Received: 24 April 2023 | Revised: 15 September 2023 | Accepted: 27 September 2023

DOI: 10.1111/1755-0998.13877

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Demogenomic inference from spatially and temporally heterogeneous samples

Nina Marchi[1,2] | Adamandia Kapopoulou[1,2] | Laurent Excoffier[1,2]

[1]CMPG, Institute for Ecology and Evolution, University of Berne, Berne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Correspondence
Nina Marchi and Laurent Excoffier, CMPG, Institute for Ecology and Evolution, University of Berne, Berne 3012, Switzerland.
Email: nina.marchi@unibe.ch and laurent.excoffier@unibe.ch

Funding information
Swiss National Science Foundation, Grant/Award Number: 310030_188883

Handling Editor: Paul A. Hohenlohe

Abstract

Modern and ancient genomes are not necessarily drawn from homogeneous populations, as they may have been collected from different places and at different times. This heterogeneous sampling can be an issue for demographic inferences and results in biased demographic parameters and incorrect model choice if not properly considered. When explicitly accounted for, it can result in very complex models and high data dimensionality that are difficult to analyse. In this paper, we formally study the impact of such spatial and temporal sampling heterogeneity on demographic inference, and we introduce a way to circumvent this problem. To deal with structured samples without increasing the dimensionality of the site frequency spectrum (SFS), we introduce a new structured approach to the existing program *fastsimcoal2*. We assess the efficiency and relevance of this methodological update with simulated and modern human genomic data. We particularly focus on spatial and temporal heterogeneities to evidence the interest of this new SFS-based approach, which can be especially useful when handling scattered and ancient DNA samples, as in conservation genetics or archaeogenetics.

KEYWORDS
archaeogenetics, conservation genetics, demogenomics, demographic inference, population genetics – empirical, site frequency spectrum

1 | INTRODUCTION

Most population genetics analyses assume that sampled individuals come from discrete and homogeneous populations without any subdivision or genetic structure (Loog, 2021). However, this is rarely the case with genomic data, as individuals are often sampled from populations that were structured in the past (Mazet et al., 2015) or from a relatively broad geographic region including differentiated populations (Chikhi et al., 2010; Peter et al., 2010), like for human samples where several small population samples were collected over a wide area (Mallick et al., 2016). On the other hand, when analysing ancient genomes, due to the scarcity of ancient remains from a particular location, one also often faces a potential temporal heterogeneity as fossils from a given archaeological site rarely have the same age, that is, are often separated by hundreds or thousands of years. The analysis of such heterogeneous samples is therefore equivalent to the analysis of samples drawn from a structured population.

Whereas such temporal and spatial heterogeneities are not impacting inferences of the genetic affinities between individuals such as principal component analysis, multidimensional scaling approaches or admixture analyses (Alexander et al., 2009; Patterson

et al., 2006), they can be more of an issue when studying populations' past history (Harding & McVean, 2004). Indeed, demographic inferences based on genomic data might be especially sensitive to such heterogeneities, which can lead to an underestimation of recent coalescent rates, an overestimation of population size or a signal of recent population growth if not properly considered (Heller et al., 2013; Orozco-terWengel, 2016; Peter et al., 2010; Rodríguez et al., 2021).

In the context of site frequency spectrum (SFS)-based demographic inferences (Excoffier et al., 2013; Gutenkunst et al., 2009; Kamm et al., 2020), one way to deal with this issue would be to consider each individual as belonging to a separate population and to estimate the past demography of each population separately. However, this approach would not only require the estimation of an overly large number of parameters (Bhaskar et al., 2015), but it would also lead to a very high-dimensional SFS, where the number of entries could be of the order of, or even exceed, the number of available SNPs (Bhaskar & Song, 2014; Terhorst & Song, 2015). For instance, the unfolded SFS of 15 samples of one diploid individual each would have 14,348,907 entries, since the derived allele can have frequencies of 0, 1 or 2 in each sample, such that there are 3^{15} possible combinations of allele frequencies among the 15 samples. In such a case, the observed SFS under a given scenario could greatly depart from its expectation under the correct model by chance, making it impossible to recover the true demography of the populations (Lapierre et al., 2017; Rosen et al., 2018; Terhorst & Song, 2015). It thus seems more reasonable to work on a SFS of lower dimension and on a simpler model with fewer parameters, which could be achieved by pooling individuals from different locations or different ages in the same population sample. However, this pooling should lead to a Wahlund effect, that is, an apparent excess of homozygotes (De Meeûs, 2018), which has been previously dealt with in SFS inference by allowing for some inbreeding in the pooled samples (Marchi et al., 2022). This approach introduces a specific inbreeding coefficient (F_{IS}) for each pooled sample, which corresponds to the probability that the two homologous alleles of an individual have a very recent common ancestor. This simple solution does not seem optimal if the pooled sample contains several individuals from the same panmictic population mixed with individuals from other panmictic units, as the various levels of relatedness among individuals might not be properly considered. Furthermore, this procedure requires assigning some unique parameters for the population despite its heterogeneity, like an average sampling age or population size, which can affect the estimation of other parameters (e.g. divergence times).

Here, we propose an alternative solution to this problem, which consists in pooling individuals into samples while explicitly considering their geographic or temporal structure, that is, by modelling their genetic structure and taking into account the exact sampling age of each sample. This explicit structure modelling seems relevant when one is interested in the demography of ancestral populations or of metapopulations in which the sampled are embedded (e.g. the divergence between cultural or continental groups).

In order to evaluate the relevance and efficiency of this latter approach, implemented into a new version of the *fastsimcoal2* program (Excoffier et al., 2021; Figure S1), we have simulated a series of population samples presenting various extent of spatial and/or temporal heterogeneity (Figure 1) and we tested the ability of different modelling strategies (i.e. an absence of structure, an implicit structure or an explicit structure; Figure 2) to best recover the parameters of the simulated models. Note that the first two recovery strategies (absence of structure and implicit structure) were already implemented in *fastsimcoal2*, whereas the explicit structure approach associated with a pooling of the SFS of specified populations is new to this paper.

2 | MATERIALS AND METHODS

2.1 | *fastsimcoal2* input file extension

The pooling of samples is made possible for *fastsimcoal2* users by an extension of the input files (*.par* or *.tpl* files; see Figure S1) where a new keyword '*sfspool*' allows one to assign samples to specific SFS pools. Assigning the same pool number to different samples indicates that the SFS should be computed by estimating allele frequencies on all members of the same pool. Note that one can still use old *fastsimcoal2* files that do not mention any *sfspool*, as in that case each sample is assigned automatically to a different SFS pool, and the new strategy is thus transparent to the user.

2.2 | Simulated models

We assessed the effect of the new strategy under four simple evolutionary scenarios showing various types and various amounts of heterogeneities between samples: a spatial structure (Figure 1a,b) and a temporal structure (Figure 1c,d). In these scenarios, we modelled the divergence of two continents of 1000 diploid individuals 1000 generations ago from an ancestral population of 10,000 diploid individuals. These continents exchanged 200 generations ago a single pulse of migrants at rate *admPROP* from Continent 2 to Continent 1 (looking forward in time). The simulated models differ by the number of populations sampled from each metapopulation (2 or 5, each population including 50 diploid individuals), the divergence time (*TDivPop*) of these populations, the admixture rate (*admPROP*) and the age (*Age*) of the samples, as reported in Table 1.

For each model, we performed 10 independent simulations (replicates): each time 500,000 unlinked DNA segments of 100 bp (i.e. 50 Mb) were simulated under an infinite site mutation model with a mutation rate of 1.25e−8 per bp per generation (Tian et al., 2022). For each replicate, we sampled 1, 2 or 5 diploid

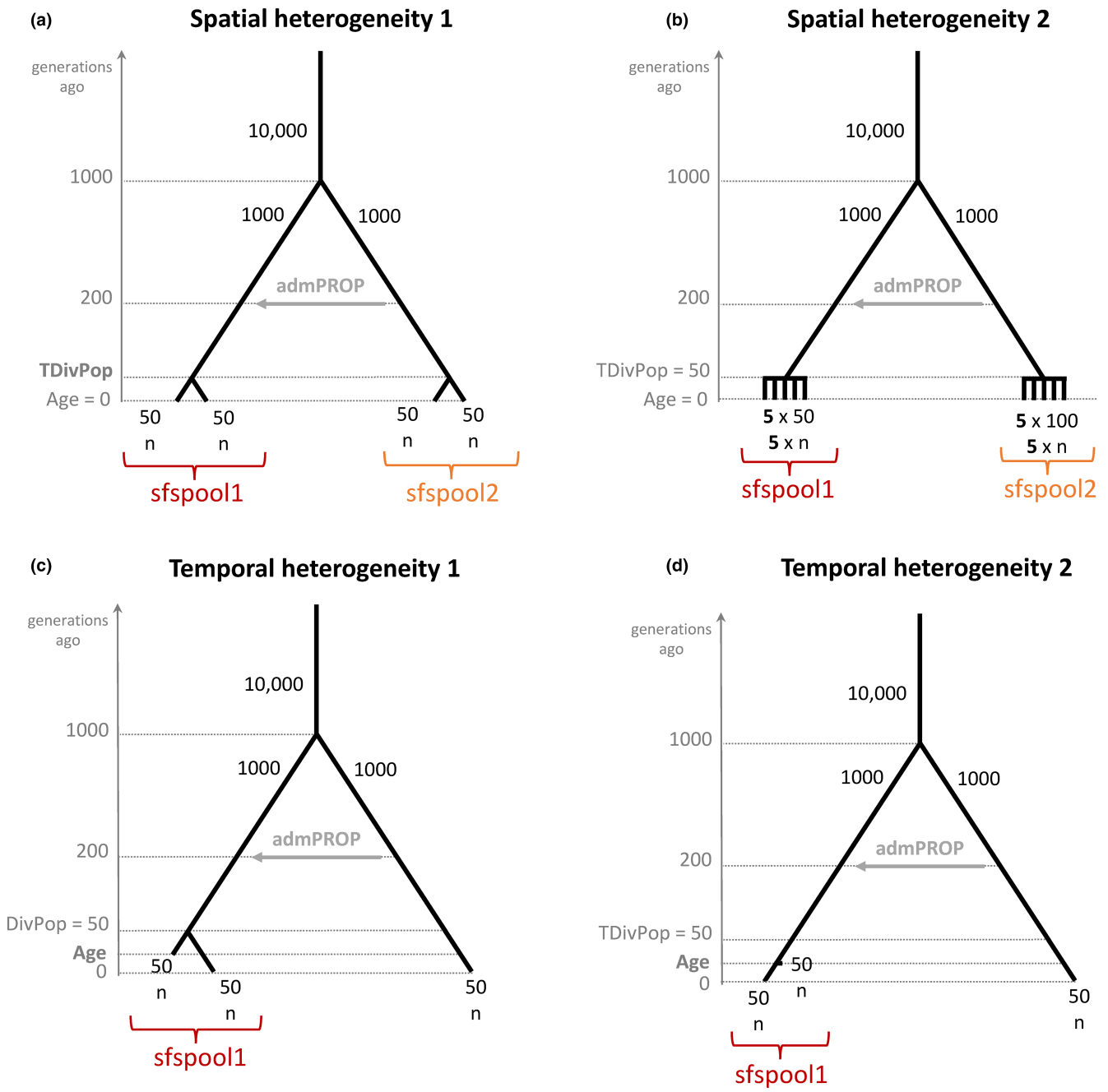

FIGURE 1 Simulated model implementing a spatial heterogeneity between a pair of populations (a) or among five populations (b), or a temporal heterogeneity between individuals sampled from separate populations (c) or from the same population at different times (d), the latter case corresponding to a population continuity scenario. Numbers along the branches of the population tree indicate the diploid population sizes, *n* to the diploid sample size, *TDivPop* refers to the divergence time of the sampled populations from the continent, *Age* to the sampling time of the individuals and *admPROP* to the admixture rate from the second to the first continent 200 generations ago. See Table 1 for simulation conditions.

individuals per population, from which we computed an unfolded SFS (the SFS recording the derived allele frequencies in each population). This simulated SFS was then considered as the 'observed' or 'true' data, and it was used in the next step to estimate model parameters and compute likelihoods. All input files and *fastsimcoal2* command lines used for the simulations and computations are available in our GitHub repository (https://github.com/CMPG/sfspool).

2.3 | Recovering simulated models

2.3.1 | Estimation strategies

To evaluate whether the pooling approach was efficient, we estimated our ability to recover some key parameter estimates (ages of the events, ancestral population and continent sizes, admixture rate) of the 'true' (simulated) models with different estimation strategies (Figure 2):

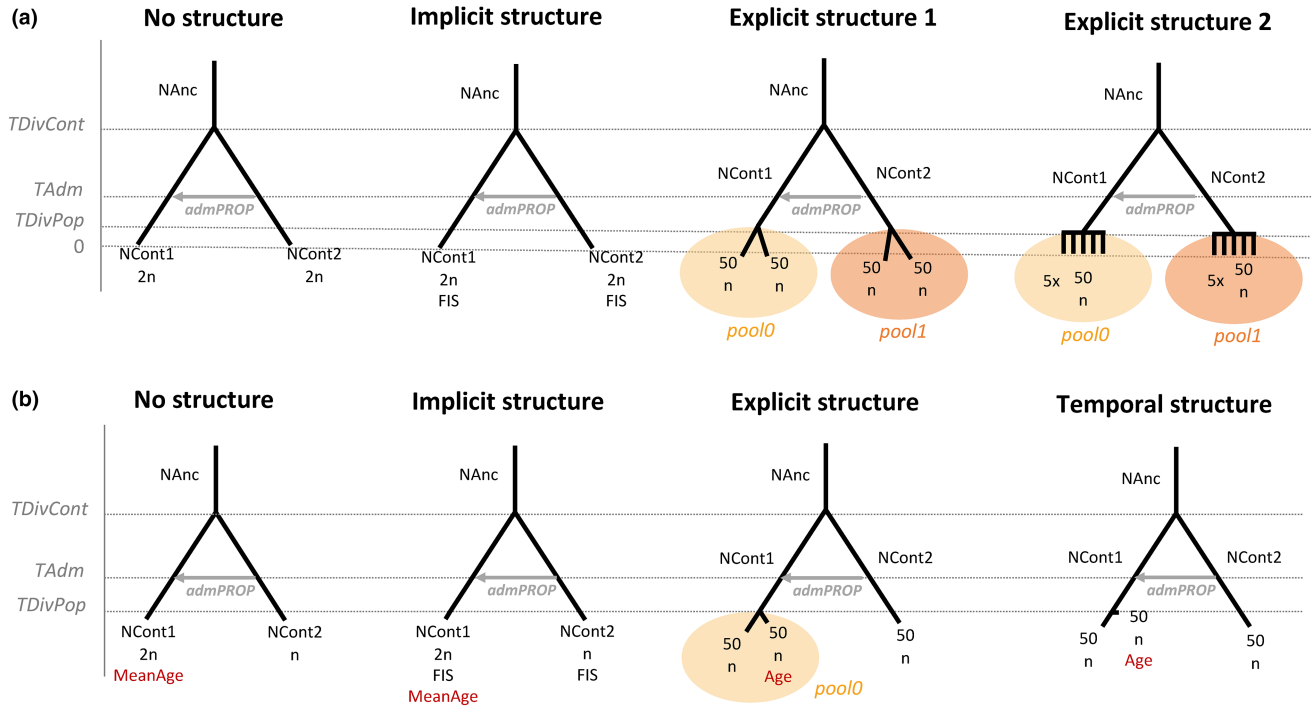

FIGURE 2 Schematic representation of the various models ('estimation strategies') used to recover the parameters of the simulated models shown in Figure 1 for a spatial (a) or a temporal (b) heterogeneity. *n* is the diploid sample size (except for Explicit Structure 2, we sampled a total of 2*n* diploid individuals in Continent 1, which could then be allocated to different sampled populations in the explicit structure strategies), and *NAnc* refers to the ancestral population size and *NCont1* and *NCont2* to the size of the continents; *TDivCont* is the time of divergence between the continents, and *TDivPop* is the divergence time of populations within continents; the ages of the samples are written in red. In some scenarios, we also modelled a single pulse of admixture at rate *admPROP* some *TAdm* generations ago.

No structure: We considered that the sampled individuals come from panmictic populations called Continent 1 and 2; that is, there is no hidden substructure (Figure 2a). For individuals sampled at different times (Figure 2b), we used the mean age of the individuals as the sampling time.

Implicit structure: We used a population 'inbreeding coefficient' (F_{IS} in Figure 2) to account for a potential hidden substructure within the continents (Wahlund effect), a strategy that was used previously to account for potential deviation from Hardy–Weinberg due to population structure (de Manuel et al., 2016; Marchi et al., 2022).

Explicit structure: We considered here that the sampled individuals belonged to different populations that split *TDivPop* generations ago from the panmictic continents (2 or 5 populations per continent in the different Explicit Structure scenarios). However, we computed the SFS at the level of the continents as in the other strategies; that is, the different populations within the same continent are considered altogether to calculate the allelic frequencies of the SFS entries. Thus, with this pooling strategy, we explicitly allowed for some sample heterogeneity and modelled the potentially different ages of each sample in case of temporal heterogeneity.

Temporal sampling: We considered a model where the individuals sampled at different epochs originate from the same population, which is like assuming a genetic continuity of the populations from which the samples were drawn.

Note that for each strategy, we tested scenarios without and with admixture from Continent 2 to Continent 1 (*TADM* generations ago, at rate *AdmPROP*), even for conditions where no admixture was simulated in the true model.

2.3.2 | Parameter estimation

The (multinomial) likelihood is estimated by approximating the estimated SFS via coalescent simulations (Excoffier et al., 2021) performed under a given set of parameter values. An expectation conditional maximization (ECM) algorithm (Meng & Rubin, 1993) is then used to iteratively find parameters maximizing the likelihood. In this approach, each parameter of the model is maximized in turn, keeping the other parameters at their last estimated value, and this procedure is repeated for a predefined number of cycles. See our GitHub repository (https://github.com/CMPG/sfspool) for the input files and command lines that were used.

2.3.3 | Model comparison

We evaluate the performance of the different strategies to recover the true model by comparing their estimated parameter values and their likelihood. We computed the difference (Delta Lhood)

between the likelihood of the strategy estimated by *fastsimcoal2* from simulated data and the maximal possible value of the likelihood computed by equating estimated SFS entries to observed SFS entries. Furthermore, as the simulated data consist mostly of independent SNPs, our likelihoods are not composite likelihoods and we can use Akaike criterion (AIC; Akaike, 1974) to compare models with different numbers of parameters. We first computed AIC as $AIC_i = 2k_i - 2\ln(10)\log_{10}\left(\widehat{L}_i\right)$, where k_i is the number of degrees of freedom of the *i*-th model and $\log_{10}\left(\widehat{L}_i\right)$ is the $\log_{10}$ likelihood reported by *fastsimcoal2*. The relative likelihood of different models was then estimated as $\exp\left(\left(AIC_{min} - AIC_i\right)/2\right)$, where AIC_{min} is the minimum AIC value obtained for the best model for each replicate. The best model has thus a relative likelihood of 1, and other models have lower relative likelihoods. We consider that other models are significantly less good if their relative likelihood is smaller than 5%.

2.4 | Application to human data

2.4.1 | Genomic data

To illustrate the validity of this approach for demographic inferences, we have applied our new pooling strategy to genomic data from modern samples. We investigated the relationships between populations from Africa, Europe and Asia. To do so, we selected from the SGDP panel (Mallick et al., 2016) 24 individuals from 12 populations: the Esan, Mandenka, Mende and Yoruba populations from Western Africa; the Bulgarian, Sardinian, Spanish and Tuscan populations from Europe; the Cambodian, Han, Kinh and Thai populations from Southeast Asia. We filtered the individual vcf files based on their sequencing depth (DP ≥ 8 and DP < min(200, Q99%), where Q99% is the 99th percentile of the depth distribution). Furthermore, we filtered out low-quality variants (QUAL < 20). Following the filtering step, we merged all individual files in a single vcf file and we polarized it to the EPO ancestral allele (human_ancestor_GRCh37_e59 from ensembl_compara_59@ens-livemirror:3306). From these genomic data and using the same approach as in Marchi et al. (2022), we obtained a SFS on the neutral portion of the genome (with local recombination rate >1 cM/Mb and no mutations potentially affected by biased gene conversion (BGS), see (Pouyet et al., 2018)) including 141,504,530 sites for which we estimated a neutral mutation rate of 6.13e-09 per bp per generation using a procedure described in Marchi et al. (2022). We performed a multidimensional scaling (MDS) analysis from a matrix of pairwise nucleotide divergences π_{XY} (Nei & Li, 1979) computed between all pairs of genomes, only considering the sites used for the computation of the observed SFS, using the R *cmdscale* function from the *stats* package (Figure S2).

2.4.2 | Demogenomic modelling

We then built demographic scenarios inspired from previous work (Malaspinas et al., 2016; Massilani et al., 2020) enabling us to estimate the divergence times between and within continents, the size of the different ancestral and sampled populations and other parameters, as described in Figure S3. We tested two scenarios: (i) an explicit structure within continent that is based on the new pooling approach, where the sampled populations are assumed to have simultaneously diverged from each other some time ago (a parameter to be estimated); (ii) an implicit structure considering that all samples from a given continent belong to a unique population, where the hidden underlying genetic structure potentially leading to a Wahlund effect is dealt with by modelling an inbreeding F_{IS} coefficient. Furthermore, we tested these two scenarios in the presence or absence of an admixture from a ghost population into the Western African metapopulation.

For each scenario, we performed 200,000 coalescent simulations per likelihood estimation and 30 expectation conditional maximization (ECM) cycles to find parameters maximizing the likelihood. This procedure was repeated from 100 different initial conditions, and the parameters with the overall maximum likelihood were kept.

Note that as the polymorphic sites of the real human data are not all independent (contrary to those generated by simulations), the computed likelihoods are here composite likelihoods. We nevertheless computed the model AIC, but we are aware that the inferred relative likelihood of the least fit model might be underestimated.

TABLE 1 Parameter combinations of the simulated models.

Simulated models		*TDivPop*	*Age*	*admPROP*
Spatial heterogeneity (two populations per continent)	No heterogeneity	0	0	0
	Weak heterogeneity	10	0	0
	Medium heterogeneity	50	0	0
	Strong heterogeneity	100	0	0
Spatial heterogeneity with admixture (two or five populations per continent)	No admixture	50	0	0
	Small admixture	50	0	0.15
	Large admixture	50	0	0.30
Temporal Heterogeneity (one or two populations per continent)	Recent heterogeneoty; strong admixture	50	10	0.30

2.4.3 | Parametric bootstraps

Confidence intervals for the parameters estimated under the two best models were obtained using a parametric bootstrap approach described previously (Excoffier et al., 2013) and summarized hereafter. In brief, we first generated 100 SFS computed from the simulation of 100 chromosomes of 141 Mb (corresponding approximately to the neutral part of the genome that was used for the original parameter estimation) using the maximum-likelihood estimated parameter values. For each of the 100 SFS, we re-estimated the parameters of the model for 20 independent runs. The parameters of the run having overall largest likelihood were kept for each of the 100 simulated data sets. These 100 sets of parameters were then used to compute the 2.5% and 97.5% quantiles of the distribution of each parameter, approximately delimiting a 95% CI around the initially estimated ML values.

3 | RESULTS

3.1 | Spatial heterogeneity without admixture

We assessed the performance of various estimation strategies by computing the difference (Delta Lhood) between the estimated likelihood and the maximum likelihood given the observed SFS. In addition, we also computed the relative likelihoods of the different models from their AIC (see Section 2.3.3). When there is medium to strong spatial heterogeneity between the populations (Figure 1a, Table 1), the estimations based on an explicit structure outperformed the other estimation strategies (Figure 3). Globally, the estimation strategy that does not assume any structure at all is not performing well in that case, for any sample size (Figure 3), as it has consistently very low relative likelihood (Figure S4). When a single diploid individual is sampled per population, the implicit structure strategy (estimating F_{IS} within pooled populations) leads to Delta (Figure 3) and relative likelihoods (Figure S4) that are very close or even better than those obtained under the explicit structure strategy. The recovered parameters are also well estimated, except perhaps the continent sizes that are overestimated (Figure S5A). However, when more than one diploid individual is sampled from each population, the explicit structure strategy leads to Delta and relative likelihoods that are much better than the two strategies not explicitly considering the genetic structure of the samples (Figure 3, Figure S4). Regarding the ability of the different strategies to estimate parameter values, we find that the explicit strategy allows one to correctly recover true parameter values, while the implicit strategy leads to a clear overestimation of the ancestral population size and an underestimation of the continent sizes and divergence time (Figure S5A).

Differences in performance between estimation strategies thus depend on the actual number of individuals sampled per population, but even more on the level of genetic heterogeneity among population samples. The advantage of the explicit strategy is indeed more visible when a strong heterogeneity is simulated (*TDivPop* = 100, i.e. $F_{ST} \simeq 0.63$) than for a medium heterogeneity (*TdivPop* = 50, $F_{ST} \simeq 0.4$; Figure 3), or for a weak spatial heterogeneity (*TdivPop* = 10, $F_{ST} \simeq 0.1$), but it is present in all cases for $n \geq 2$ (Figure S4: the explicit structure strategy shows the best relative likelihood for all 10 replicates when there is a strong and medium heterogeneity, and for the majority of the replicates when there is a weak heterogeneity). Importantly, if the true model does not include any spatial heterogeneity (*TDivPop* = 0), the use of an explicit structure as an estimation strategy is not penalizing as the likelihood of this type of model is similar to that obtained under the true model without any structure (Figure 3, bottom row). This strategy also leads in that case to relative likelihoods that are often very close or equal to those of the best model (Figure S4). In other words, the estimation of additional parameters under this more complex model does not prevent us to correctly estimate parameters as our more complex model can still recognize that there is no genetic structure among the samples. Finally, note that the relative performances of the three estimation strategies are similar to those described above when more than two populations are simulated per continent (Figure 2a, Figure S6).

3.2 | Spatial heterogeneity with admixture

In the simulations and analyses presented above, the two continents were isolated since their divergence. We relaxed this assumption by examining cases where there is a single pulse of gene flow (an admixture event) from Continent 2 to Continent 1, that is, by adding two additional parameters in the estimations: the admixture time and the admixture proportion. We find that the estimation of these two additional parameters is not a burden when there is no admixture, as the likelihood of the explicit structure strategy with admixture is very similar to that without admixture (Figure S7A–D). However, when we analysed data sets with simulated admixture (Figure 2a, Table 1), the estimation strategies with admixture clearly outperform strategies without admixture (Figure S7B,C) with best relative likelihoods exclusively found for the strategy with admixture (Figure S7E,F).

3.3 | Temporal heterogeneity

Like in the case of a simple spatial sampling heterogeneity, when there is both a spatial and a temporal sampling heterogeneity (Figure 2b, Table 1), we find that a model with an explicit structure is working well for any number of sampled individuals (Figure 4, Figure S8—top row). In this case as well, when only one diploid individual is sampled from each population, the explicit and implicit structure strategies are equally good (no significant differences between their relative likelihood in Figure S8). However, when the sampled individuals come from the same population without spatial structure (Figure 2d), the temporal strategy has a significantly higher relative likelihood than the other strategies (for 9 of the 10 replicates, Figure S8—bottom row), showing that the best recovery strategy is indeed in both cases the one exactly matching the simulation framework. It implies that

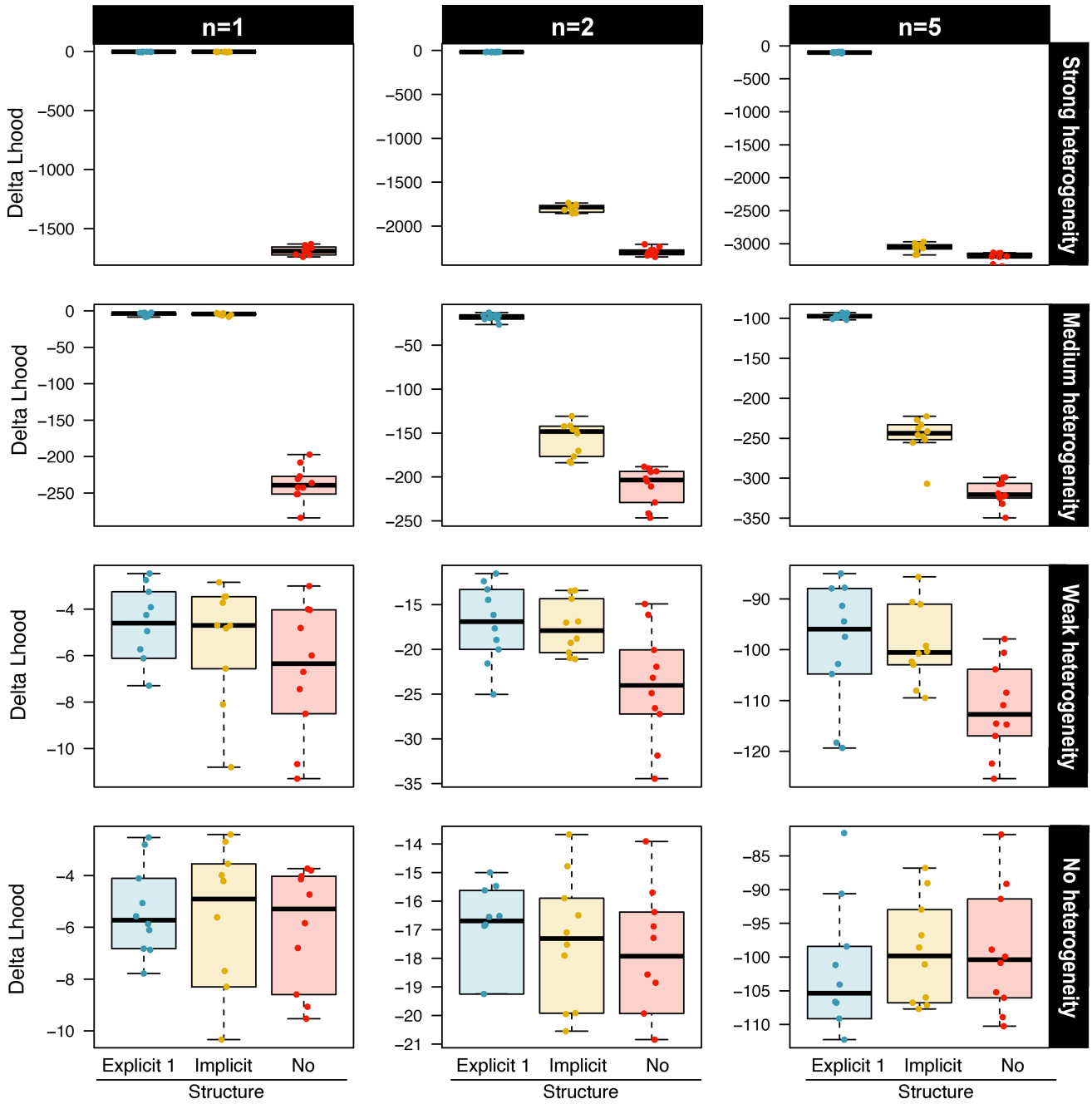

FIGURE 3 Model likelihoods of three estimation strategies for different sample sizes (*n*) and different levels of spatial heterogeneity among samples. Note that no admixture between continents was allowed for these strategies. Each estimation strategy was attempted on 10 data sets. Simulation conditions are defined in Table 1.

it should be possible to distinguish models where individuals sampled at different times from a given archaeological site come from the same population (genetic continuity) or if they come from distinct populations suggesting a population replacement.

3.4 | Analysis of modern human data

In order to illustrate the interest of the new strategy for demogenomic inferences in the presence of spatial heterogeneity, we analysed a small subset of SGDP human individuals (Mallick et al., 2016) sampled in three continents (Figures S2 and S3) with some degree of spatial heterogeneity. Using our new approach, we analysed the eight individuals sampled from four populations in each continent as a pooled sample. With this explicit structure modelling, we obtained a much better likelihood (71 $\log_{10}$ likelihood units improvement) than for a scenario where we used an implicit structure (Wahlund effect estimated by an F_{IS} coefficient), and the relative likelihood is vastly inferior (10^{-70}) for the implicit model (Table S1), suggesting that the explicit model is very

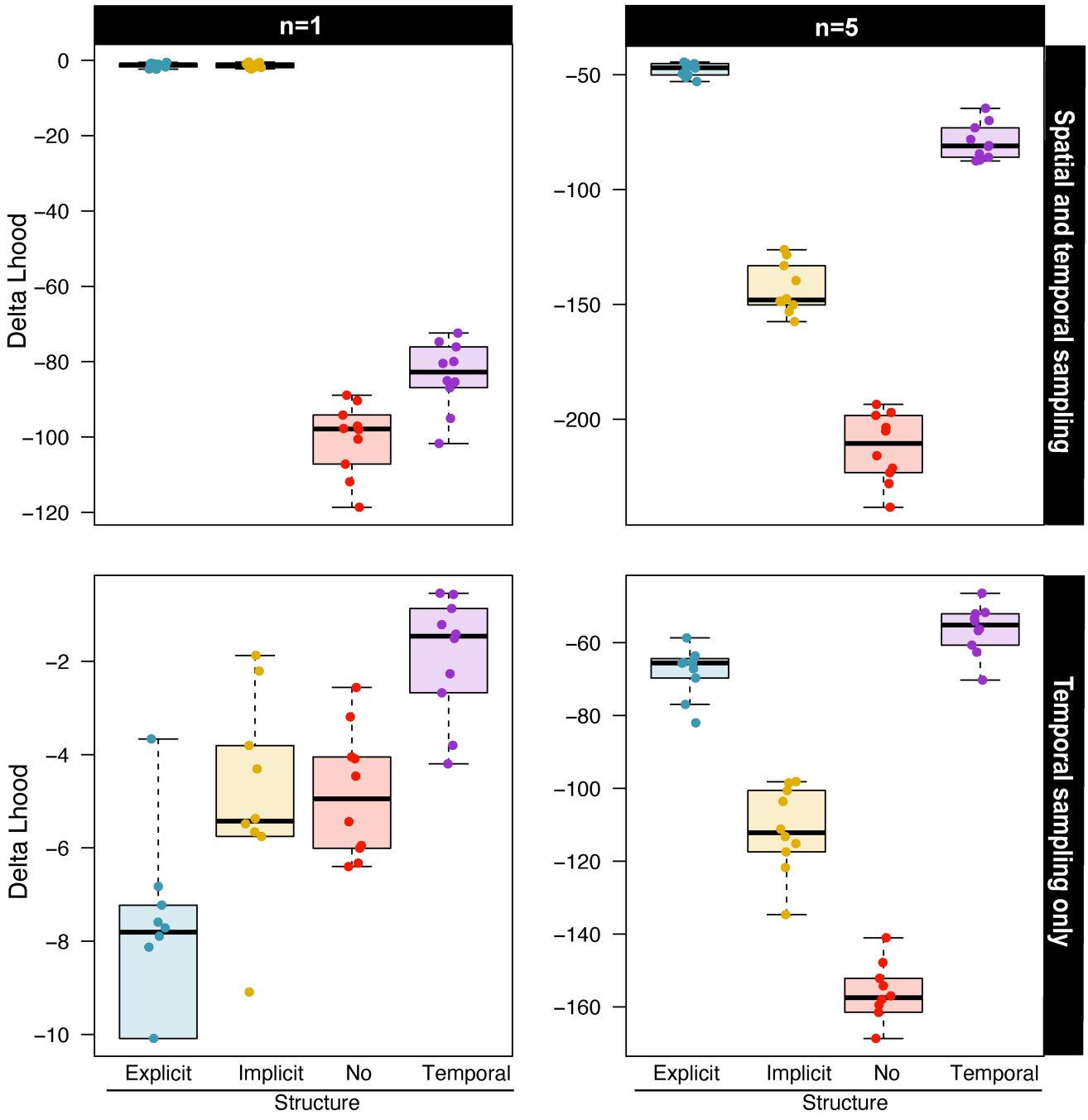

FIGURE 4 Performance of the estimation strategies at fitting the data simulated under some temporal and spatial heterogeneity due to sampling from different populations (top row) or under temporal sampling from the same population (bottom row). Different sample sizes ($n=1$ and 5) have been tested. Each estimation strategy was run on 10 data sets. For each simulated condition, we attempted to fit the data under the four models defined in Figure 2b. Simulation conditions are defined in Table 1.

significantly better supported, even though this relative likelihood might be underestimated as the SNPs from our human data set are not fully independent.

As expected, continental population sizes differ between the explicit and implicit structure modelling, but a few other parameters are substantially different between the two approaches, like those related to the bottleneck out of Africa and later drift in the ancestral Eurasian population, as well as the amount of migration between Asia and Europe. Interestingly, but unrelated to the difference between the explicit and implicit structure model, we find evidence for a relatively strong input (8–9%) from a very differentiated ghost population into the Western African metapopulation. Indeed, this ghost population would have diverged from the modern humans 640–700 kya, at about the same time than the population ancestral to both Neandertals and Denisovans. Without this admixture, we indeed found a relatively large excess of almost fixed derived mutations in Africans compared with our predictions (Figure S9), a signal that we had shown to be

potentially due to unaccounted gene flow from a distinct population (Marchi & Excoffier, 2020), which motivated us to model this ghost admixture in Africans.

4 | DISCUSSION

Sequencing costs have been dramatically reduced, but full genomes are rarely sequenced on many individuals from the same population and most studies will sequence one or a few individuals from several populations. Whereas powerful methods have been developed to infer the past demography from single or a very few individuals from the same population (Li & Durbin, 2011; Schiffels & Durbin, 2014; Sheehan et al., 2013), it remains difficult to infer the demography of several populations from a set of individuals sampled at different geographical locations or at different times. To alleviate this problem, we have introduced here a new way to deal with potential spatial and temporal heterogeneities when inferring the past demography from the site frequency spectrum computed on a set of genome samples. Our approach allows for a detailed modelling of sample locations and times while keeping the dimensionality of the SFS low. Our simulations suggest that it is a more appropriate way to account for heterogeneity than to simulate a simple Wahlund effect when we sample more than one individual per subdivision, and that allowing for the inference of population subdivision when there is none has no negative impact on our fitting of the data (Figure 3). Note, however, that the implicit structure (F_{IS}) approach might be simpler and appropriate when dealing with a set of ancient DNA samples, which are often widely temporally and spatially spaced. While we have implemented a simple scenario of populations splitting from a common source to model population subdivision in our simulations and estimations (Figures 1 and 2), alternative and potentially more complex models can easily be implemented, like in our application to human data where we simulated continent-island models (Figure S3).

The exact way to model sample heterogeneity will indeed depend on the model system that is studied, and it may be wise to test alternative models of population structure and designate the best fitting model based on the AIC (e.g. in Figures S4 and S8). For a nonmodel organism, where one has no predefined idea of the history of populations and how to group individuals, geography is often providing a natural first hint. However, simple descriptive statistics computed from the raw genetic data (e.g. PCA, MDS and F-statistics; Marchi et al., 2021) should be helpful to complement this information and more firmly decide how individuals might be grouped for further analyses. Then, it is clear that more realistic models will require more parameters to be estimated, which might make them more difficult to explore (Bhaskar & Song, 2014; Terhorst & Song, 2015), such that a compromise needs to be found between realism and tractability. However, one might not be necessarily interested in the exact way populations within groups are related, but by other parameters related to ancestral populations such as divergence times between groups, and in this case a more exact, if not perfect, modelling of sample heterogeneity might be helpful. In this respect, our application to human data is interesting, as some ancestral parameters are indeed markedly different between an implicit and an explicit modelling of heterogeneity. For instance, we find 3–6 times more gene flow between Europe and Asia (*Nm_EUtoAS*) and a smaller ancestral size of the Eurasian ancestral population (*NEura*, ~1900 vs. ~5560 individuals). Also, the bottleneck out of Africa (*iBotOOA*) is found 3.5 times less strong and have occurred slightly earlier (*TOOA*, 67.8 kya vs. 56.8 kya), and the African metapopulation size (*NaAF*) before the last glacial maximum was found larger (23,418 diploid individuals) than just after the exit out of Africa (17,783), whereas it was found smaller with the implicit model, which would lead us to conclude that the African ancestral population would have decreased rather than increased since the exit out of Africa. Finally, the ancestral European size (*NaEU*) is found to be smaller with an explicit structure modelling (8160 diploids) than without (11,901 diploids). We note, however, that most parameters have relatively wide confidence intervals and that these differences might not always be significant (Table S1). Because the explicit heterogeneity approach allows for a better modelling of intracontinental diversity, we believe that it naturally leads to a better estimation of the older part of the population history, as these parameters do not need to be tweaked to compensate the misspecification of the recent intracontinental history. Of course, our current modelling has room for improvement as we can see that for instance Asian or African diversity is not partitioned into exactly four equidistant population groups (Figure S2) as is done in our modelling.

Lastly, the detection that 8–9% of the genome of Western Africans could come from an archaic human population having diverged ~639 kya from the human lineage is in line with previous studies reporting evidence of specific archaic admixture in sub-Saharan Africans (Chen et al., 2020; Wall et al., 2019) and in very good agreement with a previous study of the conditional frequency spectrum of Western African populations (Durvasula & Sankararaman, 2020), which suggested that these populations had received 2–19% of their genome from a population having diverged about 625 kya from modern humans. Note that we did not attempt at estimating the time of admixture, since we imposed it to be occurring at the time of the diversification of West African population, here estimated at ~33 kya. Some more specific modelling of Western African populations would be necessary to refine the exact scenario of this African-specific archaic admixture, but it suggests that this African archaic population had diverged from the human lineage approximately at the same time as the ancestors of Neandertals and Denisovans.

AUTHOR CONTRIBUTIONS

L.E. and N.M. designed the simulation framework. L.E. programmed the new version of *fastsimcoal2* and simulated test data sets. N.M. analysed the simulated data and performed the parameter estimation and likelihood analyses. A.K. curated the human genomic data set and performed their bioinformatic analysis. L.E.

and N.M. analysed the results. All authors contributed to writing the manuscript.

ACKNOWLEDGEMENTS

This work was supported by a Swiss NSF grant 310030_188883 to LE. We are grateful to Sandra Da Silva Oliveira for helpful discussions on the subject and to other CMPG lab members for their comments when presenting this work at lab meetings.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY AND BENEFIT-SHARING STATEMENT

Input files for the simulation–estimation framework as well as the data used for the applications are available on a dedicated GitHub: https://github.com/CMPG/sfspool.

ORCID

Laurent Excoffier https://orcid.org/0000-0002-7507-6494

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Bhaskar, A., & Song, Y. S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics*, *42*(6), 2469–2493.

Bhaskar, A., Wang, Y. X. R., & Song, Y. S. (2015). Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, *25*(2), 268–279.

Chen, L., Wolf, A. B., Fu, W., Li, L., & Akey, J. M. (2020). Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell*, *180*(4), 677–687.

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B., & Beaumont, M. A. (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, *186*(3), 983–995.

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481.

De Meeûs, T. (2018). Revisiting FIS, FST, Wahlund effects, and null alleles. *The Journal of Heredity*, *109*(4), 446–456.

Durvasula, A., & Sankararaman, S. (2020). Recovering signals of ghost archaic introgression in African populations. *Science Advances*, *6*(7), eaax5097.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905.

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., & Sousa, V. C. (2021). fastsimcoal2: Demographic inference under complex evolutionary scenarios. *Bioinformatics*, *37*, 4882–4885. https://doi.org/10.1093/bioinformatics/btab468

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), e1000695.

Haber, M., Mezzavilla, M., Bergström, A., Prado-Martinez, J., Hallast, P., Saif-Ali, R., Al-Habori, M., Dedoussis, G., Zeggini, E., Blue-Smith, J., Wells, R. S., Xue, Y., Zalloua, P. A., & Tyler-Smith, C. (2016). Chad genetic diversity reveals an African history marked by multiple Holocene Eurasian migrations. *The American Journal of Human Genetics*, *99*(6), 1316–1324.

Harding, R. M., & McVean, G. (2004). A structured ancestral population for the evolution of modern humans. *Current Opinion in Genetics & Development*, *14*(6), 667–674.

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One*, *8*(5), e62992. https://doi.org/10.1371/journal.pone.0062992

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *115*(531), 1472–1487.

Lapierre, M., Lambert, A., & Achaz, G. (2017). Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics*, *206*(1), 139–449.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496.

Loog, L. (2021). Sometimes hidden but always there: The assumptions underlying genetic inference of demographic histories. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *376*(1816), 20190719.

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., Bergström, A., Athanasiadis, G., Cheng, J. Y., Crawford, J. E., Heupink, T. H., Macholdt, E., Peischl, S., Rasmussen, S., Schiffels, S., Subramanian, S., Wright, J. L., Albrechtsen, A., Barbieri, C., … Willerslev, E. (2016). A genomic history of aboriginal Australia. *Nature*, *538*(7624), 207–214.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), 201–206.

Marchi, N., & Excoffier, L. (2020). Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evolutionary Applications*, *20*, 1–10.

Marchi, N., Schlichta, F., & Excoffier, L. (2021). Demographic inference. *Current Biology*, *31*(6), R276–R279.

Marchi, N., Winkelbach, L., Schulz, I., Brami, M., Hofmanová, Z., Blöcher, J., Reyna-Blanco, C. S., Diekmann, Y., Thiéry, A., Kapopoulou, A., Link, V., Piuz, V., Kreutzer, S., Figarska, S. M., Ganiatsou, E., Pukaj, A., Struck, T. J., Gutenkunst, R. N., Karul, N., … Excoffier, L. (2022). The genomic origins of the world's first farmers. *Cell*, *185*(11), 1842–1859.

Massilani, D., Skov, L., Hajdinjak, M., Gunchinsuren, B., Tseveendorj, D., Yi, S., Lee, J., Nagel, S., Nickel, B., Devièse, T., Higham, T., Meyer, M., Kelso, J., Peter, B. M., & Pääbo, S. (2020). Denisovan ancestry and population history of early east Asians. *Science*, *370*(6516), 579–583.

Mazet, O., Rodríguez, W., & Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, *104*, 46–58.

Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the*

National Academy of Sciences of the United States of America, *76*(10), 5269–5273.

Orozco-terWengel, P. (2016). The devil is in the details: The effect of population structure on demographic inference [review of *the devil is in the details: The effect of population structure on demographic inference*]. *Heredity*, *116*(4), 349–350.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and Eigenanalysis. *PLoS Genetics*, *2*(12), 2074–2093.

Peter, B. M., Wegmann, D., & Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology*, *19*(21), 4648–4660.

Pouyet, F., Aeschbacher, S., Thiéry, A., & Excoffier, L. (2018). Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife*, *7*, e36317. https://doi.org/10.7554/eLife.36317

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2021). Correction to: The IICR and the non-stationary structured coalescent: Towards demographic inference with arbitrary changes in population structure. *Heredity*, *126*(4), 706.

Rosen, Z., Bhaskar, A., Roch, S., & Song, Y. S. (2018). Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, *210*(2), 665–682.

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925.

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, *194*(3), 647–662.

Terhorst, J., & Song, Y. S. (2015). Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, *112*(25), 7677–7682.

Tian, X., Cai, R., & Browning, S. R. (2022). Estimating the genome-wide mutation rate from thousands of unrelated individuals. *The American Journal of Human Genetics*, *109*(12), 2178–2184.

Wall, J. D., Ratan, A., Stawiski, E., & GenomeAsia 100K Consortium. (2019). Identification of African-specific admixture between modern and archaic humans. *The American Journal of Human Genetics*, *105*(6), 1254–1261.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Marchi, N., Kapopoulou, A., & Excoffier, L. (2024). Demogenomic inference from spatially and temporally heterogeneous samples. *Molecular Ecology Resources*, *24*, e13877. https://doi.org/10.1111/1755-0998.13877