Accuracy of demographic inferences from Site Frequency Spectrum: The case of the Yoruba population

Marguerite Lapierre[*,†], Amaury Lambert[†,‡], and Guillaume Achaz[*,†]

[*]Atelier de Bioinformatique, UMR 7205 ISyEB, MNHN-UPMC-CNRS-EPHE, Muséum National d'Histoire Naturelle, 75005 Paris, France

[†]SMILE (Stochastic Models for the Inference of Life Evolution), UMR 7241 CIRB, Collège de France, CNRS, INSERM, PSL Research University, 75005 Paris, France

[‡]Laboratoire de Probabilités et Modèles Aléatoires (LPMA), UMR 7599, UPMC-CNRS, 75005 Paris, France

Running title: Accuracy of demographic inferences

Key Words: human demography, model identifiability, coalescent theory, site frequency spectrum

Corresponding Author:

Marguerite Lapierre

Atelier de Bioinformatique

Muséum National d'Histoire Naturelle

Boîte Courrier 50, Bâtiment 139

45 rue Buffon

75005 PARIS

(33)(0)140 79 45 09

marguerite.lapierre@mnhn.fr

Abstract

Some methods for demographic inference based on the observed genetic diversity of current populations rely on the use of summary statistics such as the Site Frequency Spectrum (SFS). Demographic models can be either model-constrained with numerous parameters such as growth rates, timing of demographic events and migration rates, or model-flexible, with an unbounded collection of piecewise constant sizes. It is still debated whether demographic histories can be accurately inferred based on the SFS. Here we illustrate this theoretical issue on an example of demographic inference for an African population. The SFS of the Yoruba population (data from the 1000 Genomes Project) is fit to a simple model of population growth described with a single parameter (*e.g.*, founding time). We infer a time to the most recent common ancestor of 1.7 million years for this population. However, we show that the Yoruba SFS is not informative enough to discriminate between several different models of growth. We also show that for such simple demographies, the fit of one-parameter models outperforms the model-flexible method recently developed by Liu and Fu. The use of this method on simulated data suggests that it is biased by the noise intrinsically present in the data.

INTRODUCTION

Inference of human population history based on demographic models for genomic data can complement archaeological knowledge, owing to the large amount of polymorphism data now available in human populations. Polymorphism data can be viewed as an imprint left by past demographic events on the current genetic diversity of a population (see, *e.g.*, review by Pool *et al.* 2010).

There are several means of analyzing this observed genetic diversity for demographic inference. The polymorphism data can be used to reconstruct a coalescence tree of the sampled individuals. The demography of the sampled population can be inferred by comparing this reconstructed tree with theoretical predictions under a constant size model (Pybus *et al.* 2000). For example, in an expanding population, the reconstructed coalescent tree will have relatively longer terminal branches than the reference coalescent tree in a population of constant size. However, methods based on a single reconstructed tree are flawed because of recombination (Lapierre *et al.* 2016), since the genealogy of a recombining genome is described by as many trees as there are recombining loci.

The genome-wide distribution of allele frequencies is a function of the average genealogies, and can thus be used as a summary statistic for demographic inference. This distribution, called the Site Frequency Spectrum (SFS), reports the number of mutated sites at any given frequency. The demographic history of a population affects the shape of its SFS (Adams and Hudson 2004; Marth *et al.* 2004). For example, an expanding population carries an excess of low-frequency variants, compared with the expectation under a constant size model. The shape of the SFS is also altered by selection, which results in an excess of low- and high-frequency variants (Fay and Wu 2000). However, selection acts mainly on the coding parts of the genome and the non-coding segments linked to them, while demography impacts the whole genome. Furthermore, unlike reconstructed trees, the SFS is not biased by recombination (Wall 1999). Quite on the contrary, by averaging the SFS over many correlated marginal genealogies, recombination lowers the variance of the SFS while

its expectation remains unchanged. Therefore, the SFS of a sample is a summary of the genetic diversity, averaged over all the genome due to recombination, that can be analyzed in terms of demography.

Several types of methods exist to infer the demography of a population based on its SFS. A specific demographic model can be tested by computing a pseudo-likelihood function for this model, based on the comparison of the observed SFS and the SFS estimated by Monte Carlo coalescent tree simulations (Nielsen 2000; Coventry *et al.* 2010; Nelson *et al.* 2012). This method can be extended to infer demographic scenarios of several populations, using their joint SFS (Excoffier *et al.* 2013). Methods based on Monte Carlo tree simulations are typically very costly in computation time. Other approaches rely on diffusion processes: they use the solution to the partial differential equation of the density of segregating sites as a function of time (Gutenkunst *et al.* 2009; Lukić *et al.* 2011).

Whereas all these methods are model-constrained, *i.e.*, they use the SFS to test the likelihood of a given demographic model, more flexible methods have been developed. Recently, Bhaskar *et al.* (2015) derived exact expressions of the expected SFS for piecewise-constant and piecewise-exponential demographic models. Liu and Fu (2015) developed a model-flexible method based on the SFS: the stairway plot. This method infers the piecewise-constant demography which maximizes the composite likelihood of the SFS, without any previous knowledge on the demography. This optimization is based on the estimation of a time-dependent population mutation rate θ. Although they show that their method infers efficiently some theoretical demographies, they do not test the goodness of fit of the expected SFS, reconstructed under the demography they infer, with the input SFS on which they apply their method.

All these methods are widely used for the inference of demography in humans and other species, but doubts remain on the identifiability of a population demography based on its SFS. It has been shown theoretically that certain population size functions are unidentifiable from the population SFS (Myers *et al.* 2008; Terhorst and Song 2015). Myers *et al.*

(2008) showed that for any given population size function $N(t)$, there exists an infinite number of smooth functions $F(t)$ such that $\xi^N = \xi^{N+F}$ where ξ^N is the SFS of a population of size function $N(t)$. However, other theoretical works have recently shown that for many types of population size functions commonly used in demography studies, such as piecewise constant or piecewise exponential functions, demography can be inferred based on the SFS, provided the sample is large enough (Bhaskar and Song 2014). These studies argued that the unidentifiability proven by Myers *et al.* (2008) relied on biologically unrealistic population size functions involving high frequency oscillations near the present. Lately, two studies (Kim *et al.* 2015; Terhorst and Song 2015) have provided bounds on the amount of demographic information contained in the SFS or in coalescent times.

In this study, we use the SFS of an African population (the Yoruba population, data from The 1000 Genomes Project Consortium 2015) as an example of a somewhat simple demography, to illustrate the risks of over-confidence in demographic scenarios inferred. Namely, we highlight two issues potentially arising even in the case of simple demographies: unidentifiability of models and poor goodness of fit of inferences. We first infer the Yoruba demography with a model-constrained method, using diverse one-parameter models of growth, and then with a model-flexible method, the stairway plot (Liu and Fu 2015). For the model-constrained method, we test four different growth models derived from the standard neutral framework used in the vast majority of population genetics studies, also compared with a more uncommon type of model based on a branching process. Individual-based models such as the branching process are widely used in population ecology (Lambert 2010): the population is modeled as individuals which die and give birth at given rates independently. These models are not commonly used in population genetics although they provide interesting features of fluctuating population sizes for example, and benefit from a strong mathematical framework.

MATERIALS AND METHODS

1 000 Genomes Project data: Variant calls from the 1 000 Genomes Project phase 3 were downloaded from the project ftp site (The 1000 Genomes Project Consortium 2015). The sample size for the Yoruba population is $n = 108$ individuals (polymorphism data available for both genome copies of each individual, *i.e.*, $2n = 216$ sequences). We kept all single nucleotide bi-allelic variants to plot the sample SFS. The number of bi-allelic sites is $S = 20\,417\,698$. The average distance between two sites is 136 bp (median 81 bp). The number of sites for which the ancestral allele is known is $S' = 19\,441\,528$. To avoid possible bias due to sequencing errors, we ignored singletons (mutations appearing in only one chromosome of one individual in the sample) for the rest of the study. The implications of ignoring singletons are examined in the discussion.

Site Frequency Spectrum definition and graphical representation: The Site Frequency Spectrum (SFS) of a sample of n diploid individuals is described as the vector $\xi = (\xi_1, \xi_2, ..., \xi_{2n-1})$ where for $i \in [1, 2n-1]$, ξ_i is the number of dimorphic (*i.e.*, with exactly two alleles) sites with derived form at frequency $i/2n$. To avoid potential orientation errors, we assumed that the ancestral form is unknown for all sites: we worked with a folded spectrum, where we consider the frequency of the less frequent (or minor) allele. In this case, the folded SFS is described as the vector $\eta = (\eta_1, \eta_2, ..., \eta_n)$ where $\eta_i = \xi_i + \xi_{2n-i}$ for $i \in [1, n-1]$ and $\eta_n = \xi_n$. The folded SFS of the Yoruba sample is plotted in Figure S1. For a better graphical representation, all SFS were transformed as follows: we plot ϕ_i normalized by its sum, where

- for unfolded SFS, $\phi_i = i\,\xi_i$ for $i \in [1, 2n-1]$
- for folded SFS, $\phi_i = \eta_i \frac{i(2n-i)}{2n}$ for $i \in [1, n-1]$ and $\phi_n = n\,\eta_n$

The transformed SFS has a flat expectation (*i.e.*, constant over all values of i) under the standard neutral model (Nawa and Tajima 2008; Achaz 2009).

Demographic models used for the model-constrained methods: We inferred the demography of the Yoruba population using five growth models (Figure 1), compared with the predictions of the standard model with constant population size. Time is measured in coalescent units of $2N$ generations, where the scaling parameter N has the same dimension as the current population size, which we will not estimate. Time starts at 0 (present time) and increases backward in time. Four models are based on the standard Kingman coalescent (Kingman 1982), amended with demography. Three of them are described with an explicit demography: either *Linear* growth since time τ, *Exponential* growth at rate $1/\tau$ or *Sudden* growth from a single ancestor to the entire population at time τ. We also use another model based on the Kingman coalescent, with an implicit demography: the *Conditioned* model. This model is based on a standard constant size model, but the Time to the Most Recent Common Ancestor (T_{MRCA}) is conditioned on being reached before time τ. The fifth model, *Birth-Death*, is not based on the standard Kingman coalescent, but on a critical branching process measured in units of $2N$ generations. Forward in time, the process starts with a founding event of one individual. Individuals give birth and die at equal rate 1. The process is conditioned on not becoming extinct before a period of time τ, and on reaching on average $2N$ individuals.

Stairway plot inference on the Yoruba SFS: We applied the model-flexible stairway plot method developed by Liu and Fu (2015) on the unfolded Yoruba SFS. Inferences are made on 200 SFS as suggested by their method. We use the script they provide to create 199 bootstrap samples of the Yoruba SFS. We also ignore the singletons for this method, and use the default parameter values suggested in their paper for the optimization.

SFS simulation with demography: We used two different method to simulate SFS under the four demographic models derived from the Kingman coalescent (*Linear*, *Exponential*, *Sudden* and *Conditioned*) or under a piecewise-constant demography reconstructed by the

stairway plot method.

Method 1: Simulate l independent topologies under the Kingman coalescent on which mutations are placed at rate θ (population mutation rate) (HUDSON *et al.* 1990). This allows to simulate the SFS of l independent loci.

Method 2: Another way to simulate SFS is using the following formula:

$$\mathbb{E}[\xi_i] = \frac{\theta}{2} \sum_{k=2}^{2n-i+1} k\, \mathbb{E}[t_k]\, \mathbb{P}(k,i) \tag{1}$$

where θ is the population mutation rate, t_k is the time during which there are k lines in the tree (hereafter named state k) and $\mathbb{P}(k,i)$ is the probability that a randomly chosen line at state k gives i descendants in the sample of size $2n$ (*i.e.*, at state $2n$) (FU 1995). For all models, the neutrality assumption ensures that

$$\mathbb{P}(k,i) = \frac{\binom{2n-i-1}{k-2}}{\binom{2n-1}{k-1}}$$

for $i \in [1, 2n-1]$ and $k \in [2, 2n-i+1]$. Using this probability allows to average over the space of topologies. This reduces considerably computation time since the space of topologies is very large, and produces smooth SFS for which only the t_k need to be simulated to obtain the expectations $\mathbb{E}[t_k]$.

The expectations $\mathbb{E}[t_k]$ are obtained as follow: for $k \in [2, 2n]$, times in the standard coalescent t_k^* are drawn from an exponential distribution of parameter $\binom{k}{2}$. For the *Linear* and *Exponential* models, and for the piecewise-constant demographies reconstructed by the stairway plot method, these times are then rescaled to take into account the given explicit demography (see, *e.g.*, HEIN *et al.* 2004, chap.4). For the *Sudden* model, we assume the coalescence of all lineages at time τ if the common ancestor has not been reached yet. For the *Conditioned* model, we keep only simulations for which $\sum_{k=2}^{2n} t_k^* \leq \tau$ where τ is the model parameter. The expectations $\mathbb{E}[t_k]$ are obtained by averaging over 10^7 simulations.

Alternatively, the expectations $\mathbb{E}[t_k]$ could also be obtained with analytic formulae provided by Polanski and Kimmel (2003).

For the *Birth-Death* model, we use the explicit formula for the SFS given in Delaporte *et al.* (2016).

We normalize the SFS computed under all these models so that their sum equals 1. This normalization removes the dependence on the mutation rate parameter θ. Consequently, the standard model has no parameters while all others have exactly one (τ).

Optimization of the parameter τ: For each demographic model, we optimize the parameter τ by minimizing the weighted square distance d^2 between the observed SFS of the Yoruba population and the predicted SFS under the model (simulated with *Method 2*). Both SFS are normalized for comparison. The distance is computed for all τ values in a given interval (no specific optimization method was used to find the minimum). With $\tilde{\eta}^{model}$ and $\tilde{\eta}^{obs}$ the folded and normalized SFS in the tested model and in the data respectively,

$$d^2(\tilde{\eta}^{model}, \tilde{\eta}^{obs}) = \sum_{i=2}^{n} \frac{(\tilde{\eta}_i^{model} - \tilde{\eta}_i^{obs})^2}{\tilde{\eta}_i{}^{model}}$$

The sum starts at $i = 2$ because we ignore $\tilde{\eta}_1^{obs}$, corresponding to singletons. To calculate the distance $d^{2'}$ between the SFS predicted by two models A and B, we weight the terms by the mean of the two models:

$$d^{2'}(\tilde{\eta}^A, \tilde{\eta}^B) = \sum_{i=2}^{n} \frac{(\tilde{\eta}_i^A - \tilde{\eta}_i^B{}^2}{)} (\tilde{\eta}_i^A + \tilde{\eta}_i^B)/2$$

nference of the Yoruba demography with ∂a∂i: We inferred the demography of the Yoruba population with the software ∂a∂i v1.7 (Gutenkunst *et al.* 2009), testing the three models of explicit demography (*Linear*, *Exponential* and *Sudden*). The demographic models were specified so that the only parameter to optimize is τ like for the distance-based inference method. Singletons were masked and the method was applied on the folded Yoruba SFS. Details on the demographic functions and parameter values used for the optimization in ∂a∂i

are provided in the Supplementary Methods. We ran the method 100 times for each model and kept the parameter value with the best maximum log composite likelihood over the 100 runs. In Figure S4, we plot the best log composite likelihood of the 100 runs.

Scaling of the coalescent time: Optimized values of the parameter $\widehat{\tau}$ for each model are expressed in coalescent time units, *i.e.*, scaled in $2N(0)$ generations. As the model population size at time zero, $2N(0)$, is unknown, to scale these coalescent time units in numbers of generations and consequently in years, we used the expected number of mutations per site M. From the dataset, we have $M^{obs} = S/L$ where S is the number of single nucleotide mutations (a k-allelic SNP accounts for $k-1$ mutations) and L is the length of the accessible sequenced genome in the 1 000 genomes project (90% of the total genome length, The 1000 Genomes Project Consortium). For the theoretical value, we get that $M^{theo} = \mu\,\widehat{T}_{tot}\,C$, where we know the mutation rate μ from the literature and the total tree length expressed in coalescent time units $\widehat{T}_{tot}$ from the SFS simulations. Here C is the coalescent factor, that is the number of generations per coalescent time unit, also corresponding to $2N_e(0)$ where $N_e(0)$ is the effective population size at present time. The total number of generations in the tree is $\widehat{T}_{tot}\,C$ from which we derive the total number of mutations per site M^{theo}. Thus, using the observed value M^{obs}, we can estimate C by $S/(\mu\,L\,\widehat{T}_{tot})$. We assumed a mutation rate of 1.2×10^{-8} per base pair per generation (Conrad *et al.* 2011; Campbell *et al.* 2012; Kong *et al.* 2012). With the coalescent factor C, we can then convert a coalescent time unit into a number of generations, or into a number of years assuming 24 years as generation time (Scally and Durbin 2012).

Graphical representation of the inferred demographies: To represent the inferred explicit demographies (models *Linear*, *Exponential* and *Sudden*), we plot the shape of the demography with the optimized value $\widehat{\tau}$ for each model. For the implicit demographies (models *Conditioned* and *Birth-Death*), as there is no explicit demographic shape, we plot

the mean trajectory of fixation of a new allele in the population: forward in time, these fixation trajectories illustrate the expansion of the descendance of the sample's ancestor in the population (see the Supplementary Methods for details).

Comparing the model-constrained and model-flexible methods to infer *Linear* growth: We applied both methods (the one-parameter inference method and the stairway plot method) on SFS simulated under *Linear* growth. To test the stairway plot method on a *Linear* growth demography, we simulate 200 independent SFS using *Method 1*, with sample size $2n = 216$, $\theta = 100$ (arbitrary value removed by normalization) and a founding time $\tau = 2.48$ (estimated for the Yoruba population, see Results). The SFS are simulated with either 10^3, 10^4 or 10^5 independent loci. We scaled the simulated SFS to obtain a total number of $S = 20\,417\,698$ variants, so that the total number of variants in the simulated SFS is the same as in the Yoruba SFS. We ran the stairway plot method on these 200 independent SFS with the default parameter values suggested in the method, and with the same mutation rate (1.2×10^{-8} per base pair per generation) and generation time (24 years) as in our study. We report the median demography of these 200 independent inferences.

To test the one-parameter inference method on these SFS simulated under the *Linear* model, we run the parameter optimization on a SFS simulated with either 10^3, 10^4, 10^5 or 10^6 loci. The search of the parameter value that minimizes the distance d^2 was optimized with a Newton-Raphson algorithm. Derivatives were calculated at $\tau \pm 0.05$ where τ is the parameter value being optimized. The optimization stopped when the optimization step of the parameter value was smaller than 10^{-3}.

Data and software availability The 1000 genomes project data used in this study is publicly available at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`. The code in Python and C written for the study is available at `https://github.com/lapierreM/Yoruba_demography`. The code in C used for the *Method 1* of SFS simulation is

available upon request to G. Achaz.

RESULTS

We inferred the demography of the Yoruba population (Africa), from the whole-genome polymorphism data of 108 individuals (data from the 1 000 Genomes Project, The 1000 Genomes Project Consortium), with SFS-based methods, either model-constrained or model-flexible.

It has been shown that human populations have been growing since their emergence in Africa, and that African populations were supposedly not affected by the Out-of-Africa bottleneck described for Eurasian populations (Marth *et al.* 2004; Atkinson *et al.* 2008; Gutenkunst *et al.* 2009; Gronau *et al.* 2011; Tennessen *et al.* 2012). Analyses using the PSMC method (Li and Durbin 2011) have shown a reduction of the African population size after the divergence with non-African populations. However, Mazet *et al.* (2016) have recently shown that these analyses could be biased by population structure. Based on this previous knowledge, for the model-constrained method, we chose to infer the Yoruba demography with simple models of growth, *i.e.*, with only one phase of growth characterized by a single parameter. These five models are: *Linear*, *Exponential* or *Sudden* growth, a *Conditioned* model where the T_{MRCA} is conditioned on being smaller than the given parameter, and a critical *Birth-Death* model based on a branching process (Figure 1). To infer the Yoruba demography, we fit the SFS predicted under each model with the observed Yoruba SFS (all SFS are folded). The SFS were normalized to remove the population mutation rate parameter θ, so that each model is characterized by one single parameter τ which has the dimension of a time duration. We fit this parameter by least-square distance between the observed SFS and the predicted SFS, and by maximum likelihood using the ∂a∂i software (Gutenkunst *et al.* 2009). For the model-flexible inference, we used the stairway plot method developed recently by Liu and Fu (2015), which infers a piecewise-constant demography based on the SFS. For this method, the number of parameters to be estimated

is determined by a likelihood-ratio test. It can range from 1 to $2n-1$ where $2n$ is the number of sequences in the sample.

The Yoruba SFS was constructed by taking into account the entire genome. Removing the coding parts of the genome to avoid potential bias due to selection does not affect the shape of the SFS substantially (Figure S2), since the coding parts represent a very small fraction of the human genome. The first bin of the observed SFS, accounting for mutations found in one chromosome of one individual in the sample (black dot in the observed SFS in Figure 3B), seemed to lie outside the rest of the distribution. This could be due to sequencing errors being considered as singletons (Achaz 2008), and thus we chose to ignore this value for the model optimization. We have also made sure that the SFS shape was not affected greatly by the sample size. We compared the SFS of a subsample of half the Yoruba individuals ($2n = 108$) with the full sample SFS ($2n = 216$) (Figure S3). This shows that the only bin of the SFS which is significantly affected by this subsampling is the first one, containing the singletons. As we ignore it in our study, it does not influence our results.

The analysis of the Yoruba SFS with the stairway plot method results in a complex demography with several bottlenecks in the last 160 000 years (Figure 2). The current effective population size $N_e(0)$ is 28 500 (time 0 does not correspond to present time as we ignored singletons, see discussion). The demographic history earlier than 160 000 years ago shows spurious patterns that should not be interpreted, according to Liu and Fu (2015).

The inference of the Yoruba demography with one-parameter models was done by minimizing the distance between observed and predicted SFS. This gave an optimized value $\widehat{\tau}$ of the parameter τ (Figure 3A and Table 1) (with $\widehat{\tau}$ in coalescent units, *Linear*: $\widehat{\tau} = 2.48$, *Exponential*: $\widehat{\tau} = 1.82$, *Sudden*: $\widehat{\tau} = 1.36$, *Conditioned*: $\widehat{\tau} = 1.89$, *Birth-Death*: $\widehat{\tau} = 2.28$). Plotting the predicted SFS with the optimized parameter value $\widehat{\tau}$ confirmed their goodness of fit with the observed Yoruba SFS (Figure 3B). Compared to the standard model without demography, the addition of just one parameter allows for a surprisingly good fit of the observed Yoruba SFS. The Yoruba demography thus seems to be compatible with a

simple scenario of growth. On the other hand, the demography inferred by the stairway plot predicts a SFS which does not fit well the observed Yoruba SFS: the distance between the observed Yoruba SFS and the expected SFS under the stairway plot demography is ten times the distance between any of the one-parameter model SFS and the data (Figure 3B and Table 1).

The best fitting SFS under each of the five demographic models all have a square distance d^2 of the order of 10^{-4} with the observed Yoruba SFS (Figure 3A and Table 1) and have highly similar shapes (Figure 3B). This suggests that the five demographic models used to infer the demography of the Yoruba are hard to distinguish based only on the observed SFS. To validate the use of a least square distance to find the best fitting SFS, we also infered the Yoruba demography using the ∂a∂i software. This model-constrained method based on the SFS uses a diffusion approximation to simulate SFS and a likelihood framework for the parameter optimization. We tested the three models of explicit demography (*Linear*, *Exponential* and *Sudden* growth) parametrized in the same way as in our method. The best parameter values found by ∂a∂i by maximum log composite likelihood are the same as by our method (with $\widehat{\tau}$ in coalescent units, *Linear*: $\widehat{\tau} = 2.48$, *Exponential*: $\widehat{\tau} = 1.82$, *Sudden*: $\widehat{\tau} = 1.36$). Moreover, the log composite likelihoods of the best fitting SFS under each model are on the same scale (the likelihoods are directly comparable because the number of parameters is the same for each model) : *Linear*: $\ln(L) = -3107$, *Exponential*: $\ln(L) = -3953$, *Sudden*: $\ln(L) = -3393$ (Figure S4). They rank the explicit demography models in the same order as the least square distance d^2 would rank them: the best model is *Linear* growth, then *Sudden* and finally *Exponential* growth.

We computed the expected T_{MRCA} based on the predicted SFS using (1): as the SFS predicted under each model are very similar, it means that they have roughly the same estimated time durations t_k while there are k branches in the coalescent tree of the Yoruba sample. From these expected t_k we can compute $T_{MRCA} = \sum_{i=2}^{2n} t_k$. This is the T_{MRCA} of the sample, but we can assume that it is the same as the T_{MRCA} of the population, because with

such a large sample size, the probability that the T_{MRCA} of the population is different from the T_{MRCA} of the sample becomes very small. Under each of four models (excluding the *Birth-Death* model for which there is no obvious common time scaling), the expected T_{MRCA} for the Yoruba population is 1.3 in coalescent units. By using the number of mutations per site in the data and the total tree length inferred from the simulations, we scaled back this T_{MRCA} in number of generations and in years, assuming a mutation rate of 1.2×10^{-8} per base pair per generation (Conrad *et al.* 2011; Campbell *et al.* 2012; Kong *et al.* 2012) and a generation time of 24 years (Scally and Durbin 2012) (see Methods). The T_{MRCA} of the Yoruba population inferred under the four demographic models is of 87 100 generations corresponding to 1.7 million years. The inferred demographic models, with scaling in coalescent units, number of generations and number of years, are shown in Figure 4. The coalescent unit of 67 000 estimated to scale the inferred coalescent times in number of years corresponds to a present effective population size $N_e(0)$ of 33 500.

The demography inferred by the stairway plot method for the Yoruba population is a piecewise-constant demography showing much more complex patterns of growth and bottlenecks than the one-parameter models (Figure 2). Moreover, the expected SFS under this inferred demography does not fit well the observed Yoruba SFS (Figure 3B). To understand what could produce such a complex demography, we simulated SFS under a *Linear* growth with the founding time $\widehat{\tau} = 2.48$ inferred for the Yoruba population. We simulated three sets of 200 SFS, with respectively 10^3, 10^4, and 10^5 loci, to obtain SFS with more or less noise (solid lines on Figure 5A). We applied the two inference methods to these SFS. The median demographies inferred by the stairway plot method are strongly affected by the noise of the SFS, as shown on Figure 5B. When the number of simulated loci is very large (median of 200 independent demographies inferred with 10^6 loci), the stairway plot gives a good approximation of the true demography, and the expected SFS under the inferred demography fits the input SFS. However, for smaller numbers of loci (median of 200 independent demographies inferred with 10^5 loci or less), the stairway plot shows complex patterns of growth and bot-

tlenecks incompatible with the true demography, and the expected SFS under the inferred demographies do not fit the input SFS. On the contrary, the one-parameter method infers a *Linear* demography with a founding time close to the true value for SFS simulated with 10^4 loci or more (Table 2).

DISCUSSION

In this study, we fit the SFS of the Yoruba population with five simple demographic models of growth described by one parameter. Surprisingly, even though these five models are quite distinct in the way they model population growth, fitting them on the Yoruba data results in strongly similar SFS, which all show an excellent goodness of fit with the observed Yoruba SFS. Fitting the same SFS with the stairway plot method (Liu and Fu 2015), a model-flexible method which infers a piecewise-constant demography, resulted in a complex demography with several bottlenecks in the last 160 000 years. The poor goodness of fit of the expected SFS under this inferred demography with the Yoruba SFS indicates that this complex demography is not to be trusted and suggests that the way the method estimates the number of change points is too flexible.

The results obtained by the model-constrained and model-flexible methods showed some similarities: the current population size $N_e(0)$ of about 30 000 inferred with the stairway plot corresponds roughly to the coalescent unit of 67 000 generations (equivalent to $2N_e(0)$ in the coalescent theory) found with the one-parameter models. Similarly, the T_{MRCA} of $\sim$1.7 million years inferred with the one-parameter models seems to match with the last time point of the stairway plot, at about 1.9 million years.

We hypothesize that the complexity of the demography inferred by the stairway plot method is caused by the irregularities of the observed Yoruba SFS. Two concurrent non-exclusive explanations can be put forward for these irregularities. First, they can be due to the sampling and thus be considered as noise that should not be interpreted as evidence for demography. Second, these irregularities could be biologically relevant and result from

a very complex demographic history. To assess the impact of noise on the stairway plot method, we tested it on simulated SFS under a *Linear* growth. These SFS were simulated with different numbers of independent loci: the more loci, the less noise in the simulated SFS. The stairway plot inference on these SFS shows that the method is strongly affected by the noise in the SFS simulated data: whereas the demography inferred for a smooth SFS (corresponding to a high number of independent loci) corresponds to the true demography approximated as piecewise constant, the demographies inferred for smaller numbers of loci show complex patterns of bottlenecks and deviate strongly from the true demography. It could be that this method captures the signal contained in these irregularities and infers a demography taking them into account, whereas the one-parameter models fit the global trend of the SFS shape and can thus infer the true demography for much smaller numbers of loci. One solution could be to constrain the number of parameters allowed for model-flexible methods: it seems that determining it by likelihood-ratio test, as it is done in the stairway plot method, is not conservative enough, as it does not prevent from overfitting the noise. If the number of parameters was forced to be small, the method might capture the global trend of the demography and avoid this issue. The SFS reconstructed under the demographies inferred by the stairway plot, however, differ strongly from the input SFS. If the issue was the overfitting of noise, we would expect the reconstructed SFS to fit the data more closely. The method is clearly biased by noise on the SFS but it remains unclear why. It would require further investigation to analyze how the different characteristics of this particular method, such as the parametrization of population size history, respond to noise, and what is responsible for this bias.

The five one-parameter demographic models all predict virtually the same SFS for the Yoruba population. Therefore, they also predict the same T_{MRCA} for the Yoruba population. This T_{MRCA} of $\sim$1.3 in coalescent units corresponds, with our scaling of coalescent time based on the number of mutations per site, to $\sim$1.7 million years. This estimation is similar to results concerning the whole human population, obtained by Blum and Jakobsson (2011)

or reviewed in Garrigan and Hammer (2006). Although the commonly admitted date of emergence of the anatomically modern human is around 200 000 years ago, Blum and Jakobsson showed that finding a much older T_{MRCA} was compatible with the single-origin hypothesis, assuming a certain ancestral effective population size. These ancient times to most recent common ancestor could also be explained by gene flow in a structured ancestral population (Garrigan and Hammer 2006).

Although all five models predict the same T_{MRCA}, the inferred demographies differ substantially between the models (Figure 3A). In the time range further beyond the T_{MRCA}, no information is carried by the sample. Thus, the inferred demographies differ in this time range (Figure 4), making the inferred founding time of the Yoruba population unreliable.

Our results with one-parameter models are reproducible with another model-constrained method, ∂a∂i, which uses different approaches both for the theoretical SFS simulations (diffusion approximation) and the parameter optimization (composite likelihood). This shows that, for models having the same number of parameters, a distance-based approach finds the same ranking of models as a likelihood framework, while being computationally less intensive. Furthermore, the distance-based approach allows for intuitive evidence on the fact that these different models actually all perform very well to fit the Yoruba SFS: the small differences of distance between the best SFS predicted by each model and the observed SFS could be due only to the noise in the observed SFS and thus do not mean that one model is better than another.

Among the five tested demographic models, two pairs of models seem to predict particularly similar SFS (pairs of models with the two smallest values of d^2 in Table 1). First, the *Linear* (L) and *Exponential* (E) growth models predict almost identical SFS for the Yoruba population ($d^2(\tilde{\eta}^L, \tilde{\eta}^E) = 2.2 \times 10^{-5}$). Figure 4 shows that, in the time range where information is conveyed by the mean coalescent tree of the population, *i.e.*, between present time and the T_{MRCA}, these two demographies are very similar. This explains why their SFS are almost indistinguishable, and shows that in this parameter range, it is impossible to dis-

tinguish linear from exponential growth. Second, the SFS predicted under the two models with implicit demography, *Conditioned* (C) and *Birth-Death* (BD), are so similar that they are undistinguishable in Figure 3B ($d^2(\tilde{\eta}^C, \tilde{\eta}^{BD}) = 3.5 \times 10^{-6}$). This raises a question on how these two models, based on different processes — a Wright-Fisher model or a branching process — compare and in particular why their SFS are so similar.

As we compute the distance statistic to optimize the models on normalized SFS, the information of the magnitude of the SFS (often referred to as θ, the population mutation rate) is lost. However, as the inferred SFS under the five demographic models all have the same shape, the constant θ by which they should be multiplied to fit the real, not normalized, Yoruba SFS would be the same for all five models. Thus, this information would not allow to choose which model infers the most realistic value of θ.

The outlying first bin of the Yoruba SFS, corresponding to singletons, was removed from our inference because it can be affected by sequencing errors. As the relatively low to moderate coverage of the 1000 Genomes project could also result in an underestimation of doubletons and tripletons, we optimized τ masking also these values. It did not change the estimation of $\widehat{\tau}$ and thus had no effect on the inferred demographies. As the first bin of the SFS accounts for the mutations that occur in the terminal branches of the coalescent tree, a large part of the excess of singletons can be due to very recent and massive growth. Recent studies with deep sequencing coverage have shown that there is a large abundance of rare variants in human populations (Coventry *et al.* 2010; Nelson *et al.* 2012; Gazave *et al.* 2014). As the dataset we used for this study had a limited sample size and low-coverage, we focused on the inference of demography in the more distant past. Thus, because of both sequencing errors and incompatibility with our one-parameter models, singletons were not taken into account. Our inferences concern the population before this recent and massive growth. It should also be noted that Liu and Fu (2015) emphasize that the strength of their method is in capturing recent demographic history. Thus, ignoring singletons, although it is an existing feature of their software, might not be the most appropriate use of the stairway

plot.

For non-African human population, the SFS based on the 1 000 Genomes Project data are not monotonous: their shape is more complex than the SFS of the Yoruba population. Thus, one-parameter models cannot capture the complexity of the demographic histories underlying these types of observed SFS. Even for the Yoruba population, capturing the recent growth event, by taking into account the singletons, would have required adding another parameter. The stairway plot method shows more flexibility and could capture the signal for more complex demographic histories, provided that the number of independent loci is very large so that there is no bias due to noise.

Overall, this study shows that even in the case of a simple demography, the scenario inferred by the stairway plot, a model-flexible method, can show spuriously complex patterns of growth and decline and can predict SFS poorly fitting with the initial SFS data. This might be explained by overfitting of the method to the noise present in the observed SFS, which can be expected for a reasonable number of loci. We also show that simple models described by one parameter can have an excellent goodness of fit to the data and avoid the issue of noise overfitting. The results indicate that the demography of the Yoruba population is compatible with simple one-parameter models of growth, and that the expected T_{MRCA} of this population can be estimated at $\sim$1.7 million years. However, the SFS is not sufficient to determine which model better characterizes the Yoruba demographic growth, and estimations of the founding time of the population, that depend on the chosen model, are thus unreliable. More generally, this study illustrates the issue of non-identifiability of demographies based on the SFS of a finite sample.

Our comparison of a model-constrained method using one parameter models with a model-flexible method using a potentially large number of parameters highlights the importance of the model complexity. How many parameters should we use to "properly" characterize a demography? We argue that low complexity models should be tested first. For model-flexible methods, the number of parameters is usually unbounded and determined

by successive likelihood ratio tests. This statistical framework implies that a certain risk is taken at each successive step, and that with the repetition of steps, errors can potentially be made. For example, these errors can lead to spurious inferences in noisy data (*i.e.*, any real data). We recommend (visually) monitoring the improvement in goodness of fit when adding new parameters on statistical grounds. Examination of the intermediate steps of fitting would likely prevent an unnecessary increase in the model complexity.

ACKNOWLEDGMENTS

We thank Cécile Delaporte for preliminary work on this project and Simon Boitard, Michael Blum, Konrad Lohse and three anonymous reviewers for useful comments on the manuscript. G.A. and M.L acknowledge support from the grant ANR-12-NSV7-0012 Demochips from the Agence Nationale de la Recherche (France). M.L. is funded by the PhD program 'Interfaces pour le Vivant' of UPMC Univ Paris 06. G.A., A.L. and M.L. thank the *Center for Interdisciplinary Research in Biology* for funding.

LITERATURE CITED

Achaz, G., 2008 Testing for neutrality in samples with sequencing errors. Genetics *179*(3): 1409–1424.

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics *183*(1): 249–258.

Adams, A. M. and R. R. Hudson, 2004 Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics *168*(3): 1699–1712.

Atkinson, Q. D., R. D. Gray, and A. J. Drummond, 2008 mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. Molecular biology and evolution *25*(2): 468–474.

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

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

Blum, M. G. and M. Jakobsson, 2011 Deep divergences of human gene trees and models of human origins. Molecular biology and evolution *28*(2)**:** 889–898.

Campbell, C. D., J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, L. Vives, B. J. O'Roak, P. H. Sudmant, J. Shendure, M. Abney, C. Ober, and E. E. Eichler, 2012 Estimating the human mutation rate using autozygosity in a founder population. Nat Genet *44*(11)**:** 1277–1281.

Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Cassals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, and P. Awadalla, 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics *43*(7)**:** 712–714.

Coventry, A., L. M. Bull-Otterson, X. Liu, A. G. Clark, T. J. Maxwell, J. Crosby, J. E. Hixson, T. J. Rea, D. M. Muzny, L. R. Lewis, and others, 2010 Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature communications **1:** 131.

Delaporte, C., G. Achaz, and A. Lambert, 2016 Mutational pattern of a sample from a critical branching population. Journal of mathematical biology**:** 1–38.

Excoffier, L., I. Dupanloup, E. Huerta-Snchez, V. C. Sousa, and M. Foll, 2013 Robust Demographic Inference from Genomic and SNP Data. PLoS Genet *9*(10)**:** 1–17.

Fay, J. C. and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics *155*(3): 1405–1413.

Fu, Y.-X., 1995 Statistical properties of segregating sites. Theoretical population biology *48*(2): 172–197.

Garrigan, D. and M. F. Hammer, 2006 Reconstructing human origins in the genomic era. Nature Reviews Genetics *7*(9): 669–680.

Gazave, E., L. Ma, D. Chang, A. Coventry, F. Gao, D. Muzny, E. Boerwinkle, R. A. Gibbs, C. F. Sing, A. G. Clark, and others, 2014 Neutral genomic regions refine models of recent rapid human population growth. Proceedings of the National Academy of Sciences *111*(2): 757–762.

Gronau, I., M. J. Hubisz, B. Gulko, C. G. Danko, and A. Siepel, 2011 Bayesian inference of ancient human demography from individual genome sequences. Nature genetics *43*(10): 1031–1034.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genet *5*(10): 1–11.

Hein, J., M. Schierup, and C. Wiuf, 2004 *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA.

Hudson, R. R. and others, 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology *7*(1): 44.

Kim, J., E. Mossel, M. Z. Rácz, and N. Ross, 2015 Can one hear the shape of a population history? Theoretical population biology **100:** 26–38.

Kingman, J. F. C., 1982 The coalescent. Stochastic processes and their applications *13*(3): 235–248.

Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. W. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson, 2012 Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*(7412): 471–475.

Lambert, A., 2010 Population genetics, ecology and the size of populations. Journal of mathematical biology *60*(3): 469–472.

Lapierre, M., C. Blin, A. Lambert, G. Achaz, and E. P. Rocha, 2016 The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Molecular biology and evolution: msw048.

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

Liu, X. and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nature genetics *47*(5): 555–559.

Lukić, S., J. Hey, and K. Chen, 2011 Non-equilibrium allele frequency spectra via spectral methods. Theoretical population biology *79*(4): 203–219.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry, 2004 The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. Genetics *166*(1): 351–372.

Mazet, O., W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi, 2016 On the importance of being structured: instantaneous coalescence rates and human evolution[mdash]lessons for ancestral population size inference[quest]. Heredity *116*(4): 362–371.

Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the

allelic spectrum? Theor Popul Biol *73*(3): 342–8.

Nawa, N. and F. Tajima, 2008 Simple method for analyzing the pattern of DNA polymorphism and its application to SNP data of human. Genes & genetic systems *83*(4): 353–360.

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, and others, 2012 An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*(6090): 100–104.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics *154*(2): 931–942.

Polanski, A. and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics *165*(1): 427–436.

Pool, J. E., I. Hellmann, J. D. Jensen, and R. Nielsen, 2010 Population genetic inference from genomic sequence variation. Genome research *20*(3): 291–300.

Pybus, O. G., A. Rambaut, and P. H. Harvey, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics *155*(3): 1429–1437.

Scally, A. and R. Durbin, 2012 Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet *13*(10): 745–753.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, and others, 2012 Evolution and functional impact of rare coding variation from deep sequencing of human exomes. science *337*(6090): 64–69.

Terhorst, J. and Y. S. Song, 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.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature *526*(7571)**:** 68–74.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genetical Research **74:** 65–79.

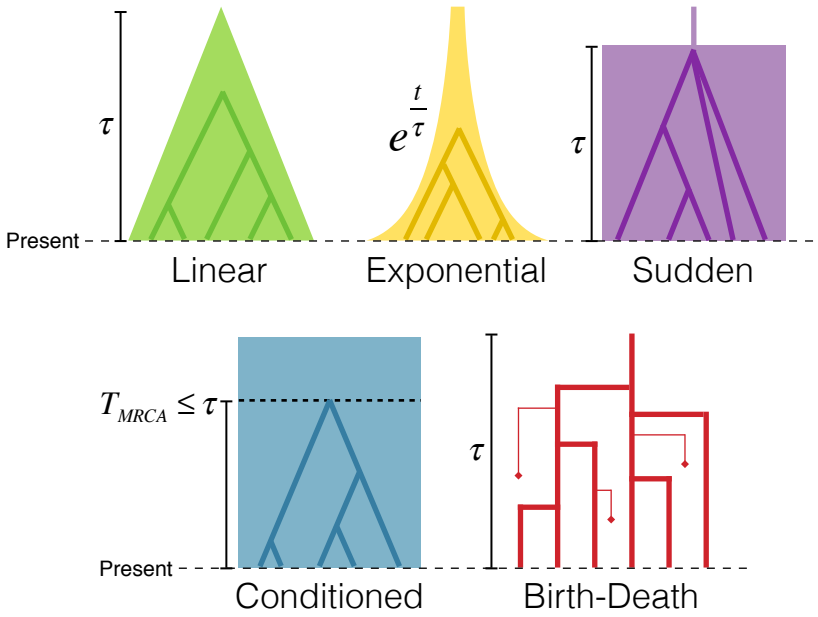

Figure 1: The five demographic models. Each model has one single time parameter τ.

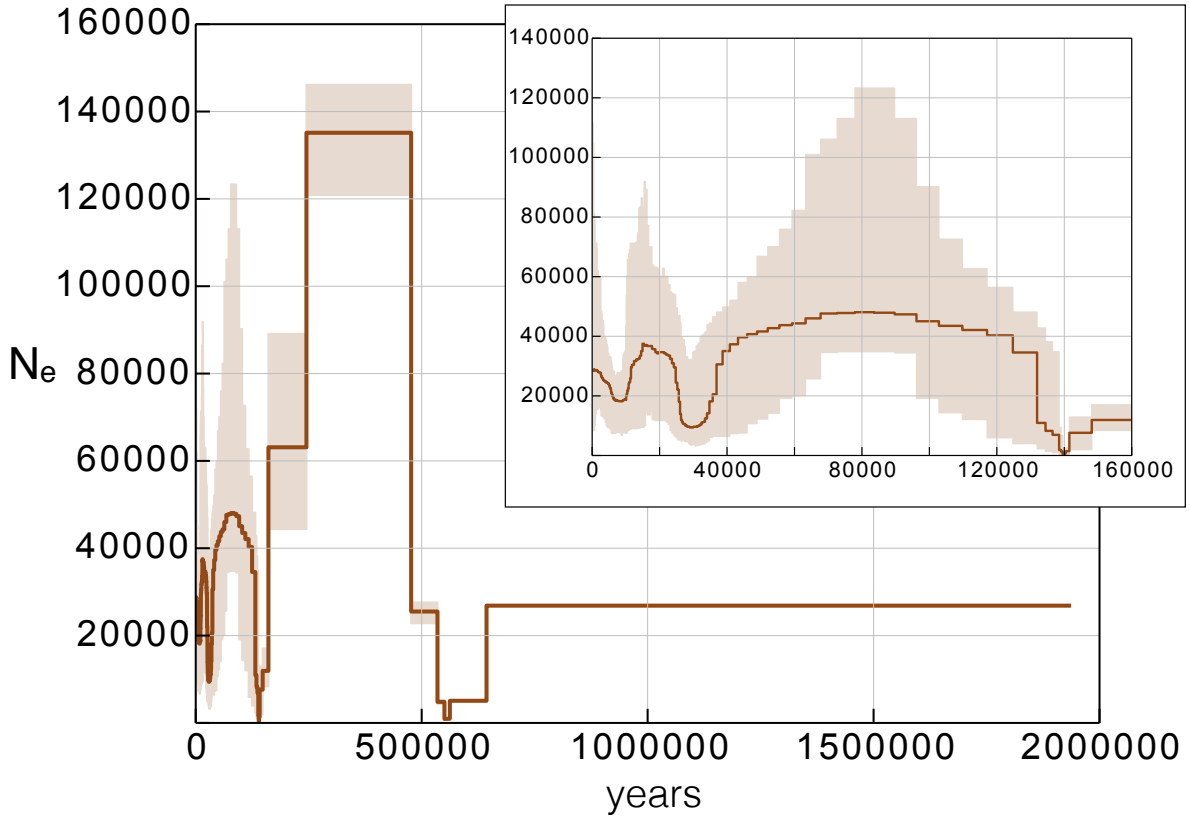

Figure 2: Stairway plot inference of the Yoruba demography. The inferred effective size N_e of the Yoruba population is plotted from present time (0) to the past. The inset is a zoom between 0 and 160 000 years. The thick brown line is the median N_e, the light brown area is the $[2.5, 97.5]$ percentiles interval. The inference is based on 200 bootstrap samples of the unfolded Yoruba SFS. The singletons are not taken into account for the optimization of the stairway plot.

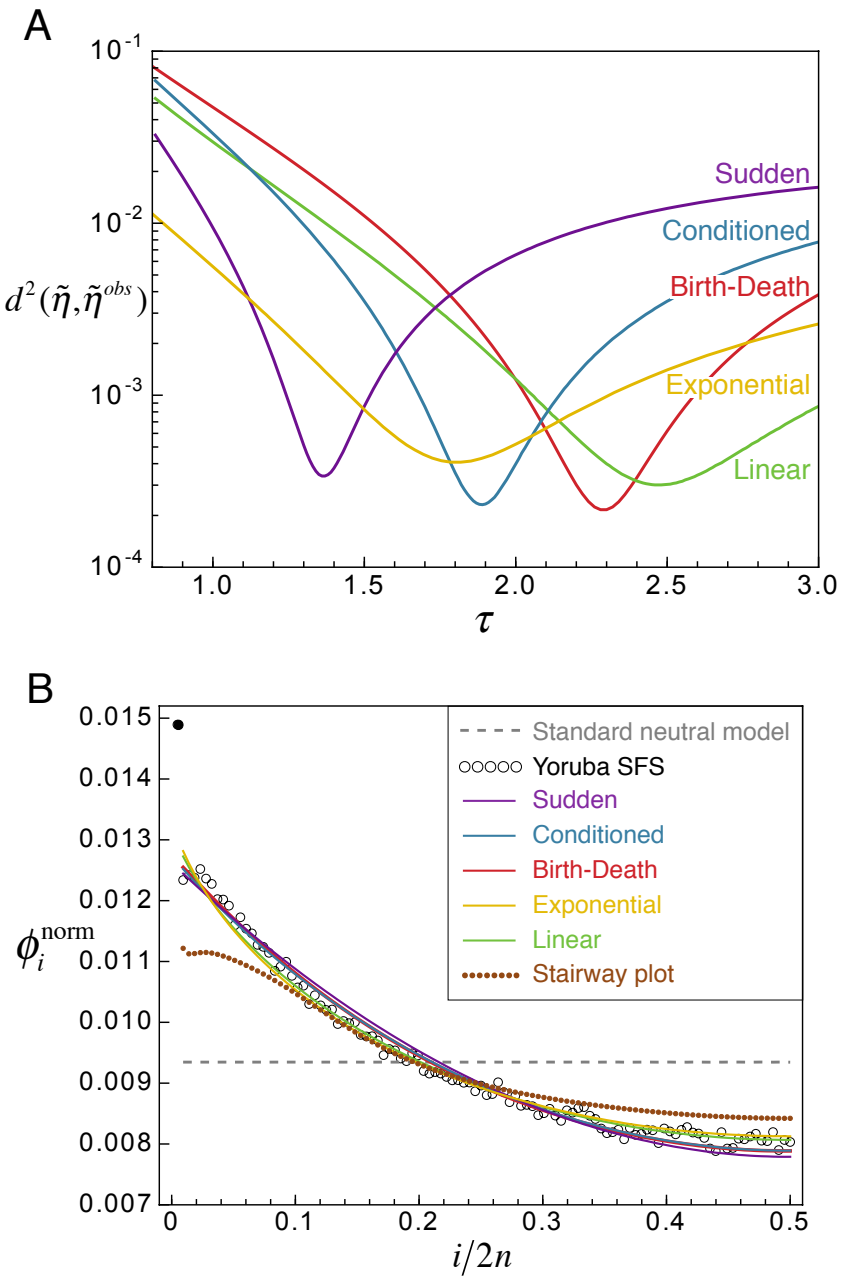

Figure 3: Inference of the Yoruba demography with one-parameter models. A) Weighted square distance $d^2(\tilde{\eta}, \tilde{\eta}^{obs})$ between the normalized Yoruba SFS $\tilde{\eta}^{obs}$ and the normalized predicted SFS $\tilde{\eta}$ under each of the five models, depending on the value of the parameter τ (Purple: *Sudden*, Blue: *Conditioned*, Red: *Birth-Death*, Yellow: *Exponential*, Green: *Linear*). B) Predicted SFS under each of the five models, with the optimized value $\widehat{\tau}$ of the parameter, and under the demography inferred by the stairway plot (brown dotted line). The Yoruba SFS is shown in empty circles. The first dot, colored in black, accounting for the singletons, was not taken into account for the optimization of τ to avoid potential bias due to sequencing errors. The grey dashed line is the expected SFS under the standard neutral model without demography. Colors match the plot above (the predicted SFS under the models *Birth-Death* and *Conditioned* are indistinguishable). The SFS are folded, transformed and normalized (see Methods).

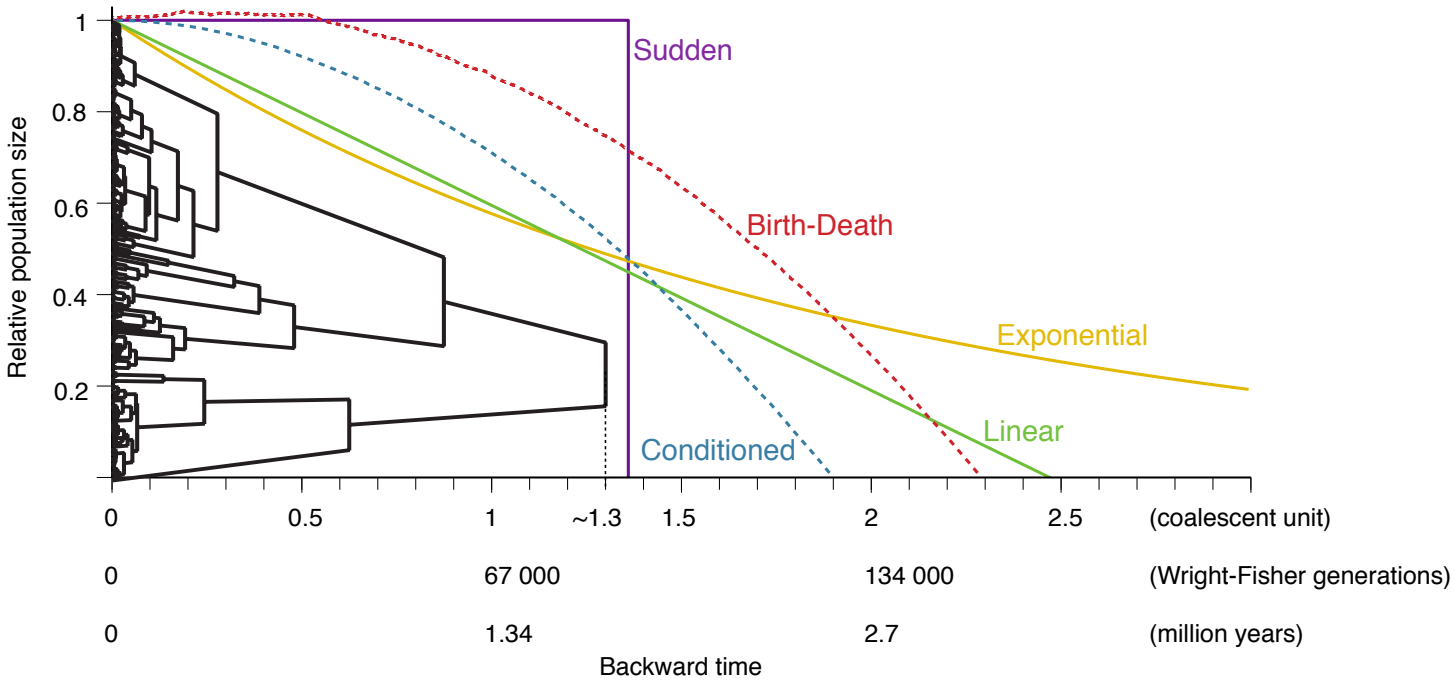

Figure 4: Demographic histories and reconstructed tree estimated from the Yoruba SFS. The tree shown has internode durations t_k during which there are k lineages consistent with the SFS (the topology was chosen uniformly among ranked binary trees with $2n$ tips). Time is given in coalescent units, and scaled in number of generations and in millions of years. The demographic histories (solid lines: explicit models, dashed lines: implicit models) are plotted with their optimized $\widehat{\tau}$ values. See the supplementary methods for details on the demographic histories plotted for the models with implicit demographies (*Birth-Death* and *Conditioned*)

.

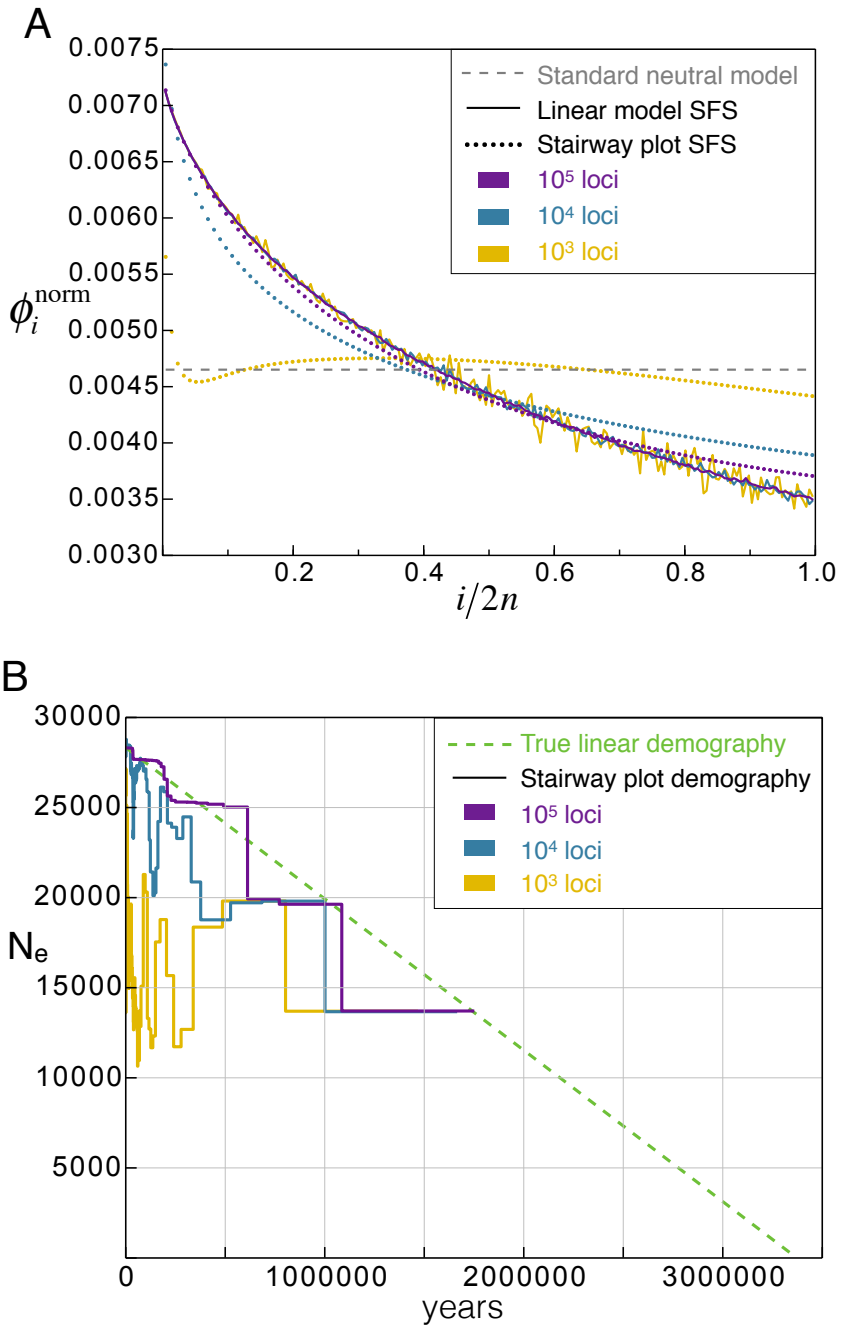

Figure 5: Stairway plot inference of a linear demography SFS with noise. A) Solid lines: mean of 200 SFS simulated independently under the *Linear* growth model, with either 10^5 loci (purple), 10^4 loci (blue) or 10^3 loci (yellow). Dotted lines: expected SFS under the demography reconstructed by the stairway plot method for different number of loci (same colors than solid lines). The grey dashed line is the expected SFS under the standard neutral model without demography. The SFS are transformed and normalized (see Methods). B) Stairway plot demographic inference: Median of 200 independent demographies inferred with 200 independently simulated SFS for each number of loci (colors match the plot above). The true demography is the green dashed line. The inferred effective size N_e is plotted from present time (0) to the past.

	Data	*Linear*	*Exponential*	*Sudden*	*Conditioned*	*Birth-Death*
Linear	3.0×10^{-4}	0				
Exponential	4.1×10^{-4}	2.2×10^{-5}	0			
Sudden	3.4×10^{-4}	3.5×10^{-4}	5.5×10^{-4}	0		
Conditioned	2.3×10^{-4}	1.6×10^{-4}	5.5×10^{-4}	3.7×10^{-5}	0	
Birth-Death	2.2×10^{-4}	1.7×10^{-4}	3.1×10^{-4}	4.1×10^{-5}	3.5×10^{-6}	0
Stairway plot	2.9×10^{-3}	3.1×10^{-3}	3.3×10^{-3}	2.8×10^{-3}	2.8×10^{-3}	2.9×10^{-3}

Table 1: Least-square distance d^2 between pairs of observed Yoruba SFS and optimized SFS under the five demographic models or the stairway plot method.

Number of loci	5% percentile	Mean $\widehat{\tau}$	95% percentile
10^3	2.569	2.713	2.893
10^4	2.463	2.503	2.540
10^5	2.473	2.485	2.498
10^6	2.478	2.483	2.487

Table 2: Inference of the founding time $\widehat{\tau}$ under the *Linear* model on SFS with noise. Mean, 5% and 95% percentile of the founding time inferred with a *Linear* model. The SFS on which the inference is made are simulated with a founding time τ of 2.48, with different number of loci, using the method with topology reconstruction.