5129): 624–26. https://doi.org/10.1038/217624a0.

———. 1983. *The Neutral Theory of Molecular Evolution*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511623486.

Klieman, Kairn A. 2003. *“The Pygmies Were Our Compass”: Bantu and Batwa in the History of West Central Africa, Early Times to c. 1900 C.E.* Social History of Africa. Portsmouth, NH: Heinemann.

Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania press.

Lang, Jürgen. 2009. *Les langues des autres dans la créolisation: théorie et exemplification par le créole d’empreinte wolof à l’île Santiago du Cap Vert*. Tübingen: Narr.

Laurent, Romain, Zachary A. Szpiech, Sergio S. da Costa, Valentin Thouzeau, Cesar A. Fortes-Lima, Françoise Dessarps-Freichey, Laure Lémée, et al. 2022. "The Admixture Histories of Cabo Verde." Preprint. Genetics. https://doi.org/10.1101/2022.04.11.487833.

Lawson, Daniel J., Lucy van Dorp, and Daniel Falush. 2018. "A Tutorial on How Not to Over-Interpret STRUCTURE and ADMIXTURE Bar Plots." *Nature Communications* 9 (1): 3258. https://doi.org/10.1038/s41467-018-05257-7.

Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. "Inference of Population Structure Using Dense Haplotype Data." Edited by Gregory P. Copenhaver. *PLoS Genetics* 8 (1): e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Long, J C. 1991. "The Genetic Structure of Admixed Populations." *Genetics* 127 (2): 417–28. https://doi.org/10.1093/genetics/127.2.417.

Lopez, Marie, Jeremy Choin, Martin Sikora, Katherine Siddle, Christine Harmant, Helio A. Costa, Martin Silvert, et al. 2019. "Genomic Evidence for Local Adaptation of Hunter-Gatherers to the African Rainforest." *Current Biology* 29 (17): 2926-2935.e4. https://doi.org/10.1016/j.cub.2019.07.013.

Lopez, Marie, Athanasios Kousathanas, Hélène Quach, Christine Harmant, Patrick Mouguiama-Daouda, Jean-Marie Hombert, Alain Froment, et al. 2018. "The Demographic History and Mutational Load of African Hunter-Gatherers and Farmers." *Nature Ecology & Evolution* 2 (4): 721–30. https://doi.org/10.1038/s41559-018-0496-4.

Mathias, Rasika Ann, Margaret A. Taub, Christopher R. Gignoux, Wenqing Fu, Shaila Musharoff, Timothy D. O'Connor, Candelaria Vergara, et al. 2016. "A Continuum of Admixture in the Western Hemisphere Revealed by the African Diaspora Genome." *Nature Communications* 7 (1): 12522. https://doi.org/10.1038/ncomms12522.

Micheletti, Steven J., Kasia Bryc, Samantha G. Ancona Esselmann, William A. Freyman, Meghan E. Moreno, G. David Poznik, Anjali J. Shastri, et al. 2020. "Genetic Consequences of the Transatlantic Slave Trade in the Americas." *The American Journal of Human Genetics* 107 (2): 265–77. https://doi.org/10.1016/j.ajhg.2020.06.012.

Moran, P. A. P. 1958. "Random Processes in Genetics." *Mathematical Proceedings of the Cambridge Philosophical Society* 54 (1): 60–71. https://doi.org/10.1017/S0305004100033193.

Nei, Masatoshi. 1978. "Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals." *Genetics* 89 (3): 583–90. https://doi.org/10.1093/genetics/89.3.583.

Ni, Xumin, Kai Yuan, Chang Liu, Qidi Feng, Lei Tian, Zhiming Ma, and Shuhua Xu. 2019. "MultiWaver 2.0: Modeling Discrete and Continuous Gene Flow to Reconstruct Complex Population Admixtures." *European Journal of Human Genetics* 27 (1): 133–39. https://doi.org/10.1038/s41431-018-0259-3.

Ongaro, Linda, Marilia O. Scliar, Rodrigo Flores, Alessandro Raveane, Davide Marnetto, Stefania Sarno, Guido A. Gnecchi-Ruscone, et al. 2019. "The Genomic Impact of European Colonization of the Americas." *Current Biology* 29 (23): 3974-3986.e4. https://doi.org/10.1016/j.cub.2019.09.076.

Pagel, Mark. 2009. "Human Language as a Culturally Transmitted Replicator." *Nature Reviews Genetics* 10 (6): 405–15. https://doi.org/10.1038/nrg2560.

Patin, Etienne, Guillaume Laval, Luis B. Barreiro, Antonio Salas, Ornella Semino, Silvana Santachiara-Benerecetti, Kenneth K. Kidd, et al. 2009. "Inferring the Demographic History of African Farmers and Pygmy Hunter–Gatherers Using a Multilocus Resequencing Data Set." Edited by Anna Di Rienzo. *PLoS Genetics* 5 (4): e1000448. https://doi.org/10.1371/journal.pgen.1000448.

Patin, Etienne, Marie Lopez, Rebecca Grollemund, Paul Verdu, Christine Harmant, Hélène Quach, Guillaume Laval, et al. 2017. "Dispersals and Genetic Adaptation of Bantu-Speaking Populations in Africa and North America." *Science* 356 (6337): 543–46. https://doi.org/10.1126/science.aal1988.

Patin, Etienne, Katherine J. Siddle, Guillaume Laval, Hélène Quach, Christine Harmant, Noémie Becker, Alain Froment, et al. 2014. "The Impact of Agricultural Emergence on the Genetic History of African Rainforest Hunter-Gatherers and Agriculturalists." *Nature Communications* 5 (1): 3163. https://doi.org/10.1038/ncomms4163.

Patterson, K. David. 1988. "Epidemics, Famines, and Population in the Cape Verde Islands, 1580-1900." *The International Journal of African Historical Studies* 21 (2): 291. https://doi.org/10.2307/219938.

Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History." *Genetics* 192 (3): 1065–93. https://doi.org/10.1534/genetics.112.145037.

Pemberton, Trevor J., Paul Verdu, Noémie S. Becker, Cristen J. Willer, Barry S. Hewlett, Sylvie Le Bomin, Alain Froment, Noah A. Rosenberg, and Evelyne Heyer. 2018. "A Genome Scan for Genes Underlying Adult Body Size Differences between Central African Hunter-Gatherers and Farmers." *Human Genetics* 137 (6–7): 487–509. https://doi.org/10.1007/s00439-018-1902-3.

Perry, George H., and Nathaniel J. Dominy. 2009. "Evolution of the Human Pygmy Phenotype." *Trends in Ecology & Evolution* 24 (4): 218–25. https://doi.org/10.1016/j.tree.2008.11.008.

Perry, George H., Matthieu Foll, Jean-Christophe Grenier, Etienne Patin, Yohann Nédélec, Alain Pacis, Maxime Barakatt, et al. 2014. "Adaptive, Convergent Origins of the Pygmy Phenotype in African Rainforest Hunter-Gatherers." *Proceedings of the National Academy of Sciences* 111 (35). https://doi.org/10.1073/pnas.1402875111.

Perry, George H., and Paul Verdu. 2017. "Genomic Perspectives on the History and Evolutionary Ecology of Tropical Rainforest Occupation by Humans." *Quaternary International* 448 (August): 150–57. https://doi.org/10.1016/j.quaint.2016.04.038.

Phillipson, David W. 2005. *African Archaeology*. 3rd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511800313.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. "Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites." *Molecular Biology and Evolution* 16 (12): 1791–98. https://doi.org/10.1093/oxfordjournals.molbev.a026091.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59. https://doi.org/10.1093/genetics/155.2.945.

Pudlo, Pierre, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P. Robert. 2016. "Reliable ABC Model Choice via Random Forests." *Bioinformatics* 32 (6): 859–66. https://doi.org/10.1093/bioinformatics/btv684.

Quint, Nicolas. 2000. *Le cap-verdien: origines et devenir d'une langue métisse: étude des relations de la langue cap-verdienne avec les langues africaines, créoles et portugaise*. Paris, France: L'Harmattan.

Ragsdale, Aaron P., and Simon Gravel. 2019. "Models of Archaic Admixture and Recent History from Two-Locus Statistics." Edited by Joshua M. Akey. *PLOS Genetics* 15 (6): e1008204.

https://doi.org/10.1371/journal.pgen.1008204.

Ragsdale, Aaron P., Timothy D. Weaver, Elizabeth G. Atkinson, Eileen Hoal, Marlo Möller, Brenna M. Henn, and Simon Gravel. 2022. "A Weakly Structured Stem for Human Origins in Africa." Preprint. Genomics. https://doi.org/10.1101/2022.03.23.485528.

Ramos Pérez, Demetrio, and María Lourdes Díaz-Trechuelo López Spínola. 1992. *América en el siglo XVIII: la Ilustración en América*. 2. ed. Historia general de España y América, T. 11,2. Madrid: Ed. Rialp.

Raynal, Louis, Jean-Michel Marin, Pierre Pudlo, Mathieu Ribatet, Christian P Robert, and Arnaud Estoup. 2019. "ABC Random Forests for Bayesian Parameter Inference." Edited by Oliver Stegle. *Bioinformatics* 35 (10): 1720–28. https://doi.org/10.1093/bioinformatics/bty867.

Robert, Christian P., Kerrie Mengersen, and Carla Chen. 2010. "Model Choice versus Model Criticism." *Proceedings of the National Academy of Sciences* 107 (3). https://doi.org/10.1073/pnas.0911260107.

Rousset, François. 1997. "Genetic Differentiation and Estimation of Gene Flow from *F* -Statistics Under Isolation by Distance." *Genetics* 145 (4): 1219–28. https://doi.org/10.1093/genetics/145.4.1219.

Schlebusch, Carina M., and Mattias Jakobsson. 2018. "Tales of Human Migration, Admixture, and Selection in Africa." *Annual Review of Genomics and Human Genetics* 19 (1): 405–28. https://doi.org/10.1146/annurev-genom-083117-021759.

Schweinfurth, G.A. 1873. *The Heart of Africa: Three Years' Travels and Adventures in the Unexplored Regions of Central Africa, from 1868 to 1871*. 2 vols. The Heart of Africa. S. Low, Marston, Low, and Searle. https://books.google.fr/books?id=WK5LAQAAIAAJ.

Sisson, S. A., Y. Fan, and M. A. Beaumont, eds. 2018. *Handbook of Approximate Bayesian Computation*. 1st ed. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC. https://doi.org/10.1201/9781315117195.

Soares, Maria Jo  o. 2011. "The British Presence on the Cape Verdian Archipelago (Sixteenth to Eighteenth Centuries)." *African Economic History* 39: 129–46. https://www.jstor.org/stable/23718980.

Swadesh, Morris. 1971. *The Origin and Diversification of Language*. Chicago: Aldine, Atherton.

Szpiech, Zachary A., Angel C.Y. Mak, Marquitta J. White, Donglei Hu, Celeste Eng, Esteban G. Burchard, and Ryan D. Hernandez. 2019. "Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity." *The American Journal of Human Genetics* 105 (4): 747–62. https://doi.org/10.1016/j.ajhg.2019.08.011.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. 2005. "Estimation of Individual Admixture: Analytical and Study Design Considerations." *Genetic Epidemiology* 28 (4): 289–301. https://doi.org/10.1002/gepi.20064.

Tavaré, Simon, David J Balding, R C Griffiths, and Peter Donnelly. 1997. "Inferring Coalescence Times From DNA Sequence Data." *Genetics* 145 (2): 505–18. https://doi.org/10.1093/genetics/145.2.505.

The 1000 Genomes Project Consortium, Corresponding authors, Adam Auton, Gonçalo R. Abecasis, Steering committee, David M. Altshuler, Richard M. Durbin, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Thouzeau, Valentin, Philippe Mennecier, Paul Verdu, and Frédéric Austerlitz. 2017. "Genetic and Linguistic Histories in Central Asia Inferred Using Approximate Bayesian Computations." *Proceedings of the Royal Society B: Biological Sciences* 284 (1861): 20170706. https://doi.org/10.1098/rspb.2017.0706.

Tishkoff, Sarah A., Floyd A. Reed, Françoise R. Friedlaender, Christopher Ehret, Alessia Ranciaro, Alain Froment, Jibril B. Hirbo, et al. 2009. "The Genetic Structure and History of Africans and African Americans." *Science* 324 (5930): 1035–44. https://doi.org/10.1126/science.1172257.

Vansina, J. 1995. "New Linguistic Evidence and 'The Bantu Expansion.'" *The Journal of African History* 36 (2): 173–95. https://doi.org/10.1017/S0021853700034101.

Veeramah, Krishna R, Daniel Wegmann, August Woerner, Fernando L Mendez, Joseph C Watkins, Giovanni Destro-Bisol, Himla Soodyall, Leslie Louie, and Michael F Hammer. 2012. "An Early Divergence of KhoeSan Ancestors from Those of Other Modern Humans Is Supported by an ABC-Based Analysis of Autosomal Resequencing Data." *Molecular Biology and Evolution* 29 (2): 617–30.

Verdu, Paul. 2009. "Anthropologie Génétique Des Populations Humaines d'Afrique Centrale: Histoire Du Peuplement Pygmée."

———. 2012. "Perspectives de La Génétique Humaine Sur l'origine et La Diversité Des Populations Pygmées d'Afrique Centrale." *Journal Des Africanistes* 82 (1–2): 5371. https://doi.org/10.4000/africanistes.4269.

———. 2014. "Population Genetics of Central African Pygmies and Non-Pygmies." In *Hunter-Gathers of the Congo Basin: Cultures, Histories, and Biology of African Pygmics*, 31–54. Routledge.

———. 2016. "African Pygmies." *Current Biology* 26 (1): R12–14. https://doi.org/10.1016/j.cub.2015.10.023.

———. 2019. "Do You Consent to Participate in the Research Study?" In *The Secret Lives of Anthropologists: Lessons from the Field*. Vol. 1. Abingdon, Oxon ; New York, NY: Routledge.

Verdu, Paul, and Frédéric Austerlitz. 2015. "Post Marital Residence Behaviours Shape Genetic Variation in Hunter-Gatherer and Agricultural Populations from Central Africa." *Hunter Gatherer Research* 1 (1): 107–24. https://doi.org/10.3828/hgr.2015.6.

Verdu, Paul, Frederic Austerlitz, Arnaud Estoup, Renaud Vitalis, Myriam Georges, Sylvain Théry, Alain Froment, et al. 2009. "Origins and Genetic Diversity of Pygmy Hunter-Gatherers from Western Central Africa." *Current Biology* 19 (4): 312–18. https://doi.org/10.1016/j.cub.2008.12.049.

Verdu, Paul, Noémie S.A. Becker, Alain Froment, Myriam Georges, Viola Grugni, Lluis Quintana-Murci, Jean-Marie Hombert, et al. 2013. "Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies." *Molecular Biology and Evolution* 30 (4): 918–37. https://doi.org/10.1093/molbev/mss328.

Verdu, Paul, and Giovanni Destro-Bisol. 2012. "African Pygmies, What's Behind a Name?" *Human Biology* 84 (1): 1–10. https://doi.org/10.3378/027.084.0105.

Verdu, Paul, Ethan M. Jewett, Trevor J. Pemberton, Noah A. Rosenberg, and Marlyse Baptista. 2017. "Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population." *Current Biology* 27 (16): 2529-2535.e3. https://doi.org/10.1016/j.cub.2017.07.002.

Verdu, Paul, and Noah A Rosenberg. 2011. "A General Mechanistic Model for Admixture Histories of Hybrid Populations." *Genetics* 189 (4): 1413–26. https://doi.org/10.1534/genetics.111.132787.

Versluys, Tom M. M., Alex Mas-Sandoval, Ewan O. Flintham, and Vincent Savolainen. 2021. “Why Do We Pick Similar Mates, or Do We?” *Biology Letters* 17 (11): 20210463. https://doi.org/10.1098/rsbl.2021.0463.

Wakeley, John, Léandra King, Bobbi S Low, and Sohini Ramachandran. 2012. “Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingman’s Coalescent.” *Genetics* 190 (4): 1433–45. https://doi.org/10.1534/genetics.111.135574.

Weir, B. S., and C. Clark Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Evolution* 38 (6): 1358. https://doi.org/10.2307/2408641.

Wright, Sewall. 1931. “Evolution in Mendelian Populations.” *Genetics* 16 (2): 97–159. https://doi.org/10.1093/genetics/16.2.97.

Zaitlen, Noah, Scott Huntsman, Donglei Hu, Melissa Spear, Celeste Eng, Sam S Oh, Marquitta J White, et al. 2017. “The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium.” *Genetics* 205 (1): 375–83. https://doi.org/10.1534/genetics.116.192138.

Summary

Genetic admixture is an essential mechanism of the biological evolution of numerous species, including *Homo sapiens*. In this Habilitation à Diriger des Recherche, we investigate Histories of Admixture in human populations, from population genetics theory to data analysis, via anthropological sampling on the field, and development of statistical methodologies. Chapter 1 focuses on how diverse anthropological categorizations of individuals into populations are at the root of anthropological-genetics study-designs, and how they influence sampling protocols and population-genetics methodological choices. To illustrate this, we describe how we built, since 2005, interdisciplinary projects aiming at reconstructing the evolutionary history of Central African populations, and the genetic and linguistic admixture histories of Cabo Verde, respectively. We then elaborate on ethical and deontological issues in anthropological sciences, emerging at the crossroad between anthropological data collection, data privacy protection, and Open Science. Finally, we provide perspectives about categorization challenges in the human paleo-genomics era. Chapter 2 summarizes the methods deployed and results obtained about the demographic and evolutionary history of Central African populations categorized in so-called "Pygmies" and neighboring "non-Pygmies" based on complex historical and ethnological criteria. We show how asymmetric and sex-specific admixture histories and admixture-related natural selection processes shaped observed genetic and phenotypic diversities in the Congo Basin. Chapter 3 presents a theoretical general mechanistic model of historical admixture in hybrid populations. We analytically show that complex admixture processes leave an identifiable signature in the distribution of genetic admixture fractions across individuals. Chapter 4 presents the forward-in-time genetic-data simulator and summary-statistics calculator implemented in the software *MetHis*. It is specifically designed to be coupled with machine-learning Approximate Bayesian Computation Random Forest scenario-choice, and Neural Network posterior parameter estimation, for inferring highly complex admixture histories from genetic data. In Chapter 5, we deploy this novel framework to reconstruct the detailed genetic admixture histories of Cabo Verde, the first European settlement-colony in Sub-Saharan Africa established in the 1460's. We show how shifting socio-cultural relationships between enslaved and non-enslaved communities, between the onset of the Trans-Atlantic Slave-Trade in the 15th century and the abolition of slavery in the mid-19th century, influenced diverse admixture processes between Europeans and Africans throughout the archipelago. Finally, we present a population genetics and linguistics interdisciplinary paradigm, and the associated statistical methodologies, to infer jointly the genetic and linguistic admixture histories of Cabo Verdean Kriolu-speakers, using genetic and linguistic data collected from the same individuals. Altogether, this interdisciplinary work between cultural anthropology, linguistics, and population genetics demonstrates how complex histories of admixture, deeply influenced by equally complex socio-cultural processes, have shaped the genetic and cultural diversity of human populations.

Résumé

Les métissages génétiques sont un mécanisme essentiel de l'évolution biologique de nombreuses espèces, dont *Homo sapiens*. Dans cette Habilitation à Diriger des Recherches, nous étudions les Histoires de Métissages dans les populations humaines, de la théorie de la génétique des populations à l'analyse de données, en passant par l'échantillonnage anthropologique sur le terrain et le développement de méthodes statistiques. Le Chapitre 1 se concentre sur la manière dont les diverses catégorisations anthropologiques des individus en populations sont à la base de la conception des études en anthropologie génétique, et comment elles influencent les protocoles d'échantillonnage et les choix méthodologiques en génétique des populations. Pour illustrer cela, nous décrivons comment nous avons construit, depuis 2005, des projets interdisciplinaires visant à reconstituer, respectivement, l'histoire évolutive des populations d'Afrique Centrale, et les histoires de métissages génétiques et linguistiques au Cap Vert. Nous approfondissons ensuite les questions éthiques et déontologiques en sciences anthropologiques, émergeant au carrefour entre la collecte de données anthropologiques, la protection de la confidentialité des données et la Science Ouverte. Enfin, nous proposons des perspectives sur les défis de catégorisation à l'ère de la paléogénomique humaine. Le Chapitre 2 résume les méthodes déployées et les résultats obtenus sur l'histoire démographique et évolutive des populations d'Afrique Centrale catégorisées en dits « Pygmées » et voisins « non-Pygmées » selon des critères historiques et ethnologiques complexes. Nous montrons comment des processus de métissages asymétriques et liés au genre, et des processus de sélection naturelle liés aux métissages, ont façonné les diversités génétiques et phénotypiques observées dans le Bassin du Congo. Le Chapitre 3 présente un modèle théorique général de l'histoire des métissages dans les populations hybrides. Nous montrons analytiquement que les processus de métissages complexes laissent une signature identifiable dans la distribution des proportions de métissages génétiques entre individus. Le Chapitre 4 présente le simulateur de données génétiques « foward-in-time » et le calculateur de statistiques résumées implémentés dans le logiciel *MetHis*. Il est spécialement conçu pour être couplé au méthodes « d'Approximate Bayesian Computation » en apprentissage-machine par Forêt Aléatoire pour le choix de scénarios et par Réseau de Neurones pour l'estimation des paramètres a posteriori, et qui permettent de reconstruire des histoires très complexes de métissages à partir de données génétiques. Dans le Chapitre 5, nous déployons cette nouvelle méthode afin de reconstruire les histoires détaillées des métissages génétiques au Cap Vert, première colonie européenne de peuplement en Afrique subsaharienne fondée dans les années 1460. Nous montrons comment l'évolution des relations socioculturelles entre les communautés d'esclaves et de non-esclaves entre le début de la traite esclavagiste transatlantique au XVe siècle et l'abolition de l'esclavage au milieu du XIXe siècle, a influencé divers processus de métissages entre Européens et Africains dans tout l'archipel. Enfin, nous présentons un paradigme interdisciplinaire de génétique et linguistique des populations, et les méthodes statistiques associées, pour reconstruire conjointement les histoires de métissages génétiques et linguistiques des locuteurs kriolu capverdien, en utilisant des données génétiques et linguistiques collectées auprès des mêmes individus. Dans l'ensemble, ce travail interdisciplinaire entre anthropologie culturelle, linguistique et génétique des populations montre comment des histoires complexes de métissages, profondément influencées par des processus socioculturels eux-aussi complexes, ont façonné la diversité génétique et culturelle des populations humaines.