OPEN ACCESS Freely available online

PLoS one

Analytical Framework for Identifying and Differentiating Recent Hitchhiking and Severe Bottleneck Effects from Multi-Locus DNA Sequence Data

Ori Sargsyan*

Theoretical Biology and Biophysics and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America

Abstract

Hitchhiking and severe bottleneck effects have impact on the dynamics of genetic diversity of a population by inducing homogenization at a single locus and at the genome-wide scale, respectively. As a result, identification and differentiation of the signatures of such events from DNA sequence data at a single locus is challenging. This paper develops an analytical framework for identifying and differentiating recent homogenization events at multiple neutral loci in low recombination regions. The dynamics of genetic diversity at a locus after a recent homogenization event is modeled according to the infinite-sites mutation model and the Wright-Fisher model of reproduction with constant population size. In this setting, I derive analytical expressions for the distribution, mean, and variance of the number of polymorphic sites in a random sample of DNA sequences from a locus affected by a recent homogenization event. Based on this framework, three likelihood-ratio based tests are presented for identifying and differentiating recent homogenization events at multiple loci. Lastly, I apply the framework to two data sets. First, I consider human DNA sequences from four non-coding loci on different chromosomes for inferring evolutionary history of modern human populations. The results suggest, in particular, that recent homogenization events at the loci are identifiable when the effective human population size is 50000 or greater in contrast to 10000, and the estimates of the recent homogenization events are agree with the ''Out of Africa'' hypothesis. Second, I use HIV DNA sequences from HIV-1-infected patients to infer the times of HIV seroconversions. The estimates are contrasted with other estimates derived as the mid-time point between the last HIV-negative and first HIV-positive screening tests. The results show that significant discrepancies can exist between the estimates.

Citation: Sargsyan O (2012) Analytical Framework for Identifying and Differentiating Recent Hitchhiking and Severe Bottleneck Effects from Multi-Locus DNA Sequence Data. PLoS ONE 7(5): e37588. doi:10.1371/journal.pone.0037588

Editor: David Caramelli, University of Florence, Italy

Received March 6, 2012; **Accepted** April 21, 2012; **Published** May 25, 2012

Funding: This work was supported by the United States Department of Energy through the LANL/LDRD Program and by the National Institute of Health [grant number 5R01AI08752002]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: sargsyan@lanl.gov

Introduction

Hitchhiking and severe bottleneck effects have similar signatures on the population genome by reseting the molecular clock. However, their impacts at the genome level are on different scales. The hitchhiking effect has a local signature because recombination breaks down linkage disequilibrium between sites on the genome; consequently, the locus completely linked to a site under a positive selection becomes homogenous in the population [1]. In contrast, after relatively quick recovery of a population from a severe bottleneck, it becomes genome-wide homogeneous. Identifying and differentiating recent such events at a single locus can be challenging because both processes have similar signature on the genetic diversity at single locus. Thus, multi-locus DNA sequence data can be a powerful source for this purpose.

After a recent homogenization event at a neutral locus, the accumulated genetic diversity at the locus and the elapsed time are positively correlated when assuming constant molecular clock. To quantify the relation between genetic diversity at the neutral locus in a low recombination region and the time elapsed since a recent homogenization event, Griffiths [2], Tajima [3], and Perliz and Stephan [4] used Wright-Fisher reproduction model with constant population size and infinite-sites model [5] for the dynamics of genetic diversity at the locus. They derived analytical expressions for the expected number of polymorphic sites in a sample of DNA sequences from such a locus. Although this framework is computationally efficient for inferring the elapsed time, it is applicable only for a single locus.

Simulation based inference methods have been developed for the same problem to include an exponential population growth model and full polymorphism data in samples of DNA sequences [6], [7]. Although such methods have flexibility to include more complex evolutionary scenarios, they are computationally more intense.

I consider the same setting as in [2], [3], and [4] to develop an analytical framework for identifying and differentiating recent homogenization events at multiple neutral loci in low recombination regions. The loci are considered to be evolving independently, for example, when the loci are on different chromosomes or on same chromosome but far apart. I derive an analytical expression for the probability distribution of the number of polymorphic sites in a sample of DNA sequences. Based on this, I described likelihood-ratio based tests for identifying and differentiating recent homogenization events at multiple loci. I apply the framework to two data sets. First, I use human DNA sequence

data to infer evolutionary history and origin of modern human populations. Second, I use HIV DNA sequences sampled from HIV-1-infected patients to infer the times of HIV seroconversions.

Methods

The population genetic model

Genetic diversity at a neutral locus, in a low recombination region, affected by a recent homogenization event is a result of mutations accumulated at the locus since the homogenization event. To model the dynamics of genetic diversity at such a locus after the homogenization event, this paper combines the infinite-sites mutation model and the Wright-Fisher reproduction model with constant population size. The parameters in the model represent the effective population size N, the elapsed time T since the last homogenization event, mutation rate μ per generation per sequence, and the (effective) generation time g.

Variation in a sample of DNA sequences drawn from a population evolving according to this model can be described as a combination of genealogical and mutation processes. The genealogical process traces ancestral lineages of the sample back in time until the recent homogenization event at time T and stops earlier if the most recent common ancestor of the sample is more recent than the homogenization event. When N is large and the time in this process is measured in N_ℓ generations, the genealogical process can be approximated by a coalescent process derived from the standard coalescent [8–12]. Here N_ℓ is a scaled population size at the locus, determined by N and the type of the chromosome on which the locus is located: N_ℓ is equal to N for the case of a haploid population; for a diploid population with $N/2$ males and $N/2$ females, N_ℓ is equal to $N/2$, $3N/2$, $2N$, or $N/2$ if the locus is on the Y, the X, the autosomal chromosome, or on the mitochondrial DNA, respectively. In this process the ancestral lineages of the sample are traced until time $t=T/(gN_\ell)$ and mutations are added on the branches of the genealogy as independent Poisson processes with rates equal to $\theta/2$, $\theta\equiv 2N_\ell\mu$. In the infinite-sites model, each mutation occurs at a nucleotide site that has not been mutated before.

Results

Probability distribution of the number of polymorphic sites in a sample of DNA sequences

Under the model described above, the probability distribution of the number of polymorphic sites S_n in a sample of n DNA sequences can be represented as

$$\mathrm{IP}_n(s|t,\theta)\equiv \mathrm{IP}(S_n=s|\theta,t)=\mathrm{IE}\left(\frac{(\theta L_n(t)/2)^s}{s!}e^{-\theta L_n(t)/2}\right), \quad (1)$$

where $L_n(t)$ is the total length of the genealogy of n sequences. This equation suggests that the probability can be expressed through the derivatives of the moment generating function $g_n(u,t)$ of $L_n(t)$, defined as $g_n(u,t)\equiv \mathrm{I}Ee^{-uL_n(t)}$:

$$\mathrm{IP}_n(s|t,\theta)=\frac{(-1)^s(\theta/2)^s}{k!}\frac{\partial^s g_n(u,t)}{\partial u^s}\bigg|_{u=\theta/2}. \quad (2)$$

Griffiths [2] derived an analytical expression for $g_n(u,t)$, but it can not be easily used to derive expressions for the derivatives of $g_n(u,t)$. In the following lemma, I derive an expression for $g_n(u,t)$, which allows easily to derive analytical expressions for the derivatives of $g_n(u,t)$. Note that this expression also allows to invert the moment generating function $g_n(u,t)$ and to derive an analytic expression for the density function of $L_n(t)$. The latter result is presented in the lemma of the Text S1.

Lemma 1 *The moment generating function $g_n(u,t)$ can be represented as*

$$g_n(u,t)=\phi_n(u,t)+\binom{n}{2}\sum_{i=2}^{n}\sum_{\substack{j=0\\ j\neq 1}}^{i-1}\frac{\mu_{i,j}^{(n)}(i+j-1)}{-(u+\frac{i+j-1}{2})}\xi(i,j,t,u), \quad (3)$$

where

$$\xi(i,j,t,u)\equiv \exp(-ti(u+(i-1)/2))-\exp(-tj(u+(j-1)/2)).$$

The coefficients $\{\mu_{i,j}^{(n)}\}$ are determined by the following recurrence relations with initial conditions:

$$\mu_{i,j}^{(n)}=\frac{(n-1)(n-2)}{(n-j)(n-i)}\mu_{i,j}^{(n-1)},\ 2\le i\le n-1, j=0, \text{ or } 2\le j\le i-1; \quad (4)$$

$$\mu_{n,k}^{(n)}=\sum_{\substack{j=0\\ j\neq 1}}^{k-1}\mu_{k,j}^{(n)}-\sum_{i=k+1}^{n-1}\mu_{i,k}^{(n)},\quad 2\le k\le n-1; \quad (5)$$

$$\mu_{2,0}^{(2)}=\frac{1}{2}, \quad (6)$$

$$\mu_{n,0}^{(n)}=\frac{(-1)^n}{n(n-1)}. \quad (7)$$

The prove of Lemma 1 is provided in the Text S1.

Note that the coefficients $\{\mu_{i,j}^{(n)}\}$ satisfy the following identities

$$\frac{1}{n(n-1)}=\sum_{\substack{j=0\\ j\neq 1}}^{n-1}\mu_{n,j}^{(n)}; \quad (8)$$

$$\frac{1}{n(n-1)}=\sum_{i=2}^{n}\mu_{i,0}^{(n)}; \quad (9)$$

$$\mu_{i,0}^{(n)}=\frac{(-1)^i\binom{n-2}{i-2}}{n(i-1)}. \quad (10)$$

The identities are used in the proof of Lemma 1, and also for identifying numerical instability issues with computation of $\mu_{i,j}^{(n)}$ based on (4)–(7) when using decimal approximations instead of exact computations. The proof of the identities can be done by combining mathematical induction with (4)–(7), the details not shown.

Expression (3) is used to derive expressions for the derivatives of $g_n(u,t)$, but for computational purposes they are modified to derive numerically stable expressions. The following procedure is applied

to the expressions to solve the instability issue: for each i, $i=2,\ldots,n$, the terms with factor $\exp(-ti(u+(i-1)/2))$ are combined together and the common term is factored out. For example, a numerically stable expression for $g_n(u,t)$ is

$$g_n(u,t)=-n(n-1)\left(Y_1(n,u)+\sum_{i=2}^{n}Y_2(i,n,u)\exp(-it(u+(i-1)/2))\right)+\exp\left(-nt\left(u+\frac{n-1}{2}\right)\right), \quad (11)$$

where

$$Y_2(i,n,u)=\begin{cases}\sum_{j=0j\neq 1}^{i-1}\mu_{i,j}^{(n)}D(i,j,u)-\sum_{k=i+1}^{n}\mu_{k,i}^{(n)}D(i,k,u) & \text{if } i<n\\ \sum_{j=0j\neq 1}^{i-1}\mu_{i,j}^{(n)}D(i,j,u) & i=n,\end{cases}$$

$$D(i,j,u)=\frac{i+j-1}{2u+i+j-1};$$

$$Y_1(n,u)=-\sum_{i=2}^{n}\mu_{i,0}^{(n)}D(i,0,u).$$

The numerical instability of the expression (3) is illustrated in Figure 1.

To derive a numerically stable expression for $\mathrm{IP}_n(s|t,\theta)$ by using (2), first expressions are derived for the derivatives of $g_n(u,t)$ with respect to u by using Lemma 1 and the identity

$$\frac{\partial^k}{\partial u^k}\frac{\exp(-bu)}{u+a}\equiv(-1)^k k!\exp(-bu)\sum_{i=0}^{k}\frac{b^{k-i}}{(k-i)!(a+u)^{i+1}}. \quad (12)$$

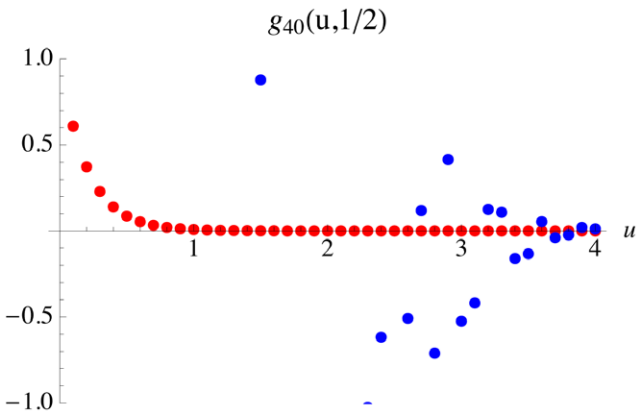

Figure 1. Illustration of numerical instability of the expression (3). The moment generating function $g_{40}(u,\frac{1}{2})$ is plotted for the same range of values of u in red and blue dots by using the expressions (11) and (3), respectively. The numerical instability of the expression (3) is obvious because the values of $g_{40}(u,\frac{1}{2})$ must be between 0 and 1 for any positive u.
doi:10.1371/journal.pone.0037588.g001

After applying the numerical stabilization procedure (described above) to these expressions, a numerically stable expression for the probability distribution is

$$\mathrm{IP}_n(s|t,\theta)=\frac{n(n-1)}{\theta}\left(\frac{\sum_{i=2}^{n}\mu_{i,0}^{(n)}\frac{i-1}{\left(\frac{i-1}{\theta}+1\right)^{s+1}}+\sum_{i=2}^{n}B(s,i,n,t,\theta)\exp(-i(i-1+\theta)t/2)}{}\right)+\frac{(\theta nt/2)^s}{s!}\exp(-nt(n-1+\theta)/2), \quad (13)$$

where

$$B(k,i,n,t,\theta)=\begin{cases}\sum_{r=i+1}^{n}\mu_{r,i}^{(n)}C(k,i,r,t,\theta)-\sum_{j=0j\neq 1}^{i-1}\mu_{i,j}^{(n)}C(k,i,j,t,\theta)\\ \text{if } 2<n \text{ and } i<n\\ -\sum_{j=0j\neq 1}^{i-1}\mu_{i,j}^{(n)}C(k,i,j,t,\theta)\\ \text{if } n=2 \text{ or } i=n;\end{cases}$$

$$C(k,i,j,t,\theta)=\frac{i+j-1}{\left(\frac{i+j-1}{\theta}+1\right)^{k+1}}\sum_{m=0}^{k}\frac{\left(\frac{it(i+j-1+\theta)}{2}\right)^m}{m!}.$$

I have implemented the formula (13) and all the other formulas in this paper in a program written in *Mathematica* [13]. The program is used to carry out all the calculations and visualizations in this paper. The program uses *Mathematica*'s ability of doing exact computations with fractions, as a result avoiding numerical instability issues. The program is available from the author on request.

Note that when θ is large, the following approximation holds for the probability distribution $\mathrm{IP}_n(s|t,\theta)$:

$$\mathrm{IP}_n(s|t,\theta)\approx\frac{(nt\theta/2)^s}{s!}\exp(-nt\theta/2). \quad (14)$$

The right side of the approximation corresponds to the probability distribution of the number of polymorphic sites in a sample of n sequences under a "simple" model where the ancestral lineages of the sample are traced until time t without coalescence. Note that population size N is not a factor in this model because the right side of the above approximation can be represented as

$$\frac{(n\mu T/g)^s}{s!}\exp(-n\mu T/g).$$

Mean and variance of the number of polymorphic sites in a sample of DNA sequences

In previous studies [2–4], expressions have been derived for the mean number of polymorphic sites in a sample of DNA sequences from a locus affected by a recent homogenization event. An expression for the variance is also derived in [4], but this expression is implicit because it includes integral expressions. Using a similar approach as in the previous section, I derive a numerically stable expressions for computing the mean and variance of the number of polymorphic sites in a sample of DNA sequences from such a locus. The conditional probability distribution of S_n when $L_n(t)$ is given is Poisson with a mean of

$\theta L_n(t)/2$. The mean and variance of S_n can be expressed as follows:

$$\mathrm{IE}(S_n)=\frac{\theta}{2}\mathrm{IE}L_n(t)$$

and

$$\mathrm{V}ar(S_n)=\frac{\theta}{2}\mathrm{IE}L_n(t)+\frac{\theta^2}{4}\left(\mathrm{IE}L_n^2(t)-(\mathrm{IE}L_n(t))^2\right).$$

Expressions for the first and second moments of $L_n(t)$ are derived by taking the first and second derivatives of (3) with respect to u and evaluating them at $u=0$. After applying the numerical stabilization procedure (described in the previous section) to these expressions, the first and second moments of $L_n(t)$ can be computed using the following formulas:

$$\mathrm{IE}L_n(t)=\binom{n}{2}\left(\sum_{i=2}^{n}\mu_{i,0}^{(n)}\frac{4}{i-1}+\sum_{i=2}^{n}A(i,n,t)\exp(-i(i-1)t/2)\right)+ nt\exp(-n(n-1)t/2),$$

where $A(i,n,t)$ is defined as

$$A(i,n,t)=\begin{cases}-\sum_{j=2}^{i-1}\mu_{i,j}^{(n)}Q(i,j,t)+\sum_{k=i+1}^{n}\mu_{k,i}^{(n)}Q(i,k,t)-\mu_{i,0}^{(n)}Q(i,0,t)\\ \text{if } 2<i<n\\ \sum_{k=i+1}^{n}\mu_{k,i}^{(n)}Q(i,k,t)-\mu_{i,0}^{(n)}Q(i,0,t)\\ \text{if } i=2\\ -\sum_{j=2}^{i-1}\mu_{i,j}^{(n)}Q(i,j,t)-\mu_{i,0}^{(n)}Q(i,0,t)\\ \text{if } i=n;\end{cases}$$

$$Q(i,j,t)=\frac{4}{i+j-1}+2it;$$

$$\mathrm{IE}L_n^2(t)=\binom{n}{2}\left(\sum_{i=2}^{n}\mu_{i,0}^{(n)}\frac{16}{(i-1)^2}+\sum_{i=2}^{n}H(i,n,t)\exp(-i(i-1)t/2)\right)+ 4n^2t^2\exp(-n(n-1)t/2),$$

where $H(i,n,t)$ is

$$H(i,n,t)=\begin{cases}-\sum_{j=2}^{i-1}\mu_{i,j}^{(n)}R(i,j,t)+\sum_{k=i+1}^{n}\mu_{k,i}^{(n)}R(i,k,t)-\mu_{i,0}^{(n)}R(i,0,t)\\ \text{if } 2<i<n,\\ \sum_{k=i+1}^{n}\mu_{k,i}^{(n)}R(i,k,t)-\mu_{i,0}^{(n)}R(i,0,t)\\ \text{if } i=2,\\ -\sum_{j=2}^{i-1}\mu_{i,j}^{(n)}R(i,j,t)-\mu_{i,0}^{(n)}R(i,0,t)\\ \text{if } i=n,\end{cases}$$

$$R(i,j,t)=2i^2t^2+\frac{8it}{i+j-1}+\frac{16}{(i+j-1)^2}.$$

Three tests for identifying and differentiating recent homogenization events at multiple loci

Using the probabilistic framework developed above, three likelihood-ratio based tests are considered in this section for identifying and differentiating recent homogenization events at independently evolving multiple neutral loci in low recombination regions.

Test I. To identify a recent homogenization event at a locus based on the number of polymorphic sites in a sample of DNA sequences, the hypothesis $H_0: T=\infty$ versus $H_a: T<\infty$ is considered. The null hypothesis $T=\infty$ represents a case in which ancestral population was evolving according to the Wright-Fisher model with constant population size. The null hypothesis can be tested by defining minus twice of the log of the likelihood-ratio statistics as

$$\Lambda_\infty(s)\equiv-2\log\left(\frac{\mathrm{I}P_n(s|t=\infty,\theta)}{\sup_{0<t\le\infty}\mathrm{I}P_n(s|t,\theta)}\right),$$

and comparing it with a χ^2 distribution with d.f. $=1$.

Note that $\mathrm{I}P_n(s|t=\infty,\theta)$ corresponds to the probability distribution of the number of polymorphic sites in a sample of DNA sequences when the genealogy of the sample is modeled by the standard coalescent and assumed the infinite-sites model for mutations. Tavaré [14] derived an expression for $\mathrm{I}P_n(s|t=\infty,\theta)$, which also follows from (13) by taking t to ∞:

$$\mathrm{IP}_n(s|t=\infty,\theta)=\lim_{t\to\infty}\mathrm{IP}(s|t,\theta)=\frac{n(n-1)}{\theta}\left(\sum_{i=2}^{n}\mu_{i,0}^{(n)}\frac{i-1}{\left(\frac{i-1}{\theta}+1\right)^{s+1}}\right).(15)$$

Test II. Suppose we know, for example, from other studies, that a recent homogenization event occurred at time T_0 and we want to identify if this event had impact on a locus of interest. Symbolically, the following hypothesis can be stated

$$H_0: T=T_0 \text{ versus } H_a: T\neq T_0.$$

The null hypothesis can be tested by comparing minus twice of the likelihood-ratio statistics

$$\Lambda_0(s)=-2\log\left(\frac{\mathrm{IP}_n(s|t_0,\theta)}{\sup_{0<t\le\infty}\mathrm{IP}_n(s|t,\theta)}\right),$$

with a χ^2 distribution with d.f. $=1$, where $t_0=T_0/(gN_\ell)$.

Based on this approximation, for each s a $(1-\alpha)\%$ confidence interval

$$(\bar{t}_\alpha(s),\ \underline{t}_\alpha(s)),$$

of t is determined by solving the equation

$$-2\log\left(\frac{\mathrm{IP}_n(s|t,\theta)}{\mathrm{IP}_n(s|\hat{t},\theta)}\right)=\chi_\alpha^2$$

with respect to t; χ_α^2 is the α critical value of the χ^2 distribution. Note that when s is 0, then $\underline{t}_\alpha=0$ and $\bar{t}_\alpha$ is the solution of the above equation. A $(1-\alpha)\%$ confidence interval of T is $(gN_\ell(\bar{t}_\alpha(s),\ gN_\ell\underline{t}_\alpha(s))$. One can use a similar approach to estimate a $(1-\alpha)\%$ confidence interval for T when inferring the elapsed time of a recent severe bottleneck event based on DNA sequence

data from independently evolving multiple neutral loci. Another approach for this case is described below.

Test III. For a case of m independent neutral loci, let the loci be labeled from 1 to m, and s_i be the number of polymorphic sites in a sample of n_i sequences at locus i, $\boldsymbol{s}=(s_1,\ldots,s_m)$. To test if the multiple loci are affected by the same recent homogenization event, the following hypothesis is considered:

$$H_0 : T_1 = T_2 = \ldots = T_m \text{ versus } H_a : T_i \neq T_j$$
$$\text{for some } i \neq j;\ i,j = 1,\ldots,m,$$

where T_i is the time elapsed since a recent homogenization event at locus i. The null hypothesis can be tested by comparing the statistics

$$\Lambda_m(\mathbf{s}) = -2\log\left(\frac{\sup\{\Pi_{i=1}^m \mathrm{I}P_{n_i}(s_i|T/(gN_i),\theta_i) : T>0\}}{\Pi_{i=1}^m \sup\{\mathrm{I}P_{n_i}(s_i|T/(gN_i),\theta_i) : T>0\}}\right),$$

with a χ^2 distribution with d.f. $=m-1$, where N_i is the scaled population size at locus i; $\theta_i = 2N_i\mu_i$ is the scaled mutation rate at locus i, and μ_i is the mutation rate per generation per sequence at locus i.

Inferring the time of a recent severe bottleneck event based on polymorphism data at multiple loci

The following steps can be taken to infer the time T of a recent severe bottleneck event from DNA sequence data at independently evolving multiple neutral loci in low recombination regions. The likelihood function for such a data set can be computed as a product of likelihood functions from each locus by using formula (13). Thus, in case of m independent loci, and s_i polymorphic sites in a sample of n_i sequences at locus i, $i=1,\ldots,m$, the maximum likelihood estimator $\hat{\boldsymbol{T}}$ of T can be derived by solving the equation

$$\sum_{i=1}^{m} \frac{\dot{\mathrm{I}}P_{n_i}\left(s_i|\hat{\boldsymbol{T}}/(gN_i),\theta_i\right)}{\mathrm{I}P_{n_i}\left(s_i|\hat{\boldsymbol{T}}/(gN_i),\theta_i\right)} = 0$$

with respect to $\hat{\boldsymbol{T}}$, where $\dot{\mathrm{I}}P_{n_i}(s_i|t,\theta_i)$ is the derivative of $\mathrm{I}P_{n_i}(s_i|t,\theta_i)$ with respect to t. It is assumed that the scaled mutation rate θ_i and the scaled population size N_i at locus i, $i=1,\ldots,m$, are known.

To estimate a $(1-\alpha)\%$ confidence interval of T, the Central Limit Theorem based approximation can be used when the following conditions hold: (1) The number m of the loci is large; (2) the loci are on same type of chromosomes (as a result $N_i = N_j$); (3) samples of DNA sequences from each locus have the same size ($n_i = n_j$); (4) the lengths of the sequences from the loci are equal ($L_i = L_j$). Thus, the $(1-\alpha)\%$ confidence interval of T can be computed as

$$\left(\hat{\boldsymbol{T}} - z_\alpha\sqrt{J_m(\hat{\boldsymbol{T}})},\ \hat{\boldsymbol{T}} + z_\alpha\sqrt{J_m(\hat{\boldsymbol{T}})}\right),$$

where z_α is the α critical value from the standard normal distribution; $J_m(\cdot)$ is observed Fisher information, which can be computed using the formula

$$J_m(\hat{\boldsymbol{T}}) \equiv -\sum_{i=1}^{m} \frac{\partial^2 \log(\mathrm{I}P_n(s_i|T/(gN_\ell),\theta)}{\partial T^2}\Big|_{T=\hat{\boldsymbol{T}}}$$

$$= -\sum_{i=1}^{m} \frac{1}{(gN_\ell)^2}\left(\frac{\ddot{\mathrm{I}}P_n(s_i|\hat{\boldsymbol{T}}/(gN_\ell),\theta)}{\mathrm{I}P_n(s_i|\hat{\boldsymbol{T}}/(gN_\ell),\theta)} - \frac{\dot{\mathrm{I}}P_n^2(s_i|\hat{\boldsymbol{T}}/(gN_\ell),\theta)}{\mathrm{I}P_n^2(s_i|\hat{\boldsymbol{T}}/(gN_\ell),\theta)}\right).$$

For evaluating the above expression, numerically stable expressions for the first and second derivatives of $\mathrm{I}P_n(s|t,\theta)$ with respect to t can be derived by using (13) and the numerical stabilization procedure.

Application of the method for inferring recent homogenization events from human genome

Anthropological and archeological data strongly support "Out of Africa" hypothesis for the origin and evolutionary history of modern humans [15–19]. The hypothesis underlies two major events: *Homo sapiens* (ancestors of modern humans) emerged in Africa between 150,000 and 200,000 years ago (kya) and dispersed to other regions of the world sometimes before 50,000 years before present (yr B.P.). Studies based on mitochondrial and Y-chromosome support this hypothesis [20–28]. However, studies based on DNA sequence data from coding and non-coding loci on autosomal and X-chromosome show that the most recent common ancestor of β-globin gene [29], the X chromosome gene for the pyruvate dehydrogenase E1 α-subunit [30], and the non-coding loci 22q11.2 [31], 17q23 [32], Xq13.3 [33] are much older than 200,000 yr B.P. These inferences are based on the framework of the standard coalescent, in which the effective human population size and the mutation rate per nucleotide site per generation are considered to be ≈ 10000 [34] and $\approx 2.3\times10^{-8}$ [35–37], respectively.

In contrast to this approach, I use the framework developed in this paper to analyze some of data sets used in the studies mentioned above. I apply the framework to DNA sequences from four non-coding loci (22q11.2, 17q23, Xq13.3, YAP) in low recombination regions on chromosomes 22, 17, X, and Y to identify and differentiate recent homogenization events associated

Table 1. Summary of the DNA sequence data sets from loci 22q11.2, 17q23, Xq13.3, and YAP.

Locus	seq. length ($\approx$kb)[a]	$S_n(n)$[b] African	$S_n(n)$ Non-African	$S_n(n)$ Combined
22q11.2	10	54(40)	44(88)	75(128)
17q23	20	57(10)	37(12)	63(22)
Xq13.3	10	24(23)	17(46)	33(69)
YAP	2.6	3(8)	1(7)	3(15)

[a]Sequence length in kilobases.
[b]The number of polymorphic sites in a sample of n sequences.
doi:10.1371/journal.pone.0037588.t001

Table 2. Estimates for the elapsed times T since a recent homogenization event for each of the four loci.

Loci	$\hat{T}^a$ $(N=10^4)$	$\hat{T}$(95%,CI)^b $(N=5\times10^4)$ African	$\hat{T}_S{}^c$ $(N=\infty)$	$\hat{T}^a$ $(N=10^4)$	$\hat{T}$(95% CI)^b $(N=5\times10^4)$ Non-African	$\hat{T}_S{}^c$ $(N=\infty)$
22q11.2	1400	220(120, 380)	117	640	72 (44, 116)	43
17q23	800	320 (220, 500)	247	360	160 (105, 260)	134
Xq13.3	525	127 (71, 240)	90	120	41 (22, 75)	32
YAP	350	195(40, ∞)	125	94	55(0, 450)	30
Combined	1300	241(183, 316)	145	360	92 (67, 125)	55

[a]The estimates of the elapsed times are in 1000 years Before Present. The estimate of T are based on formula (13) when N is equal to 10000.
[b]The estimate of T based on formula (13) when N is equal to 50000.
[c]Under the "simple" model an estimator for T is denoted as $\hat{T}_S$. It is equal to $\frac{S_n\times g}{n\times\mu}$ based on the data at a single locus; for data sets from m independent loci, it is equal to $\hat{T}_S=\frac{g\sum_{i=1}^{m}S_{n_i}}{\sum_{i=1}^{m}n_i\mu_i}$.
doi:10.1371/journal.pone.0037588.t002

with the "Out of Africa" hypothesis. The data sets are published in [25,31–33], respectively, and their summary is in Table 1. First, I consider commonly accepted estimates for values of the parameters in the model: the effective human population size N to be 10000; the mutation rate $\hat{\mu}$ per nucleotide site per generation to be 2.3×10^{-8}; the human (effective) generation time g to be 20 years. Mutation rate μ per generation per sequence at each locus is computed as $l\times\hat{\mu}$, where l is the length of the DNA sequences at the locus. After applying Test I for this set of parameter values to each of the four data sets, the power of detecting a recent homogenization event at any of the loci is very weak (the p-values close to 1, data not shown). In this case the maximum likelihood estimates for the elapsed times of recent homogenization events at the loci are much older than 200,000 yr B.P. (Table 2). Thus, these estimates disagree with the "Out of Africa" hypothesis.

To explore another possibility, I also consider human effective population size N to be 50000 based on the following observations: (1) Some studies [38–40] estimated effective human population size to be a few times larger than 10000. (2) Maximums of the likelihood functions of the data sets favor the case $N=50000$ over the case $N=10000$ for all the data sets. Thus, I consider the values of $\hat{\mu}$ and g to be the same as above but $N=50000$. After applying Test I to the data sets from each locus, the likelihood-ratio tests rejected the null hypotheses at 5% significance level, the results are in Table 3. Clearly, the results suggest that the standard coalescent framework is inadequate to describe the data sets for this set of parameter values, and recent homogenization events have impact on the four loci. The maximum likelihood estimates (see Table 2) of the elapsed times agree with the times for the two major events.

In this case, the likelihood functions of the data sets would not change dramatically as N gets larger than 50000 because they behave in large θ regime. The maximum likelihood estimates of T, when $N=\infty$, are in Table 2. These estimates show that considering the human effective population size greater than 50,000, the estimates for the elapsed times would not change dramatically.

Table 3. The values of minus twice of the log of likelihood-ratio statistics for the data sets from each of the four loci.

Locus	$\Delta_\infty(\cdot)$(p-value) African	$\Delta_\infty(\cdot)$(p-value) Non-African
22q11.2	19.8 (8.5e−6)	48.6 (0.3e−11)
17q23	11.2 (0.8e−3)	19.8 (0.8e−5)
Xq13.3	20.3 (7e−6)	46.6 (0.8e−11)
YAP	2(0.16)	4.6(0.03)

doi:10.1371/journal.pone.0037588.t003

For this set of parameter values, I use Test III to differentiate the recent homogenization events at the four loci. The results of the tests are in Table 4. These results suggest that the four loci have not been affected by the same homogenization event, p-values are less than 0.05 for the data sets from African and Non-African populations. The locus Xq13.3 is significantly younger than the locus 17q23, in particular for Non-African population, which suggest that the locus Xq13.3 has been affected by a recent positive selection or a recent bottleneck occurred to Non-African female population. Using Tajima's test [41] and Fu's and Li's tests [42], Zhao at el. [31] also observed that the diversity at locus Xq13.3 significantly deviates from the Wright-Fisher neutral model.

Application of the method for inferring the times of HIV seroconversions in HIV-1-infected patients

Usually, after few weeks of HIV infection, plasma viraemia in infected patient declines rapidly as a result of a primary immune response, which coincides with HIV seroconversion [43], [44]. In particular, HIV envelop gene at this time point shows no diversity [45]. To examine the utilities of the framework developed in this paper, I use DNA sequence data from HIV-1 envelop genes published in [46] to infer the times of HIV seroconversions in nine HIV-1-infected patients. The sequences are sampled from the patients at the first HIV-positive screening tests. The sequences are $\approx$650 nucleotide long; a summary of the data is in Table 5.

For consistency of the data sets with the infinite-sites mutation model and with no intra-locus recombination, the following conditions are checked: (a) Each polymorphic site is a result of a single mutation event, that is only two nucleotide states are possible at each polymorphic site. (b) All pairs of sites in sample of DNA sequences pass the four-gamete test [47–49]. Seven of the nine data sets (except data sets from patients 2 and 5) satisfy conditions (a) and (b). The data sets from patients 2 and 5 are inconsistent with the conditions (a) and (b), respectively. However, the two data sets are not excluded from the analysis because inconsistencies in these data sets are a result of two mutations and some recombination events, respectively.

I consider the following values for the parameters in the model: population size N equal to the viral load at the sampling time

Table 4. The values of minus twice of log of likelihood-ratio statistics for Test III.

Compared loci[a]	$\Delta_i()$ (p-value) African	$\Delta_i()$ (p-value) Non-African
(22q11.2, Xq13.3)	1.7 (0.19)	2 (0.15)
(17q23, 22q11.2)	1.3 (0.25)	5.6 (0.02)
(Xq13.3, 17q23)	5.8 (0.02)	12.5 (0.0004)
(YAP, 17q23)	0.3 (0.6)	0.9 (0.3)
(YAP, Xq13.3)	0.2 (0.65)	0.08 (0.8)
(YAP, 22q11.2)	0.01 (0.9)	0.06 (0.8)
(Xq13.3, 17q23, 22q11.2)	6 (0.05)	14 (0.001)
(YAP, Xq13.3, 17q23, 22q11.2)	53 (1.8e−11)	14 (0.003)

[a]Sets of compared loci.

doi:10.1371/journal.pone.0037588.t004

point, mutation rate $\hat{\mu}$ per nucleotide site per generation equal to 3×10^{-5}, the number of nucleotides at the locus l is equal to 650. All the insertions and deletions are excluded from the analysis. For this set of parameter values, I applied Test I to the data from each patient; the null hypotheses are rejected at 5% significance level in favor of recent homogenisation events. For each patient the maximum likelihood and 95% confidence interval estimates of T (in coalescent units) are in Table 6. These estimates can be converted in years by using the equation $T=t\times g\times N$, in which the effective HIV generation time g is considered to be equal to 1 or 2 days [50–52].

These estimates are contrasted with the estimates provided by Shankarappa et. al [46]; they estimated the time of HIV seroconversion for each of the patients as the mid-time point between the last HIV-negative and first HIV-positive screening tests. The comparison between the estimates (see Figure 2) shows that for some of the data sets the estimates are significantly in disagreement.

The observed disagreements are robust with respect to N (data not shown): when N is larger than the viral load, the likelihood functions do not change because of large θ regime. I have also applied the above estimation method by considering N equal to the one tenth of the viral loads. The result show that the observed discrepancies also hold for this case. Note that the viral load represents approximately $1/5000th$ of the total amount of the virus in an HIV-infected person since there is a total of 5 liters of blood in the body of an average adult.

Discussion

The analytical method developed in this paper is a trade-off between computational efficiency and complexity of the underlying evolutionary model. Using multi-locus DNA sequence data, the method allows identification and differentiation of the signatures of recent severe bottleneck and hitchhiking effects in a computationally efficient way. However, the method uses the number of polymorphic sites instead of full polymorphism data in samples of DNA sequences, and it is constrained by the assumptions of the constant size Wright-Fisher reproduction model and the infinite-sites model. In contrast, coalescent based simulation methods can be implemented at the cost of computational feasibility to include full polymorphism data [7], various demographic scenarios [6], and finite-sites mutation models [53]. However, before using computationally more expensive methods, the method could be a helpful guide for analyzing multi-locus DNA sequences data.

To illustrate the behavior of the likelihood function for small and large θ, I used the program to plot the likelihood function of t (see Figure 3) for a sample of 15 DNA sequences with 25 polymorphic sites when $\theta=.96$ and $\theta=19.2$. The behavior of the likelihood function can be explained based on the process that traces ancestral lineages of the sample back in time. When tracing n lineages back in time, coalescent and mutation events occur one at a time with rates $n(n-1)/2$ and $n\theta/2$, respectively. Thus, when θ is large, mutation events occur more often than coalescent events back in time, so for a given number of polymorphic sites the recent homogenization event is more likely to be before the most recent common ancestor of the sample. This also explains the approximation (14). In opposite to this, when θ is small, the sample is more likely to have the most recent common ancestor before the recent homogenization event. Similarly, as t gets larger the sample is more likely to have the most recent common ancestor before the homogenization event, hence the likelihood function has a limit (see (15)).

Table 5. Summary of Shankarappa et al's [46] data.

Patient number[a]	seroconverion time (in years)[b]	sample size[c]	number of polymorphic sites[d]	viral load[e]
1	0.28	7	7	6637
2	0.42	21	33	68706
3	0.35	10	9	598
5	0.25	22	41	7798
6	0.21	19	21	4709
7	0.2	19	32	6251
8	0.29	7	5	4045
9	0.25	10	31	145545
11	0.21	8	13	478

[a]I used the same notation for the patients as in [46].
[b]In [46] seroconversion times in the patients are estimated as the mid-time point between the last HIV negative screening test and the first HIV positive screening test.
[c]For each patient, the samples of DNA sequences are drawn from HIV populations in HIV patients at the time of the first HIV positive screening test.
[d]The number of polymorphic sites in the samples.
[e]For each patient viral load per milliliter is measured at the time of the first HIV positive screening test.

doi:10.1371/journal.pone.0037588.t005

Table 6. The estimates of the seroconversion times ($\hat{t}$ in coalescence units) in the nine patients.

Patient	$\hat{t}^a$ (95%*CI*)	$\Delta_\infty(\cdot)$	*p*-value
1	0.008 (0.003, 0.016)	36.6	1.2e−9
2	0.0012 (0.0008, 0.0017)	162.1	0
3	0.093 (0.04, 0.21)	13.4	2.5e−4
5	0.013 (0.009, 0.018)	75.3	0
6	0.013 (0.008, 0.02)	65.6	5.5e−16
7	0.015 (0.01, 0.02)	64.4	9.9e−16
8	0.009 (0.003, 0.02)	33.5	7.1e−9
9	0.001 (0.00075, 0.0015)	88.9	0
11	0.23 (0.08, 0.58)	6.2	0.012

[a]The maximum likelihood estimates of T in coalescent units.
doi:10.1371/journal.pone.0037588.t006

Computational efficiency of the method gives an advantage to explore various values for the parameters in the model for assessing the impact of parameter values on the inference. The application of the method to the human data shows that when the effective human population size N is equal to 10000 or greater than 50000, the inferences about evolutionary history of modern human populations are dramatically different. The HIV data analysis shows that the observed discrepancies between estimates for HIV seroconversions in the patients can be a result of the assumption that the effective HIV generation time is the same for all the patients. To have a better assessment for this assumption, frequent HIV screening tests can be used to assess the times of HIV seroconversion in HIV patients, and then to apply this method for exploring variability of effective HIV-1 generation times between HIV patients.

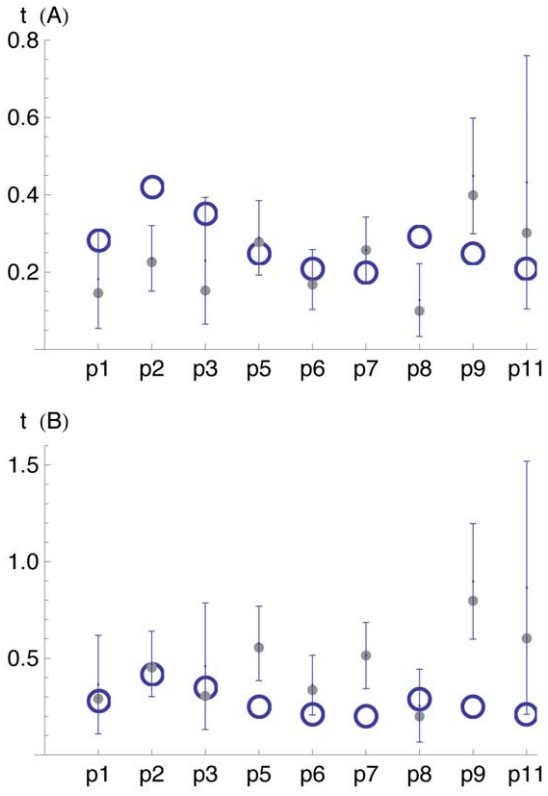

Figure 2. Comparison of two estimates of the seroconversion time for each of the nine patients. The effective generation time g in (A) and (B) are considered to be $g=1$ day and $g=2$ days, respectively. Maximum likelihood and 95% confidence interval estimates of the time of HIV seroconversion in years since the first HIV positive screening test are shown in full dots and error bars, respectively. Empty circles represent the mid-point estimates of the seroconversion times [46].
doi:10.1371/journal.pone.0037588.g002

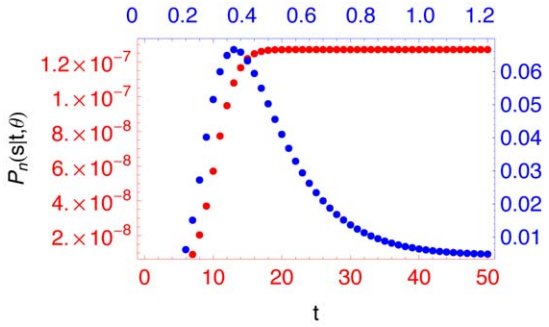

Figure 3. The likelihood function of t for two values of θ. For a sample of 15 DNA sequences with 25 polymorphic sites at a locus, the likelihood function of the elapsed time (t) is plotted for the values of $\theta=0.96$ and $\theta=19.2$ in red and blue, respectively.
doi:10.1371/journal.pone.0037588.g003

As the analysis of the human DNA sequences data shows the method developed in this paper does not have enough power to give an estimate for the effective human population size. Although the method suggest that very large values of N as maximum likelihood estimates for some of the human data sets when T and N are considered unknown, this does not mean that the "simple" model ($N=\infty$) is an appropriate model for explaining the data sets because site frequency spectrum of a sample of DNA sequences under the simple model consists only singletons, and Zhao et al. [31] observed excess number of singletons and doubletons for all the data sets. Note that under the model considered in this paper the behavior of the expected site frequency spectrum in samples of DNA sequences changes continuously respect to θ, for example when the effective population size N changes continuously. The two extreme ends of the expected site frequency spectrum under this model are described by the standard coalescent and by the "simple" model, respectively for small and very large values of θ. Under the standard coalescent the expected site frequency spectrum represents a wide range for frequencies of alleles. Thus, as θ (N) increases the expected number of low-frequency alleles increases.

Supporting Information

Text S1
(PDF)

Acknowledgments

I would like to thank the anonymous reviewer #1 for his/her helpful suggestions. This work is dedicated to Professor Simon Tavaré on the occasion of his 60th birthday.

Author Contributions

Conceived and designed the experiments: OS. Performed the experiments: OS. Analyzed the data: OS. Contributed reagents/materials/analysis tools: OS. Wrote the paper: OS.

References

1. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
2. Griffiths R (1981) Transient distribution of the number of segregating sites in a neutral infinite-sites model with no recombination. J Appl Prob 18: 42–51.
3. Tajima F (1989) The effect of change in population size on DNA polymorphism. Genetics 123: 597–601.
4. Perlitz M, Stephan W (1997) The mean and variance of the number of segregating sites since the last hitchhiking event. J Math Biol 36: 1–23.
5. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7: 256–76.
6. Jakobsson M, Hagenblad J, Tavaré S, Säll T, Halldén C, et al. (2006) A unique recent origin of the allotetraploid species Arabidopsis suecica: Evidence from nuclear DNA markers. Mol Biol Evol 23: 1217–31.
7. Galtier N, Depaulis F, Barton NH (2000) Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155: 981–7.
8. Kingman JFC (1982) On the genealogy of large populations. Journal of Applied Probability 19A: 27–43.
9. Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13: 235–48.
10. Kingman JFC (1982) Exchangeability and the evolution of large populations. In: Koch G, Spizzichino F, eds. Exchangeability in Probability and Statistics, North Holland Publishing Company. pp 97–112.
11. Hudson R (1983) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37: 203–17.
12. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–60.
13. Wolfram Research, Inc. (2007) Mathematica. Version 6.0, Champaign, IL.
14. Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 26: 119–64.
15. Stringer C (2002) Modern human origins: progress and prospects. Phil Trans R Soc Lond B 357: 563–79.
16. Mellars P (2004) Neanderthals and the modern human colonization of Europe. Nature 432: 461–5.
17. Forster P (2004) Ice ages and the mitochondrial DNA chronology of human dispersals: a review. Phil Trans R Soc Lond B 359: 255–64.
18. Forster P, Matsumura S (2005) Did early humans go north or south? Science 308: 965–6.
19. Mellars P (2006) A new radiocarbon revolution and the dispersal of modern humans in Eurasia. Nature 439: 931–5.
20. Cann R, Stoneking M, Wilson A (1987) Mitochondrial DNA and human evolution. Nature 325: 31–6.
21. Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708–13.
22. Maca-Meyer N, Gonzalez A, Larruga J, Flores C, Cabrera V (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2: 13.
23. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson A (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503–7.
24. Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6: 165–83.
25. Hammer MF (1995) A recent common ancestry for Human Y chromosomes. Nature 378: 376–8.
26. Pritchard J, Seielstad M, Perez-Lezaun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–8.
27. Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. Proc Natl Acad Sci USA 97: 7360–5.
28. Jobling M, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nature Rev Genet 4: 598–612.
29. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, et al. (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60: 772–89.
30. Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96: 3320–4.
31. Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, et al. (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. PNAS 97: 11354–8.
32. Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nat Genet 22: 59–62.
33. Kaessmann H, Heißig F, von Haeseler A, Pääbo S (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22: 78–81.
34. Takahata N (1993) Allelic genealogy and human evolution. Mol Biol Evol 10: 2–22.
35. Kondrashov AS, Crow JF (1993) A molecular approach to estimating the human deleterious mutation rate. Hum Mutat 2: 229–34.
36. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667–86.
37. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.
38. Ayala FJ (1996) HLA sequence polymorphism and the origin of humans. Science 274: 1554.
39. Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270: 1930–6.
40. Sherry ST, Harpending HC, Batzer MA, Stoneking M (1997) Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics 147: 1977–82.
41. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–95.
42. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693–709.
43. Weber J (2001) The pathogenesis of HIV-1 infection. Br Med Bull 58: 61–72.
44. Ariyoshi K, Harwood E, Chiengsong-Popov R, Weber J (1992) Is clearance of HIV-1 viraemia at seroconversion mediated by neutralising antibodies? The Lancet 340: 1257–8.
45. Holmes EC, Zhang L, Simmonds P, Ludlam C, et al. (1992) Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc Natl Acad Sci USA 89: 4835–9.
46. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. Journal of Virology 73: 10489–502.
47. Buneman P (1971) The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall D, Tautu P, eds. Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press. pp 387–95.
48. Hudson R, Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147–64.
49. Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. Networks 21: 19–28.
50. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 271: 1582–6.
51. Perelson AS, Nelson PW (1999) Mathematical analysis of HIV-1 dynamics in vivo. SIAM Rev 41: 3–44.
52. Rodrigo AG, Shpaer EG, Delwart EL, Iversen AK, Gallo MV, et al. (1999) Coalescent estimates of HIV-1 generation time in vivo. Proc Natl Acad Sci USA 96: 2187–91.
53. Yang Z (1996) Statistical properties of a DNA sample under the finite-sites model. Genetics 144: 1941–50.