MOLECULAR ECOLOGY RESOURCES

# Supplemental Note S1 for:

## 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] CNRS, Muséum National d'Histoire Naturelle, Université de Paris, UMR7206 Eco-anthropologie, Paris, France
[2] Sub-department of Human Evolution, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden
[3] CNRS, Université Paris-Dauphine, PSL University, UMR7534 Centre de Recherche en Mathématiques de la Décision, Paris, France
[4] ENS, PSL University, EHESS, CNRS, Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Etudes Cognitives, Paris, France

* Joint first authors
# Author for correspondence: Paul Verdu,
CNRS, Muséum National d'Histoire Naturelle, Université de Paris;
UMR7206 Eco-anthropologie (EA);
Address: Musée de l'Homme, 17, place du Trocadéro, 75016 Paris, France;
email: paul.verdu@mnhn.fr; tel: +33 1 44 05 73 17

**KEYWORDS**
Admixture; Approximate Bayesian Computation; Inference; Population Genetics; Machine Learning

MOLECULAR ECOLOGY
RESOURCES

MOLECULAR ECOLOGY
RESOURCES

# Supplementary Note S1

## Introduction

We henceforth present a general summary of the possibilities offered by the *MetHis* software package, beyond the proof-of-concept implementation of the *MetHis*-ABC framework to reconstruct highly complex admixture histories from SNP data presented in the main text of the article (see a schematic figure presented in **Supplementary Note S1 Figure S1.3** below). This supplementary note represents a complement to the software manual deposited on GitHub, but does not replace it.

The *MetHis* software package is composed of three separate tools specifically designed for conducting genetic data simulations in an admixed population H under any version of the two source populations general model from Verdu and Rosenberg (2011). *MetHis* is designed primarily to reconstruct complex admixture histories with machine-learning Approximate Bayesian Computation Random-Forest model-choice (Pudlo et al., 2016; Raynal et al., 2019) and Neural-Network posterior parameter estimation (Csilléry et al., 2012).
The software package and source codes are available at https://github.com/romain-laurent/MetHis, together with user manual and example files corresponding to the implementation of *MetHis* described in the main text of the manuscript.

The main tool (*MetHis* itself) is a C software to simulate independent SNPs or microsatellite markers in an admixed population H under models for which parameters are set by the user.

The *MetHis parameter generator* tool is a Python software to build lists of parameter values within prior bounds set by the user, to be used for simulations with *MetHis*.

The *MetHis summary-statistics* tool is a Python and R software to calculate summary statistics directly from the genetic data simulated with *MetHis* simulation tool.

MOLECULAR ECOLOGY
RESOURCES

## 1 | Admixture models considered in MetHis

*MetHis* allows simulating data under any versions of the two source populations version of the Verdu and Rosenberg (2011) general mechanistic model of admixture. Nine admixture specific scenarios that *MetHis* can simulate can be found in the proof-of-concept implemented in the main text of the article. However, *MetHis* is not limited to these nine models.

Note that parameters can be fixed by the user for deterministic simulations with *MetHis*, or drawn from prior distributions using *MetHis parameter generator* tool (see section **2** below), or separately as a list of parameters provided, independently of *MetHis* tools, by the user (and only fitting the input format required by *MetHis* simulator).

Let *G* be the number of generations of the admixture process set freely by the user. Note that generation 0 is the founding of the admixture process to be specified by the user, as in section 2.1.1 of the main text.

For simplicity, we describe possible models implemented in *MetHis* from a given source population S. Models from the sources can then be combined at will (provided that they satisfy that the sum of introgression parameters at each generation never exceeds 1, see section **1.4** below) as illustrated in the main text of the manuscript.

### 1.1 | n-pulse(s) of admixture

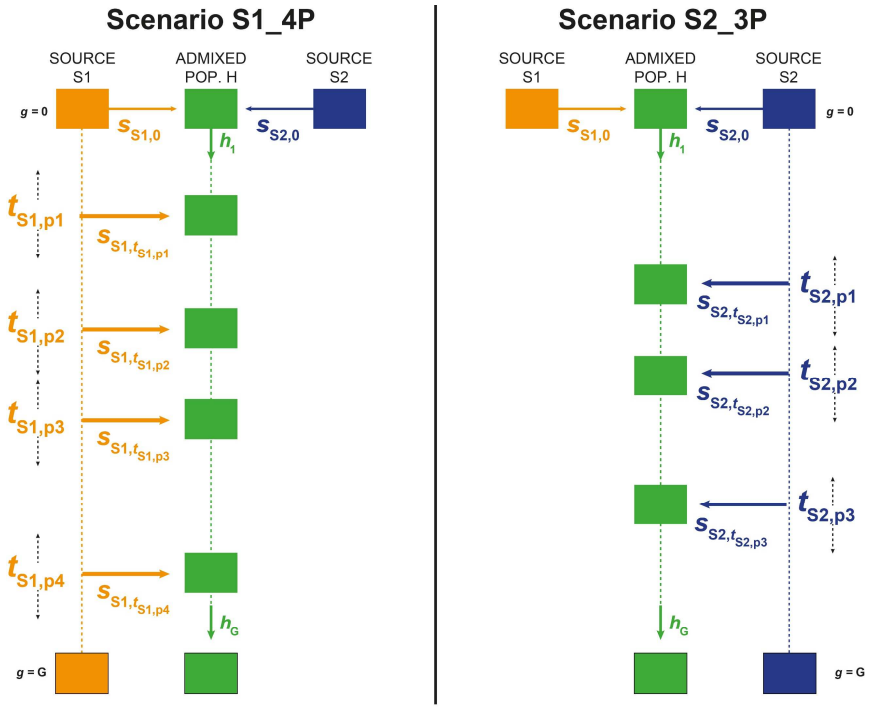
*MetHis* allows simulating *n*-pulses of admixture from either source after the foundation of population H at generation 0, with *n* superior or equal to 0. These models are parameterized in *MetHis*, for each source population S separately, by the following parameters set by the user:

- *n*, the number of pulses desired from a given source population S after the foundation of population H, between 1 and *G*. Note that $n = G$ corresponds exactly to the two-source "full-blown" version of the Verdu and Rosenberg (2011) model. Alternatively, $n = 0$ corresponds to an admixture model with no additional admixture event from source population S after the foundation admixture event at generation 0. For instance, in the main text of the manuscript, we consider several models where $n = 2$.

- For each one of the *n* pulses, $t_{S,n}$ in [1,*G*] determines the time for the *n*-th pulse from population S;

- For each one of the *n* pulses, $s_{S,tS,n}$ in [0,1] determines the rate of introgression from population S at time $t_{S,n}$.

**Supplementary Note S1 Figure 1.1**: Figure illustrates two possible *n*-pulses, *n=4* and *n=3* respectively on the left and right panels, of admixture models implementable in *MetHis*, from either source population S1 or S2.




Supplementary Note S1 Figure 1.1

## 1.2 | Recurring admixture

*MetHis* allows for the simulation of periods of recurring admixture from either source. These are parameterized by five separate parameters to be set by the user:

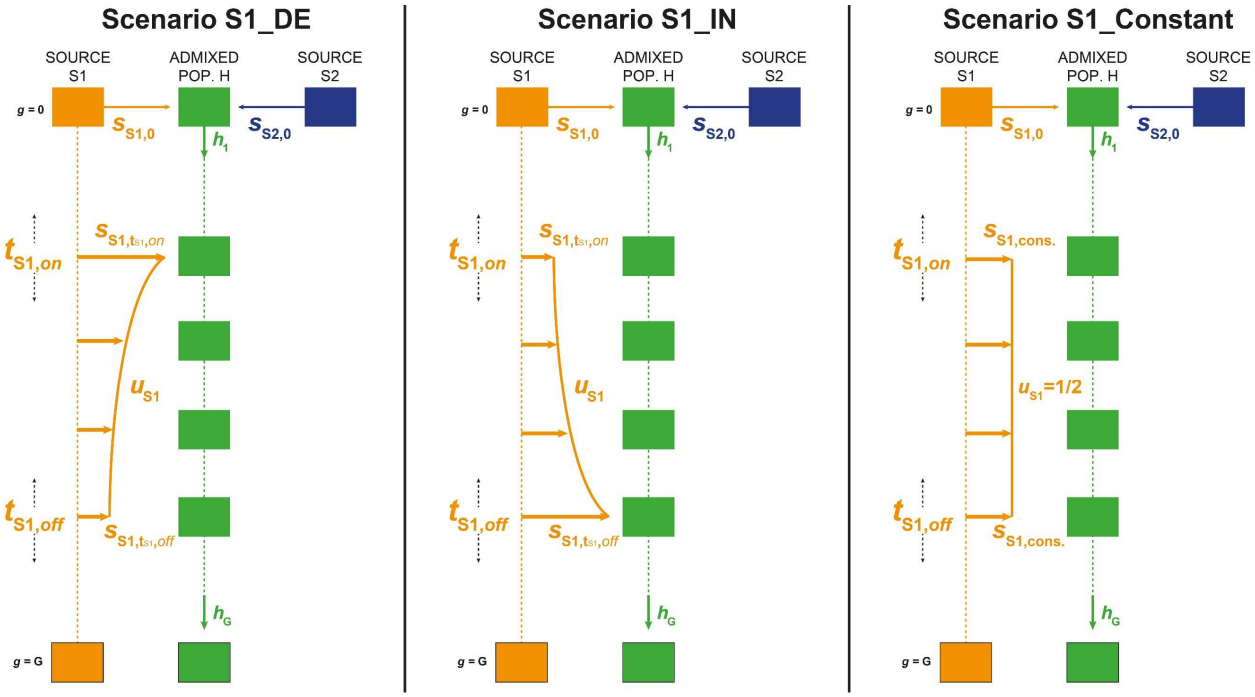
- Two "time" parameters, $t_{S,on}$ and $t_{S,off}$, with $t_{S,off} > t_{S,on}$, $t_{S,on}$ in [1,*G*-1] and $t_{S,off}$ in [2,*G*]. They determine, respectively, the onset and end of the recurring admixture period set by the user.

- Two introgression rates from population S, $s_{S,tS,on}$ and $s_{S,tS,off}$, each in [0.,1], respectively corresponding to the relative contribution of population S to the gene-pool of population H, respectively at the onset and end of the admixture period.

- The $u_S$ parameter in [0,1/2] which controls the shape of the recurring admixture in between $t_{S,on}$ and $t_{S,off}$, as described in the main text of the manuscript and detailed in **Supplementary Note S2**.

For instance, the user interested in simulating a constant recurring period of admixture (as in Verdu, & Rosenberg, (2011) special-case, and also explored in Buzbas, & Verdu, (2018)) from source population S, simply has to set: $s_{S,tS,on} = s_{S,tS,off}$, and $u_S = 1/2$. Monotonically recurring increasing or decreasing admixture can also be set easily by setting the corresponding relationship between $s_{S,tS,on}$ and $s_{S,tS,off}$, as exemplified in the main text of the article.

**Supplementary Note S1 Figure 1.2**: This figure illustrates three possible recurring admixture models implementable in *MetHis*, from one source population S1. The leftmost scenario implements a recurring decreasing admixture model from S1. The central scenario implements a recurring increasing admixture model from S1. The rightmost scenario implements a recurring constant admixture model from S1.




Supplementary Note S1 Figure 1.2

***IMPORTANT NOTE***: these models are readily implemented in *MetHis* and parameter lists under these models can be automatically generated with *MetHis parameter generator* tools. However, the user can build her/his own parameter list independently of *MetHis*, and use it as input in *MetHis* to simulate other models at will, such as, for instance, two separate periods of recurring admixture during the admixture history of population H.

### 1.3 | Effective population size in the admixed population

*MetHis* allows the user to set parameters, at each generation of the admixture process, controlling the diploid effective population size $N_g$, with $g$ in [0,$G$]. In the main text, we chose, for simplicity, to fix $N_g$. Alternatively, *MetHis* readily implements models of monotonic rectangular hyperbolic decrease or increase of $N_g$ across generations, controlled by four parameters set by the user:

- $N_0$, the diploid effective population size of the admixed population at foundation at generation 0.

- $N_1$, the diploid effective population size of the admixed population after foundation at generation 1.

- $N_g$, the diploid effective population size of the admixed population at the end of the admixture process at generation *G*.

- The $u_N$ parameter in [0,1/2] which controls the shape of the change in effective population size between generation 1 and present. This parameter is similar to the $u_S$ parameter for introgression rates under a recurring admixture scenario, and has the same properties as detailed in **Supplementary Note S2**.

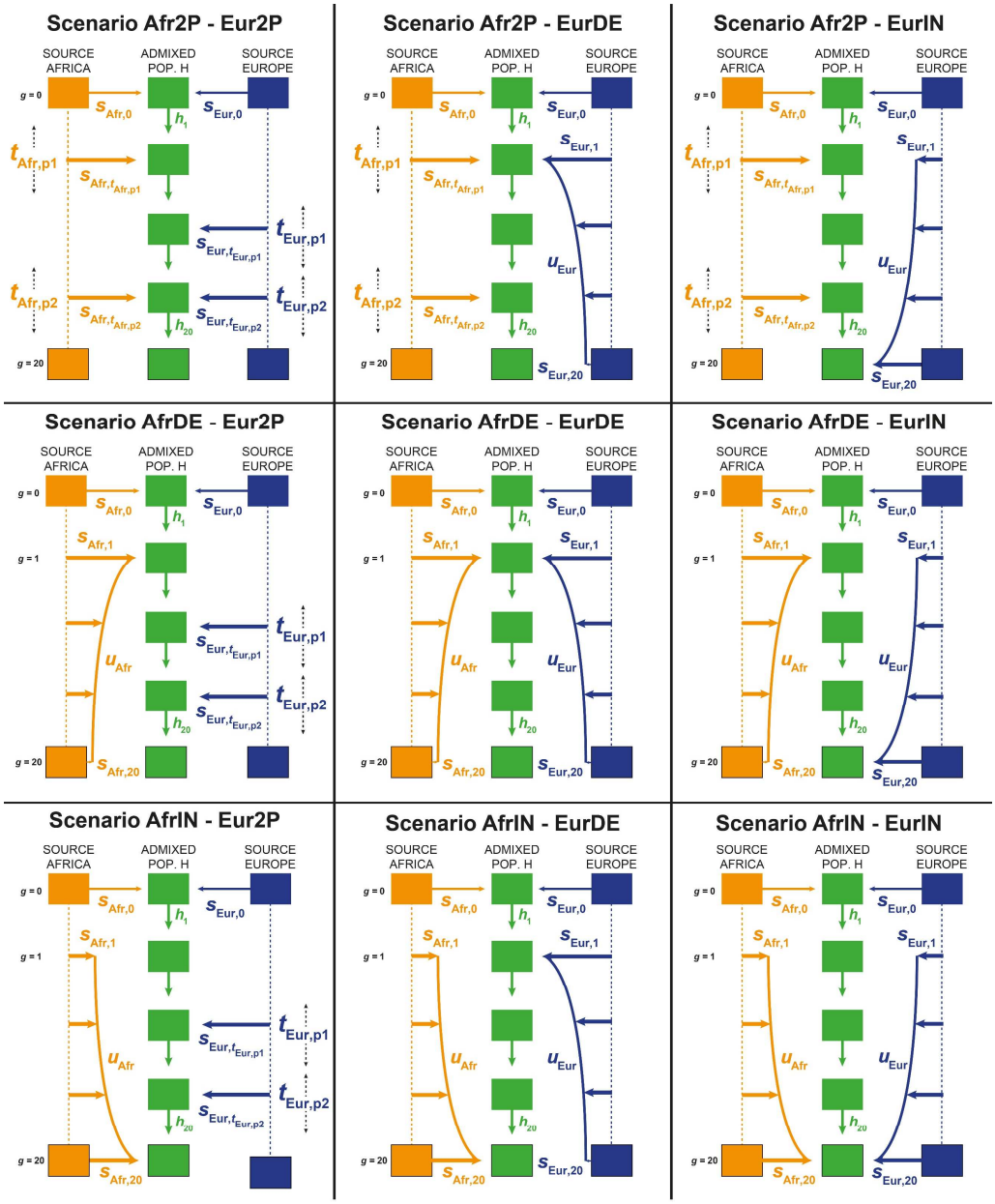
***IMPORTANT NOTE***: as above, these models are readily implemented in *MetHis* and parameter lists under these models can be automatically generated with *MetHis parameter generator* tools. However, the user can build her/his own parameter list independently of *MetHis*, and use it as input in *MetHis* to simulate other models at will. For instance, user can, independently of *MetHis*, define a bottleneck change in effective population sizes, calculate values of *Ne* at each generation following this model, and input this *Ne* list for *MetHis* simulations.

### 1.4 | Combining admixture models from both source populations

As exemplified in the main text of the article, admixture models from either source can be combined at will by the user, provided that they satisfy, at each generation of the admixture process between 1 and *G*, $s_{S1,g} + s_{S2,g} + h_g = 1$, as defined in Verdu and Rosenberg (2011).

For such an illustration, we reproduce here Figure 1 of the main text.

MOLECULAR ECOLOGY
RESOURCES




Main Text Figure 1

## 2 | Generating parameter lists with *MetHis*

*MetHis parameter generator* tool allows to readily create parameter lists to be used as input for *MetHis* simulations.

This parameter generator considers models described in the above sections **1.1**, **1.2**, **1.3**, and **1.4**, and uses, as input, the boundaries of the parameters described previously, set by the user.

The user needs to define the number of simulations desired, the parameter generator will then draw model parameters as defined above in a uniform distribution within the parameter boundaries set by the user. *MetHis parameter generator* then automatically builds parameter lists satisfying, at each generation of the admixture process between 1 and *G*, $s_{S1,g} + s_{S2,g} + h_g = 1$, as defined in Verdu and Rosenberg (2011).

Again, note that users are invited to build their own parameter lists satisfying the above condition and input format, at will.

## 3 | Forward-in-time simulations with *MetHis*

Elements in this section are also described in the Materials and Methods section of the main text of the article.

### 3.1 | Genetic data from source populations

*MetHis*, in its current form, does not allow for simulating the source populations for the admixture processes modeled in Verdu and Rosenberg (2011). Simulating source populations can be done separately using existing genetic data simulation software such as, for instance among many others, *fastsimcoal2* sequential coalescent (Excoffier, Dupanloup, Huerta-Sanchez, Sousa, & Foll, 2013; Excoffier & Foll, 2011).

Alternatively, if genetic data is readily available from known source populations at the root of the admixture process, source populations can be simulated directly from observed allelic frequencies as described in the main text section 2.2.2.

### 3.2 | Simulating the admixed population with *MetHis*

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 $s_{S,g}$, or from population H with probability $h_g = 1 - \left(s_{S1,g} + s_{S2,g}\right)$, randomly builds a haploid gamete of independent markers for each parent, and pairs the two constructed gametes to create the new individual.

## 4 | Genetic data simulated with *MetHis*

### 4.1 | Single Nucleotide Polymorphisms and Microsatellites

*MetHis* allows for simulating any number of independent SNPs or microsatellite markers set by the user. The admixed population is founded at generation 0 by the alleles respectively present in the source populations used as input for *MetHis*.
SNPs need to be biallelic and microsatellites should be coded in numbers of repetition of the motif (decimals are allowed for motifs affected by insertions and deletions, see below **4.2**).

### 4.2 | General Stepwise Mutation Model in *MetHis*

For microsatellite data, *MetHis* implements a GSMM model with possible insertion and deletions (Estoup, Jarne, & Cornuet, 2002), with a possibly infinite range of contiguous allelic states, and fully parameterized by the user.

The bounds of the uniform prior distribution for the mean mutation rate ($\mu$) across microsatellite loci are set by the user. Then, the mutation rates for each locus are drawn independently from a Gamma distribution with mean=$\mu$ and shape=2. The length in number of repeats of all mutation events is set to follow a geometric distribution of mean parameter $p$, drawn in a uniform distribution bounded by the user. Then, the length in number of repeats for each marker separately is drawn from a Gamma distribution with mean=$p$ and shape=2. Finally, in order to simulate possible insertion and deletion that alter the microsatellite motif (e.g. di-, tri-, tetra-nucleotide, etc.), we draw the rate of a single nucleotide insertion-deletion event, independently for each marker, in a Gamma distribution with mean=$2.5 \times 10^{-8}$ and shape=2.
An example of this mutation model for tetranucleotide microsatellite markers implemented for ABC demographic inference can be found in Verdu et al. (2009).

Note that we recommend to consider only microsatellites with the same repetition-motif a priori, as microsatellites are known to have very different mutation rates across motifs.

## 5 | Sampling data simulated with *MetHis*

*MetHis* can sample any number of individuals, at most equal to $N_G$, in the admixed population at generation *G* (present), set by the user.
The user can choose to sample individuals randomly, or excluding related individuals (see main text for an example of the latter).

## 6 | Summary-statistics calculation with *MetHis*

At the end of each simulation, *MetHis summary statistics* calculation tool can be used to automatically calculate the following population genetics summary statistics. Some (but not all) of the statistics computed with *MetHis* and presented in this section are also described in the Materials and Methods section of the manuscript.

***IMPORTANT NOTE 1***: the user is not forced to use *MetHis summary statistics* calculation tools. Simulated genetic data can be used as input for alternative population genetics software.

***IMPORTANT NOTE 2***: Given the model design, and given how source populations are simulated, some of the statistics below will be, *a posteriori*, constant, or possibly uninformative. For instance, in the proof-of-concept investigated in the main text, source populations are fixed reservoirs. Thus in our case studies in the main text of the article, all of the statistics calculated only between population S1 and S2, below, are constant and thus uninformative, or only variable due to sampling. Similarly, as individuals are sampled to be unrelated and effective population sizes constant in this example (for simplicity), inbreeding coefficient statistics are uninformative.
Nevertheless, other implementations and case-studies will benefit from these statistics beyond our specific case-study, for instance when interested in changes in effective population sizes in the admixed population where the inbreeding coefficient may help segregating among simulations.

### 6.1 | Distribution of admixture fractions as a summary statistic

For each simulated dataset, we first calculated a pairwise inter-individual Allele Sharing Dissimilarity (Bowcock et al., 1994) matrix using *asd* software (https://github.com/szpiech/asd) using all pairs of sampled individuals and all markers (whether SNPs or microsatellites). Then we projected in two dimensions this pairwise ASD matrix with classical unsupervised metric Multidimensional Scaling (MDS) using the *cmdscale* function in *R*. We expect 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 individuals orthogonally onto this axis, and calculate 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 the centroid when being genetically close to 100% from one source population or the other.

*MetHis* provides, as summary-statistics, the mean, mode, variance, skewness, kurtosis, minimum, maximum, and all 10%-quantiles of the admixture distribution in population H.

### 6.2 | Within-population summary statistics

#### *6.2.1 | Within-population summary statistics for SNP data*

- SNP by SNP expected heterozygosities (Nei, 1978) within the admixed and source populations, separately.
- Mean expected heterzygosity across SNPs within the admixed and source populations, separately.
- Variance of expected heterozygosity across SNPs within the admixed and source populations, separately.
- Mean pairwise ASD (see section 5.1) within the admixed and source populations, separately.
- Variance of pairwise ASD (see section 5.1) within the admixed and source populations, separately.
- Inbreeding coefficients are calculated with the method of moments similarly as in *vcftools* (Danecek, et al., 2011) within the admixed and source populations, separately.

#### *6.2.2 | Within-population summary statistics for microsatellite data*

- Mean number of microsatellite alleles per locus within the admixed and source populations, separately.

- Mean expected heterzygosity across microsatellites within the admixed and source populations, separately.
- Mean allele size variance across microsatellites within the admixed and source populations, separately.
- Mean pairwise ASD (see section **6.1** above) within the admixed and source populations, separately.
- Variance of pairwise ASD within the admixed and source populations, separately.

### 6.3 | Between-population summary statistics

#### *6.3.1 | Between-population summary statistics for SNP data*

- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between the admixed population and source population S1.
- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between the admixed population and source population S2.
- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between source population S1 and S2.
- $f_3$ statistics (Patterson et al., 2012) considering the admixed population as sink and populations S1 and S2 as sources.
- Mean pairwise ASD (see section **6.1** above) between the admixed population and source population S1.
- Mean pairwise ASD (see section **6.1** above) between the admixed population and source population S2.
- Mean pairwise ASD (see section **6.1** above) between source population S1 and S2.

#### *6.3.2 | Between-population summary statistics for microsatellite data*

- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between the admixed population and source population S1.
- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between the admixed population and source population S2.
- Multilocus pairwise $F_{ST}$ (Weir, & Cockerham, 1984) between source population S1 and S2.
- Mean pairwise ASD (see section **6.1** above) between the admixed population and source population S1.
- Mean pairwise ASD (see section **6.1** above) between the admixed population and source population S2.
- Mean pairwise ASD (see section **6.1** above) between source population S1 and S2.

## 7 | Computational cost of simulating and calculating summary statistics with *MetHis*

Using 27 cores and the design described in the material and methods of the main text of the article (**Figure 1**, **Table 1**), we performed the 90,000 simulations with *MetHis* in four days, with 2/3 of that time for summary statistics calculations only. 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 (within reason), but rather on the effective population size in the admixed population set by the user (see section **1.3** above).

## 8 | *MetHis* outputs

The *MetHis parameter generator* tool outputs ordered lists of parameter vectors corresponding to the model of interest, in text format. These are parameter reference tables ready to be used as input for machine-learning ABC *R* packages *abcrf* (Pudlo et al., 2016; Raynal et al., 2019) and *abc* (Csilléry et al., 2012).

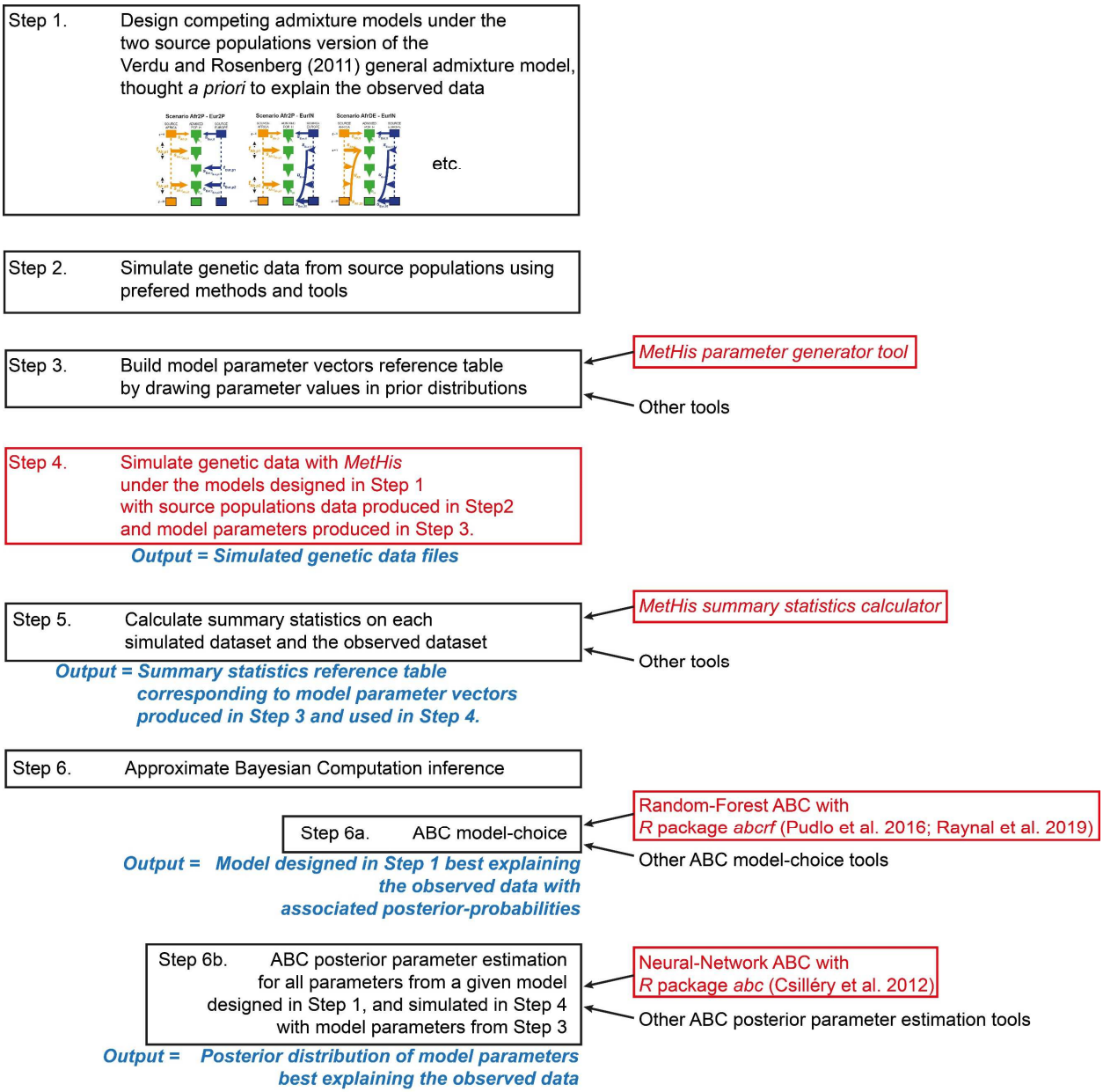
The *MetHis simulation* tool outputs individual genotype data (SNPs or microsatellite) in *vcf* format. For large amounts of simulations, as needed for instance in ABC inference, *MetHis* can be set to output only summary statistics, in which case genetic data is automatically deleted after summary statistics calculation.

The *MetHis summary statistics* tool outputs lists of vectors of summary statistics corresponding to each simulation vector of parameter (in the same order), in text format. These are summary statistics reference tables ready to be used as input for machine-learning ABC *R* packages *abcrf* (Pudlo et al., 2016; Raynal et al., 2019) and *abc* (Csilléry et al., 2012).

MOLECULAR ECOLOGY
RESOURCES

## 9 | *MetHis*-ABC framework

**Supplementary Note S1 Figure 1.3**: General pipeline for complex admixture inference using the *MetHis*-ABC framework. Steps in red are also detailed in the main text of the article.




Supplementary Note S1 Figure 1.3

## 10 | REFERENCES

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. doi:10.1038/368455a0

Buzbas, E. O., & Verdu, P. (2018). Inference on admixture fractions in a mechanistic model of recurrent admixture. *Theor Popul Biol, 122*, 149-157. doi:10.1016/j.tpb.2018.03.006

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.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

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

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet, 9*(10), e1003905. doi: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. doi:10.1093/bioinformatics/btr124

Fisher, R. A. (1922). Darwinian evolution of mutations. *Eugen Rev, 14*(1), 31-34.

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

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

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. doi:10.1093/bioinformatics/btv684

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. doi:10.1093/bioinformatics/bty867

Rosenberg, N. A., 2004 DISTRUCT: a program for the graphical display of population structure. *Molecular Ecology Notes, 4*: 137–138.

Verdu, P., Austerlitz, F., Estoup, A., Vitalis, R., Georges, M., Thery, S., . . . Heyer, E. (2009). Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol, 19*(4), 312-318. doi:10.1016/j.cub.2008.12.049

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

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.