Maximum Likelihood Estimation of Site-Specific Mutation Probabilities from Partial Phylogenetic Classification

Saharon Rosset,[*]† *R. Spencer Wells,*‡ *Ajay Royyuru,*†, *Doron M. Behar*‖ *and The Genographic Consortium*

[*] Dept. of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel; † IBM T.J. Watson Research Center, New York, USA; ‡ Missions Program, National Geographic Society, Washington DC, USA; ‖ Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa, Israel

Good estimation of mutation probabilities in hyper-variable regions of the genome (like the control region of the mitochondrial DNA) is important for many tasks in population genetics, forensic analysis and more. In this paper we develop and investigate a maximum-likelihood methodology for estimating individual (site-dependent) mutation probabilities from haplogroup-level data, that is, from samples that are placed on a phylogenetic tree into clusters representing terminal sub-trees of the full tree. Such data is available to us in abundance within the Genographic project public participation database, and we use it to estimate mutation rates for the mitochondrial DNA hyper-variable segment I. We develop an inference approach based on a combination of bootstrap and simulation, which demonstrates some interesting properties of our estimates. We discuss the potential uses of our estimate for investigating functionality in the mtDNA and for improving mtDNA classification.

Key words: Mutation probability estimation, mtDNA control region, Haplogroup classification, Generalized linear models
E-mail: saharon@tau.ac.il

© The Author 2007. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

1. Introduction

It has long been known that different regions in the genome mutate at vastly different rates (Tamura and Nei, 1993). In particular, for the mitochondrial DNA (mtDNA) two hyper-variable segments (HVS) have been identified and named HVS-I and HVS-II. Even within these segments, the mutation rates of the various sites are not fixed. Tamura and Nei (1993) show that there is strong statistical support for a Gamma "prior" distribution of mutation rates across the mtDNA control region (which contains both HVS-I and HVS-II), with a shape parameter $\alpha = 0.1$, implying many orders of magnitude difference in rates between the fastest and slowest mutating sites in these segments. Yang (1993) described methodologies for integrating this Gamma prior into maximum-likelihood phylogeny estimation.

Several authors have developed approaches to estimating site-specific in the mtDNA HVS-I. For a survey of previous approaches, see Bandelt et al. (2006). As an example, we consider here the approaches of Bandelt et al. (2006) themselves and of two other efforts, by Excoffier and Yang (1999) and Meyer and von Haeseler (2003). Both of these latter approaches are approximate *maximum likelihood* methods, attempting to reconstruct the full distribution over possible tree topologies and estimate parameters simultaneously. Because of the extreme difficulty of this task, especially assuming rate variation, even for moderately sized datasets (up to several hundred samples), as used in these two papers, they develop different approximation approaches. Excoffier and Yang (1999) generate a limited set of parsimonious candidate trees, and investigate the robustness of their estimates to their choice of topology from this set. Meyer and von Haeseler (2003), on the other hand, alternate between estimating phylogeny and mutation rates (where the phylogeny estimation step assumes known, but potentially variable, mutation rates). Bandelt et al. (2006) discuss these approaches and explore their limitations and shortcomings, which they consider to be critical. They therefore conclude that the best approach for mutation probability estimation is to construct a *best* tree (in their case, using parsimony considerations), and estimate the mutation probabilities by simple counting on this tree. They apply their methodology to about 800 samples.

Our approach fundamentally differs from these approaches and previous ones, in avoiding the construction of detailed phylogenetic trees. Instead we rely on partial, highly reliable phylogenetic information, in our case in the form of haplogroup (Hg) associations of the mtDNA samples we use. We develop a formal maximum likelihood inference approach, that *integrates out* the intra-Hg phylogenetic uncertainty. We show that maximum likelihood parameter estimation in our model is a binomial regression with complementary-log-log link function (a Generalized Linear Model) for estimating the site-specific mutation rates and the size parameters for each Hg-specific phylogenetic tree. The main advantage of our approach is that it allows us to practically apply our method to large datasets, and eliminate the difficulties resulting from uncertainty about the correct phylogeny. In our case, we apply it to a dataset of 16609 samples, collected in the Genographic project (Behar et al., 2007), and classified into Hgs relying mostly on coding region information.

2. Materials and methods

2.1 *Statistical method*

Assume we observe a large number of sequences of a non-recombining DNA region. These samples are all located on a phylogenetic tree relating all of them. We are not given their detailed phylogenetic relationship, but rather a *haplogroup* view of that relationship. That is, the samples are divided into groups that belong to the same Hg, where each Hg can be thought of as a terminal subtree of the full phylogenetic tree, whose internal structure is not known. This situation is illustrated in Figure 1.

We assume:

1. That the haplogroup classification of all sequences is known and accurate.

2. That the sequences are of the same length and differ only through single nucleotide polymorphisms (SNPs). We thus ignore insertions and deletions in our analysis. This is not a critical feature of our methodology, but since almost all insertions and deletions are unique events and not prone to homoplasy or back mutations, we leave them out of our analysis.

3. That the SNPs in each site of the considered DNA region are independent.

4. That there is a global molecular clock, i.e., that for every site considered, the rate of mutations per time unit is the same in every part of the phylogenetic tree.

5. That every site has a fixed Poisson rate with which the mutations occur. This assumption is exactly correct if we assume an appropriately simple substitution model, in particular one where the set of mutation probabilities is

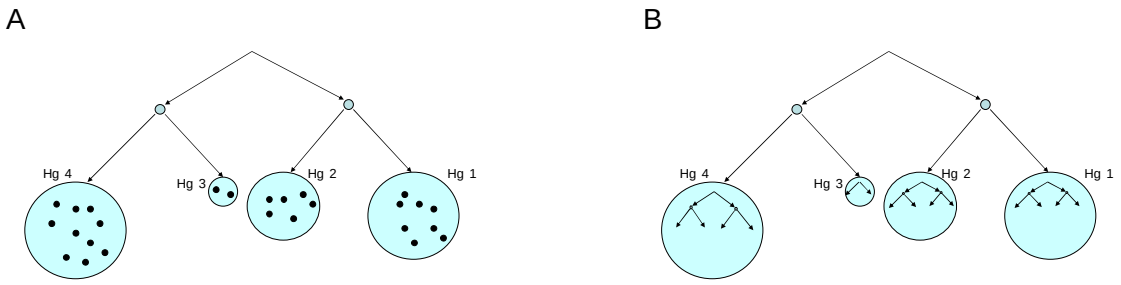

FIG. 1.—(A) Schematic of the Hg view of a phylogenetic tree and (B) the full phylogenetic tree, including the internal Hg phylogenies, which we assume we don't observe

independent of the current nucleotide (and consequently all four nucleotides are equally likely to appear). This is true of the three simplest substitution models commonly used:

- The Jukes-Cantor model (JC69, Jukes and Cantor (1969)), which assumes that all substitutions are equally likely.
- The Kimura 2-parameter model (K80, Kimura (1980)), which allows for different probability of transitions and transversions.
- The Kimura 3-parameter model (K3ST, Kimura (1981)), which allows for different probability of transitions and two different types of transversions. For a site with overall mutation rate λ per time unit, the instantaneous transition matrix for K3ST is:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \cdot & b_2\lambda & a\lambda & b_1\lambda \\ b_2\lambda & \cdot & b_1\lambda & a\lambda \\ a\lambda & b_1\lambda & \cdot & b_2\lambda \\ b_1\lambda & a\lambda & b_2\lambda & \cdot \end{pmatrix} \end{array} \quad (1)$$

with $a + b_1 + b_2 = 1$.

Assumptions 1–2 are critical for our analysis and cannot be validated. Assumption 3 can be relaxed as long as the clock changes uniformly for all sites in HVS-I. The methodology we develop will allow us to do hypothesis testing to examine the validity of assumption 4. Assumption 5 is important to make our model formally correct, but slight violations of it (e.g., in substitution models that allow slightly different *marginal* probabilities for the different nucleotides) should not affect the practical validity of our methodology.

Given a rooted phylogenetic tree T, let $t(T)$ be the total time length of all branches on the tree. Subject to our assumptions, the number of mutations on this tree in a site i in total time $t(T)$ is distributed Poisson($\lambda_i \cdot t(T)$), where λ_i is the rate parameter for this site. In our case, we are not given the full tree T but a set of K haplgroups, representing terminal sub-trees $T_1, \ldots, T_K$ whose lengths $t_1, ..., t_k$ and internal structure are not known, with n samples sorted into $n_1, \ldots, n_K$ samples in each Hg respectively.

Assume first we were able to observe the number of mutations m_{ik} in each site i in each Hg k, then the total log-likelihood of the data would be:

$$\begin{aligned} l \quad (\mathbf{m}; \boldsymbol{\lambda}, \mathbf{t}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} [\log(\lambda_i t_k) m_{ik} - \lambda_i t_k] - \\ -h(\mathbf{m}) &= \sum_{i=1}^{I} \log(\lambda_i) \sum_{k=1}^{K} m_{ik} + \\ &+ \sum_{k=1}^{K} \log(t_k) \sum_{i=1}^{I} m_{ik} - \sum_{i,k} \lambda_i t_k - h(\mathbf{m}) \quad (2) \end{aligned}$$

where $h(\mathbf{m}) = \sum_{i=1,k=1}^{I,K} \log(m_{ik}!)$ is of no consequence for maximum likelihood estimation of the parameters $(\boldsymbol{\lambda}, \mathbf{t})$. This maximum likelihood estimation problem is a straight-forward Poisson regression with a (canonical) *log* link function. In fact, it is easy to show that the decomposition in Eq. (2) implies that maximum likelihood estimation of all λ_i's can be done by simple counting (up to multiplication by an overall constant factor).

Given Hg-level classification only, however, we do not observe the m_{ik}'s, but only observe the state of site i in all n_k samples (leaves) in Hg k. If not all of these are identical, we know for certain that $m_{ik} \geqslant 1$, i.e., site i has mutated at least once somewhere on the phylogenetic tree describing our haplgroup k samples. Without knowledge of the actual Hg-specific phylogeny we cannot make any further conclusions on m_{ik} in this case. If all of the n_k samples have identical nucleotide in position i, we

conclude that this site has not mutated anywhere on the Hg's phylogenetic tree, i.e., $m_{ik} = 0$. This conclusion is not guaranteed to be correct, however we can argue that with overwhelming probability it will be.

To demonstrate this concept, consider a simple phylogenetic tree like the one in Figure 2, where we assume a mutation from *red triangle* to *black circle* has occurred on the top right branch. The shapes at the bottom describe the states of the leaves (observed samples), if no other mutations have occurred at this site. If all the leaves of the tree were to have the same nucleotide (all *triangle* or all *circle*) at site i, it would require that either the mutation reverted back from *circle* to *triangle* on a cut of the subtree below it (such as both branches marked with $**$) or the same exact mutation (*triangle* to *circle*) simultaneously happened on a set of branches completing a cut of the full tree (such as the branch marked with $\times$). If none of these highly unlikely events (requiring multiple "coordinated" mutations) occur, all leaves would not have the same nucleotide at site i, given the shown *triangle* to *circle* mutation.

We can demonstrate the low probability of missing a mutation in our approach, by comparing it to another probability, that of not observing a mutation on a coalescent tree because it has mutated back *on the same link* and thus is completely unobservable for us. Assuming for simplicity that all polymorphisms are binary, consider for example the two links marked with $**$, assume they both have length t. It is easily seen that the probability that site i mutated and reverted on either one of them is $2 \cdot \exp(-2\lambda_i t)(\lambda_i t)^2/2 + O((\lambda_i \cdot t)^3)$. The probability that the *triangle* to *circle* mutation reverted back on both of them simultaneously is similarly $\exp(-2\lambda_i t)(\lambda_i t)^2 + O((\lambda_i \cdot t)^4)$, i.e., slightly smaller. If we do not assume both links have the same length, then the first probability is potentially *much bigger* than the second. Thus, under reasonable assumptions, that reversion back is most likely on binary splits, our total chance of setting $m_{ik} = 0$ when the true value is $m_{ik} > 0$, is on the same order of magnitude as twice the chance that the coalescent tree contains mutations that reverted back on the same link, which are inherently unobservable.

It should be clarified that by setting $m_{ik} = 0$ we *are not* implying that the site i has never mutated in this haplgroup k anywhere in the world, but rather that it has not happened on the phylogenetic (coalescent) tree of the n_k samples we observe in our dataset. This is the tree whose total branch length t_k is one of the parameters we will be estimating.

Thus, we are assuming that while we cannot observe our Poisson mutation counts m_{ik}, we can observe the binary variables $b_{ik} = \mathbb{I}\{m_{ik} = 0\}$. It is easy to verify that these variables are distributed as $b_{ik} \sim \text{Bernoulli}(\exp(-\lambda_i \cdot t_k))$. If we now write the partial likelihood of the observed data $\mathbf{b}$ only we get:

$$\begin{aligned} l(\mathbf{b}; \boldsymbol{\lambda}, \mathbf{t}) &= \sum_{i=1}^{I}\sum_{k=1}^{K}[-\lambda_i t_k b_{ik} + \\ &+ \log(1 - \exp(-\lambda_i t_k))(1 - b_{ik})] \end{aligned} \quad (3)$$

and maximum likelihood estimation of the parameters $(\boldsymbol{\lambda}, \mathbf{t})$ is still a Generalized Linear Model (GLM) (McCullagh and Nelder, 1989), if a slightly less standard one: a binomial regression with a complementary log-log (CLL) link function, since:

$$\log(-\log(P(b_{ik} = 1))) = \log(\lambda_i) + \log(t_k) \quad (4)$$

This procedure yields maximum likelihood estimates of both the Hg coalescent tree lengths $\hat{t}_k,\ k = 1, ..., K$ (without information about the actual phylogeny), and the site-specific instantaneous mutation rates $\hat{\lambda}_i,\ i = 1, ..., I$. However, note that this maximum likelihood solution is defined only up to a multiplication of all the $\hat{t}_k$s by a constant and division of all the $\hat{\lambda}_i$s by the same constant (the Bernoulli probabilities in (4) would not be affected). Thus, to complete our estimation we need to resolve this remaining degree of freedom, for example through calibration of the total mutation rate $\sum_i \lambda_i$ to an external accepted number. Following Forster et al. (1996) we use $1/20180$ mutations per year in the limited HVS-I (16090 to 16395) as our calibration number.

To summarize our modeling approach:

1. We are given HVS-I sequences as data, we assume that these sequences are correctly classified into Hgs and that we get the full, correct HVS-I sequence for every sample.

2. We make assumptions 1-5 above, under which the likelihood of the Hg-site specific mutation counts m_{ik} is Poisson (2).

3. Since we do not know the intra-Hg phylogeny of our samples, we cannot observe m_{ik}, however we can (with overwhelming probability) observe $b_{ik} = \mathbb{I}\{m_{ik} = 0\}$.

4. Maximum likelihood estimation of the site-specific mutation probabilities and Hg-specific coalscent tree lengths is now a binomial regression with a *complementary log-log* (CLL) link function.

2.1.1 *Saturation and sub-sampling* Since our method relies on high-quality Hg classification, and then only considers the binary b_{ik}'s, it can happen

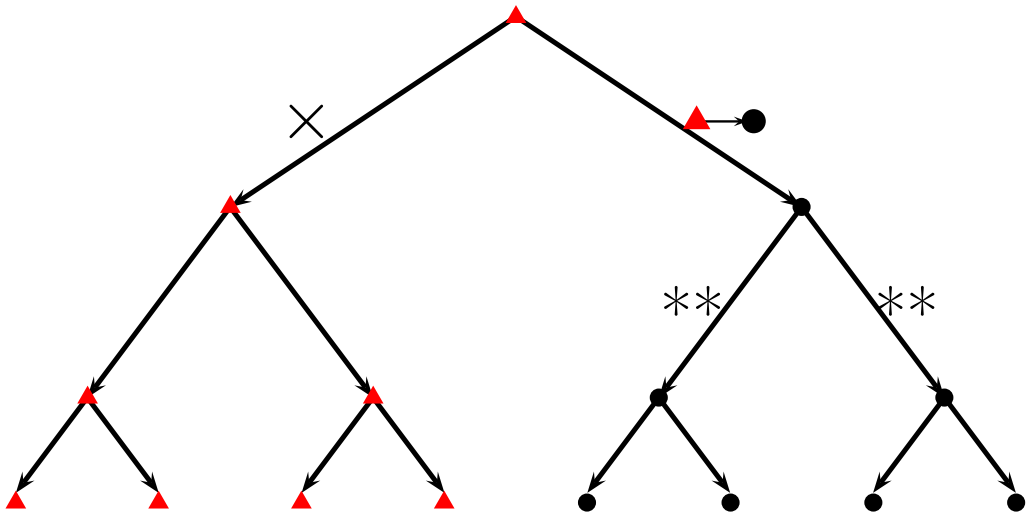

FIG. 2.—Demonstration of our reasoning, that we know whether any mutations have occurred in a specific site.

that a specific site i gives $b_{ik} = 0\ \forall k$, i.e., it is polymorphic in *all* Hgs. This is especially likely if some of the λ_i are much larger than others, and if all Hgs contain a large number of samples. This is indeed the case for the Genographic dataset we use below for our experiments.

In the event that $b_{ik} = 0\ \forall k$ the rate λ_i is unestimable in our methodology. Even if $b_{ik} \neq 0$ for a small number of Hgs, the estimate of λ_i may still suffer from stability problems. Ideally, we would like a balance between Hgs for which $b_{ik} = 1$ and ones for which $b_{ik} = 0$, especially for our fastest mutating sites.

In this situation, we propose to counter this problem by sub-sampling the large database multiple times, and generating a *distribution* of estimates generated by applying our estimation approach to sub-samples from the original larger sample. In fact, we advocate using a boostrap-based sub-sampling approach, known as the m out of n bootstrap (Bickel et al., 1997), where $m < n$ samples are sampled *with return* from the database of size n. As Bickel et al. (1997) and others discuss, this is an alternative bootstrap approach, which can lead to similar insights to the standard bootstrap, and is superior in certain situations when the standard (n out of n) bootstrap is not effective for various reasons. Our setting is different from theirs, in that the estimation itself cannot be performed from the full datatset, not only the bootstrap-based inference. Thus we are taking advantage of the m out of n bootstrap for both estimation and inference.

In our approach, we empirically try different values of m, giving rise to distributions of estimators of the mutation probabilities. We evaluate them based on their empirical spread (variance) and their bias in estimating the true probabilities. We discuss strategies for estimating these quantities in the next subsection.

2.2 *Statistical Inference*

The goal of inference is to interpret and understand the performance of our estimation procedure and validate the underlying assumptions. Our first inference goal here is to get an idea of the relationship between our estimates and the "real" values. The second is to test the hypothesis of site independence underlying our method (and much of the analysis of genetic information).

2.2.1 *Bias and variance estimation based on a simulation-bootstrap hybrid*

A key question regarding our methodology is, how reliable are our mutation probability estimates? Asymptotic theory can be used to derive approximate confidence intervals for the maximum likelihood estimates we derive (see McCullagh and Nelder (1989) for details). However, our modeling problem seems to be far from "asymptopia" and these intervals are not reliable. Also, CLL-link binomial regression has inherent bias (McCullagh and Nelder, 1989, chap. 15) We try, therefore, to investigate the error in our estimates through a combination of resampling-based and simulation approaches.

The parametric bootstrap (Efron and Tibshirani, 1994) allows us to investigate properties of our estimators through a plug-in approach: generate multiple datasets from the model we estimated, re-estimate the model from these datasets and investigate the consistent error (bias) and instability (variance) of these estimators. The main problem with application of the parametric bootstrap in our case is the implicit assumption it makes, that our estimated model is "close" to the true model, and generates data with similar properties. This assumption is clearly violated in our case in one respect: we are able to estimate probabilities only for sites in HVS-I which are polymorphic in our data (292 out of 553). However, the other 261 sites clearly do not have probability 0 of mutating. Rather, it is the luck of the draw which determines which portion of the slowly-mutating sites in HVS-I are polymorphic in our data. If we now draw a parametric bootstrap sample, using our estimated probabilities, we expect that many of the sites that are polymorphic in our data would never mutate in this bootstrap sample, and the number of polymorphic sites in every bootstrap sample would be much smaller than the number in our original dataset.

On the other hand, we have at our disposal information about the "prior" distribution of the mutation rates in HVS-I. Tamura and Nei (1993) originally showed that a Gamma prior with shape parameter roughly $\alpha = 0.1$ is appropriate for the distribution of mutation rates in the control region of the mtDNA (including HVS-I). Later authors, including Excoffier and Yang (1999) and others, have suggested different values as α may be more appropriate for HVS-I. We re-estimate this parameter from our Hg-level data, using a methodology in the spirit of Tamura and Nei (1993), as follows.

As discussed above, we assume that the sites which are non-polymorphic in all our Hgs have never mutated. Furthermore, sites which are polymorphic in one Hg only can reasonably be assumed to have mutated only once, since the fact that they are non-polymorphic in all other Hgs is indicative of their low mutation probability. While this assumption may not be completely accurate, it is "close enough" to obtain a rough estimate of α. So, assuming we know how many sites have mutated $0, 1$ and > 1 times in our complete data, we can now estimate α by a "method of moments" requiring that the empirical distribution matches the posterior probabilities for these three groups under the Negative Binomial distribution. As we show below, this method leads us to an estimate of $\alpha = 0.25$ for the shape parameter based on our data.

For simulating our process and estimating its variance, we can now simulate a set of "true" probabilities by drawing a sample of size 553 from our hypothesized distribution:

$$Gamma(\alpha, \beta)$$

and use these to generate multiple data sets, for which we know the correct probabilities, then examine our algorithm's performance on these.

To generate simulated data (that is b_{ik}'s) which is like our actual data, we also need the t_k's, i.e., the Hg tree sizes. For this purpose, we can take advantage of the parametric bootstrap, and use our estimated t_k's to generate the simulation datasets (we could then also quantify the bias our method suffers in estimating these quantities, although this is not the main focus of this paper).

We can then apply our estimation methodology to multiple samples drawn via this simulation-bootstrap hybrid methodology and obtain estimates of the bias inherent in this methodology for data "like" the genetic data we have.

To summarize our bias estimation methodology, given an estimation methodology E, and a dataset D, it proceeds as:

1. Apply E to D to obtain estimates $\hat{\lambda}_i,\ i = 1, ..., I$, and $\hat{t}_k,\ k = 1, ..., K$. If E contains m of n boosting embeded in it, apply it to multiple bootstrap samples according to this methodology.

2. Draw a sample of "true" probabilities $p_i,\ i = 1...I$ from $\Gamma(\alpha, \beta)$.

3. Repeat r times:

   (a) Create a new dataset D^* by drawing $b_{ik}\ \forall i, k$ using our simulation-bootstrap hybrid and Eq. (4).

   (b) Apply our methodology E to D^* to obtain estimates $\lambda_i^*,\ i = 1...I$.

4. Calculate empirically the bias of these estimates compared to the (known) p_i.

5. If E contains m of n bootstrap sampling, use bootstrap variance estimates. If not, use the simulation-bootstrap hybrid repeated samples to estimate the variance.

6. Evaluate the overall relationship between p_i and bias and variance, to generate a bias correction that is a function of the magnitude of p_i.

2.2.2 *Hypothesis testing about site independence* A fundamental question about our methodology and many other methods in phylogenetics is, to what extent are the molecular clock and site independence

assumptions we make realistic? In our maximum likelihood framework, we can actually test the site independence assumption statistically, against the alternative that mutation probabilities in one site may depend on the nucleotide value in another site (or multiple sites, potentially). Unfortunately, we cannot similarly test the lineage independence hypothesis, since change in the rate of the mutational clock is indistinguishably confounded with the tree sizes t_k.

Assume we want to test whether site r affects site s. Denote as before by b_{rk}, b_{sk} the indicator variables for sites r, s being non-polymorphic in Hg k, respectively. Given a "null" hypothesis of site independence between r, s, we can express the "alternative" that site s is more likely to be non-polymorphic if site r is non-polymorphic, by adding a parameter expressing this dependence to our formulation, as follows:

$$P(b_{rk} = 1) = \exp(-\lambda_r t_k) \text{ (as before)}$$
$$P(b_{sk} = 1 | b_{rk} = 1) = \exp(-\lambda_s t_k) \text{ (as before)}$$
$$P(b_{sk} = 1 | b_{rk} = 0) = \exp(-\lambda_s \lambda_{rs} t_k)$$
$$\text{(potential effect of site } r\text{)}$$

Under the null of no dependence, we have $\lambda_{rs} = 1$ and we go back to the formulation in (3), while under the alternative we can re-write the likelihood as:

$$\begin{aligned} l \quad & (\mathbf{b}; \boldsymbol{\lambda}, \mathbf{t}) = \qquad (5) \\ & = \sum_{i=1, i \neq s}^{I} \sum_{k=1}^{K} [-\lambda_i \cdot t_k \cdot b_{ik} + \\ & + \log(1 - \exp(-\lambda_i \cdot t_k))(1 - b_{ik})] + \\ & + [-\lambda_s {\lambda_{rs}}^{1-b_{rk}} \cdot t_k \cdot b_{sk} + \\ & + \log(1 - \exp(-\lambda_s {\lambda_{rs}}^{1-b_{rk}} \cdot t_k))(1 - b_{sk})] \end{aligned}$$

where the last part in Eq. (5) allows an extra parameter for the cross-effect between the two sites. We can then test the hypothesis $H_0 : \lambda_{rs} = 1$ via a generalized likelihood ratio test with one degree of freedom, comparing the maximum likelihood solutions of (3) and (5).

When we apply this testing methodology for all pairs of sites, we are performing a large number of tests, and we need to take into account the issue of multiple comparisons when evaluating the outcome of our tests. For that purpose, we employ the false discovery rate multiple comparisons correction at 5%, which guarantees that the expected rate of falsely rejected null hypotheses is at most 5% of all rejected hypotheses, possibly less, under some types of dependence (Benjamini and Hochberg, 1995). This correction is slightly less conservative than the standard Bonferroni correction (i.e., allows us to reject more nulls), but similar in spirit.

The main advantage of our testing methodology is that it aligns naturally with our modeling approach, and specifically that it does not require detailed phylogenetic reconstruction. It should be noted, however, that it cannot expose every type of non-independence, and it may have limited power to expose others. For example, if a specific combination of nucleotide values in two sites has a strong affinity, and hence once one site mutates into this state the other follows closely, our method can only identify this affinity if this phenomenon has happened in many of the Hg's. A detailed phylogenetic analysis could have more power to identify and characterize these relationships.

2.3 *Genographic mtDNA data*

Each mitochondrial DNA sample submitted to the Genographic project goes through the standard classification process (Behar et al., 2007):

1. Sequencing of a number of coding-region markers. The number has increased during the project, currently is at 22.

2. Sequencing of the full extended HVS-I, defined as sites 16024-16569 of the samples aligned to revised Cambridge Reference Sequence (rCRS).

3. Based on 1., determine a Hg designation *by SNPs* into one of 23 Hgs: L0/1, L2, L3xMN, M, C, D, N, N1, A, I, W, X, R, R9, R0, HV, H, V, J, T, U, K, B.

4. Based on 1. and 2., determine a haplgroup designation into one of 87 Hgs.

Table 2 of Behar et al. (2007) shows a summary of Hg distribution for the 16609 samples used in our analysis (the *Reference database*). Following assumption 1 in Section 2.1, we assume that the 23-Hg nomenclature labels are all correct. Since they are based on coding region SNPs and the careful classification protocol discussed in Behar et al. (2007), this assumption is likely to be true. It is less likely to be accurate for the 87-Hg nomenclature. However, as the 87-Hg version allows us to get much better resolution in our analysis, we also use it with the implicit assumption that its classification is accurate, and compare and discuss the results from using both nomenclatures.

Supplementary Table 4 of Behar et al. (2007) contains all the information required to calculate the b_{ik} values for the full dataset. We can see that some of the sites are completely saturated for the 23-Hg nomenclature: 16129, 16189, 16519 are polymorphic in all 23 Hgs and several other sites are poly-

morphic in at least 20 Hgs. Thus, to model the probabilities reliably from this data we have to resort to our sub-sampling methodology.

With the 87-Hg nomenclature, we clearly have a lot more information about the mutation probabilities in our data, but a less reliable Hg classification. Site 16519 is polymorphic in the most Hgs: 65 of the 87. Thus, based on this data we could estimate the probabilities directly without resorting to sub-sampling. The quality of estimates will be hampered by the uncertainty about the correctness of the Hg labels.

One issue about the data which is highly relevant to our analysis below is the problems in sequencing around the poly-cytosine (poly-C) region created by the mutation T16189C (relative to rCRS). This comes up in the dependence we identify below between sites 16182 and 16183 in our sequences, which we suspect may be due to sequencing problems. Mutations in these two sites always occur in concordance with the adjacent polymorphism T16189C that creates a poly-C stretch which causes significant reading difficulties of this region using standard sequencing procedures (Figure 3). These difficulties relate to a technical sequencing problem in which DNA strands that differ in the number of cytosine repeats are assembled and thus overlapping positions subsequent to T16189C are impossible to be appreciated since they are affected by the shift created by the variable number of cytosines in the different DNA strands. Therefore, the positions around the poly-C stretch are usually removed from analysis (Behar et al., 2007). A different question relates to our ability to correctly understand the number of adenosines that immediately precedes the poly-C region (four in the rCRS). Figure 3 shows that different numbers of adenosines are associated with the poly-C stretch. Since most of the mutations we observe in 16182 and 16183 are transversions between adenosine and cytosine it is possible that the poly-C stretch creates a technical problem here as well despite the unquestionable reads we get for these positions. We successfully used fragment length analysis techniques, similar to those used to count the number of repeats in short tandem repeats, to understand the real number of cytosine repeats in various samples and found no clear evidence for mistakes in the number of preceding adenosines (Data not shown). Nevertheless, caution mandates the questioning of the authenticity of our results for positions 16182 and 16183 and the possibility that the poly-C stretch plays a role in creating artificial dependence.

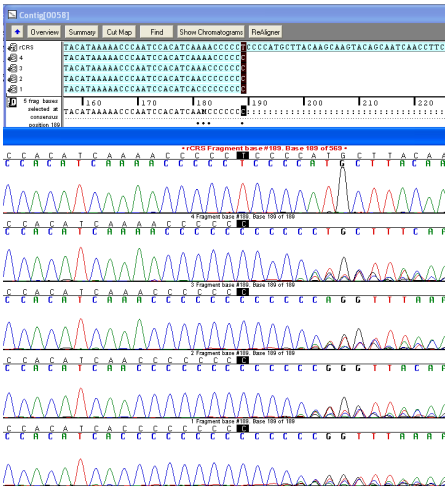

FIG. 3.—The poly-C stretch. Position 16189 is highlighted and five sequences are shown. A sequence identical to the rCRS in the presented region is shown at the top. Below it, four sequences containing the T16189C polymorphism are arranged to show 1-4 adenosines preceding the poly-C stretch. A typical chromatogram of the sequence after the poly-C stretch is also demonstrated.

3. Results

Considering the discussion above about the various Hg nomenclatures we have at our disposal and the sub-sampling approaches, we implemented four different protocols to estimate mutation probabilities from our data:

1. Sub-sampling based estimates, using 100 repeated samples of 1000 sequences out of our 16609 total sequences and the 23-Hg nomenclature.

2. Sub-sampling based estimates, using 100 repeated samples of 3000 sequences out of our 16609 total sequences and the 23-Hg nomenclature.

3. Sub-sampling based estimates, using 100 repeated samples of 4000 sequences out of our 16609 total sequences and the 23-Hg nomenclature.

4. Estimates with no sub-sampling, using the 87-Hg nomenclature.

We then used the glm function in R to calculate the maximum likelihood estimates of $(\boldsymbol{\lambda}, \mathbf{t})$ in (3). See McCullagh and Nelder (1989) for discussion of the theory of GLMs and Venables and Ripley (1994) for discussion of the glm function in S+, which is the predecessor of R.

Running the binomial regression, and applying the constraint $\sum_{i\in\{16090,\ldots,16395\}} \lambda_i = 1/20180$ from Forster et al. (1996) for calibration, we obtain maximum likelihood estimates in each setting (in the sub-sampling protocols 1–3, we actually obtain a whole distribution of estimates in each setting). We then apply our bias correction (which turns out to be small, see below) and use the empirical range of estimates from the bootstrap samples (for protocols 1–3) or the estimated variance from the simulation-bootstrap hybrid (for protocol 4) to calculate confidence intervals. Table 1 (first four columns) gives an estimate and confidence interval of mutation rates for the 48 *quickest* mutating sites in HVS-I, from several different variants of our approach. We see that the fastest mutating site, 16519, is estimated to mutate once about every 200,000-500,000 years (depending which of our estimates is used). The 10th fastest site mutates about 4 times more slowly, and the slowest site in this list mutates about 10 times more slowly. Thus, for example, two individuals whose time to most recent common mtDNA ancestor (TMRCA) is 20000 years, have a probability of about $\exp(-40000/350000) = 0.87$ to have the same nucleotide in site 16519 due to *identity by descent*. The total probability that they share the same nucleotide is of course slightly higher, since they may also have it due to homoplasy. Figure 4 shows a graphical representation of the probability estimates as they physically appear on HVS-I (using the estimates from the *3000 samples* version, as in the fourth column of Table 1). We can see the relatively uniform spread of the fastest mutating sites, perhaps with a cluster around the poly-C region in 16184-16189, and the relative dearth of fast sites after 16370, and especially in the range 16400-16519.

3.1 *Bias-variance analysis*

To quantify how biased our derived estimates are, we employe the bootstrap-simulation approach we described above. The first step is to decide on a reasonable prior distribution for the mutation probabilities. To accomplish that, we find the shape parameter α that would be most consistent with the counts of sites that have mutated $0, 1, > 1$ times, as described above. The resulting estimate is $\hat{\alpha} = 0.25$.

We then derive a sample of mutation probabilities from this prior and use the estimated t_k's from our method (Table #) to implement the bias estimation methodology. Figure 5 shows the estimated bias as a function of the true mutation probability for each one of our four estimation settings. The points are means of the estimates from 100 runs of our simulation-bootstrap algorithm, and the lines are LOESS smoothed estimates of the bias (Cleveland et al., 1992). These plots are shown on the log-scale, i.e., they represent the ratio of the mutation probability to the bias in its estimates from the different methods. We can observe that the bias has some interesting behavior, and no clear consistent pattern (although an obvious tendency to be negative and more pronounced for lower mutation probabilities). However, encouragingly we can observe that in the region of higher mutation probabilities that is of interest of us, the bias is almost invariably smaller than 0.2 in absolute value on the log scale, and therefore no bigger than roughly 20% in our probability estimates.

3.2 *Hypothesis testing*

For hypothesis testing of site independence, we utilized the 87-Hg nomenclature, since the additional information in the more detailed phylogeny is critical for our chances of identifying true dependence. We applied the generalized likelihood ratio (GLR) test described above to all pairs of sites which are polymorphic in at least 5 out of the 87 Hgs — a total of 156 sites, giving us a total of $156 \times 155 = 24180$ tests.

Effect	**Raw p val.**	**Corrected**
$16182 \Rightarrow 16183$	7.7×10^{-12}	< 0.0001
$16183 \Rightarrow 16182$	2.2×10^{-9}	< 0.0001
$16114 \Rightarrow 16526$	0.0000012	0.03
$16212 \Rightarrow 16153$	0.000027	0.66
$16266 \Rightarrow 16148$	0.000033	0.8
$16304 \Rightarrow 16163$	0.000039	0.95
$16184 \Rightarrow 16335$	0.000045	1
$16104 \Rightarrow 16111$	0.000053	1
$16327 \Rightarrow 16163$	0.000068	1
$16526 \Rightarrow 16114$	0.00009	1
$\vdots$	$\vdots$	$\vdots$

Table 2 Results of generalized likelihood ratio tests for site independence. The table shows the ten most *non-independent* pairs found in our data.

Table 2 contains the pairs of sites which gave the lowest p values for the GLR test, and their FDR-corrected p-values (Benjamini and Hochberg, 1995) [1]. We observe that after the FDR correction, we are left with only 3 cases where we can reject the site independence hypothesis at $p = 0.05$. We now analyze these cases in some more detail.

The two-way relationship $16182 \Leftrightarrow 16183$ is by far the strongest non-independence effect our

[1] Although we used the more powerful FDR scheme, the conclusions would have been the same from using the simple Bonferroni correction.

	1000 samples	3000 samples	4000 samples	87 Hg	Bandelt et al.
Locus	Est. [90% CI]	Est. [90% CI]	Est. [90% CI]	Est. [90% CI]	Est. [90% CI]
16051	0.54 [0.33-0.85]	0.5 [0.30-0.82]	0.54 [0.35-0.84]	0.48 [0.34-0.69]	0.67 [0.31-1.3]
16086	0.35 [0.12-0.7]	0.49 [0.25-0.8]	0.55 [0.31-0.87]	0.7 [0.51-0.98]	0.29 [0.078-0.74]
16092	0.56 [0.32-0.96]	0.57 [0.34-0.88]	0.54 [0.35-0.88]	0.48 [0.34-0.68]	0.57 [0.25-1.1]
16093	1.6 [0.91-2.5]	1.7 [1.1-2.3]	1.8 [1.0-3.2]	2.7 [2.0-3.7]	3.2 [2.3-4.3]
16111	0.64 [0.37-1.0]	0.58 [0.37-0.86]	0.64 [0.37-1.1]	0.62 [0.44-0.86]	0.71 [0.35-1.3]
16126	0.52 [0.28-1]	0.68 [0.45-1]	0.66 [0.44-0.9]	0.47 [0.33-0.66]	0.43 [0.16-0.94]
16129	1.9 [1.1-2.8]	1.8 [1.2-3]	1.7 [1.2-2.9]	1.2 [0.88-1.6]	1.8 [1.1-2.6]
16145	0.56 [0.31-1.2]	0.61 [0.39-0.94]	0.64 [0.44-0.95]	0.63 [0.45-0.88]	0.67 [0.31-1.3]
16148	0.34 [0.19-0.56]	0.32 [0.21-0.47]	0.3 [0.20-0.45]	0.31 [0.21-0.46]	0.38 [0.13-0.87]
16172	1.8 [1.2-2.8]	1.6 [1.1-2.6]	1.5 [0.93-2.3]	1.1 [0.83-1.5]	0.86 [0.45-1.5]
16182	0.64 [0.36-1.1]	0.68 [0.39-0.98]	0.64 [0.44-0.89]	0.59 [0.42-0.82]	0.01 [0.005-0.45]
16183	1.8 [1.1-3]	1.9 [1.2-2.9]	1.8 [1.3-2.4]	1.1 [0.82-1.5]	0 [0-0.29]
16184	0.21 [0.06-0.49]	0.32 [0.17-0.58]	0.35 [0.21-0.56]	0.47 [0.33-0.67]	0.01 [0.005-0.45]
16189	2.5 [1.6-3.7]	2.4 [1.7-3.4]	2.2 [1.3-3.8]	2.4 [1.8-3.3]	2.2 [1.5-3.1]
16192	1.1 [0.6-1.7]	0.94 [0.6-1.4]	0.88 [0.63-1.3]	0.92 [0.67-1.3]	1.4 [0.89-2.2]
16209	0.41 [0.21-0.68]	0.43 [0.26-0.68]	0.46 [0.28-0.73]	0.42 [0.29-0.6]	0.43 [0.16-0.94]
16213	0.26 [0.11-0.57]	0.32 [0.18-0.55]	0.34 [0.2-0.55]	0.24 [0.16-0.37]	0.52 [0.22-1.1]
16218	0.28 [0.12-0.54]	0.35 [0.19-0.53]	0.36 [0.23-0.52]	0.4 [0.28-0.58]	0 [0-0.29]
16223	0.46 [0.18-0.91]	0.57 [0.34-0.93]	0.64 [0.38-0.98]	0.57 [0.41-0.8]	0.86 [0.45-1.5]
16234	0.52 [0.21-0.95]	0.68 [0.42-1.2]	0.68 [0.41-1.1]	0.64 [0.46-0.9]	0.43 [0.16-0.94]
16239	0.36 [0.20-0.6]	0.35 [0.21-0.55]	0.32 [0.21-0.48]	0.37 [0.26-0.54]	0.19 [0.03-0.6]
16249	0.5 [0.25-0.81]	0.54 [0.31-0.88]	0.54 [0.36-0.8]	0.43 [0.3-0.61]	0.38 [0.13-0.87]
16256	0.54 [0.32-1]	0.64 [0.41-1.0]	0.62 [0.4-1.0]	0.77 [0.56-1.1]	0.86 [0.45-1.5]
16260	0.21 [0.06-0.48]	0.28 [0.15-0.43]	0.26 [0.16-0.44]	0.39 [0.27-0.57]	0.19 [0.03-0.6]
16261	0.65 [0.33-1.1]	0.64 [0.42-1.0]	0.6 [0.41-0.86]	0.51 [0.36-0.73]	1.0 [0.59-1.7]
16265	0.45 [0.22-0.83]	0.44 [0.28-0.64]	0.44 [0.31-0.64]	0.49 [0.35-0.7]	0.48 [0.19-1]
16266	0.34 [0.13-0.67]	0.5 [0.25-0.85]	0.5 [0.30-0.86]	0.64 [0.46-0.89]	0.38 [0.13-0.87]
16270	0.48 [0.31-0.7]	0.32 [0.22-0.5]	0.29 [0.19-0.43]	0.2 [0.13-0.32]	0.24 [0.06-0.67]
16274	0.7 [0.39-1.2]	0.81 [0.47-1.3]	0.81 [0.56-1.1]	1.2 [0.9-1.7]	0.76 [0.38-1.4]
16278	1.1 [0.7-1.7]	0.93 [0.55-1.5]	0.86 [0.6-1.2]	0.93 [0.67-1.3]	1.1 [0.66-1.9]
16290	0.17 [0.05-0.42]	0.3 [0.13-0.52]	0.31 [0.17-0.52]	0.36 [0.25-0.53]	0.38 [0.13-0.87]
16291	0.65 [0.38-1.1]	0.66 [0.42-0.98]	0.68 [0.45-0.95]	0.71 [0.51-0.99]	1.0 [0.59-1.7]
16292	0.42 [0.22-0.8]	0.43 [0.25-0.69]	0.40 [0.25-0.62]	0.41 [0.28-0.58]	0.67 [0.31-1.3]
16293	0.31 [0.18-0.59]	0.31 [0.19-0.55]	0.29 [0.16-0.46]	0.37 [0.26-0.54]	0.76 [0.38-1.4]
16294	0.74 [0.44-1.1]	0.72 [0.42-1.0]	0.75 [0.44-1.1]	0.62 [0.44-0.86]	0.29 [0.08-0.74]
16295	0.32 [0.13-0.57]	0.36 [0.23-0.58]	0.33 [0.21-0.52]	0.3 [0.2-0.45]	0.48 [0.19-1]
16298	0.41 [0.23-0.7]	0.36 [0.23-0.57]	0.32 [0.22-0.47]	0.22 [0.14-0.34]	0.57 [0.25-1.1]
16304	0.49 [0.31-0.79]	0.4 [0.26-0.68]	0.4 [0.27-0.59]	0.36 [0.25-0.52]	0.57 [0.25-1.1]
16311	2.3 [1.5-3.5]	2.4 [1.6-3.9]	2.6 [1.6-5.8]	2.6 [1.9-3.5]	2.8 [2-3.8]
16319	0.8 [0.4-1.6]	0.81 [0.51-1.3]	0.82 [0.54-1.3]	0.56 [0.4-0.79]	0.48 [0.19-1]
16320	0.53 [0.29-0.86]	0.43 [0.3-0.64]	0.40 [0.27-0.6]	0.37 [0.25-0.54]	0.8 [0.41-1.4]
16325	0.66 [0.28-1.1]	0.65 [0.43-0.94]	0.6 [0.38-0.82]	0.55 [0.39-0.78]	0.33 [0.10-0.8]
16355	0.41 [0.19-0.73]	0.45 [0.27-0.75]	0.46 [0.25-0.77]	0.52 [0.37-0.73]	0.38 [0.13-0.87]
16362	2.4 [1.4-4.1]	2.4 [1.7-3.1]	2.2 [1.6-3.0]	2.2 [1.6-3]	1.8 [1.2-2.7]
16390	0.49 [0.25-0.87]	0.54 [0.33-0.9]	0.52 [0.33-0.76]	0.59 [0.42-0.83]	
16399	0.38 [0.20-0.69]	0.39 [0.24-0.57]	0.41 [0.25-0.64]	0.49 [0.35-0.7]	
16519	3.6 [2.4-6.1]	2.9 [1.9-4.9]	3.0 [1.7-4.7]	4.4 [3.1-6.2]	
16527	0.31 [0.11-0.62]	0.36 [0.21-0.55]	0.32 [0.24-0.47]	0.38 [0.27-0.56]	

Table 1 Mutation rate estimates (in mutations per million years) and 90% confidence intervals for the fastest sites in HVS-I from some versions of our method and Bandelt et al. (2006)

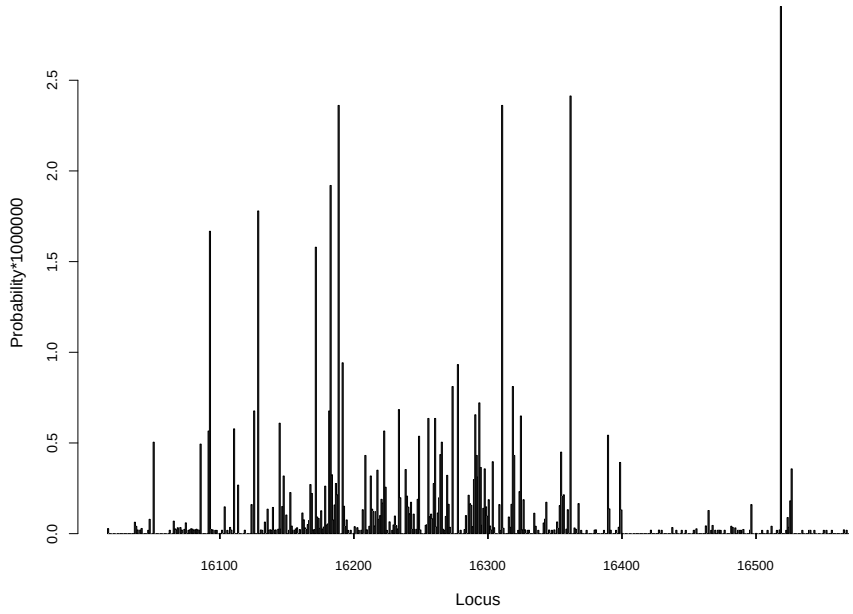

FIG. 4.—Graphical representation of mutation rates along HVS-I

methodology identifies in our sequences. As we discussed above, it is unclear to what extent sequencing ambiguity persists in these position as a result of its proximity to the poly-C region. However, since most of the mutations we observe in these two sites are between $A \Leftrightarrow C$, i.e., transversions, it seems possible that the poly-C sequence plays a role in creating artificial dependence.

The remaining significant effect is the pair $16114 \Rightarrow 16526$. Examining our raw sequences, this significant Hg-level relationship does not seem to follow from easily detectable sequence-level relationships, i.e., we do not observe a consistent tendency for mutations in site 16526 and 16114 to co-appear. We therefore lean towards attributing this discovery to chance and not to a real dependence.

So while our hypothesis testing framework did identify three significant non independence relationships in our data, further analysis of these suggests that uncertainty about sequencing issues persists for two of them, while the third is probably due to pure chance.

Our results are encouraging in that they support the validity of site-independence assumptions in analyzing mtDNA HVS-I data. Any dependence that exists is not strong enough to discover with our testing methodology, using our very large database and most detailed (87-Hg) phylogenetic protocol.

4. Discussion

The mutation dynamics of the human genome in general and mtDNA in particular have experience a surge of interest in recent years (Torroni et al., 2006). Many papers deal with the real or apparent "slow-down" effect in the molecular clock for older time periods (e.g., Ho et al. (2005)). Since we share Bandelt et al. (2006)'s opinion that there is no convincing evidence for a molecular clock slow-down rather than saturation causing these apparent effects, we view this issue as unrelated to our analysis in this paper.

4.1 *The advantage of not relying on detailed phylogeny*

The previous approaches for estimating individual mutation probabilities in HVS-I we mentioned above were all based on a reconstruction of the full phylogenetic tree through a maximum likelihood approach (Excoffier and Yang, 1999), quartet puzzling (Meyer and von Haeseler, 2003) or maximum parsimony (Bandelt et al., 2006).

In our case, if we were able to obtain a full phylogeny (like in part B of Figure 1), we would be able to observe the actual m_{ik} values (at least up to uncertainty about repeated mutations on tree branches), use equation (2) for modeling, and most likely get better quality results than our modeling

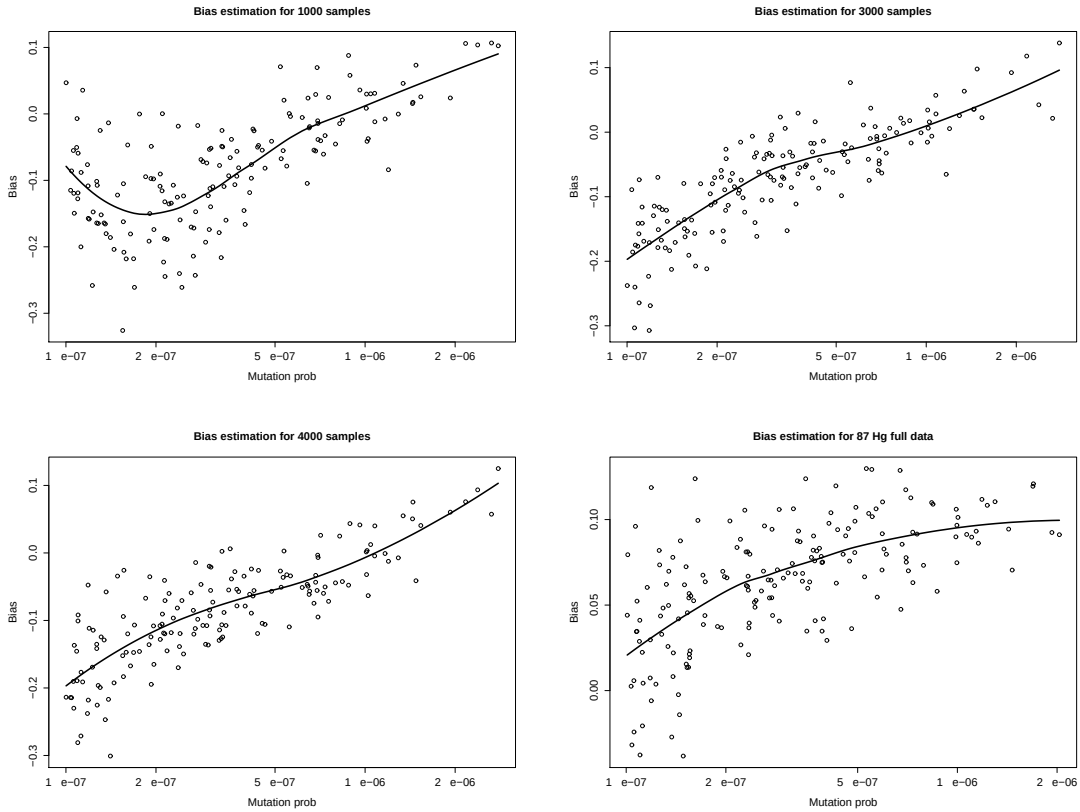

FIG. 5.—Smoothed bias estimation curves for our various estimation protocols, using our simulation-bootstrap hybrid. The smoothing was done using LOESS (Cleveland et al., 1992).

based on equation (3). However, the fundamental idea behind our approach, is that reliable Hg classification on a tree whose general structure is known (such as the human mtDNA tree) is a much simpler task than identifying the complete phylogeny of a large set of samples. Building detailed phylogenies for large samples presents significant computational and, more importantly, statistical difficulties. The resulting phylogenies may be highly underdetermined and uncertain (Felsenstein, 2003). Use of maximum likelihood methodology like Excoffier and Yang (1999) would also require parametric assumptions about the mutation rates.

For example, the data we use here is comprised of 16609 HVS-I samples of mtDNA. The Hg classification is primarily based on a set of coding-region SNPs, and is therefore very reliable. On the other hand, relying on HVS-I to build detailed, reliable phylogenies within Hgs, with hundreds, or even thousands of samples per Hg, is an overwhelming task.

4.2 *Comparison to previous estimates*

We briefly compare our estimates to those from Bandelt et al. (2006), which are most comparable to ours in terms of the large amount of data used (873 samples in their case, 16609 in our) and largely subsume the previous efforts. Since they used the limited definition of HVS-I as 16051-16365, we concentrate on the region that is common to our study and theirs. As can be seen in Table 1, the estimates are similar in spirit. In particular, since the explicit estimates given by Bandelt et al. (2006) are based on simple counting, they have a Poisson distribution under our assumptions. We can thus use standard Poisson inference methodology to build confidence intervals for them (Johnson and Kotz, 1969), which we do in Table 1. We also normalize their estimates to be on the same scale as ours, by constraining their sum to be the same as the sum of our estimates for the same range (16051-16365). We observe that the confidence intervals from their estimates are slightly smaller than ours for the fastest sites, but get much larger than ours as the rates decrease. For example, if we consider the first four rows in Table 1, we see that in rows 1–3, where the rates are relatively small, the confidence intervals from all variants of our methodology are smaller than those based on Bandelt et al. (2006). In row 4, which corresponds to 16093, one of the fastest sites in HVS-I (and coincidentally, one of the sites where the rate estimate of Bandelt et al. (2006) most disagrees with ours), the confidence interval based on Bandelt et al. (2006) is smaller than those our methods generate. We can infer that our approach, which uses less phylogenetic information but a much larger number of samples overall, has advantages for estimating fast — but not the fastest — sites compared to Bandelt et al. (2006). Qualitatively, our estimates and theirs seem to agree well, and the confidence intervals almost invariably overlap. A graphical representation of the confidence interval relationships in five randomly selected sites can be seen in Figure 6.

4.3 *Potential uses of our estimates*

Reliable mutation rate estimates are clearly important for several widely accepted reasons, for example:

- Since the mutation dynamics of the genome are a critical component of evolution, the availability of good methodology for estimating mutation rates is part of the critical infrastructure needed to study evolution. Of particular interest in this context might be our investigation of the site-independence assumption.

- Understanding the function of various regions in the genome and the mutual influence between different regions, which may be caused either by a functional relation or a physical or chemical one, is one of the key challenges of the field of Genomics, and indeed one of the most important scientific questions of our time (The International HapMap Consortium, 2005; Hardison, 2003). Creating a better understanding of the mutation mechanisms and potential dependencies in those may be an important step in this process, as it may help separate non-genic areas which have function (and are therefore preserved) from ones that do not, and discover the relationships between regions within our genome. For example, the relatively paucity of polymorphisms in the region 16400-16500 observed in Figure 4 might suggest a functional role that is not fully understood yet.

- Mutation rates can be used to improve phylogeny estimation algorithms and sequence quality checking (Bandelt et al., 2002). It should be clarified, however, that these rates are *not* very useful for time estimation on known phylogenies. As Rosset (2007) has shown, under a simple substitution model like the one we assume here, the individual rates are of no consequence for time estimation, only their sum. This is a direct consequence of the fact that the sum of independent Poisson random variables is still Poisson distributed. Under more complex models, the individual rates may have a minor effect on time estimates.

We have also recently used our estimates reported here to improve the accuracy of the

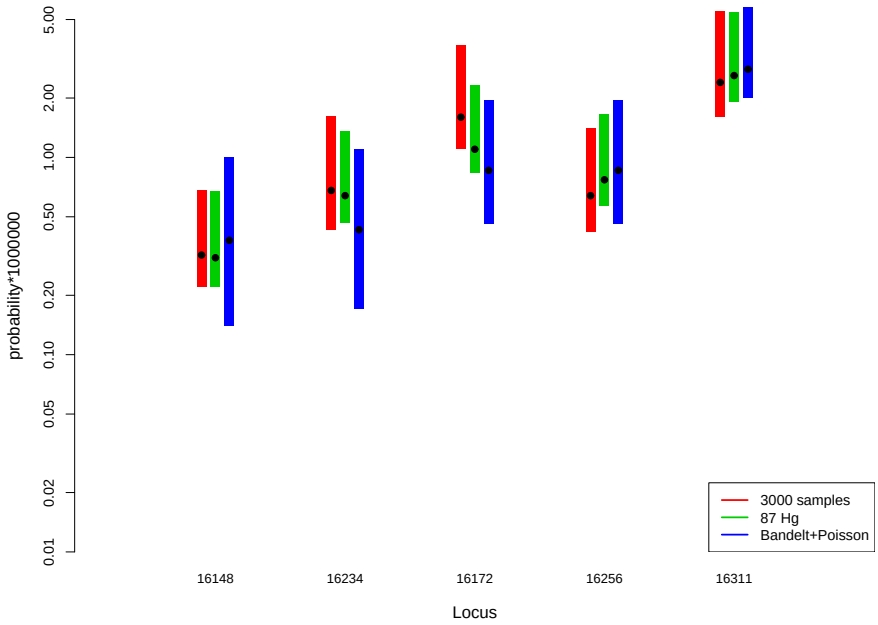

FIG. 6.—Comparison of the estimates (black dot) and confidence intervals from two of our variants and Bandelt et al. (2006). Note that the y-axis is on a logarithmic scale.

mtDNA Hg classification protocol in the Genographic project Behar et al. (2007).

An interesting by-product of our mutation probability estimation methodology is the estimates we derive of t_k, the total length of the coalescent tree of the samples we have in each Hg (it should be reiterated that this is not the TMRCA of the Hg, but the sum of the lengths of all branches in the coalescent tree). These can be used for inference on the age and demographic history of the Hg's. Table 3 gives some estimates of t_k, derived from our calculations based on the 87-Hg protocol. Detailed discussion of these results is beyond the scope of this paper, but we can clearly see the difference between Hg M* (255 samples, estimate of t_k is about 6 million years) and Hg V (471 samples, estimate of t_k is only 1.7 million years), implying that our samples from M* are much more diverse than those from V, a difference that demonstrates the older age of the polyphyletic Hg M* and its more ancient expansion.

Hg	# samples	Total tree length
A	361	4628667
B	301	5624497
C	229	3089925
D	147	2692974
H	6232	36186219
M*	255	5878315
V	471	1726071

Table 3 The value of t_k (coalescent tree size) for a subset of haplgroups in our data. The full list is available in Supplementary Table 1

4.4 *Extensions of the methodology*

In this paper we have discussed and demonstrated the application of our methodology to single nucleotide polymorphisms in the mtDNA HVS-I. This is a natural application because these sites are highly polymorphic, large amounts of data are available, and Hg-classification is relatively easy to obtain. The natural question is, what other domains would comply with these same conditions?

An interesting application may be to short tandem polymorphisms on the Y chromosome (Y-STRs), which comply with all three conditions. The mutation probabilities (and more generally, mechanisms) of these patterns have been under intense study for several years, but progress is difficult to make, unless some highly non-realistic assumptions are to be made (for more details, see for example Zhivotovsky (2001); Calabrese and Sainudiin. (2004)). Our methodology would be directly applicable to Y-STR if we could assume that the mutation probability of each Y-STR does not depend on

its state (repeat count). In that case, our approach can be directly applied to calculate this probability. Some length-dependence can even be accommodated within our method by including an independent variable for length, but the exact details of how this can be done are a topic for further research.

Literature Cited

Bandelt, H., L. Quintana-Murci, A. Salas, and V. Macaulay. 2002. The Fingerprint of Phantom Mutations in Mitochondrial DNA Data. Am. J. Hum. Genet. **71**.

Bandelt, H. J., Q. P. Kong, M. Richards, and V. Macaulay. 2006. Estimation of mutation rates and coalescence times: some caveats. In: Bandelt, H. J., V. Macaulay, and M. Richards, editors, Human mitochondrial DNA and the evolution of Homo sapiens. Springer, Berlin, 47–90.

Behar, D. M., S. Rosset, J. Blue-Smith, O. Balanovsky, S. Tzur, D. Comas, R. J. Mitchell, L. Quintana-Murci, C. Tyler-Smith, and R. S. Wells. 2007. The Genographic Project Public Participation Mitochondrial DNA Database. PLoS Genetics **3**. Doi:10.1371/journal.pgen.0030104.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Series B **57**:289–300.

Bickel, P., F. Gotze, and W. van Zwet. 1997. Resampling fewer than n observations: gains, losses and remedies for losses. Statistica Sinica **7**:1–31.

Calabrese, P., and R. Sainudiin. 2004. Models of Microsatellite Evolution. In: Nielsen, R., editor, Statistical Methods in Molecular Evolution. Springer.

Cleveland, W., E. Grosse, and W. Shyu. 1992. Local regression models. In: Chambers, J., and T. Hastie, editors, Statistical Models in S, chapter 8. Wadsworth & Brooks/Cole.

Efron, B., and R. Tibshirani. 1994. An Introduction to the Bootstrap. Chapman & Hall/CRC.

Excoffier, L., and Z. Yang. 1999. Substitution rate variation among sites in the mitochondrial hypervariable region I of humans and chimpanzees. Molecular Biology and Evolution **16**:1357–1368.

Felsenstein, J. 2003. Inferring Phylogenies. Sinaur Associates.

Forster, P., R. Harding, A. Torroni, and H. Bandelt. 1996. Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59**:935–945.

Hardison, R. 2003. Comparative Genomics. PLoS Biology **1**.

Ho, S., M. Phillips, A. Cooper, and A. Drummond. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. **22**:1561–1568.

Johnson, N., and S. Kotz. 1969. Discrete Distributions. Houghton Mifflin Company, Boston.

Jukes, T., and C. Cantor. 1969. Evolution of protein molecules. In: Munro, H., editor, Mammalian Protein Metabolism, volume 3, chapter 24. Academic Press, New York, 21–132.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences **78**:454–458.

McCullagh, P., and J. Nelder. 1989. Generalized Linear Models. Chapman and Hall, London.

Meyer, S., and A. von Haeseler. 2003. Identifying site-specific substitution rates. Molecular Biology and Evolution **20**:182–189.

Rosset, S. 2007. Efficient Inference on Known Phylogenetic Trees Using Poisson Regression. Bioniformatics **23**:e142–e147.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

The International Hapmap Consortium, 2005. A Haplotype Map of the Human Genome. Nature **437**:1299–1320.

Torroni, A., A. Achilli, V. Macaulay, M. Richards, and H.J. Bandelt 2006. Harvesting the fruit of the human mtDNA tree. Trends in Genetics **22**:339.

Venables, W., and B. Ripley. 1994. Modern Applied Statistics with S-Plus. New York: Springer.

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

Zhivotovsky, L. A. 2001. Estimating Divergence Time with the Use of Microsatellite Genetic Distances: Impacts of Population Growth and Gene Flow. Molecular Biology and Evolution **18**:700–709.