2. AGENE : Anthropologie Génétique

3.1 Ly et al (2019) illustre bien notre démarche interdisciplinaire qui consiste à évaluer l'impact de pratiques culturelles (ici les systèmes de parenté, comme la matrilinéarité) sur l'évolution de la diversité génétique humaine. Cette étude est fondée sur des données génétiques et ethno-démographiques collectées en Asie du Sud Est, l'un des terrains que nous avons développés.

Ly G, Laurent R, Lafosse S, Monidarin C, Diffloth G, Bourdier F, Evrard O, Toupance B, Pavard S, Chaix R. 2019. From matrimonial practices to genetic diversity in Southeast Asian populations: the signature of the matrilineal puzzle. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 374:20180434. doi:10.1098/rstb.2018.0434

3.2 Fortes-Lima et al (2021) représente un autre aspect de notre recherche : la reconstruction de l'histoire démographique des populations à partir de données génomiques grâce à des méthodes bio-informatiques de statistiques computationnelles basées sur des simulations, notamment l'Approximate Bayesian Computation.

Fortes-Lima CA, Laurent R, Thouzeau V, Toupance B, Verdu P. 2021. Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. Molecular Ecology Resources 21:1098-1117. doi:10.1111/1755-0998.13325

3.3 Le plateau P2GM est un équipement mutualisé d'acquisition de données de génétique et de paléogénétique moléculaires développé et coordonné par deux membres de notre équipe. Il est utilisé par diverses équipes de l'UMR, du MNHN, et d'autres EPST. Il couvre une grande diversité de matériel biologique : le bois des collections, prélèvement non invasif des primates non humains, salive et sang des humains modernes, ADN des plantes cultivés, l'ADN ancien humain et non humain.

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Ly G *et al*. 2019 From matrimonial practices to genetic diversity in Southeast Asian populations: the signature of the matrilineal puzzle. *Phil. Trans. R. Soc. B* **374**: 20180434.
http://dx.doi.org/10.1098/rstb.2018.0434

Accepted: 17 April 2019

One contribution of 17 to a theme issue 'The evolution of female-biased kinship in humans and other mammals'.

Subject Areas:
genetics, behaviour

Keywords:
residence rule, matrilineal puzzle, inbreeding, human genetics, matrilocal, patrilocal

Author for correspondence:
Raphaëlle Chaix
e-mail: chaix@mnhn.fr

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4526243.

THE ROYAL SOCIETY PUBLISHING

From matrimonial practices to genetic diversity in Southeast Asian populations: the signature of the matrilineal puzzle

Goki Ly[1], Romain Laurent[1], Sophie Lafosse[1], Chou Monidarin[2], Gérard Diffloth[3], Frédéric Bourdier[4], Olivier Evrard[5], Bruno Toupance[1], Samuel Pavard[1] and Raphaëlle Chaix[1]

[1]Unité Eco-anthropologie (EA), Muséum National d'Histoire Naturelle, CNRS, Université de Paris, 17 place du Trocadéro, 75016 Paris, France
[2]Rodolphe Merieux Laboratory and Faculty of Pharmacy of University of Health Sciences, Phnom Penh, Cambodia
[3]Siem Reap, Cambodia
[4]Unité 201 Développement et Sociétés (DEVSOC), IEDES/IRD, Panthéon Sorbonne, Paris, France
[5]Unité Patrimoines Locaux et Gouvernance (PALOC), Muséum National d'Histoire Naturelle, CNRS, IRD, 75006 Paris, France

GL, 0000-0003-1215-4608; BT, 0000-0002-8244-1824; SP, 0000-0002-6803-8123

In matrilineal populations, the descent group affiliation is transmitted by women whereas the socio-political power frequently remains in the hands of men. This situation, named the 'matrilineal puzzle', is expected to promote local endogamy as a coping mechanism allowing men to maintain their decision-making power over their natal descent group. In this paper, we revisit this 'matrilineal puzzle' from a population genetics' point of view. Indeed, such tendency for local endogamy in matrilineal populations is expected to increase their genetic inbreeding and generate isolation-by-distance patterns between villages. To test this hypothesis, we collected ethno-demographic data for 3261 couples and high-density genetic data for 675 individuals from 11 Southeast Asian populations with a wide range of social organizations: matrilineal and matrilocal populations (M), patrilineal and patrilocal populations (P) or cognatic populations with predominant matrilocal residence (C). We observed that M and C populations have higher levels of village endogamy than P populations, and that such higher village endogamy leads to higher genetic inbreeding. M populations also exhibit isolation-by-distance patterns between villages. We interpret such genetic patterns as the signature of the 'matrilineal puzzle'. Notably, our results suggest that any form of matrilocal marriage (whatever the descent rule is) increases village endogamy. These findings suggest that male dominance, when combined with matrilocality, constrains inter-village migrations, and constitutes an underexplored cultural process shaping genetic patterns in human populations.

This article is part of the theme issue 'The evolution of female-biased kinship in humans and other mammals'.

1. Introduction

In matrilineal populations (which represent about 12–17% of the world's populations), descent group affiliation is transmitted through the mother, whereas in patrilineal populations (about 45% of the populations), it is transmitted through the father [1,2]. These descent systems are not the symmetrical opposite of each other because, in both cases, authority and socio-political power (beyond the household) lie in the hands of men [3–6]. Indeed, while women play key roles within the domestic unit in terms of provisioning, childrearing and household organization, men are usually more empowered than women to control public sphere affairs [4,5,7,8]. Hence, it has been argued that matrilineal

societies are caught in what has been named by A. Richards as a 'matrilineal puzzle' [5]. Indeed, in these populations, a man has to conciliate his loyalties to his conjugal and to his natal kin [8]: he is expected to exert, at the same time, his authority of husband and father over his spouse and children, and his authority of brother and uncle over his sisters and their children, who are members of his own clan and to whom material and immaterial forms of inheritance are transmitted. In addition, by following the matrilocal residence rule that exists in 70% of these matrilineal populations [2], men are supposed to move out to settle with their wife, possibly in a different village. The further away they marry, the more challenging it is for them to exert their authority over their sisters and their children. In addition, in these populations, the husband has to share authority with his brothers-in-law and other men in charge of his wife's lineage or clan, thus generating the problem of organizing relationships between the in-marrying husband and the male members of his wife's descent group [9].

The matrilineal puzzle has been discussed for over 60 years by different schools of anthropologists. Recent works by evolutionary anthropologists have outlined the evolutionary paradoxes of these matrilineal systems [10–12]. Indeed, in these societies, frequently men invest more in their sisters' children than in their own children, which violates the expectation of Hamilton's rule [13]. In addition, according to the Trivers–Willard hypothesis, it would be more beneficial for parents to transmit wealth to the sex that is most capable of converting it into large reproductive success, typically males [14]. The relative benefit of transmission to males over females depends on the nature of heritable wealth. For instance, livestocks and productive lands are both usually considered more beneficial to men than women because of their greater impact on male's capacity to acquire partners and to increase their reproductive success [10,12]. More recently, the matriliny as daughter-biased investment (MDBI) hypothesis proposes that daughter-biased investment could be an adaptive strategy if the risk of paternity uncertainty (usually high in matrilineal societies) outweighs the benefits of wealth transmission to sons [15]. On the other hand, the expendable male hypothesis suggests that the matrilineal puzzle may not be a puzzle in the evolutionary sense at all, and proposes that matriliny may emerge if females are capable of meeting the subsistence needs of their families while males invest little in children (their own, or their sisters'), this latter condition reconciling these systems with Hamilton's rule [7].

Here, we propose to come back to the original sense given to the matrilineal puzzle by A. Richards and other structural-functionalists [5,8], who were referring to the conflict in authority, and in particular to the constant 'pull-father-pull-mother's brother' stretch existing in these matrilineal populations. Interestingly, through the study of many matrilineal populations, these anthropologists have described several 'solutions' which may appease such tensions: (i) a handful of these matrilineal populations do not follow the matrilocal residence rule but follow a duolocal residence rule—the husband does not live with his wife but visits her regularly while staying with his sisters [8,16]; (ii) more often, the residence rule is avunculocal with, for example, fraternal extended families exerting full authority over the community, while men's sisters are loaned away to other communities and their children are reclaimed at puberty [5,17]; (iii) in the case of the matrilineal populations exhibiting a matrilocal residence rule, the eldest brother may be exempted from such a rule, thus exerting his authority over his sisters and their children [5,17]; (iv) in addition, matrilateral cross-cousin marriages are frequent in these populations, contributing to strengthen the authority of men who have contracted matrilocal marriages—by marrying their daughters to their sister's sons, they bring in their spouse's village, their nephews as sons-in-law, who come from their own descent group and natal village [5,18]; (v) finally, a very frequent 'solution' to the matrilineal puzzle is the strong preference for local endogamy observed in these populations—according to Murdock [19], 17 out of 24 matrilocal populations (70%) were found to be endogamous (as opposed to only 7 out of 101 patrilocal populations). This preference for marrying a woman from the same village, or from a nearby village, may allow men to stay close to the members of their maternal descent group and to exert their authority as brothers/uncles over them, as well as control their descent-group affairs.

In this study, we propose to explore the potential impact of the matrilineal puzzle on the genetic evolution of these populations. A number of studies in the past 20 years have shown that social organizations shape the uniparental genetic diversities of human populations [20–29]. Fewer studies have explored the evolutionary implications of descent and residence rule on autosomal data [30–32]. Here, we propose to focus on the matrilineal puzzle, whose impact on human genetic diversity has been left untouched by population geneticists. Our working hypothesis is that the preference for local endogamy observed in matrilineal populations should increase the genetic inbreeding level in these matrilineal populations, in comparison to populations where such preference does not exist, in particular populations with patrilineal descent. Indeed, when local endogamy increases, we expect not only the proportion of consanguineous marriages to be higher (owing to the small size of the matrimonial market and its enrichment in relatives), but also the genetic drift to increase, both leading to higher genetic inbreeding [33]. In addition, we expect such preference for marrying within the same village, or in a nearby village, in matrilineal populations to generate isolation-by-distance patterns between villages [34]. Such isolation-by-distance patterns are less expected in patrilineal populations, because there is weaker pressure for local endogamy, leading to long distance gene flow.

To test such a hypothesis, we collected ethno-demographic data for 3261 couples as well as high density autosomal single nucleotide polymorphism (SNP) data for 675 individuals from 11 mainland Southeast Asian populations exhibiting a wide variety of social organizations, with different descent and residence rules, but living in similar tropical environments and having similar lifestyles based on rice farming. More precisely, we compared three populations (M) with matrilineal descent and matrilocal residence (Jarai, Tampuan and Kacho'), to four populations (P) with patrilineal descent and patrilocal residence (Khmu', Lamet, Ta-oih and Pacoh). This dataset has been completed by four populations (C) with cognatic descent and either matrilocal residence (Khmer and Bunong) or multilocal residence with final settlement in the wife's village (Brao and Kreung). We grouped these four cognatic populations into a single group of cognatic populations with predominant matrilocal residence. These populations were included to disentangle the effect of descent from the effect of residence on migrations of men. In particular, we investigated whether matrilocality by itself could generate a similar level of constraint on male migration as when it is associated with matrilineal descent. Indeed, it

may be that under any form of matrilocal marriage, and independently of matrilineality, men find themselves, at least initially, in a position of subjection in their wife's village (while possibly losing a position of leadership in their village of origin) [5], a situation that can be lessened by marrying a woman from the same village [6]. Consequently, in populations following a matrilocal residence rule but with no matrilineal descent group (i.e. cognatic populations), we could expect a similar preference for endogamous marriages as in matrilineal populations. The populations under study are presented in table 1 (their geographical location is shown in electronic supplementary material, figure S1).

2. Results

(a) Estimation of village endogamy

We estimated the village endogamy rate (as a proxy for local endogamy) for each population from ethno-demographic information collected by the research team for 3261 couples (figure 1). The village endogamy rate was defined as the proportion of couples for which both spouses were born in the same village. These rates were compared among social organizations using a generalized linear mixed model. As expected under the matrilineal puzzle hypothesis, village endogamy was significantly higher in M than P populations (0.87 versus 0.73 respectively, p-value = 0.029). In C populations, village endogamy was similar to M populations (0.84, p-value = 0.62) and higher than in P populations, although the difference was not statistically significant (p-value = 0.066). This suggests that the matrilocal residence rule alone (with no matrilineal descent groups but cognatic descent) may generate a similar level of constraints on migrations of men as when this residence rule is associated with matrilineal descent. In addition, social organization was estimated to explain 43% of the variance in village endogamy rate among populations.

We observed variation in the village endogamy rate within groups (electronic supplementary material, table S1). In particular, the Kacho' population had a significantly higher village endogamy rate compared to the other M populations (Tampuan and Jarai). Among the P populations, the Khmu' had a significantly higher village endogamy rate than the Pacoh and Ta-oih. The Pacoh had significantly lower village endogamy rate than the Khmu' and Ramet.

(b) Estimation of inbreeding coefficients

We tested whether M and C populations exhibited higher inbreeding levels in comparison to P populations as a result of their higher village endogamy rate. The FEstim software in the FSuite pipeline [35,36] was used to estimate the inbreeding coefficient of each individual (figure 2*a*) and to infer the genealogical relationship between the parents (parental mating type) of each individual (electronic supplementary material, table S2). M and C populations had similar mean inbreeding coefficients (0.018 and 0.017 respectively, p-value = 0.91, figure 2*b*), and both had higher mean inbreeding coefficient compared to P populations (0.011, both p-values < 0.05). In addition, social organization was estimated to explain 27% of the variance in inbreeding coefficients among populations.

Mating type inference showed that M populations had a higher proportion of individuals whose parents were related (90.2%: 77.7% of second cousins, 12.1% of first cousins, and 0.5% of double first cousins) compared to C populations (81.8%: 70.6%, 9.8% and 1.5% for the same mating types, χ^2 test p-value = 0.012) and compared to P populations (60.0%: 61.6% of second cousins, 7.6% of first cousins and 0.5% of avuncular relationship, p-value < 0.01, electronic supplementary material, table S2). In addition, C populations also had a higher proportion of individuals whose parents were related compared to P populations (p-value < 0.01).

We observed variation within groups in terms of inbreeding coefficient (figure 2 and electronic supplementary material, table S3). In particular, the Kreung population had a significantly higher inbreeding level than other C populations ($F = 0.026$ compared to 0.014 on average for the other C populations). This may relate to the fact that their effective size is lower (only significantly so compared to the Khmer) than the effective population size estimated for the other cognatic populations (see table 1 and electronic supplementary material , tables S4 and S5). In addition, the Khmu' population had a significantly higher inbreeding coefficient than other P populations ($F = 0.022$ compared to 0.0067 on average for the other P populations). Contrary to the case of the Kreung, this does not seem to be linked to differences in effective population size among P populations.

To further investigate the influence of village endogamy on inbreeding level and confirm that village endogamy is the social component that explains the differences in inbreeding level between social organizations, we measured the correlation coefficient between these two parameters at the population level (figure 3). The village endogamy rate was indeed significantly correlated with the inbreeding coefficient (Spearman's $\rho = 0.73$, p-value = 0.015).

(c) Isolation-by-distance patterns

Finally, we explored the patterns of isolation-by-distance at the village level within each population (figure 4). We performed this analysis in populations with at least four sampled villages (after excluding villages with less than five sampled individuals). This filtering step excluded the Kacho', Kreung and Ta-oih from this analysis. We observed significant isolation-by-distance patterns in the two M populations included in this analysis (Tampuan and Jarai, both p-values < 0.05). Among the C populations, such isolation-by-distance pattern was found in the Khmer (p-value = 0.034) but not in the Bunong or Brao (both p-values > 0.05). P populations did not exhibit any significant isolation-by-distance pattern (all p-values > 0.05).

3. Discussion

In this study, we observed that Southeast Asian matrilineal and matrilocal populations (M), but also cognatic populations with predominant matrilocal residence (C), have higher levels of genetic inbreeding than patrilineal and patrilocal populations (P). In addition, M populations exhibit isolation-by-distance patterns between villages. We hypothesize that such findings are the signature of the higher local endogamy resulting from what has been called the 'matrilineal puzzle', which takes root in the male dominance over socio-political power [5]: in matrilineal and matrilocal societies men are supposed to settle with their wife's family, possibly in a different village, while remaining actively involved in decision-making within their own descent groups. This becomes challenging as the

Table 1. Description of the studied populations with sampling information.

population	descent rule	residence rule	abbreviation (group)	sampled villages	DNA samples after QC	DNA samples after removing siblings	ethno-demographic interviews	effective population size (95% IC)
Tampuan	matrilineal[a]	matrilocal[d]	M	8	101	96	65	13 052 (12 040 – 14 064)
Jarai	matrilineal	matrilocal		6	92	89	56	13 641 (12 592 – 14 690)
Kacho'	matrilineal	matrilocal		3	32	30	27	13 407 (12 301 – 14 512)
Bunong	cognatic[b]	matrilocal	C	5	88	84	45	11 850 (10 895 – 12 804)
Khmer	cognatic	matrilocal		5	92	92	44	15 503 (14 376 – 16 630)
Brao	cognatic	multilocal[e]		6	49	49	39	12 023 (11 052 – 12 995)
Kreung	cognatic	multilocal		3	51	50	36	10 920 (9939 – 11 901)
Khmu'	patrilineal[c]	patrilocal[f]	P	8	72	68	65	13 868 (12 798 – 14 938)
Ramet	patrilineal	patrilocal		4	26	26	40	15 051 (13 855 – 16 248)
Ta-oih	patrilineal	patrilocal		4	28	28	34	14 050 (12 922 – 15 179)
Pacoh	patrilineal	patrilocal		5	64	63	44	12 141 (11 140 – 13 143)
total				57	695	675	495	

[a]Matrilineal descent: descent group affiliation is transmitted to the children through the mother.
[b]Cognatic descent: recognition of descent from both sides of the family in the absence of any specified lines of descent.
[c]Patrilineal descent: descent group affiliation is transmitted to the children through the father.
[d]Matrilocal residence: the husband moves to his wife's natal village after marriage.
[e]Multilocal residence: the couple lives alternatively in the husband's and wife's natal villages before settling definitively in one place, which is most often, in the case of Brao and Kreung, the wife's village.
[f]Patrilocal residence: the wife moves to her husband's natal village after marriage.

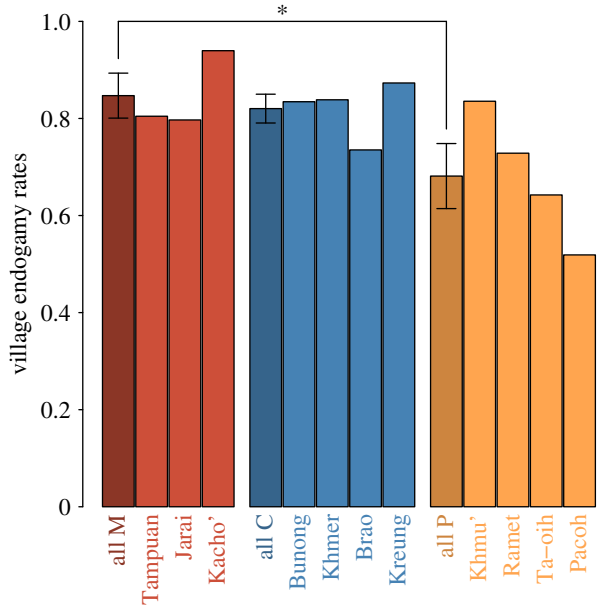

Figure 1. Mean village endogamy rate for each population. The first dark bar in each group represents the mean village endogamy rate in this group (with standard error). Asterisk indicates statistical significance (p-value < 0.05) assessed by generalized linear mixed model. (Online version in colour.)

geographical distance between the natal villages of the wife and the husband increases. Consequently, a preference for local marriages has been observed in these societies, as summarized by Murdock: 'where residence is matrilocal, a man in marrying rarely settles in a new community. He merely takes his possessions from his parents' home and moves, so to speak, across the street' [19, p. 214].

Several lines of evidence support the hypothesis that the matrilineal puzzle is responsible for the different genetic diversity patterns observed in the studied populations. First, we confirmed from our ethno-demographic dataset that M populations had higher village endogamy rates than P populations (on average 87% and 73% respectively). In addition, C populations had similar village endogamy (84%) to M populations. Previous ethnographic works on the matrilineal populations under study also confirmed the existence of a preference for local endogamy. Indeed, the Jarai marry according to a 'the closest, the safest' rule [18] and in the Tampuan population [37], the councils of elders are reluctant to integrate into their villages a man coming from a distant village, a preference still prevailing these days, that may contribute to increase the rate of village endogamy.

Secondly, the village endogamy rate was shown to be a good predictor of the genetic inbreeding levels in these populations. However, the preference for cousin marriages in these populations as reported by the ethnographic literature appears to be a poor predictor of their estimated genetic inbreeding levels: preferences for cousin marriages were reported for most M and P populations [18,37–40] but not for the C populations [38,41–43]. More generally, the percentages of populations with a preference for cousin marriages estimated in a worldwide population sample are higher for both matrilineal and patrilineal populations than for cognatic populations (42.3%, 40% and 19.7% respectively, [2]). These percentages do not fit with our observation that M and C populations have higher genetic inbreeding levels than P populations. Last but not least, a detailed ethnographic study in the Jarai population had shown that the preference for village endogamy was stronger than the preference for cousin marriages: the number of marriages within the same village exceeded the total number of preferential marriages, in particular between ego and mother's brother's daughter, and between ego and father's sister's daughter [18]. Consequently, the matrilineal puzzle, and its consequences in terms of endogamy, is a more likely candidate than the prevalence of cousin marriages to explain the observed differences in genetic patterns between M, C and P populations.

Note also that, despite the fact that M populations are famous for their high paternity uncertainty rate ([12] and references therein), we do not think this process could contribute to the observed genetic differences between M, C and P populations. Indeed, we would expect such paternity uncertainty to decrease, rather than increase, the genetic inbreeding level in these populations in comparison to patrilineal populations; for example, a child born from first cousins may have lower genetic inbreeding coefficient than expected because his parents may share the same grandmother but different grandfathers.

In patrilineal and patrilocal populations, the matrilineal puzzle does not occur as most men settle with their wife in their natal village whether they marry a woman from the same or from another village, with no risk of losing their position of influence or leadership, or their control over their descent-group affairs. This leads to comparatively lower local endogamy rates and lower inbreeding levels in these patrilineal populations, as well as an absence of isolation-by-distance patterns. In our dataset, we noticed one exception to this general observation: the Khmu', a patrilineal and patrilocal population, exhibits a matrilineal-like rate of village endogamy and genetic inbreeding level. The reasons for these differences from the other P populations remain to be investigated. One explanation could be that, although Khmu' follow a general patrilineal descent and patrilocal residence, there is in some families a period of matrilocal residence (up to three years) [40]. Despite the fact that this matrilocal residence is not permanent, this may generate a 'nascent' matrilineal puzzle, encouraging men to marry a woman from the same village.

As discussed in the introduction, local endogamy is probably not the only coping mechanism to the matrilineal puzzle; for example, one of the brothers could escape from the matrilocal rule, allowing him to stay in his natal village (with his wife coming from the same or from a different village) and deal with his descent-group matters [5]. However, our ethno-demographic data do not support such an alternative coping mechanism to the matrilineal puzzle in the Southeast Asian populations under study, since the matrilineal and matrilocal populations followed their residence rule more strictly than the patrilineal and patrilocal populations under study [26].

Lastly, the design of this study, which includes cognatic populations with predominant matrilocal residence, allows us to disentangle the effect of descent from the effect of residence on the 'matrilineal puzzle'. Indeed, as pointed out above, the cognatic populations (C) under study exhibit a similar rate of village endogamy and genetic inbreeding compared to the matrilineal populations (M). Altogether, these populations with matrilocal or predominant matrilocal residence exhibit higher village endogamy and higher genetic inbreeding than patrilocal populations (both p-values < 0.05, estimated by

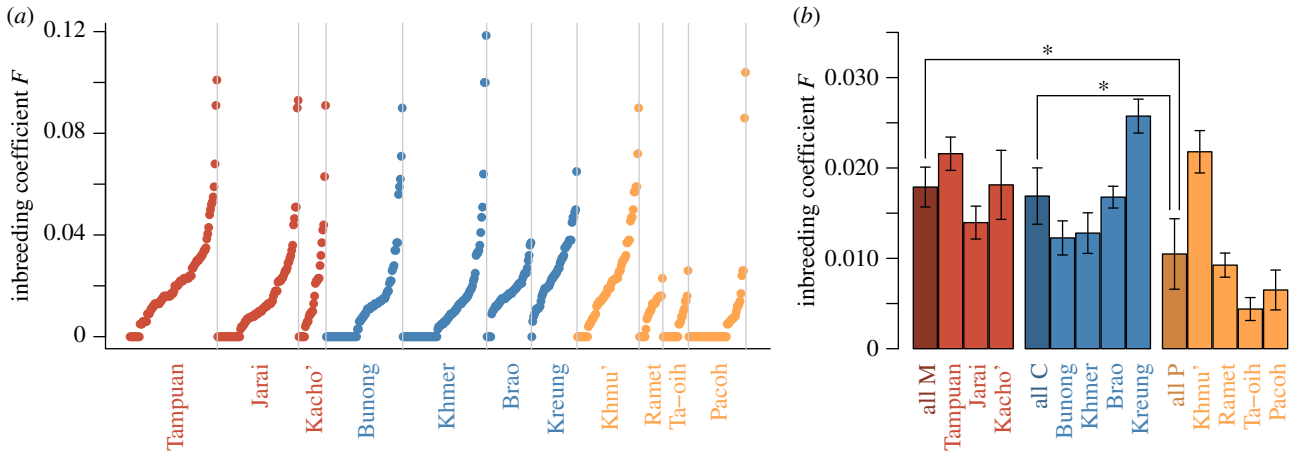

Figure 2. Genetic inbreeding coefficients. (*a*) Each point represents the inbreeding coefficient of an individual. Coefficients are sorted in ascending order in each population. (*b*) The first dark bar in each group represents the mean inbreeding level in this group. All values are represented with standard error. Asterisks indicate statistical significance (p-value < 0.05) assessed by linear mixed model. (Online version in colour.)

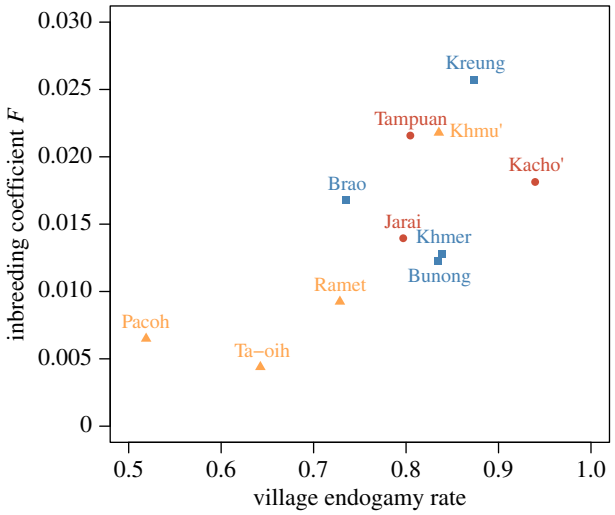

Figure 3. Relationship between village endogamy rate and mean genetic inbreeding coefficient by population. M populations are in red (circles), C populations in blue (squares) and P populations in yellow (triangles). (Online version in colour.)

linear mixed model). Consequently, the matrilineal descent rule is probably not the main component exerting a constraint over male migrations. As pointed out by A. Richards, under any form of matrilocal marriage (whatever the descent rule is), men find themselves, at least initially, in a position of subjection in their wives' villages (while possibly losing a position of leadership in their natal village), an 'irksome' situation that can be avoided by marrying a woman from the same village [5]. As such, the 'matrilineal puzzle' could be renamed the 'matrilocal puzzle' in order to express the fact that it seems to affect not only matrilineal populations but also all matrilocal populations.

Our interdisciplinary study has a number of limitations. We could expect populations facing the matrilineal puzzle to exhibit not only a higher rate of village endogamy but also smaller distances between villages when couples are exogamous. However, our ethno-demographic dataset did not allow us to measure the geographical distance between spouses' natal villages. Consequently, we used the rate of village endogamy as a proxy for local endogamy in this study. In addition, our ethno-demographic dataset may suffer from certain sampling biases. For example, only individuals having their four grandparents from the same population were sampled, a criterion often used in population genetic studies, that may have biased our estimation of village endogamy (however, only slightly, as the proportion of interethnic marriages in these populations is low). There may be some other biases in such endogamy estimation; for example, the information regarding birth places as remembered by the interviewees for some of their relatives, especially their grandparents, may be erroneous, potentially biasing our estimation of village endogamy upwards (but equally so for M, C and P populations). Some matrilineal populations are famous for their duolocal residence mode, with husbands living with their sisters and visiting their wives [8]. However, such duolocal residence was not observed for any couple in our ethno-demographic survey, and was not reported in the ethnographic literature available on the populations under study [18,37–49], so we do not believe that the undetected occurrence of this residence mode could have biased our estimated endogamy rates. The cognatic populations included in this study were not fully matrilocal but include two multilocal populations with final settlement in the wife's village. Replication of this study in fully matrilocal populations is warranted.

Despite these limitations, our study not only suggests that the matrilineal puzzle is still in action in present-day Southeast Asia but also that such a puzzle shapes genetic diversity patterns in human populations, thus identifying a new cultural factor contributing to genetic diversity patterns among human populations. It has previously been shown that genetic inbreeding levels are greatly influenced by the prevalence of consanguineous marriages in human populations [33,50–55]. Our study shows that the association of matrilocality with local endogamy, which takes root in male dominance, may also contribute to some extent to higher inbreeding levels in human populations, thus revealing an additional layer of complexity to the interactions between socio-cultural factors and human genetic diversity patterns.

It remains to be investigated whether our result can be generalized to other matrilocal populations. It is likely to be so, as high endogamy has been reported by anthropologists as a general feature of matrilocal populations [6,19]. From our study, we can predict that other matrilocal populations will have higher genetic inbreeding levels than populations sharing the same environment, the same way of life, belonging to the same linguistic family (the criteria we used to select

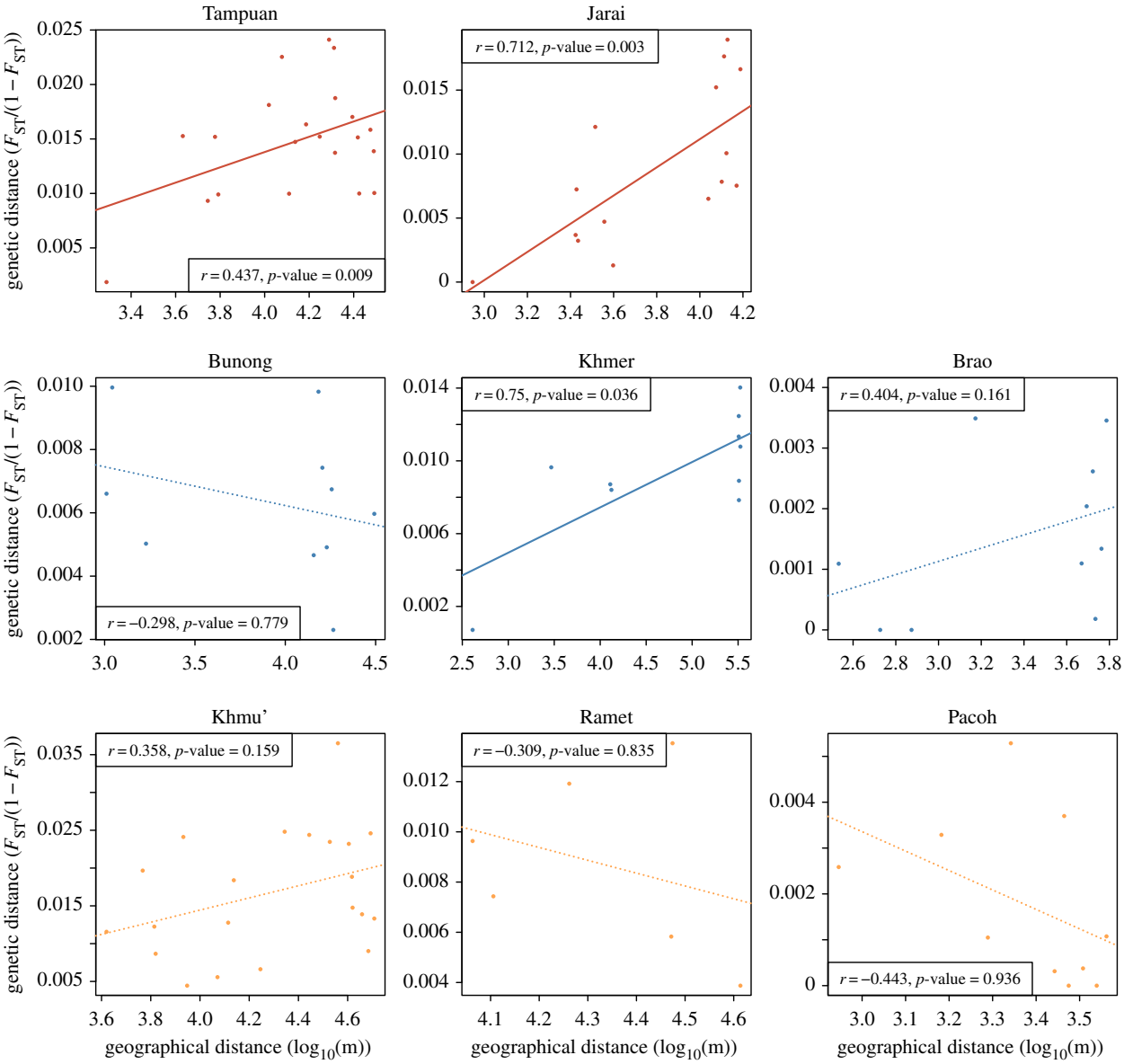

Figure 4. Relationship between genetic distance and geographical distance between villages in each population. M populations are in red (top row), C populations in blue (middle row) and P populations in yellow (bottom row). The Pearson's correlation coefficient and the *p*-value estimated by Mantel test for each population are given. Continuous lines represent regression lines associated with a significant Mantel test (*p*-value < 0.05). Dotted lines represent regression lines associated with a non-significant Mantel test. (Online version in colour.)

our populations to study) but having different social organizations. If such a prediction holds, high inbreeding could become an informative marker of matrilocality for ancient DNA studies trying to decipher the social organization of past human populations.

4. Methods

(a) Data collection

(i) Sampled populations

Twelve populations from Cambodia and Laos were sampled in 57 villages during three field missions carried out between 2011 and 2012: the Tampuan, Jarai, Kacho', Bunong, Khmer, Brao and Kreung from Cambodia and the Khmu', Ramet, Ta-oih, Pacoh and Prai from Laos (table 1). The populations were chosen for their differences in social organization. Most of them have been the focus of ethnographic works, describing in detail their social organization. The Tampuan, Jarai, Kacho' and Prai have matrilineal descent and matrilocal residence [18,37,38,48], the Khmu', Ramet, Ta-oih and Pacoh have patrilineal descent and patrilocal residence [38–40,46,47], the Bunong and the Khmer have cognatic descent and matrilocal residence [41,43,44,49], the Brao and Kreung have cognatic descent and multilocal residence [38,42,43,45]. Our previous analysis of ethno-demographic data collected in these populations [26] has shown that the four cognatic populations actually had comparable percentages of matrilocal couples (estimated to 43–48% of the exogamous couples), and that these matrilocal couples outnumbered the percentages of patrilocal couples (estimated to 13–38% of the exogamous couples). This shows that the final settlement of couples in the two multilocal populations is most often located in the wife's natal village. Consequently, we grouped these four populations into a single group of cognatic populations with predominant matrilocal residence. We refer in this paper to the matrilineal and matrilocal populations, to the cognatic populations with predominant matrilocal residence, and to the patrilineal and patrilocal populations respectively as M, C and P populations. All these populations belong to the Austro-Asiatic linguistic family, except the Jarai that speak an Austronesian language.

Note that in these Southeast Asian populations, the village is an important social unit [18,37,40]. In the case of the matrilineal

and patrilineal populations under study, each village usually comprises families belonging to several clans. Some of the 57 villages that were integrated in this study were ancient and stable in time, with, for example, a great tree marking symbolically its location. Others had been moving, either in line with a traditional practice [37] or because of recent political changes in Cambodia and Laos in the past two generations [37,40]. The members of each village usually know the history of the village [37] and our ethno-demographic data collection allowed us, while in the field, to exclude from our sampling, villages for which social integrity had been lost due to political events in the past two generations.

(ii) Ethno-demographic interviews

We interviewed 532 individuals, conjointly with their spouse, and collected ethno-demographic information (place of birth, village of residence; see [26] for details) about them and their family members (parents, grandparents, siblings, children and their respective spouses). This procedure allowed us to gather ethno-demographic information for 3530 couples.

(iii) DNA samples

Seven hundred and fifty-three individuals with all four grandparents from the same population were studied. We collected two saliva samples for each individual (4 ml each). Samples were kept in equivalent volume of lysis buffer with 800 µl of 10% SDS and 20 µl of proteinase K (20 mg ml^{-1}). DNA was extracted from saliva samples using a standard ethanol precipitation protocol [56]. All participants provided written informed consents and the study was approved by the National Ethic Comities for Health Research in Cambodia and Laos as well as by the Comité Opérationnel pour l'Ethique (CNRS, France).

(iv) SNP genotyping

Samples were genotyped on Illumina Omni1 (529 individuals) and Omni2.5 SNP arrays (224 individuals). SNPs present on both chips were retained, leading to a dataset of 701 163 SNPs for 753 individuals. After quality control (electronic supplementary material, figure S2), the dataset contained 598 764 SNPs for 743 individuals.

We used the method described in Conomos *et al*. [57] in order to check if any siblings were present in the dataset. We removed 24 individuals in order to get a sibling-free dataset (which will be used when estimating the genetic inbreeding coefficients). This dataset contained 598 764 SNPs genotyped for 719 individuals.

In addition, we prepared a dataset excluding first and second-degree relationships in order to estimate effective sizes, F_{ST}, isolation-by-distance patterns, and allelic frequencies (necessary for the estimation of genetic inbreeding coefficients). To do so, first- and second-degree relationships were inferred using KING v. 2.1.6 [58]. Two hundred and thirty individuals were removed to generate this first- and second-degree relationships-free dataset, containing 598 764 SNPs genotyped for 489 individuals.

(b) Data analysis

(i) Selection of populations with similar effective population sizes

Firstly, we checked that all the studied populations have comparable effective sizes since this parameter is known to influence inbreeding levels [59], with smaller effective population sizes associated with higher inbreeding. We estimated the effective population size of each population using the method described in Auton & McVean [60] (electronic supplementary material, table S4). Effective population sizes were compared between populations by a Welch's *t*-test with a Bonferroni correction for multiple testing. Among all studied populations, Prai was the only population with a significantly lower effective population size compared to all other populations (7182 compared to 13 228 (s.d. $\pm$ 1403) on average for the other populations; *p*-values < 0.05; electronic supplementary material, tables S4 and S5). Consequently, the Prai population was not included in the analyses presented below (however, similar results and conclusions were reached when this population was included; see electronic supplementary material, figure S3 for a graphical summary of these results). The final dataset included 11 populations, with 598 764 SNPs genotyped for 675 individuals for the sibling-free dataset and 466 individuals for the first- and second-degree relationships-free dataset. The ethno-demographic dataset included 495 ethno-demographic interviews providing information for 3261 couples (table 1).

(ii) Village endogamy estimation

A full description of the post-marital residence patterns for each population under study is provided in our previous study [26]. Here, for each population, we estimated the proportion of couples for which both spouses were born in the same village (village endogamy rate). We then used logistic regression to assess the influence of social organization (M, P and C) on the probability that individuals marry partners from the same village with the 'glmer' function in 'lme4' package v. 1.1-9 in R [61]. We incorporated population as a fixed effect and village of residence and family as random effects in this model in order to account for potential sampling bias. *P*-values were estimated with the 'lsmeans' function in the 'lsmeans' package v. 2.27-2 in R.

(iii) Genetic inbreeding coefficients estimation

We used the FEstim software [36] integrated in the FSuite pipeline [35] to estimate the inbreeding coefficient and the parental mating types of each individual. Genetic maps were retrieved from the shapeit homepage [62]. The --hotspots option with hg19 build was used when creating the 100 submaps. Allele frequencies were estimated separately for each population (using the first- and second-degree relationships-free dataset).

Then, we used a mixed linear model to assess the influence of social organization (M, P and C) on inbreeding coefficients. We incorporated population as a fixed effect and village of residence and family as random effects in this model in order to take into account potential sampling bias. *P*-values were estimated with the 'lsmeans' function in the 'lsmeans' package v. 2.27-2 in R.

Spearman's correlation coefficient between village endogamy rate and mean inbreeding coefficient was estimated at the population level.

(iv) Isolation-by-distance pattern

Fixation indices (F_{ST}) between villages within each population were estimated using Genepop 4.7 [63]. Only villages with a minimum of five individuals were included in this analysis. Populations with less than four villages filling this condition were removed from the analysis. As such, Kacho', Kreung and Ta-oih were excluded from this analysis. The dataset was then pruned using Plink 1.9 [64] –indep-pairwise option with a window size of 50 SNPs, sliding by five SNPs and a pairwise r^2 threshold of 0.5 to create a dataset of 252 680 SNPs in low linkage disequilibrium. Negative F_{ST} were changed to 0. A linear regression model was fitted with genetic distance between villages estimated by $F_{ST}/(1 - F_{ST})$ as the dependent variable and geographical distance in metre (decimal logarithmic value) as the explanatory variable for each population. Statistical significance of the correlation between genetic distances and geographical distances was evaluated using a Mantel test with 10 000 permutations.

All statistical analyses were performed in R v. 3.2.2 [61].

Ethics. All participants provided written informed consents and the study was approved by the National Ethic Comities for Health

Research in Cambodia and Laos as well as by the Comité Opérationnel pour l'Ethique (CNRS).

Data accessibility. Genotyping data has been deposited at the European Genome-phenome Archive (EGA, https://ega-archive.org), which is hosted by the EBI and the CRG, under accession number EGAS00001003727.

Authors' contributions. R.C., S.P. and B.T. initiated the project; R.C., S.P., R.L., S.L., O.E., F.B., G.D. and C.M. contributed to the ethno-demographic and genetic sampling; S.L. and C.M. carried out the laboratory work; G.L. and R.L. analysed the data; G.L., R.C. and S.P. wrote the paper. All authors gave final approval for publication.

Competing interest. The authors declare no competing interests.

Funding. This work was supported by the ANR SoGen (JC09_441218) grant.

Acknowledgements. We thank all the volunteers for their participation, as well as Yves Buisson from the Institut de la Francophonie pour la Médecine Tropicale (IFMT) in Laos for their help in facilitating fieldwork, Hervé Pedry for his advice regarding regression models and Michael Houseman, Evelyne Heyer and Paul Verdu for helpful discussions.

References

1. Godelier M. 2004 *Métamorphoses de la parenté*. Paris, France: Fayard.
2. Murdock GP, White DR. 1969 Standard cross-cultural sample. *Ethnology* **8**, 329–369. (doi:10.2307/3772907)
3. Fox R. 1972 *Anthropologie de la parenté. Une analyse de la consanguinité et de l'alliance*. Paris, France: Gallimard.
4. Mathieu N-C. 2007 *Une maison sans fille est une maison morte*. Paris, France: Maison des sciences de l'homme.
5. Richards AI. 1950 Some types of family structure amongst the Central Bantu. In *African systems of kinship and marriage* (eds A Radcliffe-Brown, D Forde), pp. 207–251. London, UK: Oxford University Press.
6. Kloos P. 1963 Matrilocal residence and local endogamy: environmental knowledge or leadership. *Am. Anthropol.* **65**, 854–862. (doi:10.1525/aa.1963.65.4.02a00050)
7. Mattison SM, Quinlan RJ, Hare D. 2018 The expendable male hypothesis. *bioRxiv*. (doi:10.1101/473942)
8. Schneider DM, Gough K. 1961 *Matrilineal kinship*. Berkeley, CA: University of California Press.
9. Schneider DM. 1961 The distinctive features of matrilineal descent groups. In *Matrilineal kinship* (eds DM Schneider, K Gough), pp. 1–29. Berkeley, CA: University of California Press.
10. Mattison S. 2011 Evolutionary contributions to solving the 'matrilineal puzzle'. *Hum. Nat.* **22**, 64–88. (doi:10.1007/s12110-011-9107-7)
11. Fortunato L. 2012 The evolution of matrilineal kinship organization. *Proc. R. Soc. B* **279**, 4939–4945. (doi:10.1098/rspb.2012.1926)
12. Holden CJ, Sear R, Mace R. 2003 Matriliny as daughter-biased investment. *Evol. Hum. Behav.* **24**, 99–112. (doi:10.1016/S1090-5138(02)00122-8)
13. Hamilton WD. 1964 The genetical evolution of social behaviour, I. *J. Theor. Biol.* **7**, 1–16. (doi:10.1016/0022-5193(64)90038-4)
14. Trivers RL, Willard DE. 1973 Natural selection of parental ability to vary the sex ratio of offspring. *Science* **179**, 90–92. (doi:10.1126/science.179.4068.90)
15. Holden C, Mace R. 2003 Spread of cattle led to the loss of matrilineal descent in Africa: a coevolutionary analysis. *Proc. R. Soc. Lond. B* **270**, 2425–2433. (doi:10.1098/rspb.2003.2535)
16. Pasternak B. 1976 *Introduction to kinship and social organization*. Englewood Cliffs, NJ: Prenctice-Hall.
17. Gough K. 1961 Variation in residence. In *Matrilineal kinship* (eds DM Schneider, K Gough), pp. 545–576. Berkeley, CA: University of California Press.
18. Dournes J. 1972 *Coordonnées, structures jörai familiales et sociales*. Paris, France: Muséum national d'Histoire naturelle (Travaux et Mémoires de l'Institut d'Ethnologie n°77).
19. Murdock GP. 1949 *Social structure*. London, UK: Macmillan.
20. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2001 Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nat. Genet.* **29**, 20–21. (doi:10.1038/ng711)
21. Hamilton G, Stoneking M, Excoffier L. 2005 Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc. Natl Acad. Sci. USA* **102**, 7476–7480. (doi:10.1073/pnas.0409253102)
22. Tumonggor MK, Karafet TM, Downey S, Lansing JS, Norquest P, Sudoyo H, Hammer MF, Cox MP. 2014 Isolation, contact and social behavior shaped genetic diversity in West Timor. *J. Hum. Genet.* **59**, 494–503. (doi:10.1038/jhg.2014.62)
23. Besaggio D, Fuselli S, Srikummool M, Kampuansai J, Castrì L, Tyler-Smith C, Seielstad M, Kangwanpong D, Bertorelle G. 2007 Genetic variation in Northern Thailand hill tribes: origins and relationships with social structure and linguistic differences. *BMC Evol. Biol.* **7**, 1–10. (doi:10.1186/1471-2148-7-S2-S12)
24. Gunnarsdóttir ED, Nandineni MR, Li M, Myles S, Gil D, Pakendorf B, Stoneking M. 2011 Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat. Commun.* **2**, 226–228. (doi:10.1038/ncomms1235)
25. Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, Biswas S, Thangaraj K, Singh L, Reddy BM. 2006 Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genet.* **2**, 420–424. (doi:10.1371/journal.pgen.0020053)
26. Ly G *et al.* 2018 Residence rule flexibility and descent groups dynamics shape uniparental genetic diversities in South East Asia. *Am. J. Phys. Anthropol.* **165**, 480–491. (doi:10.1002/ajpa.23374)
27. Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, Austerlitz F, Heyer E. 2007 From social to genetic structures in Central Asia. *Curr. Biol.* **17**, 43–48. (doi:10.1016/j.cub.2006.10.058)
28. Heyer E, Chaix R, Pavard S, Austerlitz F. 2012 Sex-specific demographic behaviours that shape human genomic variation. *Mol. Ecol.* **21**, 597–612. (doi:10.1111/j.1365-294X.2011.05406.x)
29. Verdu P *et al.* 2013 Sociocultural behavior, sex-biased admixture, and effective population sizes in central African pygmies and non-pygmies. *Mol. Biol. Evol.* **30**, 918–937. (doi:10.1093/molbev/mss328)
30. Ségurel L *et al.* 2008 Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet.* **4**, e1000200. (doi:10.1371/journal.pgen.1000200)
31. Guillot EG, Hazelton ML, Karafet TM, Lansing JS, Sudoyo H, Cox MP. 2015 Relaxed observance of traditional marriage rules allows social connectivity without loss of genetic diversity. *Mol. Biol. Evol.* **32**, 2254–2262. (doi:10.1093/molbev/msv102)
32. Marchi N, Mennecier P, Georges M, Lafosse S, Hegay T, Dorzhu C, Chichlo B, Ségurel L, Heyer E. 2018 Close inbreeding and low genetic diversity in Inner Asian human populations despite geographical exogamy. *Sci. Rep.* **8**, 1–10. (doi:10.1038/s41598-018-27047-3)
33. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012 Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292. (doi:10.1016/j.ajhg.2012.06.014)
34. Wright S. 1943 Isolation by distance. *Genetics* **28**, 114–138.
35. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. 2014 FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* **30**, 1940–1941. (doi:10.1093/bioinformatics/btu149)
36. Leutenegger A-L, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA. 2003 Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.* **73**, 516–523. (doi:10.1086/378207)
37. Bourdier F. 2006 *The mountain of precious stones: Ratanakiri, Cambodia: essays in social anthropology*. Phnom Penh, Cambodia: Center for Khmer Studies.
38. LeBar FM, Hickey GC, Musgrave JK. 1964 *Ethnic groups of mainland Southeast Asia*. New Haven, CT: Human Relations Area Files Press.

39. Izikowitz KG. 1951 *Lamet hill peasants in French Indochina*. Göteborg, Sweden: Etnografiska Museet.
40. Evrard O. 2006 *Chroniques des cendres*. Paris, France: IRD Edition.
41. Ebihara M. 1977 Residence patterns in a Khmer peasant village. *Ann. N. Y. Acad. Sci.* **293**, 51–68. (doi:10.1111/j.1749-6632.1977.tb41805.x)
42. Baird IG. 2008 Various forms of colonialism?: the social and spatial reorganisation of the Brao in southern Laos and northeastern Cambodia. PhD thesis, University of British Columbia.
43. UNDP Cambodia. 2010 *Kreung ethnicity documentation of customary rules indigenous people in Pu-Trou village*. Phnom Penh, Cambodia: UN Development Programme.
44. Martel G. 1975 *Lovea. Village des environs d'Angkor. Aspects démographiques, économiques et sociologiques du monde rural cambodgien dans la province de Siem-Réap*. Paris, France: Ecole française d'Extrême-Orient.
45. Matras-Troubetzkoy J. 1983 *Un village en foret. L'essartage chez les brou du Cambodge*. Paris, France: Peeters.
46. Lindell K, Samuelsson R, Tayanin D. 1979 Kinship and marriage in northern Kammu villages: the kinship model. *Sociologus* **29**, 60–84.
47. Schmutz J. 2013 The Ta'oi language and people. *MonKhmer Stud.* **42**, i–xiii.
48. Dessaint WY. 1981 The T'in (Mal), dry rice cultivators of Northern Thailand. *J. Siam Soc.* **69**, 107–137.
49. Ledgerwood JL. 1995 Khmer kinship—the matriliny matriarchy myth. *J. Anthropol. Res.* **51**, 247–261. (doi:10.1086/jar.51.3.3630360)
50. Bittles AH, Hamamy H. 2010 Endogamy and consanguineous marriage in Arab populations. In *Genetic disorders among Arab populations* (ed. AS Teebi), pp. 85–108. Berlin, Germany: Springer.
51. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L. 2015 High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.* **5**, 17453. (doi:10.1038/srep17453)
52. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF. 2010 Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**, e13996. (doi:10.1371/journal.pone.0013996)
53. McQuillan R *et al*. 2008 Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372. (doi:10.1016/j.ajhg.2008.08.007)
54. Yang X, Al-Bustan S, Feng Q, Guo W, Ma Z, Marafie M, Jacob S, Al-Mulla F, Xu S. 2014 The influence of admixture and consanguinity on population genetic diversity in Middle East. *J. Hum. Genet.* **59**, 615–622. (doi:10.1038/jhg.2014.81)
55. Leutenegger A-L, Sahbatou M, Gazal S, Cann H, Génin E. 2011 Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *Eur. J. Hum. Genet.* **19**, 583–587. (doi:10.1038/ejhg.2010.205)
56. Quinque D, Kittler R, Kayser M, Stoneking M, Nasidze I. 2006 Evaluation of saliva as a source of human DNA for population and association studies. *Anal. Biochem.* **353**, 272–277. (doi:10.1016/j.ab.2006.03.021)
57. Conomos MP, Reiner AP, Weir BS, Thornton TA. 2016 Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148. (doi:10.1016/j.ajhg.2015.11.022)
58. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010 Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873. (doi:10.1093/bioinformatics/btq559)
59. Jacquard A. 1968 Evolution des populations d'effectif limité. *Popul. (French Edn)* **23**, 279–300. (doi:10.2307/1527488)
60. Auton A, McVean G. 2007 Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219–1227. (doi:10.1101/gr.6386707.scheme)
61. R Development Core Team. 2015 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/.
62. Delaneau O, Marchini J, Zagury J-F. 2012 A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181. (doi:10.1038/nmeth.1785)
63. Rousset F. 2008 Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106. (doi:10.1111/j.1471-8286.2007.01931.x)
64. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7. (doi:10.1186/s13742-015-0047-8)

Received: 4 October 2019 | Revised: 11 December 2020 | Accepted: 7 January 2021

DOI: 10.1111/1755-0998.13325

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Complex genetic admixture histories reconstructed with Approximate Bayesian Computation

Cesar A. Fortes-Lima[1,2] | Romain Laurent[1] | Valentin Thouzeau[3,4] | Bruno Toupance[1] | Paul Verdu[1]

[1]UMR7206 Eco-anthropologie, CNRS, Muséum National d'Histoire Naturelle, Université de Paris, Paris, France

[2]Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3]UMR7534 Centre de Recherche en Mathématiques de la Décision, CNRS, Université Paris-Dauphine, PSL University, Paris, France

[4]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS, Paris, France

Correspondence
Paul Verdu, Musée de l'Homme, 17, place du Trocadéro, 75016 Paris, France.
Email: paul.verdu@mnhn.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: 15-CE32-0009-01

Abstract

Admixture is a fundamental evolutionary process that has influenced genetic patterns in numerous species. Maximum-likelihood approaches based on allele frequencies and linkage-disequilibrium have been extensively used to infer admixture processes from genome-wide data sets, mostly in human populations. Nevertheless, complex admixture histories, beyond one or two pulses of admixture, remain methodologically challenging to reconstruct. We developed an Approximate Bayesian Computation (ABC) framework to reconstruct highly complex admixture histories from independent genetic markers. We built the software package MetHis to simulate independent SNPs or microsatellites in a two-way admixed population for scenarios with multiple admixture pulses, monotonically decreasing or increasing recurring admixture, or combinations of these scenarios. MetHis allows users to draw model-parameter values from prior distributions set by the user, and, for each simulation, MetHis can calculate numerous summary statistics describing genetic diversity patterns and moments of the distribution of individual admixture fractions. We coupled MetHis with existing machine-learning ABC algorithms and investigated the admixture history of admixed populations. Results showed that random forest ABC scenario-choice could accurately distinguish among most complex admixture scenarios, and errors were mainly found in regions of the parameter space where scenarios were highly nested, and, thus, biologically similar. We focused on African American and Barbadian populations as two study-cases. We found that neural network ABC posterior parameter estimation was accurate and reasonably conservative under complex admixture scenarios. For both admixed populations, we found that monotonically decreasing contributions over time, from Europe and Africa, explained the observed data more accurately than multiple admixture pulses. This approach will allow for reconstructing detailed admixture histories when maximum-likelihood methods are intractable.

KEYWORDS
admixture, Approximate Bayesian Computation, inference, machine-learning, population genetics

Cesar A. Fortes-Lima and Romain Laurent are joint first authors

1 | INTRODUCTION

Hybridization between species and admixture between populations are powerful mechanisms influencing biological evolution. Genetic admixture patterns have thus been extensively studied to reconstruct past population migrations or range-shifts and understand admixture-related adaptation such as heterosis or post-admixture selection (Brandenburg et al., 2017; Hellenthal et al., 2014; Skoglund et al., 2015).

A long history of statistical developments in population genetics provided tools to identify and describe admixture patterns from genetic data (Bernstein, 1931; Cavalli-Sforza & Bodmer, 1971; Chakraborty & Weiss, 1988; Falush et al., 2003; Long, 1991; Patterson et al., 2012). They enabled inferring the ancestral origins of admixed populations or investigation of adaptive introgression in numerous species (e.g., Martin et al., 2013; Patin et al., 2017; Stryjewski & Sorenson, 2017).

1.1 | Maximum-likelihood methods to reconstruct admixture histories

Two classes of maximum-likelihood (ML) methods have been extensively deployed to infer admixture histories from genetic data. They rely on the moments of allelic frequency spectrum divergences among populations (Lipson et al., 2013; Patterson et al., 2012; Pickrell & Pritchard, 2012), and on admixture linkage disequilibrium (LD) patterns—the distribution of LD within the admixed chunks of DNA in the genomes of admixed individuals inherited from members of the source populations (Guan, 2014; Chimusa et al., 2018; Gravel, 2012; Hellenthal et al., 2014; Loh et al., 2013; Moorjani et al., 2011). Notably, Gravel (2012) developed an approach to fit the observed curves of admixture LD decay to those theoretically expected under admixture models involving one or two pulses of historical admixture. These approaches significantly improved our understanding of past admixture histories using genetic data (e.g., Baharian et al., 2016; Martin et al., 2013).

Despite these major achievements, ML methods for admixture history inference suffer from inherent limitations acknowledged by the authors (Gravel, 2012; Hellenthal et al., 2014; Lipson et al., 2013). First, most ML approaches can only consider one or two pulses of admixture in the history of the admixed population. Nevertheless, admixture processes are often expected to be much more complex, and it is not yet clear how ML methods behave when they consider only simplified versions of the true admixture history underlying the observed data (Gravel, 2012; Hellenthal et al., 2014; Lipson et al., 2013; Loh et al., 2013; Medina et al., 2018; Ni et al., 2019). Second, it is possible to statistically compare ML values obtained from fitting models with different parameters to the observed data, as a guideline to find the "best" model. Nevertheless, formal statistical comparison of the success or failure of competing models to explain the observed data is often out of reach of ML approaches (Foll et al., 2015; Gravel, 2012; Ni et al., 2019). Finally, admixture-LD methods, in particular, rely on fine mapping of local ancestry segments in individual genomes and thus require substantial amounts of genomic data, and, sometimes, accurate phasing, which remain difficult in numerous empirical data sets from most non-model organisms.

1.2 | Approximate Bayesian Computation demographic inference

Approximate Bayesian Computation (ABC) approaches (Beaumont et al., 2002; Tavaré et al., 1997) represent a promising alternative to infer complex admixture histories from genetic data. Indeed, ABC has been successfully used previously to formally test alternative demographic scenarios hypothesized to be underlying observed genetic patterns, and to estimate, a posteriori, the parameters of the winning models, when ML methods could not operate (Boitard et al., 2016; Fraimout et al., 2017; Verdu et al., 2009).

ABC scenario-choice and posterior-parameter estimation rely on comparing observed summary statistics to the same set of statistics calculated from simulations produced under competing demographic scenarios (Beaumont et al., 2002; Blum & François, 2010; Csilléry et al., 2012; Pudlo et al., 2016; Sisson et al., 2018; Wegmann et al., 2009). Each simulation, and corresponding vector of summary statistics, is produced using model-parameters drawn randomly from prior distributions explicitly specified by the user. This makes ABC a priori particularly well-suited to investigate highly complex historical admixture scenarios for which likelihood functions are very often intractable, but for which genetic simulations are feasible (Buzbas & Verdu, 2018; Gravel, 2012; Pritchard et al., 1999; Verdu & Rosenberg, 2011).

1.3 | An ABC framework for reconstructing complex admixture histories

In this paper, we show how ABC can be successfully applied to reconstruct, from genetic data, highly complex admixture histories beyond models with a single or two pulses of admixture classically explored with ML methods. To do so, we propose a novel forward-in-time genetic data simulator and a set of parameter-generator and summary statistic calculation tools, embedded in an open source C software package called MetHis. It simulates genetic data from independent SNP or microsatellite loci under any two source-populations versions of the Verdu and Rosenberg (2011) general model of admixture; and is adapted to conduct ABC inferences with existing machine-learning ABC tools implemented in R (R Development Core Team, 2020).

We show that our MetHis-ABC framework can accurately distinguish major classes of complex historical admixture models, involving multiple admixture-pulses, recurring increasing or decreasing admixture over time, or combinations of these models, and provides conservative posterior parameter inference under the chosen models. Furthermore, we introduce the quantiles and higher moments of

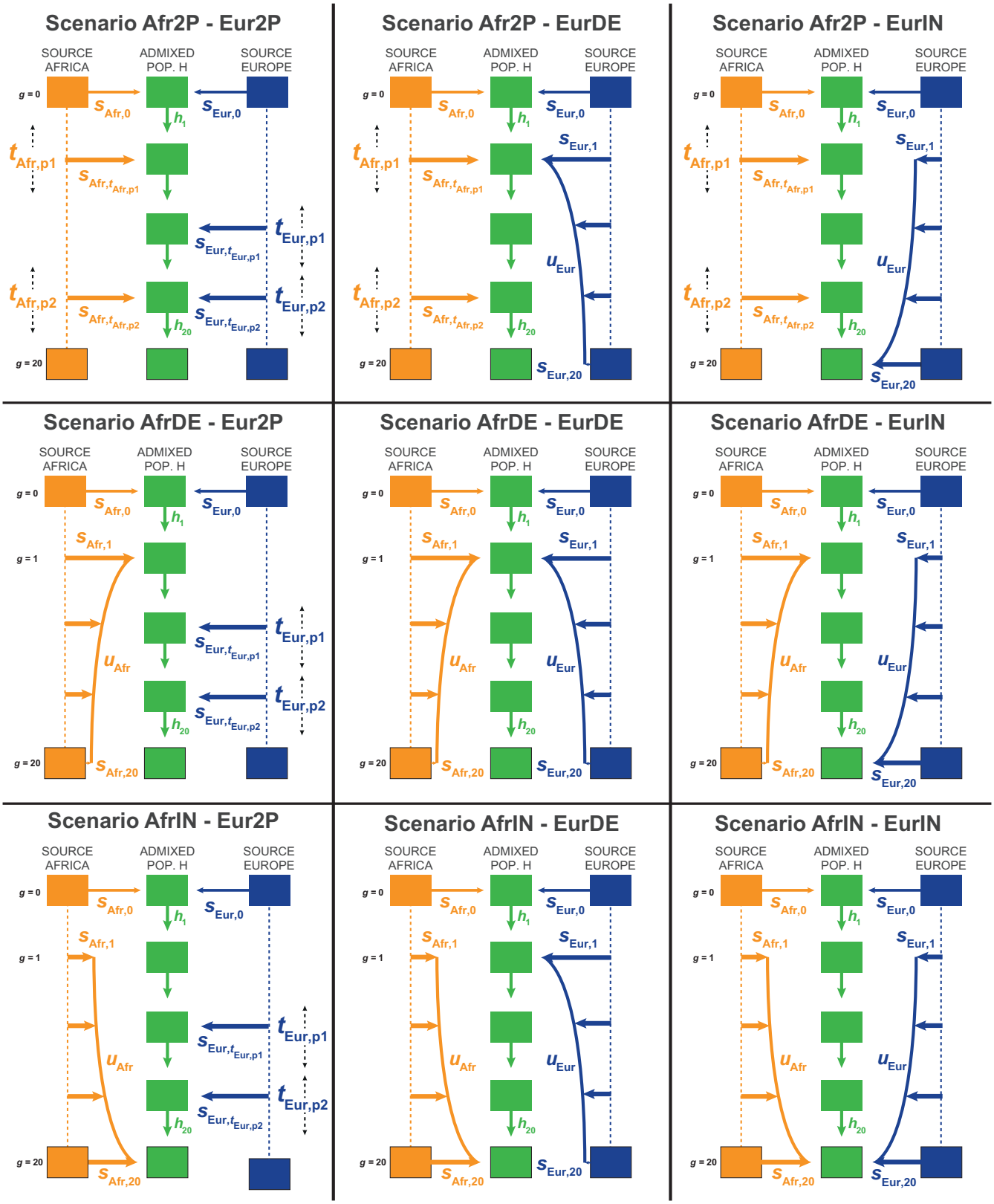

FIGURE 1 Nine competing scenarios for reconstructing the admixture history of African American ASW or Barbadian ACB populations descending from West European and West sub-Saharan African source populations during the Transatlantic Slave Trade. "EUR" represents the Western European and "AFR" represents the West Sub-Saharan African source populations for the admixed population H. See Table 1 and Section 2 for descriptions of the parameters of the scenarios

the distribution of admixture fractions in the admixed population as highly informative summary statistics for ABC scenario-choice and posterior-parameter estimation.

We exemplify our approach by reconstructing the complex admixture histories underlying observed genetic patterns separately for the African American (ASW) and Barbadian (ACB) populations. Both populations are known to be admixed populations of European and African descent in the context of the Transatlantic Slave Trade (TAST), whose detailed histories of admixture remain largely unknown (e.g., Baharian et al., 2016; Martin et al., 2017). In this case-study, we find that the ACB and ASW populations' admixture histories are much more complex than previously inferred, and further reveal the diversity of histories undergone by these admixed populations during the TAST in the Americas.

2 | MATERIALS AND METHODS

We evaluated how ABC scenario-choice and posterior parameter estimation performed for reconstructing highly complex historical admixture processes from genetic data. To do so, we chose to work under the two source-populations version of the general mechanistic model of Verdu and Rosenberg (2011) briefly presented in Figure S1. We introduce a novel software, METHIS, for genetic data simulation and summary statistic calculation for machine-learning ABC inference under this general model (Note S1).

We conducted our proof of concept considering nine competing scenarios of complex admixture histories involving multiple admixture pulses, recurring decreasing or increasing admixture, and combinations of these processes (Figure 1, Table 1). We explored

TABLE 1 Parameter prior distributions for simulation with METHIS and Approximate Bayesian Computation historical inference

Parameter names	Parameter description	Prior distribution	Condition	Scenarios
$s_{Afr,0}$ $s_{Eur,0} = 1 - s_{Afr,0}$	Afr. source introgression rate at founding of H	*Uniform* [0,1]	—	All Scenarios
$t_{Afr,p1}$ $t_{Afr,p2}$	Times of the Afr. source introgression pulses p1 and p2	*Uniform* [0,20]	$t_{Afr,p1} \neq t_{Afr,p2}$	Afr2P Scenarios
$s_{Afr,t\ Afr,p1}$ $s_{Afr,t\ Afr,p2}$	Afr. source introgression rates of pulses Afr,p1 and Afr,p2	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	Afr2P Scenarios
$t_{Eur,p1}$ $t_{Eur,p2}$	Times of the Eur. source introgression pulses p1 and p2	*Uniform* [0,20]	$t_{Eur,p1} \neq t_{Eur,p2}$	Eur2P Scenarios
$s_{Eur,t\ Eur,p1}$ $s_{Eur,t\ Eur,p2}$	Eur. source introgression rates of pulses Eur,p1 and Eur,p2	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	Eur2P Scenarios
$s_{Afr,1}$	Afr. source introgression rate at the first generation after founding	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrDE Scenarios
$s_{Afr,20}$	Afr. source introgression rate in the present	*Uniform* [0, $s_{Afr,1}$/3]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrDE Scenarios
u_{Afr}	Steepness of the decrease in Afr. source introgression rates	*Uniform* [0,0.5]	—	AfrDE Scenarios
$s_{Eur,1}$	Eur. source introgression rate at the first generation after founding	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurDE Scenarios
$s_{Eur,20}$	Eur. source introgression rate in the present	*Uniform* [0, $s_{Eur,1}$/3]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurDE Scenarios
u_{Eur}	Steepness of the decrease in Eur. source introgression rates	*Uniform* [0,0.5]	—	EurDE Scenarios
$s_{Afr,1}$	Afr. source introgression rate at the first generation after founding	*Uniform* [0, $s_{Afr,20}$/3]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrIN Scenarios
$s_{Afr,20}$	Afr. source introgression rate in the present	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	AfrIN Scenarios
u_{Afr}	Steepness of the increase in Afr. source introgression rates	*Uniform* [0,0.5]	—	AfrIN Scenarios
$s_{Eur,1}$	Eur. source introgression rate at the first generation after founding	*Uniform* [0, $s_{Eur,20}$/3]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurIN Scenarios
$s_{Eur,20}$	Eur. source introgression rate in the present	*Uniform* [0,1]	For all g, $h_g = 1 - s_{Afr,g} - s_{Eur,g}$ in [0,1]	EurIN Scenarios
u_{Eur}	Steepness of the increase in Eur. source introgression rates	*Uniform* [0,0.5]	—	EurIN Scenarios

Parameter list corresponds to the nine competing historical admixture scenarios described in Figure 1 and Section 2.

the recent admixture history of two enslaved-African descendant populations in the Americas with genome-wide independent SNPs. Beyond this work, the METHIS-ABC framework can readily be used to study numerous histories of complex admixture using independent SNP or microsatellite markers (Note S1).

2.1 | Nine competing complex admixture scenarios

2.1.1 | Founding of the admixed population H

For all scenarios (Figure 1, Table 1), we chose a fixed time for the founding (generation 0, forward-in-time) of the target admixed population H occurring 21 generations before present, with admixture proportions $s_{Afr,0}$ and $s_{Eur,0}$ from either source population S respectively, African and European in our case, with $s_{Afr,0} + s_{Eur,0} = 1$, and $s_{Afr,0}$ in [0,1]. This duration corresponds roughly to the first arrival of European permanent settlers in the Americas in the late 15th century, considering 20 or 25 years per generation and the sampled generation born in the 1980s. Note that simulations with a parameter $s_{Afr,0}$ close to 0, or alternatively 1, corresponded to the founding of the population H from one source population only, therefore delaying the first "real" genetic admixture event to the next admixture event. Following founding, we considered three alternative scenarios for the admixture contribution of each source population S separately.

2.1.2 | Admixture-pulse(s) scenarios

For a given source population S, African or European, scenarios *S-2P* considered two possible pulses of admixture into population H occurring respectively at time $t_{S,p1}$ and $t_{S,p2}$ distributed in [1,20] with $t_{S,p1} \neq t_{S,p2}$, with associated admixture proportion $s_{S,tS,p1}$ and $s_{S,tS,p2}$ in [0,1] satisfying, at all times t, $\sum_{S \in (Afr,Eur)} s_{S,t} \leq 1$ (Figure 1, Table 1). Note that for one of either $s_{S,t}$ values close to 0, the two-pulses scenarios were equivalent to single-pulse scenarios after the founding of H. Furthermore, for both $s_{S,t}$ values close to 0, scenarios *S-2P* were nested with scenarios where only the founding admixture pulse 21 generations ago was the source of genetic admixture. Alternatively, $s_{S,t}$ parameter values close to 1 considered a virtual complete replacement of population H by population S at that time, thus obliterating all previous admixture events.

2.1.3 | Recurring decreasing admixture scenarios

For a given source population S, scenarios *S-DE* considered a recurring monotonically decreasing admixture from population S at each generation between generation 1 (after founding at generation 0) and generation 20 (sampled population) (Figure 1, Table 1). In these scenario, $s_{S,g}$, with g in [1,20], were the discrete numerical solutions of a rectangular hyperbola function over the 20 generations of the admixture process until present, as described in Note S2. In brief, this function is determined by the parameter u_S, the "steepness" of the curvature of the decrease, in [0,1/2], $s_{S,1}$, the admixture proportion from population S at generation 1 (after founding), in [0,1], and $s_{S,20}$, the last admixture proportion in the present, in [0,$s_{S,1}$/3]. Note that we chose the boundaries for $s_{S,20}$ in order to reduce the parameter space and nestedness among competing scenarios, by explicitly forcing scenarios *S-DE* into substantially decreasing admixture processes. Furthermore, note that parameter u_S values close to 0 created pulse-like scenarios of intensity $s_{S,1}$ occurring immediately after founding, followed by constant recurring admixture of intensity $s_{S,20}$ at each generation until present. Alternatively, parameter u_S values close to 1/2 created scenarios with linearly decreasing admixtures between $s_{S,1}$ and $s_{S,20}$ from population S at each generation after founding.

2.1.4 | Recurring increasing admixture scenarios

For a given source population S, scenarios *S-IN* mirrored the *S-DE* scenarios by considering instead a recurring monotonically increasing admixture from population S (Figure 1, Table 1). Here, $s_{S,g}$, with g in [1,20], were the discrete numerical solutions of the same function as in the *S-DE* decreasing scenarios (see above), flipped over time between generation 1 and 20. In these scenarios, $s_{S,20}$ was defined in [0,1] and $s_{S,1}$ in [0,$s_{S,20}$/3], and u, in [0,1/2], parametrized the "steepness" of the curvature of the increase. Note that *S-IN* scenarios were nested with pulse-like scenarios over the parameter space of u values, analogously to the nestedness of *S-DE* and pulse-like scenarios described above.

2.1.5 | Combining admixture scenarios from either source populations

We combined these three scenarios to obtain nine alternative scenarios for the admixture history of population H (Figure 1, Table 1), with the only condition that, at each generation g in [1,20], parameters satisfied $s_{Afr,g} + s_{Eur,g} + h_g = 1$, with h_g, in [0,1] being the remaining contribution of the admixed population H to itself at generation g.

Four scenarios (Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, and AfrIN-Eur2P) considered a mixture of pulse-like and recurring admixture from each source. Three scenarios (Afr2P-Eur2P, AfrDE-EurDE, and AfrIN-EurIN), considered symmetrical classes of admixture scenarios from either source. Two scenarios (AfrIN-EurDE and AfrDE-EurIN) considered mirroring recurring admixture processes. Importantly, this scenario design considered nested historical scenarios in specific parts of the parameter space.

2.2 | Forward-in-time simulations with METHIS

Simulation of independent genetic markers under highly complex admixture histories is often not trivial under the coalescent and using

classical existing software. Indeed, the coalescent generally assumes a different pedigree for each independent locus instead of a single pedigree having, in reality, produced all observed gene genealogies (see Wakeley et al., 2012). In this context, and because pedigrees are rarely known a priori, we developed METHIS, a C open-source software package available at https://github.com/romain-laurent/MetHis. METHIS simulates independent SNPs or microsatellite markers in an admixed population H under any version of the two source-populations general model from Verdu and Rosenberg (2011), and calculates summary statistics of interest to the study of complex admixture processes (Note S1).

2.2.1 | Simulating the admixed population, effective population size, and sampling individuals

At each generation, METHIS performs simple Wright-Fisher (Fisher, 1922; Wright, 1931) forward-in-time simulations, individual-centered, in a panmictic population of diploid effective size N_g. For a given individual in the population H at the following generation (g + 1), METHIS independently draws each parent from the source populations with probability $s_{S,g}$ (Figure 1, Table 1), or from population H with probability $h_g = 1 - \sum_{S \in (Afr,Eur)} s_{S,g}$, randomly builds a haploid gamete of independent markers for each parent, and pairs the two constructed gametes to create the new individual.

Here, we decided to neglect mutation over the 21 generations of admixture considered. This was reasonable when studying relatively recent admixture histories and considering independent genotyped SNP markers. For users interested in microsatellite variation and longer admixture histories, METHIS readily implements a standard general stepwise mutation model allowing for insertion or deletion (Estoup et al., 2002), with parameters set by the user (Note S1).

To focus on the admixture process itself without excessively inflating the parameter space, we considered, for each of the nine competing scenarios, the admixed population H with constant effective population size N_g = 1000 diploid individuals. Nevertheless, note that METHIS readily allows the user to parametrize, instead, stepwise or continuous changes in effective population size over time (Note S1).

After each simulation, we randomly drew individual samples matching sample-sizes in our observed data set (see Section 2.4.3). We sampled individuals until our sample set contained no individuals related at the first degree cousin within each population and between population H and either source population, based on explicit parental flagging during the last two generations of the simulations. Note that this is done to best mimic, a priori, the observed case-study data sets, but excluding related individuals is an option set by the user in METHIS (Note S1).

2.2.2 | Simulating source populations

METHIS, in its current form, does not allow simulating the source populations for the admixture process modeled in Verdu and Rosenberg (2011). Simulating source populations can be done separately using existing genetic data simulation software such as fastsimcoal2 sequential coalescent (Excoffier et al., 2013; Excoffier & Foll, 2011).

Another possibility to simulate source populations emerges if genetic data is already available for the known source populations, as it is the case in our case studies of enslaved-African descendants in the Americas (see Section 2.4.3). We considered here that the African and European source populations were very large populations at the drift-mutation equilibrium, accurately represented by the Yoruban YRI and British GBR data sets here investigated (see Section 2.4.3). Therefore, we first built two separate data sets each comprising 20,000 haploid genomes of 100,000 independent SNPs, each SNP being randomly drawn in the site frequency spectrum (SFS) observed for the YRI and GBR data sets respectively. These two data sets were used as fixed gamete reservoirs for the African and European sources separately, at each generation of the forward-in-time admixture process. From these reservoirs, we built an effective individual gene pool of diploid size N_g, by randomly pairing gametes avoiding selfing. These virtual source populations provided the parental pool for simulating individuals in the admixed population H with METHIS, at each generation. Thus, while our gamete reservoirs were fixed, the parental genetic pools were randomly built anew at each generation. Again, note that this is not necessary to the implementation of METHIS for investigating complex admixture histories; source populations can be simulated separately by the user at will.

2.3 | Summary statistics

METHIS is designed to work in an ABC inference framework and, thus, can calculate numerous summary statistics. A complete list of summary statistics can be found in Note S1. Below are the summary statistics considered in our case studies, in particular introducing the distribution of admixture fractions in population H, as summary statistics for ABC inference.

2.3.1 | The distribution of admixture fractions as a set of summary statistics

Most methods developed to estimate individual admixture fractions from genetic data (e.g., Alexander et al., 2009), are computationally intensive, and are thus difficult to iterate over large sets of simulated genetic data. This explains why they have not been routinely used in ABC in the past, despite being theoretically highly informative for admixture inference (Gravel, 2012; Verdu & Rosenberg, 2011).

Here, we propose, and implement in METHIS, an efficient way to use estimated individual admixture fractions as summary statistics for ABC inference, based on allele sharing dissimilarity (ASD) (Bowcock et al., 1994) and multidimensional scaling (MDS). For each simulated data set, we first calculated a pairwise interindividual ASD matrix using our implementation of the ASD software (https://github.com/szpiech/asd), using all pairs of sampled individuals and all markers. Then we

projected in two dimensions this pairwise ASD matrix with classical unsupervised metric MDS using the "cmdscale" function in R. We expected individuals in population H to be dispersed along an axis joining the centroids of the proxy source populations on the two-dimensional MDS plot. We projected population H's individuals orthogonally onto this axis, and calculated each individual's relative distance to each centroid. We considered this measure as an estimate of individual average admixture level from either source. Note that by doing so, some individuals might show "admixture fractions" higher than one, or lower than zero, as they might be projected on the other side of a source-population's centroid when being genetically close to 100% from this source population. Under an ABC framework, this was not a difficulty since this may happen also with the real data a priori, and the goal of ABC is to use summary statistics that mimic the observed ones.

This individual admixture estimation method has been shown to be highly concordant with cluster membership fractions as estimated with STRUCTURE (Falush et al., 2003) or ADMIXTURE (Alexander et al., 2009) in real data analyses (e.g., Verdu et al., 2017). We confirmed these previous findings since we obtained a Spearman's rank correlation (calculated using the cor.test function in R), of $\rho = 0.950$ (p-value $<2.10^{-16}$) and $\rho = 0.977$ (p-value $<2.10^{-16}$) between admixture estimates based on ASD-MDS and on ADMIXTURE, for the two case-study data sets here explored (Figure S2).

We used the mean, mode, variance, skewness, kurtosis, minimum, maximum, and all 10%-quantiles of the admixture distribution in population H, as 16 separate summary statistics for ABC inference.

2.3.2 | Within population summary statistics

We calculated marker by marker heterozygosities (Nei, 1978), and we considered the mean and variance of this quantity across markers in the admixed population as two separate summary statistics for ABC inference. In addition, we considered the mean and variance of ASD values across pairs of individuals within population H.

2.3.3 | Between populations summary statistics

We calculated multilocus pairwise F_{ST} (Weir & Cockerham, 1984) between population H and each source population respectively. Furthermore, we calculated the mean ASD between individuals in population H and individuals in each source population, separately. Finally, we calculated the f_3 statistics (Patterson et al., 2012).

2.4 | Approximate Bayesian Computation

MetHis provides, as outputs, vectors of scenario parameters and corresponding vectors of summary statistics in reference tables ready to be used with the machine-learning ABC R packages abc (Csilléry et al., 2012), and abcrf (Pudlo et al., 2016; Raynal et al., 2019).

2.4.1 | Simulating by randomly drawing parameter values from prior distributions

We performed MetHis simulations under each of the nine competing scenarios (Figure 1), drawing the corresponding scenario-parameters in prior distributions detailed in Table 1 and automatically generated by MetHis parameter-generator tools (Note S1).

2.4.2 | Complex admixture scenario-choice with Random-Forest ABC

For ABC scenario-choice, we performed 10,000 independent MetHis simulations for each of the nine competing scenarios. To mimic our case study data sets (see Section 2.4.3), we simulated 100,000 SNPs and sampled 50 individuals in population H, and 90 and 89 individuals respectively in the African and European source populations. Using 27 cores and the above design, we performed the 90,000 simulations with MetHis in four days, with 2/3 of that time for summary statistics calculation only (Note S1).

We used Random-Forest ABC for scenario-choice implemented in the "abcrf" function of the abcrf package to obtain the cross-validation table and associated prior error rate using an out-of-bag approach. We considered a uniform prior probability for the nine competing models. We considered 1,000 decision trees in the forest after visually checking that error-rates converged appropriately, using the "err.abcrf function". RF-ABC cross-validation procedures using groups of scenarios were conducted using the group definition option in the "abcrf" function (Estoup et al., 2018). Finally, the relative importance of each summary statistic to the scenario-choice cross-validation was computed using the "abcrf" function.

We explored scenario-choice erroneous assignation due to scenario nestedness in the parameter space, by considering 1000 randomly chosen simulations per scenario as pseudo-observed data. We trained the RF algorithm based on the 9000 remaining simulations per scenario using the "abcrf" function as described above, which provided highly similar results as when considering 10,000 simulations per scenario (results not shown). We then used the "predict.abcrf" function to perform scenario-choice independently for each of the 1,000 simulated pseudo-observed data with known parameter vectors.

To empirically evaluate the power of the RF-ABC scenario-choice to distinguish complex admixture processes, we conducted similar cross-validations procedures based on additional 10,000 simulations per scenario for 50,000 and, separately, 10,000 SNPs, instead of 100,000 SNPs (180,000 additional simulations in total).

Furthermore, using 100,000 SNPs, we produced 90,000 additional simulations and performed cross-validations, considering a five-times smaller sample set, with 10 sampled individuals in population H (instead of 50 as previously) and 18 individuals in each source population (instead of 90 and 89).

2.4.3 | Case-study population genetics data sets

We investigated, as two separate study-cases, the admixture histories of the African American (ASW) and Barbadian (ACB) population samples from the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium, 2015). Previous studies identified, within the same database, the West European Great-Britain (GBR) and the West African Yoruba (YRI) populations as reasonable proxies for the sources of both ACB and ASW, consistent with the macro-history of the Transatlantic Slave-Trade (Baharian et al., 2016; Martin et al., 2017; Verdu et al., 2017).

Individuals in the 1000 Genomes Project were a priori sampled to be family unrelated. To avoid confounding factors due to cryptic relatedness in this sample set compared to METHIS simulations, we excluded individuals more closely related than first-degree cousins in the four populations separately using RELPAIR (Epstein, Duren, & Boehnke, 2000), as previously done (Verdu et al., 2017). We also excluded the three ASW individuals showing traces of Native American or East-Asian admixture, as reported in previous studies (Martin et al., 2017). Among the remaining individuals we randomly drew 50 individuals in the target admixed ACB and ASW, respectively, and included the remaining 90 YRI individuals and 89 GBR individuals.

We extracted biallelic polymorphic sites (SNPs as defined by the 1000 Genomes Project Phase 3) from the merged ACB+ASW+GBR+YRI data set, excluding singletons. Since METHIS could only simulate independent markers, we LD-pruned the ACB and ASW SNP-sets using the PLINK (Purcell et al., 2007) "--indep-pairwise" option with a sliding window of 100 SNPs, moving in increments of 10 SNPs, with an r^2 threshold of 0.1. Finally, we randomly drew 100,000 SNPs from the remaining SNP-set.

2.4.4 | Prior-checking of simulations' fit to the case-study data sets

We plotted prior distributions of each summary statistic and visually verified that the observed summary statistics for the ACB and ASW respectively fell within the simulated distributions. Then, we explored the first four axes of a principal component analysis (PCA) computed with the "princomp" function in R, using the 24 summary statistics and all 90,000 simulations, and visually checked that observed summary statistics were within the cloud of simulated statistics. Finally, we performed a goodness-of-fit approach using the "gfit" function from the ABC package in R, with 1,000 replicates and tolerance level 0.01.

2.4.5 | RF-ABC scenario-choice for the admixture history of ACB and ASW populations

For the ACB and ASW observed data separately, we performed scenario-choice prediction and estimation of posterior probabilities of the winning scenario using the "predict.abcrf" function in the ABCRF package, using the complete simulated reference table for training the Random-Forest algorithm (100,000 SNPs, 50 individuals in population H, 90 and 89 individuals in the African and European sources, respectively).

2.4.6 | Posterior parameter estimation with Neural-Network ABC

It is difficult to estimate jointly the posterior distribution of all model parameters with RF-ABC (Raynal et al., 2019). Furthermore, although RF-ABC performs satisfactorily well with an overall limited number of simulations under each model (Pudlo et al., 2016), posterior parameter estimation with other ABC approaches, such as simple rejection (Pritchard et al., 1999), regression (Beaumont et al., 2002; Blum & François, 2010) or Neural-Network (NN) (Csilléry et al., 2012), require substantially more simulations a priori. Therefore, we performed, for posterior parameter estimations, 90,000 additional simulations, for a total of 100,000 simulations under the best scenarios identified with RF-ABC for the ACB and ASW separately. For comparison purposes, we also performed an additional 90,000 simulations (for a total of 100,000 simulations) under the loosing scenario Afr2P-Eur2P (see Results), and conducted anew the below parameter estimation and error evaluation procedures for this scenario.

2.4.7 | Neural-Network tolerance level and number of neurons in the hidden layer

We determined empirically the NN tolerance level (i.e., the number of simulations to be included in the NN training), and number of neurons in the hidden layer. Indeed, the NN needs a substantial amount of simulations for training, and there is also a risk of overfitting posterior parameter estimations when considering too large a number of neurons in the hidden layer. However, there are no absolute rules for choosing both numbers (Csilléry et al., 2012; Jay et al., 2019).

Therefore, we tested four different tolerance levels to train the NN for parameter estimation (0.01, 0.05, 0.1, and 0.2), and a number of neurons that ranged between four and seven (the number of free parameters in the winning scenarios, see Results). For each pair of tolerance level and number of neurons, we conducted cross-validation with 1000 randomly chosen simulated data sets that we used, in turn, as pseudo-observed data with the "cv4abc" function in the package ABC. We compared the median point-estimate of each posterior parameter ($\widehat{\theta}_i$) to the true parameter value used for simulation (θ_i). The cross-validation parameter prediction error was then calculated across the 1000 separate posterior estimations for pseudo-observed data sets for each pair of tolerance level and number of neurons, and for each parameter θ_i, as $\sum_1^{1000} (\widehat{\theta}_i - \theta_i)^2 / (1000 \times Variance(\theta_i))$, using the "summary.cv4abc"

function in the package ABC (Csilléry et al., 2012). Results showed that, a priori, all numbers of neurons considered perform very similarly for a given tolerance level. Furthermore, results showed that considering the 1% closest simulations to the pseudo-observed ones reduced the average error for each number of neurons tested. Thus, we decided to opt for four neurons in the hidden layer and a 1% tolerance level for training the NN in all subsequent parameter inference, in order to avoid overfitting.

2.4.8 | Estimation of scenario-parameters' posterior distributions

We jointly estimated the posterior distributions of scenario parameters for the ACB and ASW admixed populations separately, using NN-ABC "neuralnet" method option in the function "abc", with logit-transformed ("logit" transformation option) summary statistics using a 1% tolerance level and four neurons in the hidden layer.

2.4.9 | Posterior parameter estimation error

We evaluated the posterior error of the NN-ABC approach in the vicinity of our observed data rather than randomly on the entire parameter space. To do so, we first identified the 1000 simulations closest to the real data by setting a tolerance level of 1% with the "abc" function, for the ACB and ASW respectively. Then, we performed 1000 separate NN-ABC parameter estimations, each parameterized as described above, using in turn the remaining 99,999 simulations as reference tables, and recorded the median point estimate for each parameter. We then compared each parameter estimate with the true parameter used for each one of the 1000 pseudo-observed target data and provided three types of error measurements. The mean-squared error scaled by the variance of the true parameter $\sum_1^{1000}(\hat{\theta}_i-\theta_i)^2/(1000\times Variance(\theta_i))$, as previously (Csilléry et al., 2012); the mean-squared error $\sum_1^{1000}(\hat{\theta}_i-\theta_i)^2/1000$, which allowed to compare errors for a given scenario and parameter between the ACB and ASW analyses; and the mean absolute error $\sum_1^{1000}|\hat{\theta}_i-\theta_i|/1000$, which provided a more intuitive parameter estimation error. For comparison, we conducted the above analysis using instead parameters estimated under the losing scenario Afr2P-Eur2P.

2.4.10 | 95% credibility interval accuracy

We evaluated a posteriori, if, in the vicinity of the two observed data sets respectively, the lengths of the estimated 95% confidence intervals (CI) for each parameter were accurately estimated or not (e.g., Jay et al., 2019). To do so, we calculated how many times the true parameter (θ_i) was found inside the estimated 95% CI (2.5% quantile $(\hat{\theta}_i)$; 97.5% quantile $(\hat{\theta}_i)$), among the 1000 out-of-bag NN-ABC posterior parameter estimations. For each parameter, if fewer than 95% of the true parameter values were found inside the 95% CI estimated for the observed data, we considered the length of this credibility interval as underestimated which was indicative of a nonconservative behaviour of the parameter estimation. Alternatively, if more than 95% of the true parameter values were found inside the estimated 95% CI, we considered its length as overestimated, indicative of an excessively conservative behaviour of parameter estimation. For comparison, we conducted the above analysis using instead parameters estimated under the losing scenario Afr2P-Eur2P.

2.4.11 | Comparing the accuracy of posterior parameter estimations using NN, RF, or rejection ABC

We compared four ABC posterior parameter estimation methods: NN-ABC estimation of the parameters taken jointly as a vector (as described in the above procedures), NN-ABC estimation of the parameters taken in turn separately, RF-ABC estimation of the parameters which also considers parameters in turn and separately (Raynal et al., 2019), and simple Rejection-ABC estimation for each parameter separately (Pritchard et al., 1999). For each method, we used in turn the 1000 simulations closest to the real data as pseudo-observed data and the 99,999 remaining simulations as reference tables. We considered the same parameters for the NN, and we used 500 decision trees for the RF to limit the computational cost at little accuracy cost a priori. We computed the three types of errors and the accuracies of the 95% CI for each ABC method as described above.

3 | RESULTS

3.1 | Complex admixture scenarios cross-validation with RF-ABC

We trained the RF-ABC scenario-choice algorithm using 1000 trees, which guaranteed the convergence of the scenario-choice prior error rates (Figure S3). Based on this training, the complete out-of-bag cross-validation matrix showed that the nine competing scenarios of complex historical admixture (Figure 1, Table 1) could be relatively reasonably distinguished despite the high level of nestedness of the scenarios here considered (Figure 2). Indeed, we calculated an out-of-bag prior error rate of 32.41%, considering each of the 90,000 simulations, in turn, as out-of-bag pseudo-observed target data sets, compared to a prior probability of 88.89% to erroneously select a scenario. Furthermore, we found that cross-validation probabilities of identifying the correct scenario ranged from 55.17% (prior probability = 11.11% for each competing scenario), for the two-pulses scenarios from both the African and European sources (Afr2P-Eur2P), to 77.71% for the scenarios considering monotonically decreasing recurring admixture from both sources (AfrDE-EurDE).

The probability, for a given admixture scenario, of choosing any one alternative (wrong) scenario was on average 4.05% across the

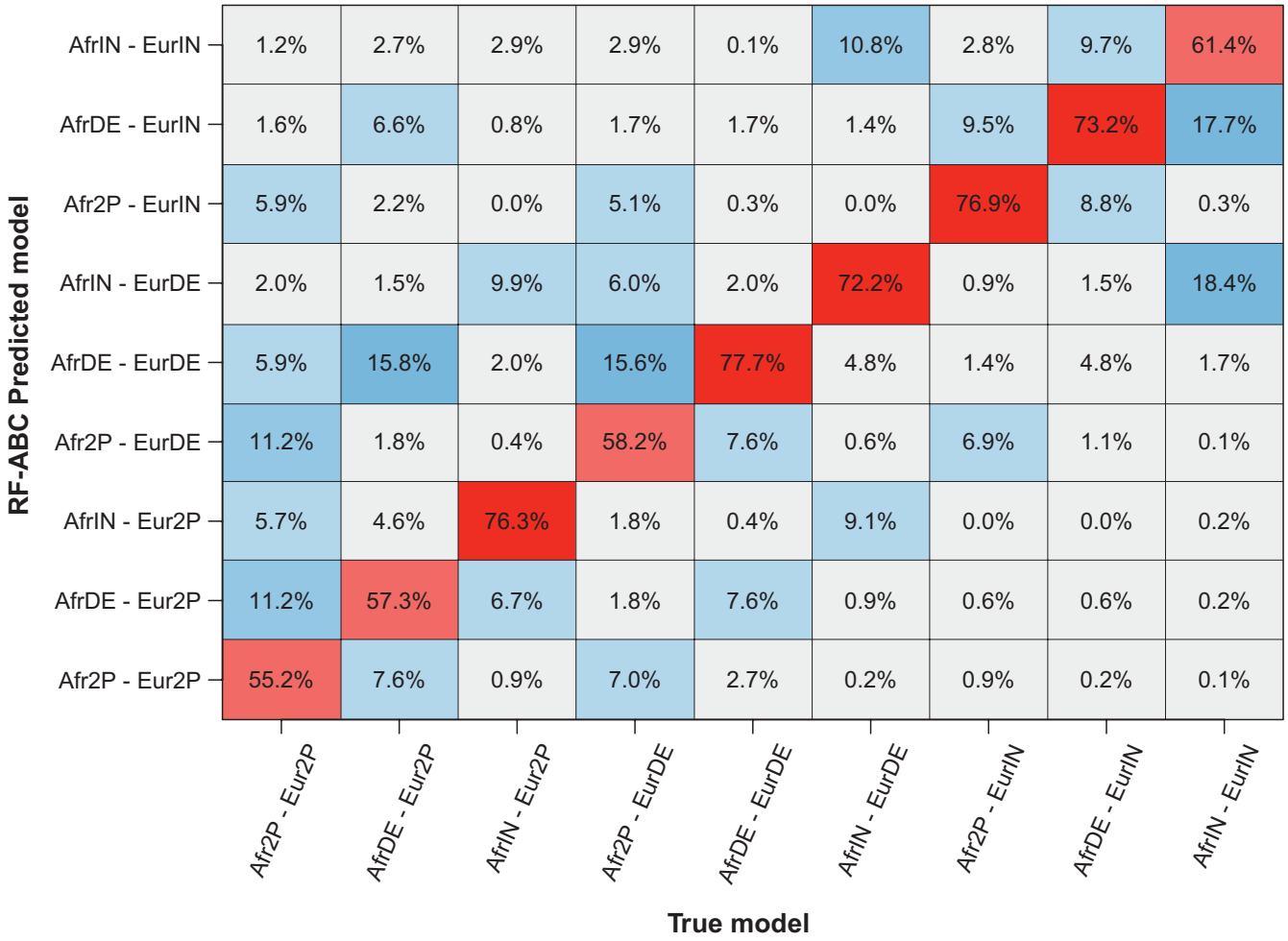

FIGURE 2 Random-Forest Approximate Bayesian Computation scenario-choice cross-validation. Heat map of the out-of-bag cross-validation results considering each of the 10,000 simulations per each of the nine competing scenarios (Figure 1, Table 1) in turn as pseudo-observed target for RF-ABC model-choice. Prior probability of correctly choosing a given scenario was 11%. Out-of-bag prior error rate was 32.41%. RF-ABC scenario-choice performed using 1000 decision trees and 24 summary statistics (see Section 2)

eight alternative scenarios, ranging from 2.79% for the AfrDE-EurDE scenario, to 5.60% for the Afr2P-Eur2P scenario (Figure 2). However, cross-validation assignment errors, for a given true scenario, were not uniformly distributed across the eight alternative scenarios. Instead, Figure 2 shows that assignment errors were relatively less frequent for classes of scenarios a priori more differentiated from the true scenario. For instance, the Afr2P-Eur2P true scenarios were less often confused (10.7%) with scenarios encompassing recurring admixture from both source populations (AfrDE-EurDE, AfrIn-EurDE, AfrDE-EurIN, AfrIN-EurIN), than with scenarios containing pulses of admixture from one source population (34.0%; AfrDE-Eur2P, Afr2P-EurDE, AfrIN-Eur2P, Afr2P-EurIN). Furthermore, note that AfrDE-EurDE scenarios were rarely confused (3.8%) with recurring scenarios containing at least one admixture increase (AfrIN-EurDE, AfrDE-EurIN, AfrIN-EurIN). Across the nine nested competing scenarios of highly complex admixture processes, these results showed a strong discriminatory power of RF-ABC scenario-choice a priori.

In cross-validation analyses of groups of scenarios (Estoup et al., 2018), monotonically recurring admixture scenarios (AfrDE-EurDE, AfrDE-EurIN, AfrIN-EurDE, AfrIN-EurIN) could be well distinguished from scenarios considering two possible pulses after the founding event (Afr2P-Eur2P, Afr2P-EurDE, Afr2P-EurIN, AfrDE-Eur2P, AfrIN-Eur2P). Indeed, we found an out-of-bag prior error rate of 13.85%, and cross-validation probabilities of identifying the correct group of scenarios of 86.08% and 86.23% for the two groups, respectively.

Detailed investigation of cross-validation results showed that inaccuracies of RF-ABC scenario-choices occurred mainly in spaces of values of parameters where scenarios were highly nested and, in fact, close biologically (Figure 2). As expected, scenario-choice increasingly mistook the AfrDE-EurDE scenarios for scenarios containing two admixture pulses (Afr2P-Eur2P, Afr2P-EurIN, AfrIN-Eur2P), as values of u_{Afr} and u_{Eur} were closer to 0, regardless of the values of introgression rates (Figure S4a). Intuitively for the S-DE

scenarios, values of the parameter *u* close to 0 corresponded to steeper decreases of recurring admixture over time, which increased scenario-choice confusion with pulse-like scenarios. Simulation with *u*-values closer to 0.5 corresponded to linearly decreasing admixture over time and could hardly be confounded with pulse-like scenarios. Furthermore, the scenario-choice increasingly confused, as expected regardless of introgression values, Afr2P-Eur2P scenarios with recurring increasing admixture scenarios (AfrIN-EurIN, AfrDE-EurIN, AfrIN-EurDE), as the time of the second admixture pulse from Europe or Africa became more recent (Figure S4b).

Most importantly, RF-ABC scenario-choice power to discriminate among complex admixture processes a priori was not strongly affected by the numbers of markers considered. Indeed, we found an out-of-bag prior error of 33.53% and 37.93% (instead of 32.41%), considering respectively 50,000 and 10,000 SNPs, instead of 100,000, together with a very similar distribution of correct and mistaken cross-validation assignments among scenarios (Figures S5a,b). Finally, dividing by five the sample sizes in population H and each source population increased, as expected, the cross-validation error rate (48.39%). Nevertheless, all scenarios continued to be correctly identified three to six times more often than expected a priori, and the distribution of erroneous predictions remained similar to previously (Figure S5c).

Altogether, these results showed that RF-ABC scenario-choice can be successfully used to distinguish highly complex admixture models even when substantially less genetic and sample data are considered. Finally, the estimated relative importance of each summary statistic for RF-ABC scenario-choice showed that the minimum, maximum, 10%-quantile, 90%-quantile, variance, and skewness of the distribution of admixture fractions among individuals in the admixed population were, among the 24 summary statistics used, the most informative statistics for our scenario-choice cross validation results (Figure S6).

3.2 | Simulating data similar to the observed data with MetHis

Using MetHis, we produced 90,000 vectors of 24 summary statistics each, overall highly consistent with the observed ones for the ACB and the ASW populations. First, each observed statistic was visually reasonably well simulated under the nine competing scenarios here considered (Figure S7). Second, the observed data each fell into the simulated sets of 24 summary statistics projected in the first four PCA dimensions (Figure S8). Finally, the observed vectors of summary statistics were not significantly different (*p*-value = 0.468 and 0.710, for the ACB and ASW respectively) from the simulated ones using a goodness-of-fit approach (Figure S9). Therefore, we successfully simulated data sets producing sets of summary statistics reasonably close to the observed ones, despite considering constant effective population sizes, using fixed virtual source population genetic pool-sets, and neglecting mutation during the admixture process.

3.3 | Random-Forest ABC scenario-choice for the history of ACB and ASW populations

We performed RF-ABC scenario-choice separately for the admixture history of the ACB and the ASW populations, to evaluate whether our MetHis-ABC method could identify subtle differences in the history of both populations having experienced the TAST under the British colonial empire (Baharian et al. 2016; Martin et al. 2017). For the ACB, Figure 3 shows that the majority of votes (53.1%) went to an admixture scenario AfrDE-EurDE with a posterior probability of the winning scenario of 60.3%. This posterior probability was above the mean posterior probability obtained when the wrong scenario was chosen for the 1000 AfrDE-EurDE simulations closest to the observed one (56.8%, *SD* = 11.6%, for 37 simulations wrongly assigned in total). The second most chosen scenario was the AfrDE-Eur2P scenario. However, this scenario was voted for 3.5 times less often than the winning scenario AfrDE-EurDE, gathering 15.1% of the 1000 votes, only slightly above the 11.11% prior probability for the nine competing scenarios (Figure 3; Table S1).

RF-ABC scenario-choice results were less decisive for the ASW (Figure 3). The AfrDE-EurDE scenario also gathered the majority of votes, albeit with lower posterior probability than for the ACB (33.5% of 1000 votes, with posterior probability = 48.0%). This posterior probability was slightly below the average posterior probability obtained when the wrong scenario was chosen for the 1000 AfrDE-EurDE simulations closest to the ASW observed data (50.7%, *SD* = 7.9%, for 192 simulations wrongly assigned). The second most chosen scenario, AfrDE-Eur2P, was only slightly less chosen with 31.7% of the votes (Figure 3, Table S1). Altogether these results denoted an ambiguity of the RF-ABC scenario-choice in the part of the space of summary statistics occupied by the ASW.

Considering only these two best scenarios to train the RF and reconducting ABC scenario-choice improved the scenario discrimination in favor of the AfrDE-EurDE scenario. While we found, again, only a slight majority of votes (51.8%) in favour of the AfrDE-EurDE scenario, the posterior probability for this scenario was substantially increased to 57.9%, thus above the average posterior probability threshold calculated previously (50.7%). This indicated that the AfrDE-EurDE scenario best explained the ASW observed genetic patterns, despite overall limited discriminatory power of our approach in the ambiguous part of the space of summary statistics occupied by this population.

3.4 | Neural-Network ABC parameter inference accuracy

For the ACB under the AfrDE-EurDE scenario (Figure 4a, Table 2), we conducted a NN-ABC posterior parameter inference considering four neurons and a tolerance level of 1% (Table S2). We found that the two recent admixture intensities from Africa and Europe ($s_{Afr,20}$ and $s_{Eur,20}$, respectively), and the steepness of the European

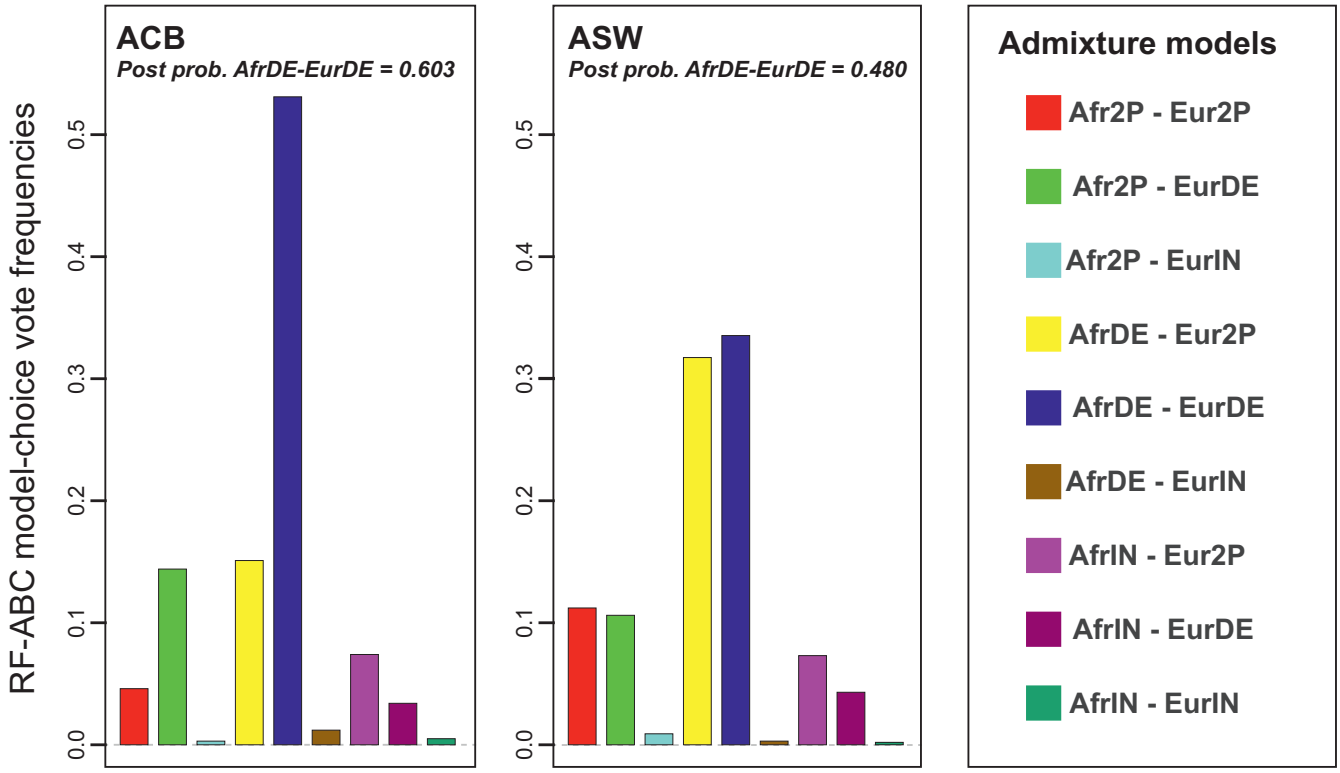

FIGURE 3 Random-Forest Approximate Bayesian Computation scenario-choice predictions for the ACB (left panel) and ASW (right panel) populations. Nine competing scenarios were compared, each with 10,000 simulations (Figure 1, Table 1), and 1,000 decision trees were considered in the scenario-choice prediction, respectively for each population

recurring introgression decrease (u_{Eur}), had sharp posterior densities clearly distinct from their respective priors. Note that the cross-validation error on these parameters in the vicinity of our real data were low (average absolute error 0.02744, 0.0044, and 0.1084, respectively for $s_{Afr,20}$, $s_{Eur,20}$, and u_{Eur}) (Table 3), and lengths of 95% CI reasonably accurate (96.4%, 94.4%, 94.1% of 1000 cross-validation true parameter values fell into estimated 95% CI, Table S3).

Furthermore, the two ancient admixture intensities from Africa and Europe at generation 1 ($s_{Afr,1}$ and $s_{Eur,1}$, respectively), also had posterior densities apparently distinguished from their prior distributions, but both had much wider 95% CI (Figure 4a, Table 2). Consistently, we found a slightly increased posterior parameter error in this part of the parameter space for both parameters, with average absolute error equal to 0.121 and 0.095, respectively for $s_{Afr,1}$ and $s_{Eur,1}$ (Table 3). Nevertheless, note that 95.8% and 94.7% of 1000 cross-validation true values for those two parameters fell into the estimated 95% CI (Table S3). This showed that information was somewhat lacking in our set of summary statistics for a more accurate point estimation of these parameters, albeit our method was reasonably conservative for these estimations.

Interestingly-, we found that accurate posterior estimation of the steepness of the African recurring introgression decrease (u_{Afr}) was difficult. Indeed, the posterior density of this parameter showed a tendency towards small values only slightly departing from the prior, indicative of a limit of our method to estimate this parameter (Figure 4a, Table 2). Finally (Figure 4a, Table 2), we found that we had virtually no information to estimate the founding admixture proportions from Africa and Europe at generation 0, as our posterior estimates barely departed from the prior, and as associated mean absolute error was high (0.2530, Table 3). Nevertheless, our method seemed to be performing reasonably conservatively for these two latter parameters (95.6% and 95.3% of 1000 cross-validation true parameter values fell into estimated 95% CI, Table S3).

For the ASW under the AfrDE-EurDE scenario, our posterior parameter estimation results were overall less accurate compared to those obtained for the ACB population, as indicated by overall larger CI and cross-validation errors (Figure 4b, Table 2, Table 3, Table S3). This was consistent with the more ambiguous RF-ABC scenario-choice results obtained for this population (Figure 3).

Note that, we conducted the above analyses under the losing scenario Afr2P-Eur2P instead, for comparison. We found, as expected, that parameters and 95% CI were very poorly estimated for all parameters under this scenario (Tables S4 and S5). This indicated, consistently, that no information was available in the ACB or ASW data for accurate and conservative estimation of the parameters of the losing scenario Afr2P-Eur2P using ABC.

3.5 | Comparing NN, RF, and Rejection ABC posterior parameter estimation accuracy

The three types of posterior parameter estimation errors (scaled mean-squared error, mean-squared error, average absolute error) were systematically lower for the two NN methods (joint or

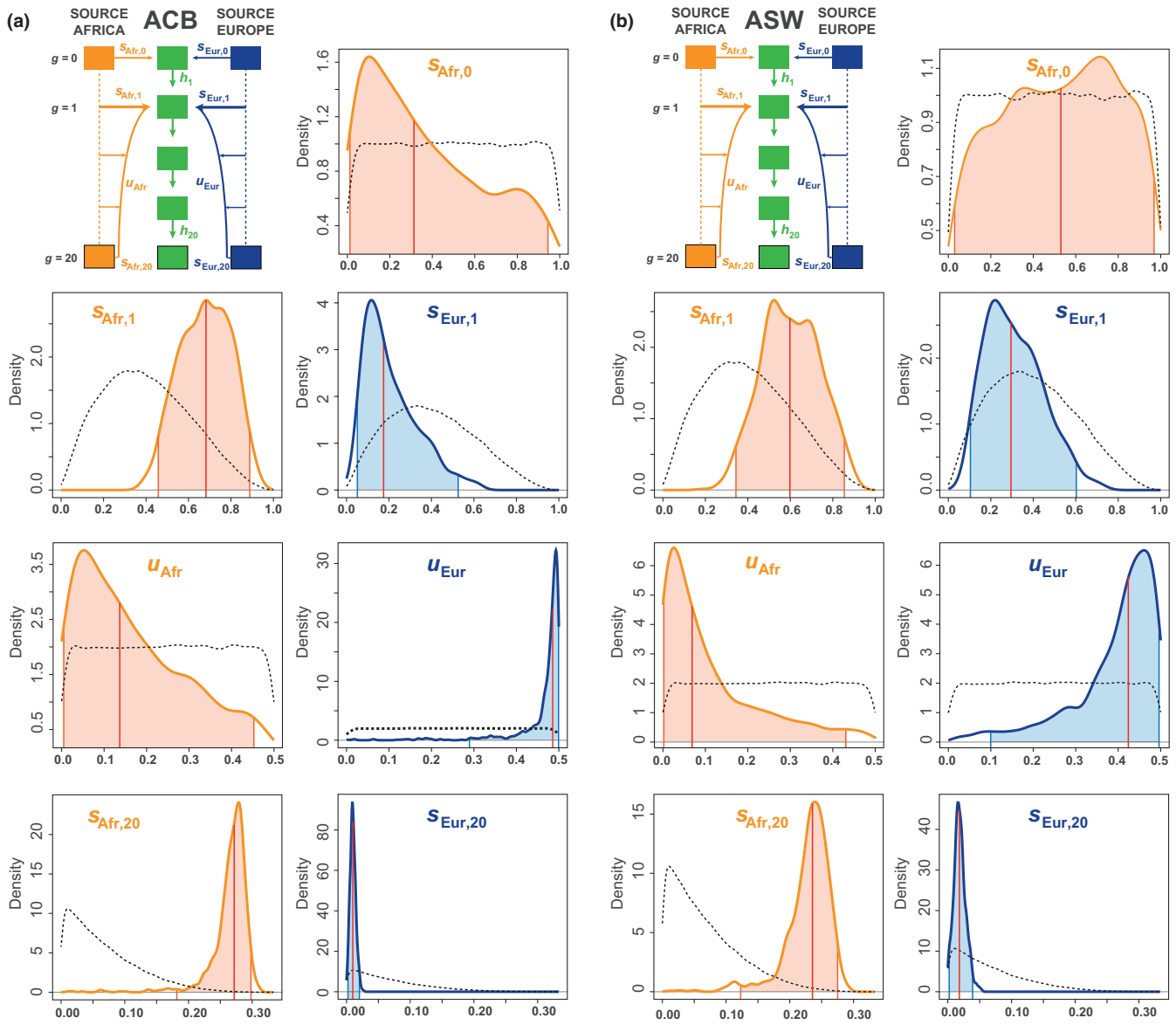

FIGURE 4 Neural-Network Approximate Bayesian Computation posterior parameters estimated densities under the winning scenario AfrDE-EurDE, for (a) the ACB and (b) the ASW populations. Median posterior point estimates are indicated by the red vertical line, 95% credibility intervals are indicated by the colored area under the posterior density-curve (Table 2). All posterior parameter estimations were conducted using 100,000 simulations under scenario AfrDE-EurDE, a 1% tolerance rate (1000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see Section 2). For all parameters separately, densities were plotted with 1000 points, a Gaussian kernel, and were constrained to the prior limits. Posterior parameter densities are indicated by a solid line; prior parameter densities are indicated by black dotted lines

independent posterior parameter estimations) than for the RF and Rejection independent posterior parameter estimations (Table 4). Altogether, these results showed that considering the NN estimation for parameters taken jointly as a vector was overall preferable, since it further allowed the joint interpretation of parameter values estimated a posteriori with little accuracy loss.

The lengths of 95% CI estimated with NN joint parameter estimation were, across all parameters, more accurate than those obtained with all other methods with, on average, 95.1% and 95.2% of true parameter values falling within the estimated 95% CI, for the ACB and ASW respectively (Table S3). Furthermore, the lengths of 95% CI estimated with NN and RF independent posterior parameter estimations were systematically underestimated, with less than 94% of the true parameter values falling into the estimated 95% CI. Finally, the lengths of 95% CI estimated with the Rejection method were also rather accurately estimated, although on average slightly overestimated compared to the NN joint estimation with, on average, 95.5% of the 1000 cross-validation true parameter values within the estimated 95% CI for the ACB, and 95.8% for the ASW.

TABLE 2 Neural-Network Approximate Bayesian Computation posterior parameter weighted distributions under the winning scenario AfrDE-EurDE, for the ACB and ASW populations

Admixed population	AfrDE-EurDE parameters	Median	Mean	Mode	95% credibility interval
ACB	$s_{Afr,0}$	0.3097	0.3747	0.1121	[0.0116; 0.9347]
	$s_{Afr,1}$	0.6797	0.6769	0.6813	[0.4577; 0.8880]
	$s_{Afr,20}$	0.2707	0.2655	0.2788	[0.1985; 0.2967]
	u_{Afr}	0.1409	0.1684	0.0508	[0.0041; 0.4507]
	$s_{Eur,1}$	0.1807	0.2160	0.1158	[0.0542; 0.5525]
	$s_{Eur,20}$	0.0100	0.0102	0.0093	[0.0018; 0.0200]
	u_{Eur}	0.4858	0.4627	0.4929	[0.1886; 0.4992]
ASW	$s_{Afr,0}$	0.5258	0.5124	0.7015	[0.0262; 0.9758]
	$s_{Afr,1}$	0.6006	0.6026	0.6081	[0.3506; 0.8581]
	$s_{Afr,20}$	0.2352	0.2286	0.2385	[0.1222; 0.2714]
	u_{Afr}	0.0662	0.1105	0.0253	[0.0025; 0.4393]
	$s_{Eur,1}$	0.2917	0.3080	0.2203	[0.1048; 0.5951]
	$s_{Eur,20}$	0.0180	0.0189	0.0157	[0.0022; 0.0389]
	u_{Eur}	0.4250	0.3966	0.4567	[0.1077; 0.4950]

All posterior parameter estimations were conducted using 100,000 simulations under the AfrDE-EurDE scenario (Figure 1, Table 1), a 1% tolerance rate (1,000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see Section 2).

TABLE 3 Neural-Network Approximate Bayesian Computation posterior parameter errors under the winning scenario AfrDE-EurDE, for the ACB and ASW populations

	ACB			ASW		
AfrDE-EurDE parameters	Av. absolute error	Mean-square error	Mean-square error/var.	Av. absolute error	Mean-square error	Mean-square error/var.
$s_{Afr,0}$	0.2530	0.0857	1.0070	0.2444	0.0805	1.0081
$s_{Afr,1}$	0.1206	0.0216	0.8533	0.1158	0.0197	0.9259
$s_{Afr,20}$	0.02744	0.0012	0.4162	0.0219	0.0007	0.4773
u_{Afr}	0.1166	0.0198	0.9974	0.1254	0.0216	0.9757
$s_{Eur,1}$	0.0952	0.0164	1.0526	0.1001	0.0157	1.0152
$s_{Eur,20}$	0.0044	0.0001	0.6452	0.0069	0.0001	0.6623
u_{Eur}	0.1084	0.0174	0.9431	0.1021	0.0153	0.8036

For each target population separately, we conducted cross-validation by considering in turn 1000 separate NN-ABC parameter inferences each using in turn one of the 1000 closest simulations to the observed ACB (or ASW) data as the target pseudo-observed simulation. All posterior parameter estimations were conducted using 100,000 simulations under the AfrDE-EurDE scenario (Figure 1, Table 1), a 1% tolerance rate (1000 simulations), 24 summary statistics, logit transformation of all parameters, and four neurons in the hidden layer (see Section 2). Median was considered as the point posterior parameter estimation for all parameters. First column provides the average absolute error; second column shows the mean-squared error; third column shows the mean-squared error scaled by the parameter's observed variance (see Section 2 for error formulas)

3.6 | Admixture histories of the African American ASW and Barbadian ACB

Figure 5 visually synthesizes the estimated posterior parameters of the complex admixture scenarios reconstructed with the MetHis-ABC framework, and associated 95% CI (Table 2).

We found a virtual complete replacement of the ACB and ASW populations at generation 1, thus consistent with our inability to accurately estimate the founding proportions from the African and European sources at generation 0. Furthermore, we found an increasingly precise posterior estimation of introgression rates forward-in-time. This is also consistent with the nature of recurrent admixture processes, where older information may be lost or replaced when more recent admixture events occur.

Interestingly, we found that the recurring introgression from the European gene pool rapidly decreased after generation 1, for both the ACB and ASW, albeit with substantial differences (Figure 5). Indeed, we found that, for the ACB, European introgression falls below 10% at generation 9 to no more than 1% in the present. Comparatively, the European contribution diminished substantially

TABLE 4 Approximate Bayesian Computation mean posterior parameter errors over all parameters under the winning Scenario AfrDE-EurDE, for the ACB and ASW populations separately, using four different methods: NN estimation of the parameters taken jointly as a vector, NN estimation of the parameters taken separately, Random-Forest (parameters taken separately), and Rejection (parameters taken separately)

	ACB			ASW		
Posterior parameter estimation ABC method	Av. absolute error	Mean-squared error	Mean-squared error/var.	Av. absolute error	Mean-squared error	Mean-squared error/var.
NN joint	0.1037	0.0232	0.8450	0.1024	0.0219	0.8383
NN independent	0.1032	0.0236	0.8294	0.1025	0.0225	0.8344
RF independent	0.1042	0.0246	0.8534	0.1036	0.0233	0.8697
Rejection independent	0.1071	0.0238	0.9299	0.1050	0.0223	0.8951

For each target population separately and for each method, we conducted an out-of-bag cross validation by considering in turn 1000 separate parameter inferences each using one of the 1000 closest simulation to the observed ACB (or ASW) data as the target pseudo-observed data set. All posterior parameter estimations were conducted using the remaining 99,999 simulations under the AfrDE-EurDE scenario (Figure 1, Table 1), a 1% tolerance rate (i.e., 1000 simulations), 24 summary statistics, logit transformation of all parameters, four neurons in the hidden layer per Neural-Network and 500 trees per Random-Forest. Median was considered as the point posterior parameter estimation for all parameters. The first column provides the average absolute error; second column shows the mean-squared error; third column shows the mean-squared error scaled by the parameter's observed variance (see Section 2 for error formulas).

less rapidly for the ASW, going below 10% only after generation 12 to roughly 2% in the present. Therefore, it seemed that neither sustained European migrations, nor the relaxation of social and legal constraints on admixture subsequent to the abolition of slavery and the end of segregation, have translated into increased European genetic contribution to the gene-pool of admixed populations in the Americas.

Finally, we found substantial recurring contributions from the African source for both admixed populations (Figure 5). For the ACB, we found a progressive decrease of the African recurring introgression until a virtually constant recurring admixture close to 28% from generation 10 onward. For the ASW, our results showed a sharper decrease of the African contribution after founding until a virtually constant recurring admixture process close to 24% from generation 5 onward. Nevertheless, the ASW occupy an ambiguous region of the parameter space, and results should be considered cautiously, as another complex admixture model might more accurately explain this data. Altogether, the signal of substantial ongoing admixture from Africa could have emerged due to the known importance of African recurring forced migrations during the TAST into the Americas, as well as from enslaved-African descendants migrations within the Americas before and after the end of slavery (Baharian et al., 2016; Fortes-Lima et al., 2018).

4 | DISCUSSION

Our novel MetHis forward-in-time simulator and summary statistic calculator coupled with RF-ABC scenario-choice could distinguish among highly complex admixture histories using genetic data. As expected, scenario-choice errors were particularly made in regions of the parameter space for which scenarios were highly nested (Robert et al., 2010), and, thus, biologically similar. Furthermore, we found that NN-ABC provided accurate and reasonably conservative posterior parameter estimation for numerous parameters of the winning scenario, using human population data as a case study. Finally, we empirically demonstrated that the moments of the distribution of admixture fractions in the admixed population were highly informative for ABC inference, as expected theoretically (Gravel, 2012; Verdu & Rosenberg, 2011).

In general, the machine-learning ABC approaches here deployed for reconstructing highly complex admixture histories provided significant improvements for population genetics demographic inferences using genetic data. First, RF algorithms are, by nature, categorization algorithms and therefore a priori conceptually particularly well suited for scenario-choice inferences as compared to, for instance, previous regression-based ABC scenario-choice algorithms (Beaumont et al., 2002). In addition, they substantially reduce the simulation costs while improving scenario-choice performances, as compared to previous ABC scenario-choice algorithms that classically require 10–100 times more simulations (Pudlo et al., 2016). Finally, RF-ABC scenario-choice allow exploring, in detail, the relative contribution of each summary statistic to the scenario-choice, in addition to being insensitive to correlations among statistics. These improvements can thus both improve the user's understanding of the general behavior and performances of the scenario-choice inference procedures applied to her/his specific study-case, and alleviate the major difficulty induced by large spaces of summary statistics encountered in previous ABC scenario-choice approaches (Sisson et al., 2018). Nevertheless, posterior parameter estimation with RF-ABC remains difficult, as it only allows estimating the quantiles of the posterior parameters independently, rather than the full posterior distributions of the parameters estimated jointly (Raynal et al., 2019).

Second, NN-ABC parameter inference also provide a promising line of future developments for posterior parameter inference based on high dimensional parameter spaces. Indeed, using NN

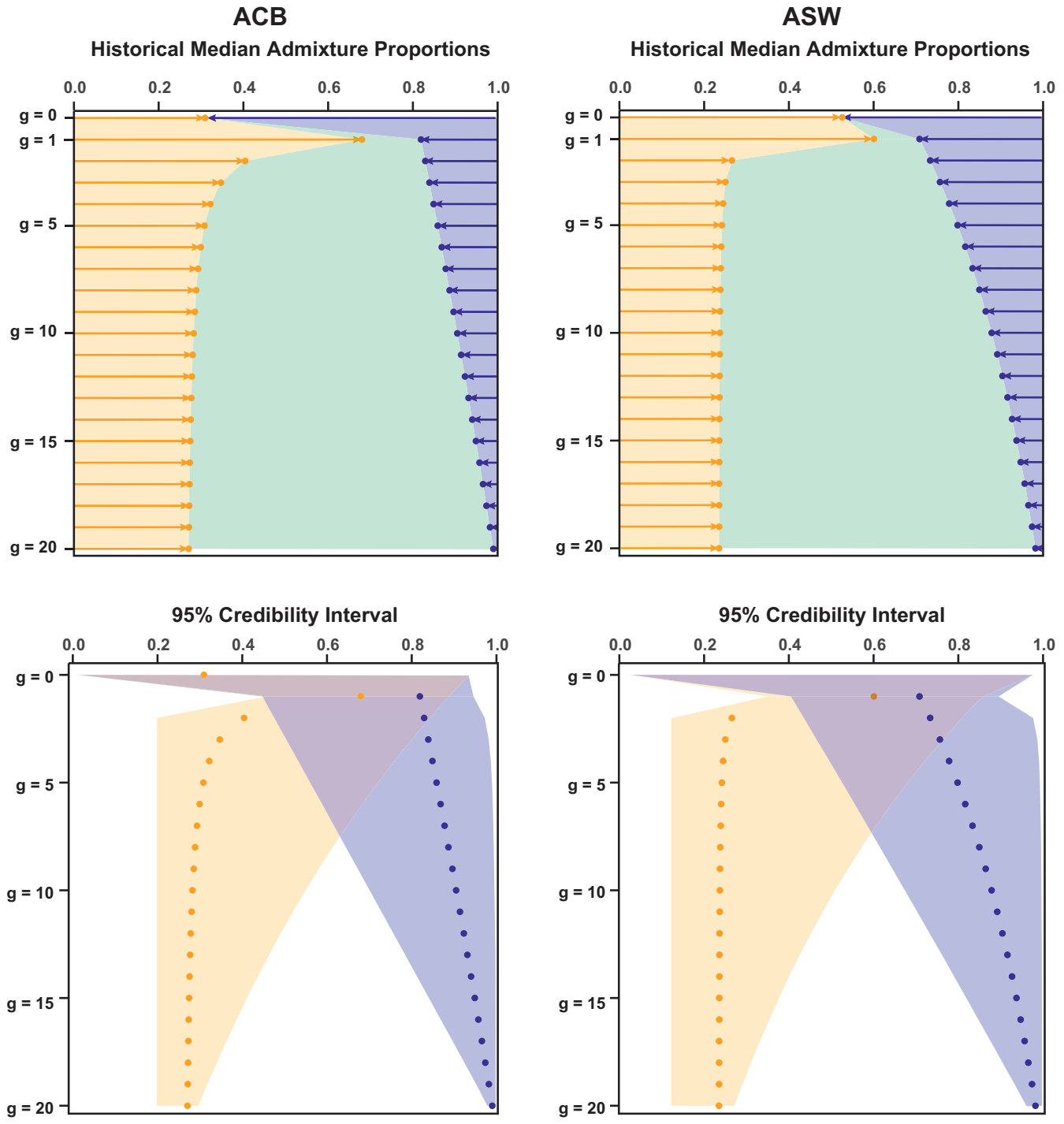

FIGURE 5 Approximate Bayesian Computation inference of the admixture history of the ACB and ASW populations respectively. Top panels are based on median point-estimates of parameters for the relative contribution of each source to the gene pool of the admixed target population ACB and ASW, at each generation. Bottom panels show 95% credibility intervals for each inferred parameter around the median point-estimates. The African introgression is plotted in orange, the European introgression in blue, and in green the remaining contribution of the admixed population to itself at the following generation. Left column presents results for the ACB under the AfrDE-EurDE winning scenario; Right column presents results for the ASW under the AfrDE-EurDE winning scenario

methods allows for the joint estimation of all model parameters by weighting the informativeness of each summary statistic about the parameters, beyond what most other ABC parameter-inference methods can do. Nevertheless, future studies will need to explore all the possibilities brought by posterior parameter inferences with NN, such as increasing the number of layers of hidden neurons, fine-tuning the NN procedures respective to the specific weighting of each summary statistics' importance for posterior

parameter estimations, and/or different NN algorithms for exploring the space of summary statistics (Csilléry et al., 2012). These will allow researchers to fully benefit from the power of this novel conceptual way of extracting information about model parameters from population genetics statistics computed from genetic data.

Altogether, our results for the two recently-admixed human populations illustrated how our MetHis-ABC framework can bring fundamental new insights into the complex demographic history of admixed populations; a framework that can easily be adapted, using MetHis (Note S1), for investigating complex admixture histories when ML methods are intractable.

We considered nine competing scenarios all deriving from the general mechanistic admixture model of Verdu and Rosenberg (2011). While the two-source version of this model can readily be simulated with MetHis, it considers $2g$-1 model parameters (with g the duration of the admixture process), plus effective population size parameters and mutation parameters. Estimating jointly all these parameters is out of reach of ML methods, and further probably out of reach of ABC posterior parameter estimation procedures. However, conducting ABC scenario-choice for disentangling major classes of relatively simplified admixture processes followed by ABC parameter estimation under the winning scenario, is flexible enough to bring new insights into the evolutionary history of admixed populations, far beyond all admixture scenarios that can be explored with existing ML methods (Gravel, 2012; Hellenthal et al., 2014).

The sample and SNP set explored here is often out-of-reach in non-model species. Nevertheless, our results considering vastly reduced SNP or sample sets demonstrated that ABC could remain remarkably accurate for disentangling highly complex admixture processes with much less genetic or sample data. This is due to the fact that ABC relies on the amount of information carried by summary statistics about model parameters, rather than on the absolute amount of genetic data investigated. Therefore, the MetHis-ABC framework remains promising to reconstruct complex admixture histories in study-cases with substantially fewer genetic and sample data, provided that the summary statistics considered by the user are, a priori, informative about model parameters, and that they are reasonably well estimated for the observed data. Altogether, large spaces of parameters and summary statistics, lack of information from summary statistics, and scenario nestedness, are well known to affect ABC performances and, thus, imperatively need to be thoroughly evaluated case by case (Csilléry et al., 2010; Robert et al., 2010; Sisson et al., 2018).

To further increase the range of applicability of our MetHis-ABC framework, our software readily implements microsatellite markers together with a general stepwise mutation model (Estoup et al., 2002), fully parameterizable by the user (Note S1). This will allow investigating numerous complex admixture histories from non-model species for which large amounts of SNP data are less frequently available, but for which microsatellite markers are readily available.

Even if prior knowledge of the date for the founding admixture event is lacking, MetHis users can simply set the founding of the admixed population in a remote past and implement a second founding event with variable date to be estimated with ABC, together with later additional admixture events and other parameters of interest. Nevertheless, it is not trivial to predict how old admixture processes can be to remain successfully investigated with ABC (Buzbas & Verdu, 2018). Indeed, ancient admixture processes could leave scarcely identifiable signatures in the observed data, if they have been obliterated by more recent admixture events. This was theoretically expected (Buzbas & Verdu, 2018), and future studies combining ancient and modern DNA samples may bring further information into the reconstruction of ancient admixture history.

Importantly, the computational cost of our study depends, for 2/3, on the calculation of all summary statistics at the end of the admixture process, as is often the case in ABC. Considering much longer admixture processes than the ones here investigated will mechanically increase computation time but will not increase summary statistics calculation time. Furthermore, note that the computational cost of simulating data with MetHis does not rely excessively on the number of generations considered (within reason), nor on the absolute number of markers used, but rather on the effective population size in the admixed population set by the user.

Although MetHis readily allows considering changes of effective population size in the admixed population at each generation as a parameter of interest to ABC inference (Note S1), we did not, for simplicity, investigate here how such changes affected our results. Future work using MetHis will specifically investigate how effective size changes may influence genetic patterns in admixed populations, a question of major interest as numerous admixed populations have experienced bottlenecks during their genetic history (e.g., Browning et al., 2018).

The current MetHis-ABC approach does not make use of admixture linkage-disequilibrium patterns in the admixed population, and only relies on independent SNP or microsatellite markers. Nevertheless, admixture-LD has consistently proved to bring massive information about complex admixture histories in populations where large genomic data sets were available (Gravel, 2012; Hellenthal et al., 2014; Malinsky et al., 2018; Medina et al., 2018; Ni et al., 2019; Stryjewski & Sorenson, 2017). However, existing methods to calculate admixture-LD patterns remain computationally intensive and require both dense marker-sets and accurate phasing, which is difficult under ABC where such statistics have to be calculated for each one of the numerous simulated data sets. In this context, RF-ABC (Pudlo et al., 2016; Raynal et al., 2019), or AABC (Buzbas & Rosenberg, 2015), methods substantially reduce the number of simulations required for satisfactory ABC inference. This makes both approaches promising for using, in the future, admixture-LD patterns to reconstruct complex admixture processes with ABC using genomic data.

Finally, future developments of the MetHis-ABC framework will focus on implementing sex-specific admixture models, as these processes are known to affect genetic diversity patterns in a specific way, and are of interest to numerous study-cases (Goldberg et al., 2014). Furthermore, the MetHis forward-in-time simulator

represents an ideal tool to further investigate admixture-related selection forces, and admixture-specific assortative mating processes, as these processes can simply be modeled by specifically parameterizing individual reproduction and survival in the simulations, unlike most coalescent-based simulators.

ACKNOWLEDGEMENTS

We thank Frédéric Austerlitz, Erkan O. Buzbas, Antoine Cools, Flora Jay, Evelyne Heyer, Margueritte Lapierre, Guillaume Laval, Nina Marchi, Etienne Patin, Noah A. Rosenberg, and Zachary A. Szpiech for useful comments and discussions. We warmly thank Olivier Hardy for help designing the microsatellite mutation model implemented in MetHis. We thank three anonymous reviewers and the editor for recommendations having improved the article. This project was funded in part by the French Agence Nationale de la Recherche project MetHis (ANR 15-CE32-0009-01). CFL was funded in part by the Sven and Lilly Lawski's Foundation (N2019-0040).

AUTHOR CONTRIBUTIONS

Cesar A. Fortes-Lima built the alpha version of the software, conducted preliminary benchmarking and data analyses, and assisted in writing the article. Romain Laurent built the beta version of the software, conducted benchmarking and data analyses and assisted in writing the article. Valentin Thouzeau conducted benchmarking and data analyses and assisted in writing the article. Bruno Toupance assisted in building the beta version of the software, conducted benchmarking and data analyses and assisted in writing the article. Paul Verdu designed and supervised the project, conducted benchmarking and data analyses and wrote the article.

DATA AVAILABILITY STATEMENT

MetHis software package is open source under the GNU General Public License v3.0, and can be downloaded with manual and example data sets from https://github.com/romain-laurent/MetHis. Genetic data used in this article were downloaded from the 1000 Genome Project Phase 3 open-access data repository (https://www.internationalgenome.org/data/).

ORCID

Cesar A. Fortes-Lima https://orcid.org/0000-0002-9310-5009
Romain Laurent https://orcid.org/0000-0003-0363-2954
Valentin Thouzeau https://orcid.org/0000-0002-2096-2675
Bruno Toupance https://orcid.org/0000-0002-8244-1824
Paul Verdu https://orcid.org/0000-0001-6828-268X

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109.

Baharian, S., Barakatt, M., Gignoux, C. R., Shringarpure, S., Errington, J., Blot, W. J., Bustamante, C. D., Kenny, E. E., Williams, S. M., Aldrich, M. C., & Gravel, S. (2016). The great migration and African-American genomic diversity. *PLoS Genetics*, *12*(5), e1006059. https://doi.org/10.1371/journal.pgen.1006059.

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

Bernstein, F. (1931). Die geographische Verteilung der Bludgruppen und ihre anthropologische Bedeutung. In *Comitato Italiano per o studio dei problemi della populazione* (pp. 227–243). : Instituto Poligraphico dello Stato.

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, *20*, 63–67. https://doi.org/10.1007/s11222-009-9116-0.

Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data - An approximate bayesian computation approach. *PLoS Genetics*, *12*(3), e1005877. https://doi.org/10.1371/journal.pgen.1005877.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., & Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, *368*(6470), 455–457. https://doi.org/10.1038/368455a0.

Brandenburg, J.-T., Mary-Huard, T., Rigaill, G., Hearne, S. J., Corti, H., Joets, J., Vitte, C., Charcosset, A., Nicolas, S. D., & Tenaillon, M. I. (2017). Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLOS Genetics*, *13*(3), e1006666. https://doi.org/10.1371/journal.pgen.1006666.

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

Buzbas, E. O., & Rosenberg, N. A. (2015). AABC: approximate approximate Bayesian computation for inference in population-genetic models. *Theoretical Population Biology*, *99*, 31–42. https://doi.org/10.1016/j.tpb.2014.09.002.

Buzbas, E. O., & Verdu, P. (2018). Inference on admixture fractions in a mechanistic model of recurrent admixture. *Theoretical Population Biology*, *122*, 149–157. https://doi.org/10.1016/j.tpb.2018.03.006.

Cavalli-Sforza, L. L., & Bodmer, W. F. (1971). The genetics of human populations. W. H. Freeman.

Chakraborty, R., & Weiss, K. M. (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 of the United States of America*, *85*(23), 9119–9123.

Chimusa, E. R., Defo, J., Thami, P. K., Awany, D., Mulisa, D. D., Allali, I., Ghazal, H., Moussa, A., & Mazandu, G. K. (2018). Dating admixture events is unsolved problem in multi-way admixed populations. *Briefings in Bioinformatics*, 144–155. https://doi.org/10.1093/bib/bby112.

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

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*, 475–479.

Epstein Michael P., Duren William L., Boehnke M. (2000). Improved Inference of Relationship for Pairs of Individuals. *The American Journal of Human Genetics*, *67*(5), 1219–1231. http://dx.doi.org/10.1016/s0002-9297(07)62952-8.

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

Estoup, A., Raynal, L., Verdu, P., & Marin, J. M. (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 of the SFdS*, *159*(3), 167–190.

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLOS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905.

Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334. https://doi.org/10.1093/bioinformatics/btr124.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.

Fisher, R. A. (1922). Darwinian evolution of mutations. *Eugenics Review*, *14*(1), 31–34.

Foll, M., Shim, H., & Jensen, J. D. (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, C., Bybjerg-Grauholm, J., Marin-Padrón, L. C., Gomez-Cabezas, E. J., Bækvad-Hansen, M., Hansen, C. S., Le, P., Hougaard, D. M., Verdu, P., Mors, O., Parra, E. J., & Marcheco-Teruel, B. (2018). Exploring Cuba's population structure and demographic history using genome-wide data. *Scientific Reports*, *8*(1), 11422. https://doi.org/10.1038/s41598-018-29851-3.

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., & Estoup, A. (2017). Deciphering the routes of invasion of *Drosophila suzukii* by means of ABC random forest. *Molecular Biology and Evolution*, *34*(4), 980–996. https://doi.org/10.1093/molbev/msx050.

Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393.

Goldberg, A., Verdu, P., & Rosenberg, N. A. (2014). Autosomal admixture levels are informative about sex bias in admixed populations. *Genetics*, *198*(3), 1209–1229.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*(2), 607–619. https://doi.org/10.1534/genetics.112.139808.

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518.

Jay, F., Boitard, S., & Austerlitz, F. (2019). An ABC method for whole-genome sequence data: Inferring paleolithic and neolithic human expansions. *Molecular Biology and Evolution*, *36*(7), 1565–1579. https://doi.org/10.1093/molbev/msz038.

Lipson, M., Loh, P. R., Levin, A., Reich, D., Patterson, N., & Berger, B. (2013). Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular Biology and Evolution*, *30*(8), 1788–1802. https://doi.org/10.1093/molbev/mst099.

Loh, P. R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, *193*(4), 1233–1254. https://doi.org/10.1534/genetics.112.147330.

Long, J. C. (1991). The genetic structure of admixed populations. *Genetics*, *127*(2), 417–428.

Malinsky, M., Trucchi, E., Lawson, D. J., & Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data. *Molecular Biology and Evolution*, *35*(5), 1284–1290. https://doi.org/10.1093/molbev/msy023.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *American Journal of Human Genetics*, *100*(4), 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., & Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, *23*(11), 1817–1828. https://doi.org/10.1101/gr.159426.113.

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the timing of multiple admixture pulses during local ancestry inference. *Genetics*, *210*(3), 1089–1107. https://doi.org/10.1534/genetics.118.301411.

Moorjani, P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L. I., Atzmon, G., Burns, E., Ostrer, H., Price, A. L., & Reich, D. (2011). The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLOS Genetics*, *7*(4), e1001373. https://doi.org/10.1371/journal.pgen.1001373.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*(3), 583–590.

Ni, X., Yuan, K., Liu, C., Feng, Q., Tian, L., Ma, Z., & Xu, S. (2019). MultiWaver 2.0: modeling discrete and continuous gene flow to reconstruct complex population admixtures. *European Journal of Human Genetics*, *27*(1), 133–139. https://doi.org/10.1038/s41431-018-0259-3.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., Froment, A., Heyer, E., Massougbodji, A., Fortes-Lima, C., Migot-Nabias, F., Bellis, G., Dugoujon, J.-M., Pereira, J. B., Fernandes, V., Pereira, L., … Quintana-Murci, L. (2017). Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, *356*(6337), 543–546. https://doi.org/10.1126/science.aal1988.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLOS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091.

Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. *Bioinformatics*, *32*(6), 859–866. https://doi.org/10.1093/bioinformatics/btv684.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., Bender D., Maller J., Sklar P., de Bakker P. I. W., Daly M. J., Sham P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. http://dx.doi.org/10.1086/519795.

R Development Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867.

Robert, C. P., Mengersen, K., & Chen, C. (2010). Model choice versus model criticism. *Proceedings of the National Academy of Sciences*, *107*(3), E5–E5. https://doi.org/10.1073/pnas.0911260107.

Sisson, S. A., Fan, Y., & Beaumont, M. A. (2018). In S. A. Sisson, Y. Fan, & M. A. Beaumont (Eds.). Handbook of approximate bayesian computation. pp 678. New York, NY: Chapman and Hall/CRC.

Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalen, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, *25*(11), 1515–1519. https://doi.org/10.1016/j.cub.2015.04.019.

Stryjewski, K. F., & Sorenson, M. D. (2017). Mosaic genome evolution in a recent and rapid avian radiation. *Nature Ecology and Evolution*, *1*(12), 1912–1922. https://doi.org/10.1038/s41559-017-0364-7.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*(2), 505–518.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Théry, S., Froment, A., Le Bomin, S., Gessain, A., Hombert, J.-M., Van der Veen, L., Quintana-Murci, L., Bahuchet, S., & Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Current Biology*, *19*(4), 312–318. https://doi.org/10.1016/j.cub.2008.12.049.

Verdu, P., Jewett, E. M., Pemberton, T. J., Rosenberg, N. A., & Baptista, M. (2017). Parallel trajectories of genetic and linguistic admixture in a genetically admixed creole population. *Current Biology*, *27*(16), 2529–2535 e2523. https://doi.org/10.1016/j.cub.2017.07.002.

Verdu, P., & Rosenberg, N. A. (2011). A general mechanistic model for admixture histories of hybrid populations. *Genetics*, *189*(4), 1413–1426. https://doi.org/10.1534/genetics.111.132787.

Wakeley, J., King, L., Low, B. S., & Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, *190*(4), 1433–1445. https://doi.org/10.1534/genetics.111.135574.

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, *182*(4), 1207–1218. https://doi.org/10.1534/genetics.109.102509.

Weir, B. S., & Cockerham, C. C. (1984). Estimating *F*-statistics for the analysis of population-structure. *Evolution*, *38*(6), 1358–1370.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*(2), 97–159.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Fortes-Lima CA, Laurent R, Thouzeau V, Toupance B, Verdu P. Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. *Mol Ecol Resour.* 2021;21:1098–1117. https://doi.org/10.1111/1755-0998.13325

3.3 Plateau technique Paléogénomique et génétique moléculaire P2GM

Le plateau technique Paléogénomique et génétique moléculaire (P2GM) de l'UMR7206 Eco-anthropologie a été intégré, dès 2016, à la plateforme analytique du Muséum National d'Histoire Naturelle (PAM). Il est mutualisé par l'UMR7206 aux laboratoires du MNHN et des autres organismes de recherche français et internationaux, et réalise des travaux d'expertise.

Présentation

Le plateau P2GM (https://www.ecoanthropologie.fr/fr/plateau-technique-6204) comprend différents espaces et équipements mutualisés afin d'analyser tous types de matériaux génétiques et de réaliser des dosages endocrinologiques. P2GM comprend notamment des espaces dédiés aux extractions, préparations de librairies, amplifications et qualifications des ADN : une salle blanche pour l'ADN ancien et dégradé et des laboratoires pour l'ADN moderne ou amplifié. Les environnements de ces espaces sont strictement contrôlés pour limiter toute contamination. Le personnel de l'UMR7206 est affecté à P2GM : deux ingénieurs d'études (Sophie Lafosse CNRS et José Utge MNHN) et deux assistantes ingénieures (Françoise Dessarps-Freichey CNRS, fin de carrière en 2020, et Amélie Chimènes CNRS) sont les responsables scientifiques et techniques des projets et responsables opérationnels des projets externes. P2GM est coordonné par Paul Verdu (CR CNRS) et Céline Bon (MC MNHN) (Figure 1).
Entre 2017 et 2022, 5 membres de l'UMR7206 (hors-P2GM) ont réalisé leurs analyses sur P2GM ainsi que plus de 20 chercheurs, ingénieurs et étudiants extérieurs à l'unité. 43 étudiants, du BTS jusqu'au post-doctorat, ont été formés en partie sur P2GM. P2GM a contribué à 29 articles scientifiques, plus de 50 communications à des congrès, et plus de 40 rapports de stages, thèses de doctorat et habilitations à diriger des recherches.

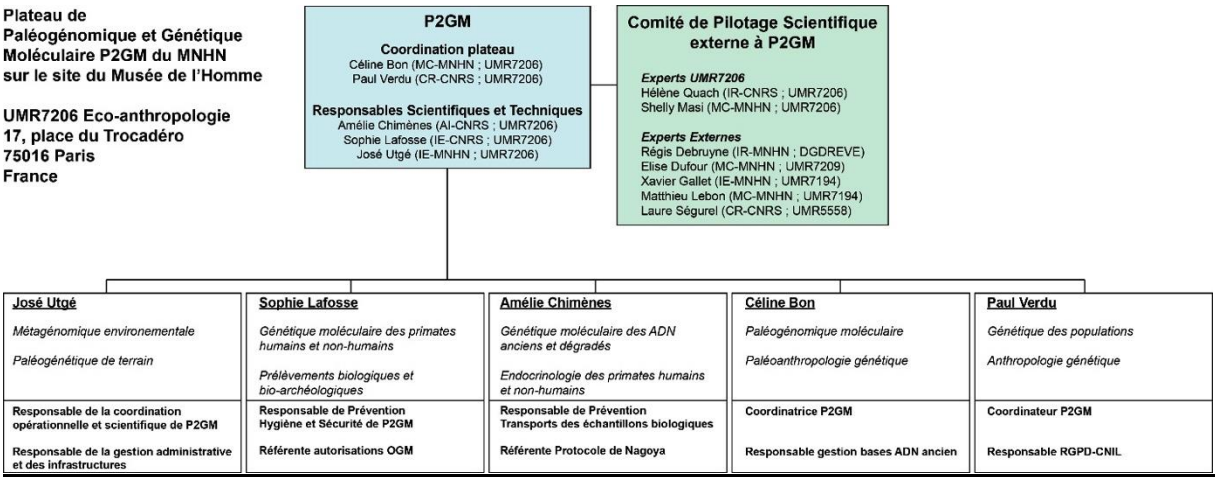

Figure 1 : organigramme de P2GM au 31 décembre 2022

Activités scientifiques

Entre 2017 et 2022, 129 projets ont été menés sur P2GM. Les six équipes de l'UMR7206 ont toutes conduit plusieurs projets sur P2GM. Ainsi, P2GM est un outil d'interdisciplinarité et d'innovation scientifique pour toute l'UMR7206, dont l'expertise en analyse des ADN, notamment anciens et dégradés, est sollicitée par d'autres UMR. Nous présentons trois projets illustrant la diversité de ces activités scientifiques.

Evolution de la diversité génétique humaine autour de la Mer Caspienne

Pour étudier les mouvements de populations humaines en Asie Centrale et dans le sud du Caucase du Néolithique à l'Âge du Fer, l'ADN de 138 échantillons a été extrait par les membres de P2GM dans la salle blanche, permettant d'éclaircir les mécanismes de néolithisation dans le sud du Caucase (Guarino-Vignon et al. Comm Biol 2022), la structure génétique des populations de la Civilisation de l'Oxus à l'Âge du Bronze (Guarino-Vignon et al. Front Genetics 2022), ainsi que celle de l'Iran à l'Âge du Fer (*in prep*). Ces données paléogénétiques ont nourri le doctorat de Perle Guarino-Vignon en collaboration entre les équipes AGene et ABBA de l'UMR7206 et d'autres unités (UMR7209, UMR5133, UMR7192).

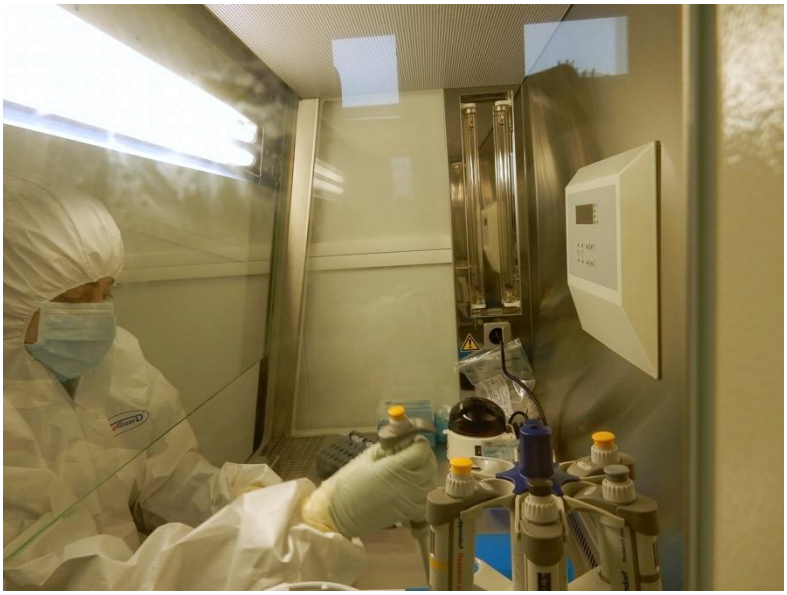

Figure 2 : Extraction d'ADN ancien humain d'Asie Centrale dans la salle blanche de P2GM

Détermination des relations de parenté entre primates sauvages

Pour comprendre la structure sociale de primates non humains dans leur environnement, P2GM a réalisé l'extraction d'ADN dégradé de 186 fèces de Gorilles de l'Ouest collectées en RCA. Ces extraits ont été génotypés pour 10 marqueurs microsatellites mis au point sur P2GM, puis analysés par des méthodes de génétique des populations avec les équipes AGene et IPE de l'UMR7206. Les résultats montrent que les mâles dispersent sur de grandes distances alors que les femelles dispersent très localement en évitant les reproductions consanguines (Masi et al. Mol Ecol Evol 2021). Ce projet se poursuit à une échelle régionale à partir de fèces collectées au Cameroun, en RDC et en RCA. Ces approches ont ensuite été appliquées à des populations de Chimpanzés du Sénégal dans plusieurs expertises menées par P2GM pour des sociétés privées (Oryx, Sylvatrop), dans le cadre d'études d'impact de projets industriels sur la conservation de la biodiversité locale.

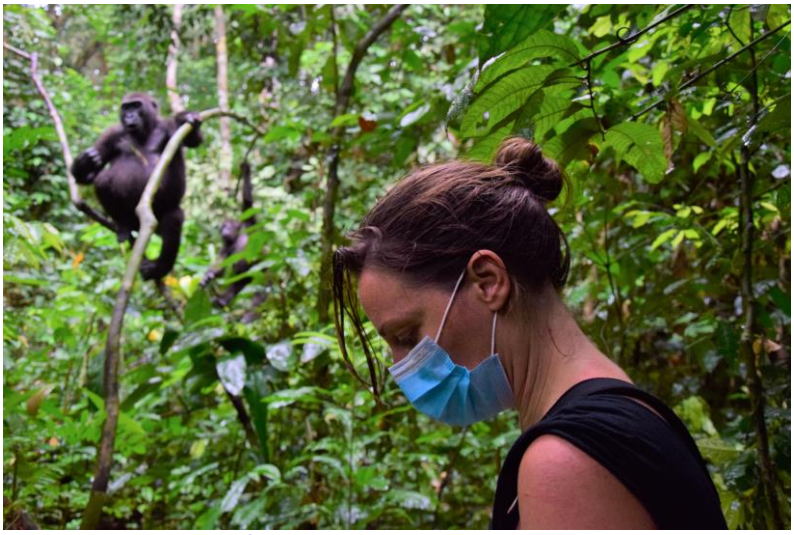

Figure 3 : Shelly Masi étudiant les Gorilles de l'Ouest en RCA ©Marcella Sana

Projet TARA

Dans le cadre du projet TARA-Ocean, P2GM et l'équipe de Chris Bowler (IBENS, ENS) ont mis au point un protocole d'analyse de sédiments sous-marins anciens afin d'étudier l'évolution de la diversité des diatomées. Les méthodes d'analyses métagénomiques en shotgun et en metabarcoding ont permis de montrer que les reconstructions paléo-environnementales à partir de sédiments peuvent être biaisées selon les méthodes utilisées (Armbrecht et al. ISME Com 2021). Cette étude a fait l'objet du post-doctorat de Linda Armbrecht et se poursuit sur P2GM avec les doctorats de Mathilde Bourreau (IBENS) et Manon Sabourdy (UMR5805) sur la réponse des micro-organismes marins aux changements climatiques passés et présents.

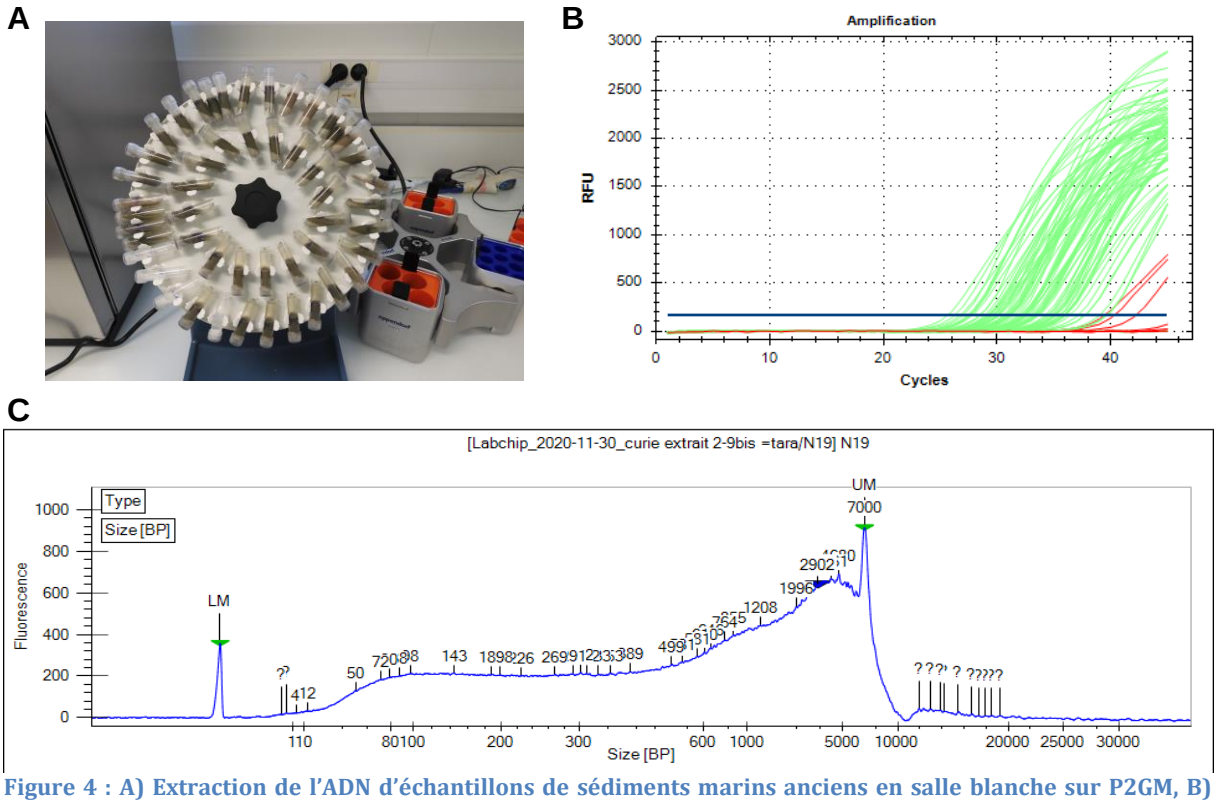

Figure 4 : A) Extraction de l'ADN d'échantillons de sédiments marins anciens en salle blanche sur P2GM, B) Amplification qPCR : les échantillons apparaissent en vert, les contrôles négatifs en rouge. C) Distribution des longueurs de fragments d'ADN dans un échantillon sédimentaire marin ancien caractérisée par électrophorèse en capillaire LabChip.

Covid19 et confinement

Durant les périodes de confinements dus au Covid 19 en 2020, P2GM s'est mobilisé pour fournir à l'AERES les équipements de protection personnelle à sa disposition. Avec l'aide de l'UMR7194, P2GM a aussi produit plus de 50L de solution hydroalcoolique pour les personnels d'astreinte au MNHN. Amélie Chimènes a été déployée à sa demande par le CNRS pour la réalisation de tests PCR COVID à l'Hôpital Grand-Est (https://www.paris-centre.cnrs.fr/fr/cnrsinfo/face-au-covid-19-scientifiques-benevoles-en-renfort-hopital-marne-la-vallee).

Figure 5 : Préparation de 25L de solution hydroalcoolique sur P2GM pour les personnels d'astreinte au MNHN en Mai 2020.

Activités de diffusion scientifique et formation

Les membres de P2GM sont impliqués dans l'enseignement et la formation permanente par la mise en place de plusieurs stages pour les personnels du MNHN, un TP dans le cours Museum « Paléogénétique des restes archéologiques », et une journée de formation permanente pour les personnels techniques de l'APHP.

En outre, P2GM a participé à de nombreuses actions de diffusion auprès du grand public :

- plus de 30 interviews dans la presse écrite et audiovisuelle
- ateliers d'extraction d'ADN à la Fête de la Science et auprès de publics scolaires
- conseil scientifique d'une pièce de théâtre (https://theatre-cite.com/programmation/a-venir/spectacle/neandertal-et-ceux-qui-dansaient)
- présentation de la paléogénétique lors des Journées Européennes de l'Archéologie
- participation à des documentaires télévisés, comme celui consacré à l'étude d'un crâne du XVIIIème siècle : Bruto (https://www.france.tv/documentaires/voyages/2604281-bruto.html).

Figure 6 : Tournage du documentaire « Bruto » sur P2GM avec la réalisatrice Carole Grand.