Slicing and dicing the genome: A statistical physics approach to population genetics.

Yosef E. Maruvka,[1] Nadav M. Shnerb,[1] Sorin Solomon,[2] Gur Yaari,[3] and David A. Kessler[1]

[1]*Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel*

[2]*Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem 91904, Israel*

[3]*Department of Ecology and Evolutionary Biology,*
Yale University, New Haven, Connecticut 06520, USA

Abstract

The inference of past demographic parameters from current genetic polymorphism is a fundamental problem in population genetics. The standard techniques utilize a reconstruction of the gene-genealogy, a cumbersome process that may be applied only to small numbers of sequences. We present a method that compares the total number of haplotypes (distinct sequences) with the model prediction. By chopping the DNA sequence into pieces we condense the immense information hidden in sequence space into a function for the number of haplotypes versus subsequence size. The details of this curve are robust to statistical fluctuations and are seen to reflect the process parameters. This procedure allows for a clear visualization of the quality of the fit and, crucially, the numerical complexity grows only linearly with the number of sequences. Our procedure is tested against both simulated data as well as empirical mtDNA data from China and provides excellent fits in both cases.

I. INTRODUCTION

The genetic variations within a population reflect its demographic history [1]. The DNA sequence of an individual sampled at present is determined by the original genetic code of its past ancestors plus the mutations accumulated along its line of descent. If two individuals are sampled, their average genetic distance depends on the number of generations since their most recent common ancestor. Genetic variability thus reflects the genealogy of the population. The genealogy, in turn, depends very much on demography: if the population is small the family trees are shorter than those of a large population, and the typical shape of the tree differs between fixed and growing populations.

The standard inference techniques [2–4] used in population genetics are based on reconstruction of the genetic tree. Given the variations in a sampled population at present, one may recover, in principle, the most probable gene genealogy and extract demographic parameters from the size and the topology of the tree. For noncoding parts of the DNA the process of neutral coalescence, conditioned on the observed sequences, generates a set of random-joining trees for the gene-genealogy. In principle one would like to identify the most probable tree using some sort of a maximum likelihood procedure. In practice an exact enumeration of all possible gene genealogies is impossible as the number of trees grows faster than exponentially with the sample size. Practicable algorithms are based on Monte-Carlo simulations that search through a subset of all possible trees and weight the outcome with the likelihood of each genealogy.

These inference methods, however, have not kept pace with the ongoing tremendous rate of improvement of sequencing techniques. The best computer programs for coalescence-based reconstruction of trees may handle about 200 samples; this should be compared with the almost 20000 mtDNA sequences already available on the internet [5]. For the scientists in this field, this growth may be seen as the fulfillment of a utopian dream, paving the way for an in depth study of the processes shaping the form of genetic variation within populations and allowing for reconstruction of the histories of different human communities. Achieving these goals requires coping with the embarrassment of riches this data presents. For this, completely new methods are needed. In this article we present such a technique and demonstrate its efficiency.

The framework presented below is based on a new paradigm and is scalable, as its nu-

merical complexity grows only linearly (as opposed to super-exponentially) with the sample size. It is clear that any attempt to make progress must give up on retrieving the family tree, with its exponentially large amount of (for our purposes, superfluous) details. This has the major advantage of moving away from the regnant combinatoric, maximum likelihood, methodology, and has the promise of transforming the search for the demographic parameters to one of simple curve fitting more common in the physical sciences. To achieve this aim, we focus on the quantity $F(\ell)$, the average number of different haplotypes (hereafter polymorphism) in arbitrary subsequence of length ℓ, which as we will see, is both *robust*, with small statistical variability, and *discriminating*, sensitive to the parameters of interest of the demographic process.

To understand why this particular summary statistic can satisfy the above criteria, let us focus on the total polymorphism F of a sample of mtDNA sequences. These elements of the genome are inherited from mother to daughter, thus they are the equivalent of a (maternally inherited) "surname": an offspring has the same surname/haplotype as its father/mother, unless a chance mutation occurs, the probability of which is denoted μ. In the standard case, n mtDNA sequences are obtained from individuals sampled at random from a total population of N_0. Our statistic is based solely on F, the overall number of different haplotypes in the sample, which is equivalent to asking how many different last names there are in a town.

The total polymorphism F is clearly related to the past demography of the population. It is quite easy to use this number in order to infer qualitative features: for example, in Korea (population $>$ 73 million) there are less than 270 different surnames, which indicates a very low "mutation" (surname change) rate in the past. It is much harder, however, to make quantitative inferences from that number. In particular, the same F may reflect slow mutation in a growing population or, equally well, faster mutation in a fixed size community, as stochastic extinction of small families is less frequent in a growing population.

Thus while the total polymorphism contains useful information, it would appear to be by itself insufficient to reveal all that we wish to know. To overcome this obstacle, we utilize an important difference between mtDNA sequences and surnames: the mutation process in mtDNA sequences (after appropriate alignment) is by random letter substitution. This means that *any subsequence of the genome may serve us as a "surname."* By way of example, for the entire sequences of Fig. 1 there are five haplotypes, whereas if we consider

Individual	Loci 1	2	3	4	5	6	7
α	A	G	C	T	A	G	C
β	A	A	C	T	A	G	C
γ	A	G	C	A	A	G	C
δ	A	G	T	T	A	G	C
ϵ	G	G	T	T	A	G	T

FIG. 1: **Defining Haplotypes:** Illustration of the definition of haplotypes depending on the subsequence considered. If we consider the whole sequence, we have 5 haplotypes. If we define a haplotype based on loci 1–3 only, we have 4 families, as individuals α and γ are identical. If we consider loci (1,4,5,6), we have 3 haplotypes as individual α, β and δ are identical.

just the first three loci there are only four distinct haplotypes. Since all these subsets share the same genealogy, the number of haplotypes of any subsequence of size ℓ should obey the same statistics as the full sequence, where the only difference is the effective mutation rate. Assuming independent point mutations with rate μ_1 per base pair per generation, the mutation rate μ_ℓ of a sequence of size ℓ is given by $\mu_\ell \simeq \mu_1 \ell$. The nonlinear dependence (see below) of F on μ_ℓ, and thereby on ℓ, then allows us to fix the demographic parameters. As our method uses all possible slices of a sequence of the L total base pairs, we indeed make use of all the information hidden in the genetic polymorphism of the population; here this data is compacted into the function $F(\ell)$. It is then relatively straight-forward to extract the desired demographic parameters from this data.

This procedure has the desirable features set out above. The first is that it is a "bulk" quantity, and so the random sample-to-sample variation inherent in the underlying stochastic process is remarkably small. We will see that even subtle fine details of the function $F(\ell)$ are very reproducible. While one might be concerned that the smaller number of mutations present for smaller ℓ renders the results in this regime more unreliable, in fact the very large number of short subsequences compensates. Thus, $F(\ell)$ represents a *robust* family of summary statistics. Our fitting process is also very much similar to that which obtains in the physical sciences, where an experimental curve is fit by a few parameters and where the effects of the various parameters on the fit can by directly visualized.

II. EXTRACTING DEMOGRAPHIC PARAMETERS FROM FAMILY STATISTICS

We consider the problem of haplotype polymorphism in the context of a standard model of a growing population. In the model, every female each generation gives rise to a random number of female offspring, (the males being irrelevant for our concerns) and is then removed from the population. The number of offspring is chosen from a Poisson distribution with mean $\lambda = 1 + \gamma$. Each individual is characterized by a genome consisting of L entries, each consisting of one of the $A = 4$ letters, $\{G, T, A, C\}$. The individual inherits its genome from the mother at birth, subject to random letter substitution with probability μ_1 per entry. Unfortunately, there is no known expression for the haplotype statistics in this model. However, the simpler limiting case of $A \to \infty$ has been solved analytically by us [6], based on earlier work by Manrubia and Zannete [7]. In particular, we present there a formula for the expected polymorphism, F^∞ of a subsample of size n drawn from a population of N_0 individuals, F^∞ being a function of n, N_0, γ, and $\mu = \mu_1 L$. In this limiting case, the expected polymorphism for subsequences of length ℓ is just $F(\ell) = F^\infty(n, N_0, \gamma, \mu_1 \ell)$. For example, for $\gamma > \mu_1 \ell$, which is the most relevant case herein,

$$F(\ell) = \frac{\nu n_c}{2+\nu} \left[{}_2F_1\left(1, 1; 3+\nu; 1\right) - (1-s)\, {}_2F_1\left(1, 1; 3+\nu; 1-s\right)\right]$$
$$\nu \equiv \frac{\mu_1 \ell}{\gamma - \mu_1 \ell} \qquad n_c \equiv \frac{2N_o\gamma}{(1+\nu)} \qquad s \equiv \frac{n}{n_c} \tag{1}$$

where ${}_2F_1$ is a hypergeometric function [8]. This analytic formula is useful since it provides a theoretical framework with which to assimilate the results of the realistic $A = 4$ model. In particular, it implies that even in this simplified limit, the dependence of F on ℓ is nontrivial, initially growing linearly and finally saturating at large ℓ.

Before applying the theoretical model results to the empirical data, it is useful to understand the extent of the difference between the $A = \infty$ analytic theory and the realistic $A = 4$ case. The $A = 4$ results were obtained by averaging over many runs of our simulation program, to be described in Appendix A below. The two cases are compared in Fig. 2, where we have taken $N_0 = 200,000$, $\mu_1 = 6.4 \cdot 10^{-6}$, and $L = 377$, the latter two parameters being the ones relevant to our empirical data. We see that for large subsequence length, the results in the two cases coincide, but there is a significant difference for the shorter subsequences. The origin of the difference lies in the presence of recurrent mutations in the

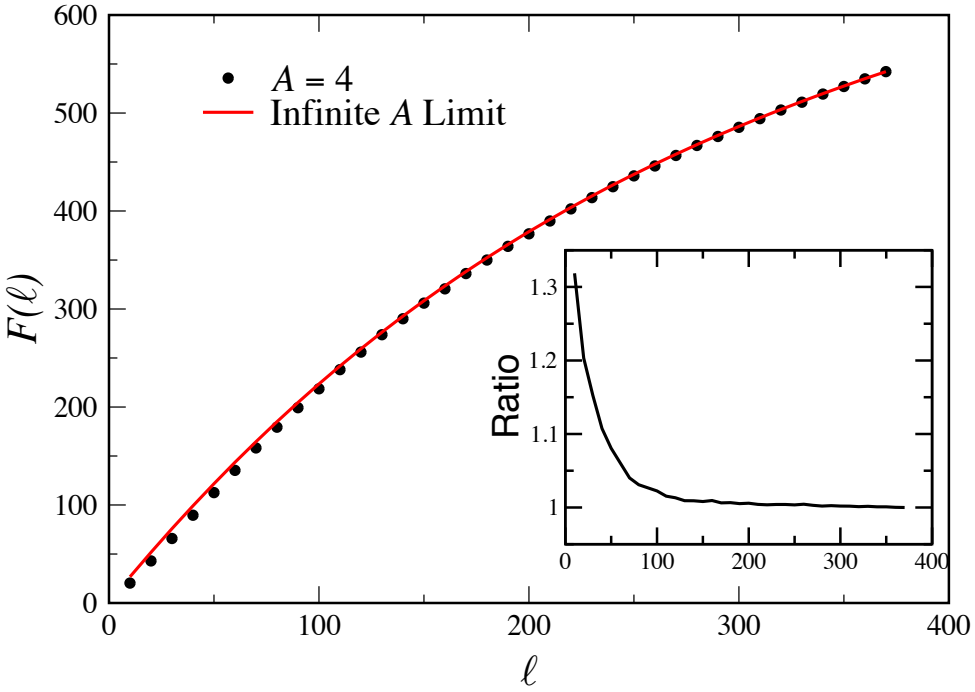

FIG. 2: **Infinite allele vs.** $A = 4$ **model:** The number of haplotypes as a function of the sequence length for the $A = 4$ model, (black dots) as computed by simulation, (averaged over 20 runs), together with the analytical result, Eq. (1) for the $A \to \infty$ limit (solid red line). Both sets of data are for uniform mutation rate, with parameters $N_0 = 200,000$, $\mu_1 = 6.4 \cdot 10^{-6}$, $L = 377$ and $n = 1212$. Inset: The ratio of the $A \to \infty$ results to the $A = 4$ results, showing a 30% effect at small ℓ.

$A = 4$ case. Since the alphabet is limited, independent mutations may lead to the same result, especially for smaller sequences where the chance of two mutations on the same locus are higher, whereas in the $A \to \infty$ limit, all mutations generate never-before-seen sequences. This leads to a systematically smaller number of haplotypes for the $A = 4$ case.

With this in hand, we now turn to our empirical data, consisting of sequences of the $L = 377$ HVR1 control region (noncoding) of mtDNA obtained from 1212 Chinese individuals [5]. As noted above, the current estimated value of the nucleotide mutation rate is $\mu_1 = 6.4 \cdot 10^{-6}$ [9]. This sequence data was then used to construct $F(\ell)$ as described in Appendix A. The results are presented in Fig 3, together with the result of the $A = 4$ model, with the values of the unknown parameters taken to be a net growth rate per generation of $\gamma = 0.0026$ and an (effective, see Appendix B) population size of $N_0 = 200,000$. We see

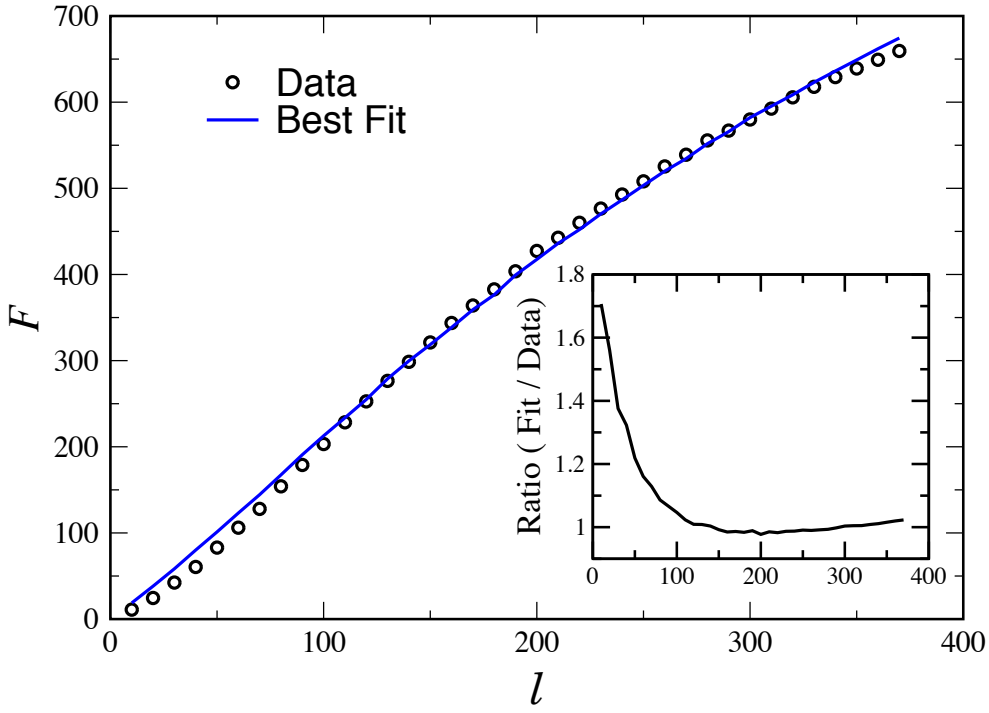

FIG. 3: **Uniform Mutation Model best fit:** The best fit (blue solid line) of the uniform mutation rate, $A = 4$ model to the Chinese empirical data, $F^{obs}(\ell)$, (black circles) consisting of $n = 1212$ samples. The parameters of the best fit are $N_0 = 4.9 \cdot 10^5$ and $\gamma = 0.0055$. The mutation rate is $\mu_1 = 6.4 \cdot 10^{-6}$. Inset: The ratio of the best fit to the data, as a function of ℓ. The failure of the model to capture the convexity of the data at small ℓ is apparent.

that the model indeed succeeds well in capturing the overall structure of the data. However, a closer examination of the small ℓ regime (see inset) indicates a systematic deviation. The model results are concave downwards, whereas the data exhibits a convexity at small ℓ. This difference is not a function of the fitting parameters; the concave nature of the model prediction persists for all choices of the fitting parameters.

There is, amazingly enough, real biological content in this subtle discrepancy between the model and the data in the small ℓ regime. It points to the presence of "hot" spots, regions of anomalously high mutation rates, along the genome. This known variability of the mutation rate, ultimately traceable to the different strengths of the hydrogen bonds of the various base pairs, has been found experimentally to be well described by a Gamma distribution, with an α parameter variously estimated as $\alpha = 0.44 - 0.6$ [10] and $\alpha = 0.28 - 0.39$ [11]. Incorporating this variability into the model changes the nature of the $F(\ell)$ curve at small

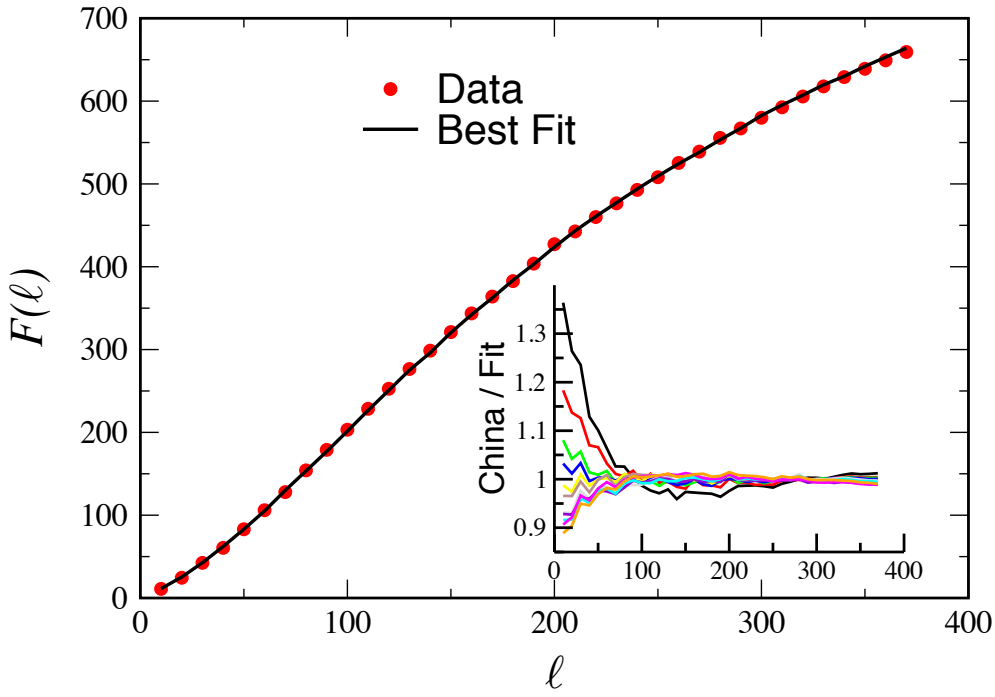

FIG. 4: **The best fit of the Chinese data:** In the main panel the number of haplotypes as a function of the sequence length obtained from the Chinese data-set, F^{obs}, is presented, along with its best fit to the simulation based function, F^{model}. The correspondence is almost perfect. The parameters of the best fit are $N_0 = 230000$, $\gamma = 0.0026$, and $\alpha = 0.5$. In the subplot, we present the ratio between the empirical $F(\ell)$ and the best fit (from an independent fit for N_0 and γ for each given α) for different variation coefficients α. If α is too small the expected number of families is smaller than the real one, and for large α it is larger. This is because small coefficients produce too many recurrence events thus reducing the number of small families, while too large coefficient do not produce enough recurrence events, and therefore overestimate the number of different haplotypes at small ℓ.

ℓ. While recurrent mutations (in the guise of finite A) only changes the small ℓ slope, but leaves the curve concave downward, the variability in the mutation rate introduces convexity into the curve, resulting in an S-shape, exactly in line with the Chinese empirical data. The exact degree of convexity then gives us a handle to measure the α parameter governing the mutation rate variability.

With our extended model, it is now straightforward in principle to fit the three parameters γ, N_0 and α. The details of this procedure are discussed in Appendix A. In Fig. 4, we

present the best fit, namely $\gamma = 0.0026$, $N_0 = 2.3 \cdot 10^5$ and $\alpha = 0.5$, together with the empirical data. We see that the fit is excellent, capturing correctly the detailed structure at small ℓ. The inset shows a more stringent test of the fit, plotting the ratio of the data to the fitting curve, for a variety of α's, with the other parameters taken to give the optimal fit for that α. We see that for the best $\alpha = 0.5$, given by the blue curve, the theory agrees with the data to within 3% for the whole range of ℓ. We note that the fit values are consistent with those obtained by other methods, [12, 13].

It is also possible to obtain confidence intervals for our fit parameters, using a parametric bootstrap technique. We have generated 5000 runs of simulated data for 1212 individuals with the parameters $\gamma = 0.0026$, $N_0 = 2 \cdot 10^5$ and $\alpha = 0.5$. For each, we applied our fitting procedure, obtaining the three fitting parameters. The distribution of fitting parameters is roughly Gaussian around their nominal values, and we were thus able to obtain the following 95% confidence intervals: $N_0 = 230,000[180,000 - 300,000]$, $\gamma = 0.0026[0.0019 - 0.0033]$ and $\alpha = 0.5[0.3 - 0.9]$. The confidence intervals on γ are a factor of two smaller than that given by the standard tree building method, BEAST (see Appendix C). In addition, the estimate given by BEAST is significantly biased, being a factor of three too small, when run on simulated data.

As a final demonstration of the power of our technique and its broader applicability, let us return to the simple infinite allele ($A \to \infty$) model. The main result of [6], from which Eq. (1) has been derived, is the following analytic formula for n_m, the number of haplotypes with m individuals:

$$n_m(N_0, \gamma, \mu) = \frac{\nu n_c \Gamma(2+\nu)}{m} U(1+\nu, 0, \frac{n_c m}{N_0}) \,, \tag{2}$$

where ν and n_c have been defined above and U is the Kummer function [8]. This result is not applicable for the HVR1 control region: as we have seen, the effect of recurrent mutation is significant, since the sequence length is too small. Using the whole human mtDNA loop with its $L = 16400$ base pairs (after alignment with ClustalW [14]) we can obtain much better statistics, with essentially no recurrent mutations, for a dataset that shares the same demographic history with the control region discussed so far. The price to be paid is that now we deal with the coding part of the DNA, so the assumption of neutrality no longer holds. Are the deviations from neutrality strong enough to destroy the predicted haplotype statistics?

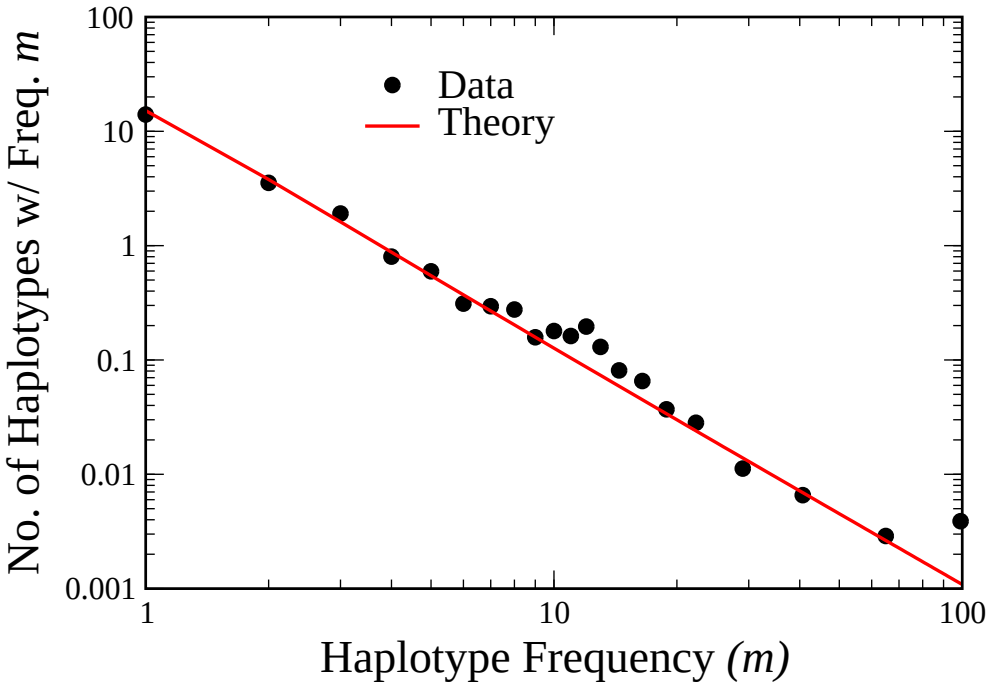

FIG. 5: **Haplotypes frequencies statistics:** n_m, the number of haplotypes that occur m times in the data is plotted against m on a log-log scale. From the alligned mtDNA sequences of the Chinese population, $L = 16419$ base pairs, a set of 200 base pairs was selected at random. The number of individuals that have the very same haplotype (with respect to these 200 bp's) was determined, and the statistics presented here reflects 1000 iterations of this procedure. The red line is the theoretical prediction, Eq. 2, with the parameters $N_0 = 230000$ and $\gamma = 0.0026$ obtained from the best fit to the HVR1 data, and $\mu_1 = 6 \cdot 10^{-7}$, which is the weighted average between the values for the control and the coding regions. About 75% of the mutations appear in the coding region.

In Fig. 5 the answer is depicted for the haplotype frequencies of the Chinese population, together with the line predicted by Eq. (2). Note that there are *no fitting parameters*: N_0 and γ are those extracted above from the control region, while the value of $\mu_1 = 6 \cdot 10^{-7}$ taken was that calculated for the average of the noncoding and coding regions [15]. The correspondence is very good, suggesting that the most of the observed single nucleotide mutations are indeed neutral or almost neutral; what seems to be significant deviations close to $m = 10$ and at very large m may indicate the effect of selection.

III. DISCUSSION

Inference techniques in the biological sciences are traditionally based on some sort of a maximum likelihood procedure, either with respect to tree-building [4] or to some chosen set of summary statistics [16–22]. This involves calculating the likelihood of observing the actual dataset as a function of the parameter set. The "right" set of parameters is identified as the one that yields the maximal likelihood to the data. The methodology presented herein obviates the need to calculate probability distributions and likelihoods and (in case of Bayesian estimation) the arbitrary choice of an *a priori* distribution of the parameters. Instead one simply constructs the function $F(\ell)$ from the empirical data and compares with the model prediction, using a standard χ^2 minimization to fix the parameters.

In point of fact, the strategy implemented here has its roots in one of the most famous summary statistics, the celebrated Ewens' sampling formula [23]. For a fixed population of effective size N_0 and mutation rate μ, Ewens' formula gives the allele frequency statistics in a sampled population, within the infinite allele ($A \to \infty$) model. If n individuals are sampled out of a total population of N_0, where $n \ll N_0$, the formula provides one with $P(m_1, m_2, m_3...|N_0)$, the probability that m_1 haplotypes occur exactly once in the sample, m_2 haplotypes are represented twice and so on. Our analytic results for the $A = \infty$ case represent a reduction of this formula to the *expected* total number of haplotypes given by Eq. (1), generalized to the case of a *growing* population. A formal solution, in terms of recursion relations, for a sampling formula for a growing population was presented by Griffiths and Tavaré [35]. This work have triggered a number of applications [24, 25] and further extensions in order to deal with subdivided populations [26–28], recombination [29, 30] and selection [31] that have to date not proven to be superior to the tree-building algorithms. We have seen in this paper that the further reduction to the average polymorphism does yield an improved method, both from a theoretical and practical point of view.

From the theoretical side, a decisive advantage in our view is that the parameter fitting is performed on a quantity that is directly measurable. This is in sharp contrast to the standard method, where the demographic parameters are inferred from the most probable tree, which is not directly accessible. Thus the quality of fit is immediately apparent, and systematic deviations (like those connected to α) are easily detected.

The practical merits of our approach are clear: first, it can handle very large datasets,

something that techniques that are based on the Kingman coalescent model cannot do, both from the practical point of view and because of the theoretical limitation of a relatively small sample with respect to the population size. Second, our technique produces an effectively unbiased estimator for the current population size and growth rate, in contradistinction to the common methods that produce a strongly biased estimator for the growth rate [32, 33]. Moreover, in spite of its simplicity, our method can infer at once demographic parameters as well as parameters that characterize the mutational process.

Our method may be applicable to other areas of biology, where exponentially complicated searches are used, from taxonomy to proteomics. Many archetypical problems, such as the identification of fitness associated with certain mutations, require the use of large datasets and depend strongly on the sensitivity of the results to the model assumptions. The scaleable method presented herein, being both robust and discriminating, is a promising platform for attacking these problems.

APPENDIX A: DETAILS OF THE NUMERICAL PROCEDURE

As noted above, our theoretical model at finite A (with or without mutation variability) is not analytically solvable, and so we have to employ simulations to compute the predictions of the model. We first present details of our simulation. The simulation, following the model, assumes nonoverlapping generations (i.e., a Wright-Fisher process). We index time backward, such that N_t denotes the size of the population t generations ago and N_0 is the current population. For convenience, we ignore the statistical fluctuations in the overall population and assume here an exponentially growing population with a fixed rate. Thus with our definition of t, the population is given by $N(t) = [N_0 e^{-\gamma t}]$, as in [32, 34], where the brackets indicate the integer part. Our numerics shows no difference between deterministic and stochastic dynamics of the total population. Nor is there a difference between our nonoverlapping generations model and an overlapping generations (i.e., Moran) version. Genealogical trees were generated for a given sample of size n out of N_0 individuals using a "balls in a box" algorithm [36]. The ancestral population of the sampled individuals $n \leq N_0$ is determined as follows: in every generation, each individual "chooses" its parent with equal probability from the N_{t-1} possible parents in the former generation. Two lineages that share the same ancestor coalesce and the process is iterated, until there is only one lineage left —

the most recent common ancestor of the sampled population.

After generating the genealogical history the gene-genealogy is produced. Every ancestral individual of the n sampled individuals is endowed with a genome of L base pairs. The number of mutations on each branch of the tree is an integer taken from a binomial distribution with average $\mu_1 L t_b$, where t_b is the length of the branch in generations. To allow for different mutation rates for different sites (in addition to the transition-transversion ratio $\kappa = 20$ [37]), each base pair admits its own mutation rate taken from a Gamma distribution with parameter α. The chance of a mutation to be attributed to a specific site is proportional to its local mutation rate.

This numerical procedure can be used to generate a set of simulated data to be compared with the analytic theory, corresponding to $A \to \infty$. This allows one to crosscheck the simulation procedure and the theory. The results of the simulation are indeed consistent with the analytic solution in this limit, (see Appendix B for details) while for $A = 4$ give a lower initial slope for F_ℓ, as detailed in Fig. 2.

The set of observables $F(\ell)$ is extracted from the raw observed data as follows. A subset of ℓ loci was chosen at random from the whole genetic sequence. The number of different haplotypes (unique sequences) for this particular subset was tallied (see Fig. 1). This procedure was repeated k times in order to obtain a better estimate for the mean number of haplotypes for this specific ℓ. Optimally, the smaller ℓ, the larger the k that should be employed in order to minimize the statistical noise. In practice, we used the relatively large value $k = 1000$ for all ℓ, unless otherwise specified. This procedure was repeated for each ℓ, thus generating the observed $F^{obs}(\ell)$.

To retrieve the demographic parameters from the observed $F^{obs}(\ell)$, a least-squares minimization procedure was implemented, defining the χ^2 function by:

$$\chi^2 \equiv \sum_{\ell} \left[F^{obs}(\ell) - F^{model}(\ell; N_0, \gamma, \alpha) \right]^2 \ ; \tag{A1}$$

the estimate for the parameters is based on minimizing this χ^2 (see Appendix B).

Since an analytic expression for $F(\ell)$ is not available, the fitting is performed by constructing the function $F^{model}(\ell)$ through simulations. $F^{model}(\ell; N_0, \gamma, \alpha)$ is defined as the average over many realizations of F obtained from simulation with this given set of parameters. We find that 30 realizations is sufficient to accurately calculate F^{model}, thus lending support to our claim that F is a robust statistic.

APPENDIX B: TESTS OF OUR SIMULATION AND FITTING PROCEDURE

In this appendix we describe a number of checks we performed on our simulation and fitting procedure. Our purpose is to verify the robustness of our statistic $F(\ell)$ with respect to the stochastic noise inherent in the birth-death-mutation process. We also test how well our fitting procedure can recover the parameters in a case where there are known *a priori*, i.e., when they are generated by a simulation. We also describe in more detail the fitting procedure itself.

1. Infinite Allele Model ($A \to \infty$)

The easiest tests to perform are in the context of the infinite allele model, $A \to \infty$, where we have analytic results to compare against. For completeness, we first record the analytic result for $F(\ell)$ for the regime $\mu_1 \ell > \gamma$:

$$F(\ell) = n_c s \, {}_2F_1(1, 1; 2 + \nu; -s) \tag{B1}$$

where now

$$\nu \equiv \frac{\gamma}{\mu_1 \ell - \gamma}\,; \qquad n_c \equiv 2N_0(\mu_1 \ell - \gamma)\,; \qquad s \equiv \frac{n}{n_c} \tag{B2}$$

which supplements Eq. 1 above. We ran a *single* simulation with $N_0 = 5 \cdot 10^6$, net growth rate $\gamma = 0.0055$, genome length $L = 500$, uniform mutation rate $\mu_1 = 4.6 \cdot 10^{-6}$, and sample size $n = 2000$. From this, we constructed our $F(\ell)$ according to the procedure outlined in the text, here using $k = 1000$ different subsets of ℓ loci for each ℓ. The result is presented in Fig. 6, together with the theoretical formula. We see that the agreement is essentially perfect, even though the theory is for the average result of an infinite number of such runs. This supports the claim that the sample-to-sample variations are indeed small. The smallness of these random variations can also be seen explicitly from the error bars, shown in Figs. 7 and 8.

The next task is to attempt to recover the demographic parameters N_0 and γ by fitting the simulation data to the theory, In order for this to succeed, it is necessary that $F(\ell)$ will be sensitive to changes in the parameters. We can get an idea of the sensitivity by showing F with N_0 raised and lowered by a factor of two with γ fixed, and similarly changing γ for fixed N_0. These are presented in Figs. 7 and 8. We see that indeed factor of two changes in

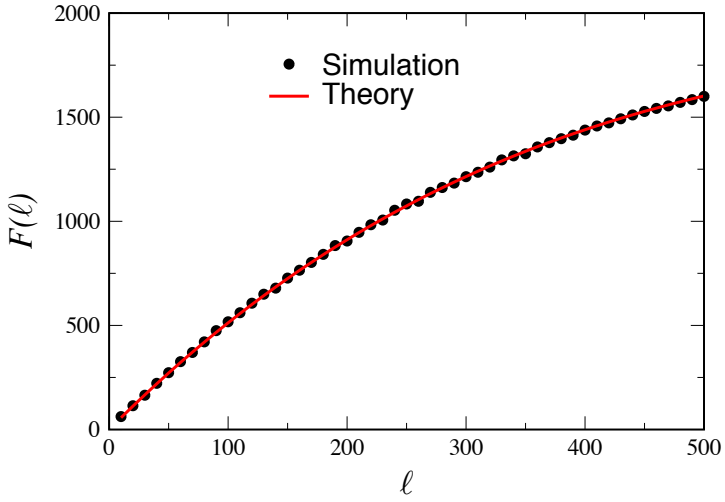

FIG. 6: F^{obs}, as extracted from a numerical simulation of the $A = \infty$ simulation (black circles), and the prediction based on Eq. 1 (red line). The parameters used in the simulation were $N_0 = 4 \cdot 10^6$, $\gamma = 0.0055$, $L = 500$, $n = 2000$, and $\mu_1 = 4.6 \cdot 10^{-6}$.

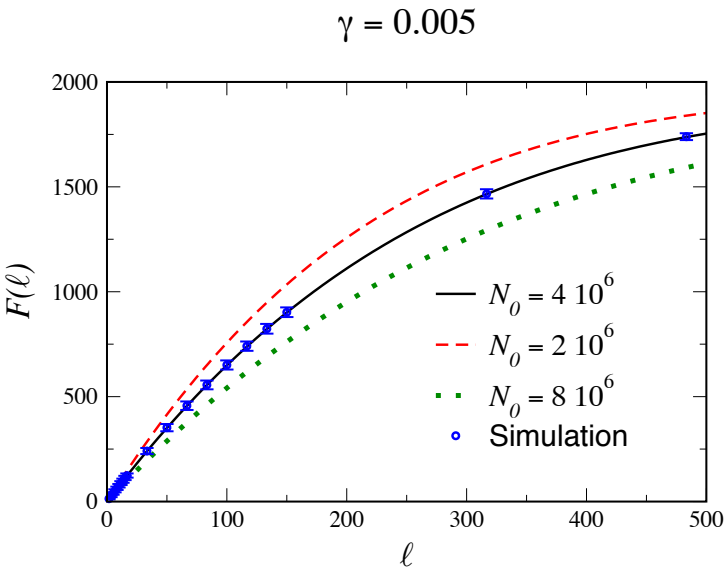

FIG. 7: Number of haplotypes $F(\ell)$ for 3 different values of N_0 with $\gamma = 0.005$, $\mu_1 = 6 \cdot 10^{-6}$, $L = 500$ from the analytic $A \to \infty$ limit. The errors bars denote the standard deviation.

either parameter have a large impact on F. The impact of γ is significantly larger, however, which will be reflected in the tighter error bounds on our estimate of this parameter.

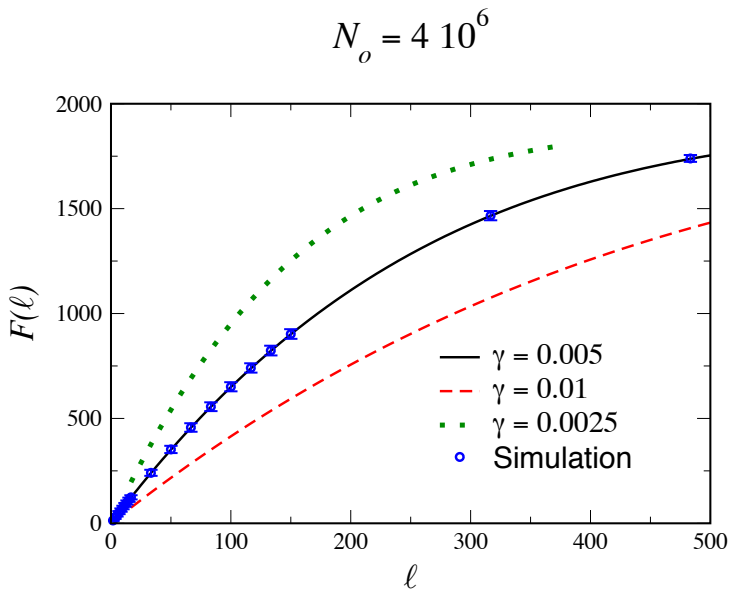

FIG. 8: Number of haplotypes F as a function of ℓ for three different values of γ with $N_0 = 4 \cdot 10^6$, $\mu_1 = 6 \cdot 10^{-6}$, $L = 500$ from the analytic $A \to \infty$ limit. The errors bars denote the standard deviation.

As described above, we employed a χ^2-minimization technique to fix the parameters. To find the minimum of χ^2 as a function of the two parameters N_0 and γ, we performed a coarse scan in this two dimensional space, producing a "landscape", whose contours are presented in Fig. 9. We see that there is a clear "valley" where χ^2 is relatively small, and in which the minimum is located. Another feature that is apparent is that the valley is roughly aligned parallel to the N_0 axis, in accord with the lower sensitivity of F to changes in N_0 we saw above. Having identified the valley, we then performed a higher resolution scan in this region, Fig. 10, locating the minimum at $N_0 = 3.9 \cdot 10^6$, $\gamma = 0.00498$. This should be compared with the true values of the run, $N_0 = 4 \cdot 10^6$ and $\gamma = 0.005$.

2. $A = 4$, uniform mutation rate

The second round of tests was performed on the $A = 4$ uniform mutation rate model. As we saw in the main text, the effect of finite A is felt most strongly at low ℓ. From a practical point of view, the major difference is that analytic predictions of the model for the mean value of F averaged over many simulations is no longer available. Rather, the model

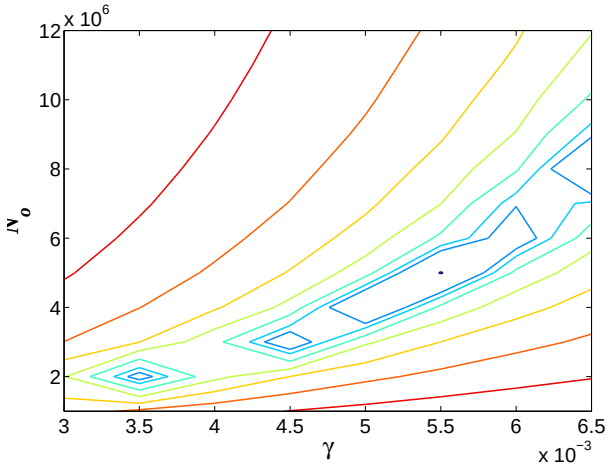

FIG. 9: A low resolution contour plot of $\ln \chi^2$ in the γ–N_0 plane. This plot measure the difference between the theoretical $F(\ell)$ for the $A = \infty$ limit, and a single simulation run of the $A = \infty$ simulation. The minimum is at $\gamma = 0.0055$, $N_0 = 5 \cdot 10^6$. The other parameters are $L = 500$, $n = 2000$, and $\mu = 0.0023$.

predictions for the mean have to be measured from actually averaging over some number of simulations. The practical issue is how many simulations are needed in order to obtain an accurate value for the average F. We have found that the low sample to sample variation obviates the need for too many different runs, with 30 sufficing. Again the χ^2 minimization is done by scanning in the parameter space, so that we have to perform this averaging for many different parameter sets. In Fig. 11 the results are presented. One specific run of the simulation provides the "data" and the fit has been performed as described above. Again the procedure yields a reliable estimation of the demographic parameters.

We used the parametric bootstrap method in order to obtain the confidence intervals for the estimation, running 20 different independent realizations of our numerical experiment using the same set of parameters used to generate the "data" above. We then examined the dispersion of the results of the fitting procedure performed on each data set. The average of the estimates from these 20 realizations were $N_0 = 1.38 \cdot 10^5$, $\gamma = 0.00136$, and their 95% confidence interval (i.e. twice the standard deviation) was $\sigma(N_0) = 25,000$ and $\sigma(\gamma) = 0.00037$. We thus see that the estimates lie within the error bars of the true values, $N_0 = 1.4 \cdot 10^5$, and $\gamma = 0.0014$.

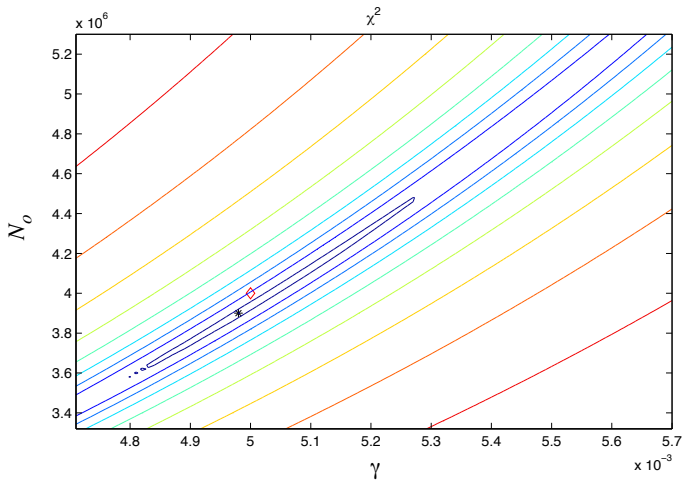

FIG. 10: A high resolution contour plot of $\ln \chi^2$ in the γ–N_0 plane. This plot measure the difference between the theoretical $F(\ell)$ for the $A = \infty$ limit, and a single simulation run of the $A = \infty$ simulation. The minimum, marked by the black dot is at $\gamma = 0.00498$, $N_0 = 3.9 \cdot 10^6$. The location of the nominal parameters, $\gamma = 0.005$, $N_0 = 4 \cdot 10^6$ is marked by a diamond. The other parameters are $L = 500$, $n = 2000$, and $\mu = 0.0023$.

APPENDIX C: FIT TO EMPIRICAL DATA

1. Effective Population

It is important to understand that the fit parameter N_0 is obviously much smaller than the current population of China. Population geneticists have introduced the concept of "effective population" to bridge the gap between the idealized models and the true situation. The effective population corrects for such effects as 1) only the female population that has not passed the reproduction years should be considered; 2) the distribution of offspring is not Poisson; and most importantly, 3) the population is not well-mixed, but exhibits a significant amount of inbreeding.

2. Error bounds for the fit to the empirical data

In the main text we gave only the 95% confidence intervals for each of the estimated parameters, obtained using the bootstrap technique as explained above. In Fig. 12 we present the full distributions of the estimated parameters. These are not far from Gaussian

and therefore one may get a reasonable estimation of the 95% confidence interval by using twice the standard deviation. A more accurate way is to find the median, 2.5% and 97.5% of each of the parameters, which are

Parameter	median	2.5%	97.5%
N_0	230000	180000	300000
γ	0.0026	0.0019	0.0033
α	0.5	0.3	0.9

3. Comparison to BEAST

In order to compare our results to more conventional methods, we used the BEAST program [38] under the assumption of a model of an exponentially growing population. The maximal efficient dataset BEAST can handle is about 200 [39]. Therefore we took 10 simulated datasets of a growing population with the same parameters we obtained for China, $N_0 = 230,000$, $\gamma = 0.0026$ and $\mu_L = 0.0024$, $L = 377$. In the notation of BEAST,

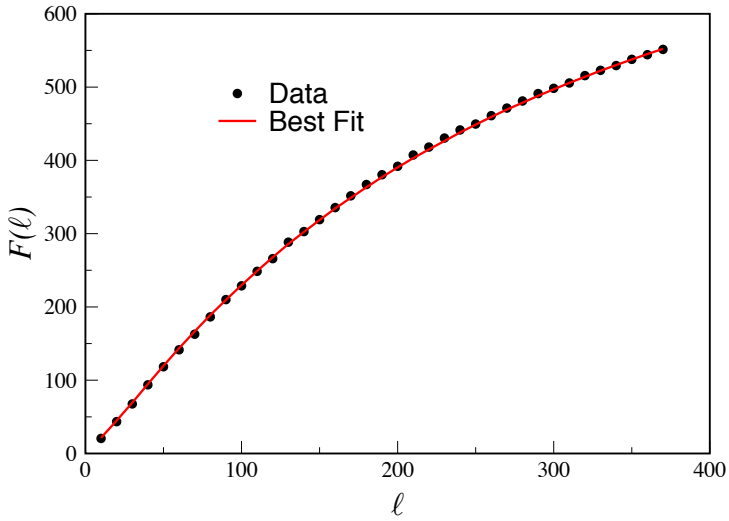

FIG. 11: The number of haplotypes in a four allele model as a function of the sequence length is presented together with its best fit to our simulation-based function. The parameters used to generate the "data" are $N_0 = 140,000$ and $\gamma = 0.0014$, where for each ℓ 50 sets of randomly picked loci were used for averaging. The parameter estimates from the fit are $N_0 = 130,000$ and $\gamma = 0.0012$. Other parameters are $\mu = 0.0023$, $L = 377$, and $n = 1000$.

the average estimations of the growth rate and current population size were $\Theta = 1.85 \pm 0.8$ (average $\pm$ 2 STD) and growth rate $g = 114.5 \pm 62$. This means an error range of 48% in the current population size and 54% in the growth rate, which is about twice the error range obtained in our method.

The way to transform the parameters to N_0 and γ are $N_0 = \Theta/\mu$ and $\gamma = g/\tau$ where τ is average time for mutation on a given site: $\tau = 1/\mu_1$. We thus get $N_0 = 2.8 \cdot 10^5 \pm 1.4 \cdot 10^5$ and $\gamma = 7.3 \cdot 10^{-4} \pm 4 \cdot 10^{-4}$. One can see that there is some bias in the estimate for the current population size. However in the growth rate there is a much larger bias, with BEAST giving on average a result roughly a factor of 3 too small.

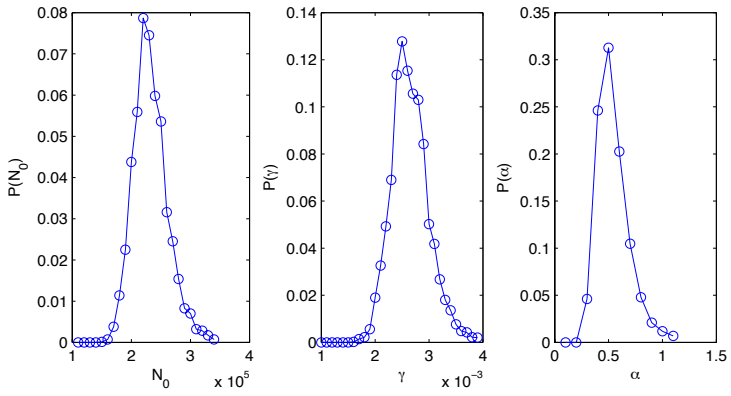

FIG. 12: The distribution of the estimated parameters is shown. The distribution was obtained by doing the fitting procedure for 5000 different gene genealogies, where all of them were generated with the same demographic parameters as the best fit of China: $N_0 = 230,000$, $\gamma = 0.0026$ and $\alpha = 0.5$. One can see that the distributions are sharply peaked close to their nominal values.

ACKNOWLEDGMENTS

This work was supported by the EU 6th framework CO3 pathfinder. NMS and YM acknowledge many useful discussions with John Wakeley on scalable approaches to population

genetics.

---

[1] Gillespie JH, (1998) *Population genetics: A concise guide*, (Johns Hopkins University Press, Baltimore).

[2] Stephens, M. (2001) Inferences under the coalescent. *Handbook of Statistical Genetics*, eds Balding DJ, Bishop MJ, Cannings C. (John Wiley & Sons, Chichester, England)m pp. 213–238.

[3] Tavaré, S. (2004) Ancestral inference in population genetics, in *L*ectures in Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXXI — 2001 (J. Picard, ed.), Lecture Notes in Mathematics, vol. 837, (Springer-Verlag, Berlin, pp. 1–188.

[4] Felsenstein J (2007) Trees of genes in populations. *Reconstructing Evolution: New Mathematical and Computational Advances*, eds Gascuel O, Steel M (Oxford University Press, Oxford), pp. 3–29.

[5] Kohl J, Paulsen I, Laubach T, Radtke A, von Haessler A (2006) HvrBase++: a phylogenetic database for primate species. *Nucleic Acids Res* 34:D700–D704.

[6] Maruvka YM, Shnerb NM, Kessler DA (2009) Universal features of surname distribution in a subsample of a growing population. *J Theor Biol* 262:245–256.

[7] Manrubia S, Zanette, DH (2002) At the boundary between biological and cultural evolution: the origin of surname distributions. *J Theor Biol* 216:461–477.

[8] Abramowitz M, Stegun I (1972) *Handbook of Mathematical Functions*. (Government Printing Office, Washington).

[9] Sigurdardo S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. *Am J Hum Genet* 66:1599–1609.

[10] Wakeley J (1993) Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. *J Mol Evol* 37:613–623.

[11] Excoffier L, Yang Z (1999) Substitution rate variation among sites in the mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* 16:1357–1368.

[12] Maruvka YE, Shnerb NM, Bar-Yam Y, Wakeley J (2009) Recovering Population Parameters

from a Single Gene Genealogy: An Unbiased Estimator of the Growth Rate. submitted.

[13] Atkinson QD, Gray RD, Drummond A (2007) mtDNA variation predicts population size in humans and reveals a major southern Asian chapter in human prehistory. *Mol Biol Evol* 25:468–474.

[14] Larkin MA, et al. (2007) ClustalW and ClustalX version 2. *Bioinformatics* 23: 2947–2948.

[15] Ho SYW, Endicott P (2008) The crucial role of calibration in molecular date estimates for the peopling of the Americas. *Am J Hum Genet* 83:142–146.

[16] Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145:505–518.

[17] Fu YX, Li WH (1997) Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol* 14:195–199.

[18] Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics* 149:1539–1546.

[19] Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 116:1791–1798.

[20] Beaumont MA, Zhang W, Balding DJ (2002) Approximate bayesian computation in population genetics. *Genetics* 162:2025–2035.

[21] Leman SC, Chen Y, Stajich JE, Noor MA, Uyenoyama MK (2005) Likelihoods from summary statistics: Recent divergence between species. *Genetics* 171:1419–1436.

[22] Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Res* 17:1505–1519.

[23] Ewens, WJ (1972) The sampling theory of selectively neutral alleles. *Theor Pop Biol* 3:87–112.

[24] Lohmueller KE, Bustamante CD, Clark AG (2009) Methods for human demographic inference using haplotype patterns from genomewide singlenucleotide polymorphism data. *Genetics* 182:217–231.

[25] Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD, (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet* 5:e1000695.

[26] Bahlo M and Griffiths RC (2000) Inference from gene trees in a subdivided population. *Theor Popul.Biol* 57:79–95.

[27] De Iorio M and Griffiths RC (2004) Importance sampling on coalescent histories, II. Subdivided population models. *Adv Appl Prob* 36:434–454.

[28] De Iorio M, Griffiths RC, Lebois R and Rousset F (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor Popul Biol* 68:41–53.

[29] Griffiths RC and Majoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* 3:479–502.

[30] Fearnhead P and Donnelly P (2001)Estimating recombination rates from population genetic data. *Genetics* 159:1299–1318.

[31] Coop G, Griffiths RC (2004) Ancestral inference on gene trees under selection. *Theor Popul Biol* 66:219–232.

[32] Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429–434.

[33] Kuhner MK, Smith LP (2007) Comparing likelihood and bayesian coalescent estimation of population parameters. *Genetics* 175:155–165.

[34] Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129:555–562.

[35] Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. *Phil Trans R Soc Lond B* 344:403–410.

[36] Wakeley J (2008) *Coalescent Theory: An Introduction.* (Roberts & Company Publishers, Greenwood Village, Colorado).

[37] Rosset S, et al. (2008) Maximum likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics* 180:1511–1524.

[38] Drummond AJ, Rambaut A (2007) "BEAST: Bayesian evolutionary analysis by sampling trees." *BMC Evolutionary Biology* 7:214.

[39] Drummond AJ, private communication.