Journal of Theoretical Biology 357 (2014) 55–61

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The limiting distribution of the effective population size of the ancestor of humans and chimpanzees

Carlos G. Schrago*

Universidade Federal do Rio de Janeiro, Instituto de Biologia, Departamento de Genética, CCS, A2-092, Rua Prof. Rodolpho Paulo Rocco, SN Cidade Universitária, Rio de Janeiro CEP 21941-617, Brazil

HIGHLIGHTS

- The variance of the effective population size of the *Homo–Pan* ancestor among studies is large.
- Estimation of evolutionary parameters from genomes contains minimum stochastic error.
- The mean ancestral effective population size of *Homo–Pan* was inferred at approximately 47,500.
- The uncertainty of the estimates was large, even under the limiting distribution.

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form
25 April 2014
Accepted 5 May 2014
Available online 14 May 2014

Keywords:
Species tree
Primate evolution
Speciation
Coalescence

ABSTRACT

The effective population size is a fundamental parameter for the understanding of microevolutionary process. Indeed, the consideration of population-level phenomena within phylogenies provides insight into the influence of the past evolutionary demography on the genetic diversity of living species. Although the effective population size of the last common ancestor of humans and chimpanzees has been extensively investigated by molecular evolutionists, variance in the estimates of this parameter among studies is large. However, with the availability of genome sequences, the estimation of evolutionary parameters may be conducted with minimum stochastic error, and the limiting distribution of the estimates may be obtained. This statistical property was utilized in the present study and coupled with analytical derivations from the coalescent theory to examine the limiting distribution of the ancestral effective population size of *Homo–Pan*. The mean ancestral effective population size of *Homo–Pan* was inferred at approximately 47,500, and the results showed that the uncertainty of the estimates was large, even under the limiting distribution. Further reductions of the estimates are feasible only if additional calibration information from the fossil record is provided and if a probabilistic model of ancestral generation time is envisioned.

1. Introduction

The age of the divergence between humans and chimpanzees is one of the most studied parameters in evolutionary biology. In fact, as the field of molecular evolution was emerging in the early 1960s, the seminal work of Sarich and Wilson (1967) challenged conventional anthropological knowledge and proposed that the age of the *Homo–Pan* split occurred at approximately 5 Ma, an estimate that was 10–25 Ma younger than the paleontological data suggested at the time (Pilbeam, 1968). As molecular evolutionary analysis developed, in addition to the age of the split, the effective population size of the ancestor of humans and chimpanzees was also a parameter of theoretical interest in order to obtain a clearer scenario of the speciation process between humans and our closest living relative (Chen and Li, 2001; McVicker et al., 2009; Takahata and Satta, 1997).

A fundamental parameter of population genetics introduced by Sewall Wright in the early 1930s (Wright, 1931), the effective population size (N_e), enables the study of natural populations with complex life histories using the standard Wright–Fisher model. N_e also permit evaluations of the rate of genetic drift and the effectiveness of natural selection (reviewed in Charlesworth, 2009): the larger the N_e, the larger the effectiveness of natural selection on loci and the smaller the power of genetic drift in eliminating new mutants. Moreover, recent developments in phylogenetic theory explicitly consider population-level phenomena within phylogenies (Degnan and Salter, 2005; Liu et al., 2009). Knowledge of the ancestral N_e is thus also crucial for making predictions in

* Tel.: +55 21 2562 6397, +55 21 4063 8278.
E-mail address: carlos.schrago@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2014.05.009
0022-5193/

phylogenetic analysis (Degnan and Rosenberg, 2009; Edwards, 2009; Liu et al., 2009).

The estimates of the ancestral N_e of humans and chimpanzees based on several studies to date have exhibited a large variance. For example, Yang (2002) reported an estimate of 12,000 Wright–Fisher (WF) individuals using data from Chen and Li (2001), who calculated that the ancestral N_e might be as great as 96,000. Recent methodological developments using genome-wide analyses have yielded N_e estimates varying from 47,000 to 65,000 (Hobolth et al., 2007; Hobolth et al., 2011). Such variation leads to contrasting evolutionary scenarios of human evolution. Indeed, if N_e of the LCA of humans and chimpanzees was as low 12,000, the hypothesis that the ancestral population of modern humans was subjected to a bottleneck (Rogers and Jorde, 1995; Tenesa et al., 2007) is weakened, whereas this hypothesis is supported if the ancestral N_e=96,000 because N_e of modern humans was inferred to be approximately 10,000 WF individuals (Kim et al., 2010; Laval et al., 2010; Takahata et al., 1995).

Nonetheless, this variation of parametric estimates is not expected when the number of sites examined is large. As the genomes of both human and chimpanzee have been sequenced, the number of sites available for analysis is large, thereby making stochastic errors negligible. If the estimators used were consistent, their variance should approach zero as the estimate approaches the true parametric value (Casella and Berger, 2002). However, estimates with zero variance are impossible in several cases of molecular evolutionary analysis (Rannala and Yang, 2007; Yang and Rannala, 2006), and it is worth investigating the limiting distribution of the parameter of interest in such cases. For example, Schrago and Voloch (2013) have recently shown that the uncertainty associated with the divergence time between *Homo* and *Pan* cannot be further reduced unless new calibration information obtained from the fossil record is considered.

Another difficulty is that molecular evolutionary analysis frequently relies on very sophisticated models that require numerical methods, such as Monte Carlo methods, to be properly solved. Surprisingly, analytical approaches are rarely used, even though the amount of data available for humans and chimpanzees is so large that the application of such strategies is possible. As an example, the study of Takahata (1986) used the variance of the evolutionary distance among four loci of mammals to estimate ancestral affective population sizes, whereas Chen and Li (2001) used the empirical frequency of the correct gene tree topology of approximately 50 genomic regions; both studies used simple derivations from coalescent theory (Hudson, 1990; Kingman, 1982).

With the aim of inferring the asymptotic value of the effective population size of the LCA of humans and chimpanzees, a large genomic data was assembled, and analytical approaches based on coalescent theory were used to infer the ancestral N_e. It was found that the estimates of the ancestral N_e of *Homo–Pan* exhibit wide confidence intervals, even when using a large number of loci under rate homogeneity. Two major sources of uncertainty were associated with the ancestral N_e estimates, namely, the uncertainty of the calibration information and the uncertainty of the ancestral generation time of the *Homo–Pan* LCA. The former will only be reduced with the availability of a richer fossil record of Hominidae, whereas the latter requires the definition of an explicit probabilistic model for the ancestral generation time.

2. Materials and methods

To investigate the limiting distribution of the effective population size of the ancestor of humans and chimpanzees, two analytical derivations from coalescent theory were used to calculate the value of the ancestral N_e: Takahata's (1986) approach, which uses the variance of the genetic distances among loci, and that of Chen and Li (2001), which uses the probability of topological matches between gene trees and the species tree. The limiting posterior distribution, given by the infinite-site theory (Rannala and Yang, 2007), of the mean coalescent time between *Homo* and *Pan* alleles and the posterior distribution of evolutionary rates were also estimated. Therefore, the rate and time estimates exhibited the smallest variance, according to the calibration priors utilized.

2.1. Dataset composition

The syntenic alignments of six primate genomes available in the Compara repository of the Ensembl database were used (ftp://ftp.ensembl.org/pub/release-67/emf/ensembl-compara). A script was written to randomly collect 5000-bp segments from the genomic alignment of *Homo*, *Pan*, *Gorilla* and *Pongo* (outgroup) and to posteriorly test the segments for the molecular clock using the likelihood ratio test implemented in the baseml program of the PAML package, with the significance level set at 5% (Yang, 2007). A total of 15,744 clock-like random genomic segments of 5000 bp each were assembled, resulting in more than 78 million bp used in total.

2.2. Estimating N_e using the variance of genetic distances

We assume that the number of loci examined as well as the total number of nucleotide sites are infinite; therefore, the genetic distance between *Homo* and *Pan* is known without error. As the proportion of sites differing between humans and chimps was only 1.2%, the Jukes–Cantor (JC) distance (Jukes and Cantor, 1969) sufficiently accounts for the multiple substitutions that occurred along the evolution of both species (Nei and Kumar, 2000).

The 15,744 genome segments studied permitted the inference of the variance of the coalescence times (genetic distances) among loci. From coalescent theory, the expected waiting time until coalescence between a pair of alleles in the ancestral population is exponentially distributed with mean equal to $2N_e$ generations ($\lambda = 1/2N_e$) and variance equal to $(2N_e)^2$ (Hudson, 1990; Kingman, 1982). Because of the large amount of segments examined, the empirical variance inferred approaches the expected variance of the coalescence. As proposed by Takahata (1986), this equivalence may be used to estimate the ancestral effective population size of a pair of species. However, because time is measured in substitutions per site (s/s) when comparing genetic distances, we must work instead with the scaled effective population size $\theta = 4N_e\mu g$, where μ is the mutation rate per year and g is the generation time. We adopt thus the variance of the exponential distribution using $\lambda = 2/\theta$, with a mean$=\theta/2$ and a variance$=\theta^2/4$ (Edwards and Beerli, 2000).

It is worth noting that the coalescent time T (measured in s/s) of a pair of orthologous genomic regions equals half the genetic distance between species ($d/2$). Thus, the average coalescence time among loci is $\overline{d}/2$ (Fig. 1). Additionally, as the speciation time is the same for all genome segments compared, the variance of the average coalescence times, var($\overline{d}/2$) is composed of (i) the variance of the coalescent process in the ancestral population, $\sigma(T)$, and (ii) the variance of the model of nucleotide substitution employed. Therefore, to use Takahata's (1986) approach, the amount (ii) must be subtracted from the empirical variance of the average genetic distance. If we define var($\overline{d}/2$) as

$$\mathrm{var}\left(\frac{\overline{d}}{2}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\hat{d}_i}{2} - \frac{\overline{d}}{2}\right)^2, \quad (1)$$

where $\hat{d}_i$ is the JC distance between *Homo–Pan* for the ith gene; the variance of the substitution model is obtained by averaging the