Published in final edited form as:
Methods Ecol Evol. 2012 August 1; 3(4): 624–627. doi:10.1111/j.2041-210X.2012.00194.x.

jPopGen Suite: population genetic analysis of DNA polymorphism from nucleotide sequences with errors

Xiaoming Liu*
Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, 1200 Herman Pressler Drive, Houston, TX 77030, USA

Summary

1. Next-generation sequencing (NGS) is being increasingly used in ecological and evolutionary studies. Though promising, NGS is known to be error-prone. Sequencing error can cause significant bias for population genetic analysis of a sequence sample.

2. We present jPopGen Suite, an integrated tool for population genetic analysis of DNA polymorphisms from nucleotide sequences. It is specially designed for data with a non-negligible error rate, although it serves well for "error-free" data. It implements several methods for estimating the population mutation rate, population growth rate, and conducting neutrality tests.

3. jPopGen Suite facilitates the population genetic analysis of NGS data in various applications, and is freely available for non-commercial users at http://sites.google.com/site/jpopgen/.

Keywords

next-generation sequencing; population genetics; sequencing error; population mutation rate; population growth rate; neutrality test

1 Introduction

The advance of next-generation sequencing (NGS) technologies has helped researchers to conduct various genetic studies efficiently in terms of both time and cost. Wider application of these technologies in ecological and evolutionary studies is expected in the near future. One disadvantage associated with these new sequencers is that their error rates are typically tenfold higher than that of Sanger sequencing (Shendure and Ji, 2008). Sequencing error can cause significant bias for population genetic analysis of a sequence sample. Given a random sample from a population, artificial polymorphisms caused by sequencing error will skew both the number and frequency spectrum of the observed SNPs. This will further skew any estimations or test statistics based on the number and/or frequency spectrum of the SNPs (Johnson and Slatkin, 2008; Achaz, 2008). The problem will be even more prominent with increased sample size because the number of sequencing errors increases linearly with sample size while that of true mutations increases slower (Liu et al., 2009, 2010).

There have been several new methods proposed to estimate population genetic parameters and test the hypothesis of strict neutrality using DNA sequences with errors (e.g. Achaz, 2008, 2009; Johnson and Slatkin, 2006, 2008, 2009; Liu et al., 2009, 2010; Hellmann et al., 2008; Knudsen and Miyamoto, 2007, 2009; Lynch, 2009). Targeting population genetic

*Corresponding author. Xiaoming.Liu@uth.tmc.edu. Fax: 001-713-500-0900.

analysis of error-prone NGS data, jPopGen Suite implements some of these new methods along with several widely used methods designed for “error-free” data.

2 Platform, interface and file format

jPopGen Suite is written in Java, which enables it to run cross-platform on a wide range of computers, as long as a proper Java Runtime Environment is installed. It uses a menu-driven graphic user interface (GUI) to specify all parameter settings and conduct all integrated analyses (Fig. 1). In addition to the GUI, the full functionality of the suite can be accessed through a command line interface, which facilitates a “batch mode” for analyzing large data sets.

The default input file format is a summary of the SNP frequency spectrum (SFS) with each row describing the observed number of a particular SNP configuration (numbers of major/ancestral alleles, minor/derived alleles and missing data). The ancestral alleles of the polymorphic sites can be either known (unfolded data) or unknown (folded data). Additionally, jPopGen Suite supports the direct input of sequence data files in PHYLIP, ALN (ClustalW2), or FASTA format.

3 Analysis methods implemented

3.1 Estimating population parameters

Thirteen methods are implemented to estimate the population mutation rate θ (θ=4$N\mu$, where N is the effective population size and μ is the mutation rate per sequence per generation), assuming a constant population size. Among them four methods are for “error-free” data: Tajima’s (Tajima, 1983) estimator based on pairwise difference between two sequences (θ_π), Watterson’s (Watterson, 1975) estimator based on the number of polymorphic sites (θ_S), Fu’s (Fu, 1994) best linear unbiased estimators (BLUEs) for unfolded and folded data and Zeng et al.’s (Zeng et al., 2006) estimator for unfolded data (θ_L). Three methods are designed for sequences with assumed known sequencing error rate: Johnson and Slatkin’s (Johnson and Slatkin, 2008) modified θ_π estimator and θ_S estimator, and Liu et al.’s (Liu et al., 2009) generalized least square (GLS) estimators with known sequencing error rate for unfolded and folded data. The remaining six methods are for data with unknown sequencing error rate: Achaz’s (Achaz, 2008) modified θ_π and θ_S based on non-singleton variants, Liu et al.’s (Liu et al., 2009) GLS estimators based on non-singleton variants, and modified θ_π, θ_S and θ_L based on variants with minor allele count larger than a user specified number m (n-1>m≥0, where n is the number of sequences in the sample), denoted as $\theta_{\pi m}$, θ_{Sm} and θ_{Lm}, respectively. That is,

$$\begin{aligned}\theta_{\pi m} &= \frac{2}{(n-m-1)(n-m)}\sum_{i=m+1}^{n-1} i(n-i)\xi_i \\ \theta_{S_m} &= \left(\sum_{i=m+1}^{n-1}\xi_i\right)\Big/\left(\sum_{i=m+1}^{n-1}\frac{1}{i}\right) \\ \theta_{Lm} &= \frac{1}{(n-m-1)}\sum_{i=m+1}^{n-1} i\xi_i,\end{aligned}$$

for unfolded data, and

$$\theta_{\pi m} = \frac{2}{n(n-2m-1)} \sum_{i=m+1}^{n-m-1} i\,(n-i)\,\xi_i$$
$$\theta_{Sm} = \left(\sum_{i=m+1}^{n-m-1} \xi_i\right) \Big/ \left(\sum_{i=m+1}^{n-m-1} \frac{1}{i}\right),$$

for folded data, where ξ_i is the number of segregating sites on which the derived allele occurs i times in the sample.

jPopGen Suite implements Liu et al.'s (Liu et al., 2010) maximum composite likelihood estimators (MCLEs) to estimate the population mutation rate θ, population exponential growth rate R, and sequencing error rate ε, simultaneously. In a population exponential growth model, $N(t)=N(0)\exp(-rt)$, where $N(t)$ is the effective population size t generations before the current time (generation 0), and the scaled population growth rate R is defined as $R=2N(0)r$. An error model assumes that there are n_a ($n_a \geq 2$) types of alleles at a nucleotide site, and when a sequencing error occurs on an allele, the allele has an equal probability $1/(n_a-1)$ to change to another type of allele. A grid search algorithm is used to estimate the three parameters.

Confidence intervals of the θ estimators and MCLEs can be inferred via coalescent simulation.

3.2 Testing the hypothesis of strict neutrality

Twelve methods are implemented for the neutrality test: Tajima's (Tajima, 1989) D test, Achaz's (Achaz, 2008) Y and Y^* tests, Fu and Li's (Fu and Li, 1993) D, D^*, F and F^* tests (correctly normalized according to Achaz, 2009), normalized Fay and Wu's (Fay and Wu, 2000) H test (Zeng et al., 2006), Zeng et al.'s (Zeng et al., 2006) E test, and three new tests via contrasting $\theta_{\pi m}$ with θ_{Sm}, $\theta_{\pi m}$ with θ_{Lm}, and θ_{Lm} with θ_{Sm}, respectively. These three new test statistics are calculated as

$$T = \left[(\theta_1 - \theta_2) - \overline{E}\,(\theta_1 - \theta_2)\right] / \sqrt{\overline{\mathrm{Var}}\,(\theta_1 - \theta_2)},$$

where θ_1 and θ_2 are two different θ estimators, $\bar{\mathrm{E}}(\theta_1-\theta_2)$ and $\overline{\mathrm{Var}}(\theta_1-\theta_2)$ are the empirical mean and variance of θ_1-θ_2, obtained via coalescent simulation (e.g. Hudson, 2002).

Coalescent simulation is also used to estimate the p-values or significance levels of the tests. One unique feature of this tool is that the sequencing error model and the population exponential growth model are incorporated into the null model. The user can specify the known or estimated population mutation rate θ, sequencing error rate ε and population exponential growth rate R for the simulation. This is particularly important for conducting the neutrality tests when sequencing errors and population growth cannot be ignored, because both may significantly skew the null distributions of the test statistics, and therefore increase the type I error rate (Johnson and Slatkin, 2008; Achaz, 2008).

4. Suggested usage

Unless constant population size can be assumed, MCLE is recommended for the first-round analysis for estimating the population mutation rate θ, the sequencing error rate ε and the population exponential growth rate R. Using coalescent simulation, the confidence interval of the above estimations can be inferred. If ε or/and R are not significantly different to 0, then θ estimators assuming constant population size or/and no sequencing errors then can be

applied for more accurate estimation of θ. We generally recommend Liu et al.'s (Liu et al., 2009) GLS estimators or Fu's (Fu, 1994) BLUE estimators for that purpose.

The selection of neutrality tests also depends on the estimated error rate. If the sequencing errors cannot be ignored, Achaz's (Achaz, 2008) Y and Y^* tests or the three new tests via contrasting $\theta_{\pi m}$ with θ_{Sm}, $\theta_{\pi m}$ with θ_{Lm}, and θ_{Lm} with θ_{Sm}, respectively, are recommended. Otherwise, the remaining seven traditional tests can be selected based on their testing power under different alternative hypotheses.

Acknowledgments

I thank Dr. Yun-Xin Fu for kindly providing the Java library for coalescent simulation and Dr. Guillaume Achaz for his assistance in correcting the normalization of the Fu and Li's tests. This research was supported by the National Institutes of Health grant 5P50GM065509 and 1U01HG005728.

References

Achaz G. Frequency Spectrum Neutrality Tests: One for All and All for One. Genetics. 2009; 183:249–258. [PubMed: 19546320]

Achaz G. Testing for neutrality in samples with sequencing errors. Genetics. 2008; 179:1409–1424. [PubMed: 18562660]

Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000; 155:1405–1413. [PubMed: 10880498]

Fu YX. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics. 1994; 138:1375–1386. [PubMed: 7896116]

Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993; 133:693–709. [PubMed: 8454210]

Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. Genome Research. 2008; 18:1020–1029. [PubMed: 18411405]

Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002; 18:337–338. [PubMed: 11847089]

Johnson PLF, Slatkin M. Accounting for bias from sequencing error in population genetic estimates. Molecular Biology and Evolution. 2008; 25:199–206. [PubMed: 17981928]

Johnson PLF, Slatkin M. Inference of microbial recombination rates from metagenomic data. PLoS Genetics. 2009; 5:e1000674. [PubMed: 19798447]

Johnson PLF, Slatkin M. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Research. 2006; 16:1320–1327. [PubMed: 16954540]

Knudsen B, Miyamoto M. Accurate and fast methods to estimate the population mutation rate from error prone sequences. BMC Bioinformatics. 2009; 10:247. [PubMed: 19671163]

Knudsen B, Miyamoto MM. Incorporating experimental design and error into coalescent/mutation models of population history. Genetics. 2007; 176:2335–2342. [PubMed: 17565962]

Liu X, Fu YX, Maxwell TJ, Boerwinkle E. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. Genome Research. 2010; 20:101–109. [PubMed: 19952140]

Liu X, Maxwell TJ, Boerwinkle E, Fu YX. Inferring Population Mutation Rate and Sequencing Error Rate Using the SNP Frequency Spectrum in a Sample of DNA Sequences. Molecular Biology and Evolution. 2009; 26:1479–1490. [PubMed: 19318520]

Lynch M. Estimation of Allele Frequencies From High-Coverage Genome-Sequencing Projects. Genetics. 2009; 182:295–301. [PubMed: 19293142]

Shendure J, Ji H. Next-generation DNA sequencing. Nature Biotechnology. 2008; 26:1135–1145.

Tajima F. Evolutionary Relationship of DNA-Sequences in Finite Populations. Genetics. 1983; 105:437–460. [PubMed: 6628982]

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

Watterson GA. Number of Segregating Sites in Genetic Models without Recombination. Theoretical Population Biology. 1975; 7:256–276. [PubMed: 1145509]

Zeng K, Fu YX, Shi S, Wu CI. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. Genetics. 2006; 174:1431–1439. [PubMed: 16951063]

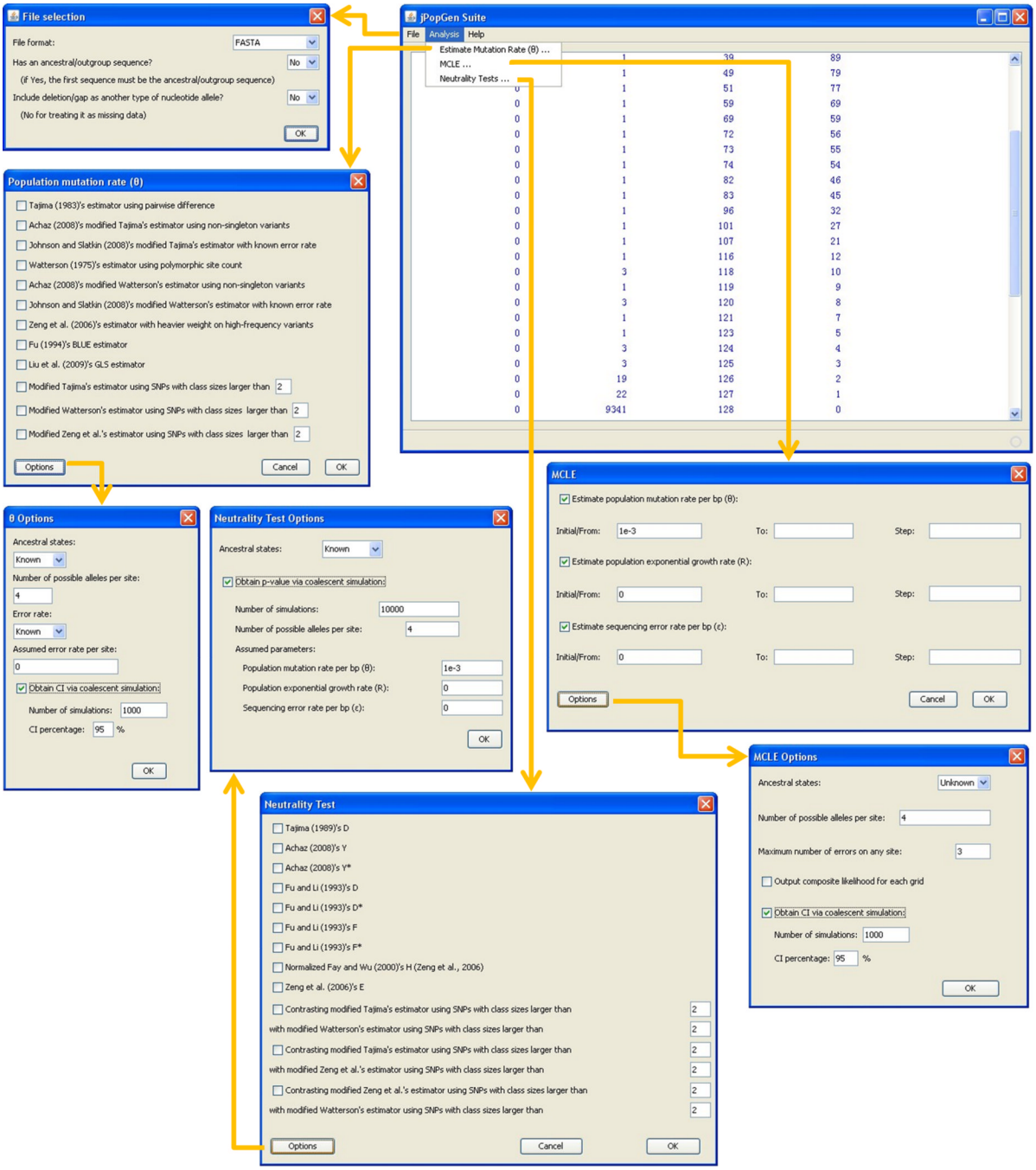

Fig 1.
The major graphic user interfaces of jPopGen Suite.