POPTREE2: Software for Constructing Population Trees from Allele Frequency Data and Computing Other Population Statistics with Windows Interface

Naoko Takezaki,*[,1] Masatoshi Nei,[2] and Koichiro Tamura[3]

[1]Life Science Research Center, Kagawa University, Ikenobe 1750-1, Kitagun, Kagawa, Japan

[2]Department of Biology and the Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

[3]Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan

***Corresponding author:** E-mail: takezaki@med.kagawa-u.ac.jp.

Associate editor: Sudhir Kumar

Abstract

Currently, there is a demand for software to analyze polymorphism data such as microsatellite DNA and single nucleotide polymorphism with easily accessible interface in many fields of research. In this article, we would like to make an announcement of POPTREE2, a computer program package, that can perform evolutionary analyses of allele frequency data. The original version (POPTREE) was a command-line program that runs on the Command Prompt of Windows and Unix. In POPTREE2 genetic distances (measures of the extent of genetic differentiation between populations) for constructing phylogenetic trees, average heterozygosities (H) (a measure of genetic variation within populations) and G_{ST} (a measure of genetic differentiation of subdivided populations) are computed through a simple and intuitive Windows interface. It will facilitate statistical analyses of polymorphism data for researchers in many different fields. POPTREE2 is available at http://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptree2/index.html.

Key words: polymorphism, allele frequency data, population tree, genetic diversity, subdivided populations, microsatellite DNA.

Introduction

The genetic relationships of different populations and closely related species can be inferred by using allele frequency data. Allele frequency data can also be used for studying other genetic features of populations. Recently, because of the high degree of genetic polymorphism, microsatellite DNA data (Estoup et al. 2002; DeSalle and Amato 2004) are widely used for genetic studies of populations in various organisms. Furthermore, single nucleotide polymorphism (SNP) data are becoming available for various nonmodel organisms and providing a large-scale data (Brito and Edwards 2009). As a consequence, there is a strong demand for a computer program to analyze such data efficiently.

POPTREE2 is a computer program that can perform evolutionary analyses of allele frequency data. It can compute various types of genetic distance and construct phylogenetic trees of populations using the neighbor-joining (NJ) method (Saitou and Nei 1987) and the unweighted pair-group method with arithmetic mean (UPGMA) (Sneath and Sokal 1973). Bootstrap tests (Felsenstein 1985) can be performed for the phylogenetic trees constructed. The distance measures computed by POPTREE2 and used for phylogeny construction are 1) D_A distance (Nei et al. 1983), 2) Nei's standard genetic distance (D_{ST}) (Nei 1972), 3) F^*_{ST} distance (Latter 1972) (added in the new version), 4) $(\delta\mu)^2$ distance (Goldstein et al. 1995), and 5) D_{SW} distance (Shriver et al. 1995). $(\delta\mu)^2$ Distance and D_{SW} distance can be used only for microsatellite DNA data, in which alleles are represented by the number of nucleotide repeats. By contrast, D_A, D_{ST}, and F^*_{ST} can be used for any kind of allele frequency data including microsatellite DNA, SNPs, and classical markers.

In addition to constructing phylogenetic trees, POPTREE2 can compute the following quantities: 1) average heterozygosity (H) and its standard error for each population, 2) number of alleles per locus for each population, 3) G_{ST}, a measure of genetic differentiation of subdivided populations for multiple alleles and loci (Nei 1973), and 4) the distance values (1)–(5) defined above.

The previous software, POPTREE version 1, was a command-line program, which runs on the Command Prompt of Windows and Unix. This version was not easy to use, and other additional programs such as TREEVIEWER (by K.T.) were necessary for graphical presentation of phylogenetic trees. Because of the recent demand, we decided to develop this new version of the program, POPTREE2. For researchers in many different fields of biology to whom this software may be useful, we would like to make an announcement of the availability of this new version. In POPTREE2, more options are available for computing genetic distances and other statistical quantities.

POPTREE2, a New Version of the Software

In POPTREE2, Windows interface with intuitive and simple design has been added. One can obtain the result of the analysis basically by 1) opening an input data file, 2) choosing the computational method to be used, and 3) running

Letter

the program. Furthermore, in this new version of the software for computing G_{ST}, H, distance measures D_{ST}, and F^*_{ST}, the original methods without sample size bias correction were added because unbiased estimates often become negative when sample size is small.

When one starts POPTREE2, the Poptree window will open. Clicking the Data input button on the upper-left corner of the window, one can see the dialog box displayed (fig. 1). The input file that contains allele frequency data for multiple populations is specified here.

After opening the file, users should choose computational methods: computation of distance values and construction of phylogenetic tree ("Distance/Phylogeny") or computation of heterozygosities and G_{ST} ("Heterozygosity/Gst") by checking one of the radio buttons in the upper section of the Poptree window (fig. 2).

For computation of distance values and construction of phylogenetic trees, users should choose the distance measure and tree construction method in pull-down menus on the right side of the radio button (fig. 2A). Bootstrap test can be performed by checking the box on the left side of "Bootstrap" and specifying the number of replications. In bootstrap tests, genetic loci are resampled with replacement in each replication. With a click of the "Run Poptree" button on the left side of the upper section of the Poptree window, a graphical presentation of the phylogenetic tree will automatically appear in the Phylogeny page (fig. 3). The distance values used for construction of the phylogenetic tree will be shown in the Output page (fig. 4A) if the Output tab on the upper section of the Poptree window (fig. 3) is clicked. Below the distance matrix, the phylogenetic tree constructed is shown in the Newick format.

The presentation of the constructed phylogenetic trees displayed in the Phylogeny page can be changed by using icons on the upper section of the window (fig. 3). The functions available can change the root position of a tree, the vertical order of two descendant clusters of a branch, tree style (rectangular or radial presentation), size of a tree in the horizontal and the vertical directions, font of the population names, and line width. It should be noted that the root position can be changed only for NJ trees (fig. 3A) because the root of a UPGMA tree (fig. 3B) is determined automatically by the method.

In addition, icons can be used to save the displayed tree in a text file in the Newick format, print it, and copy it to the clipboard of the operating system. The phylogenetic tree saved in the Newick format can be opened by other applications such as MEGA 4 (Tamura et al. 2007), in which

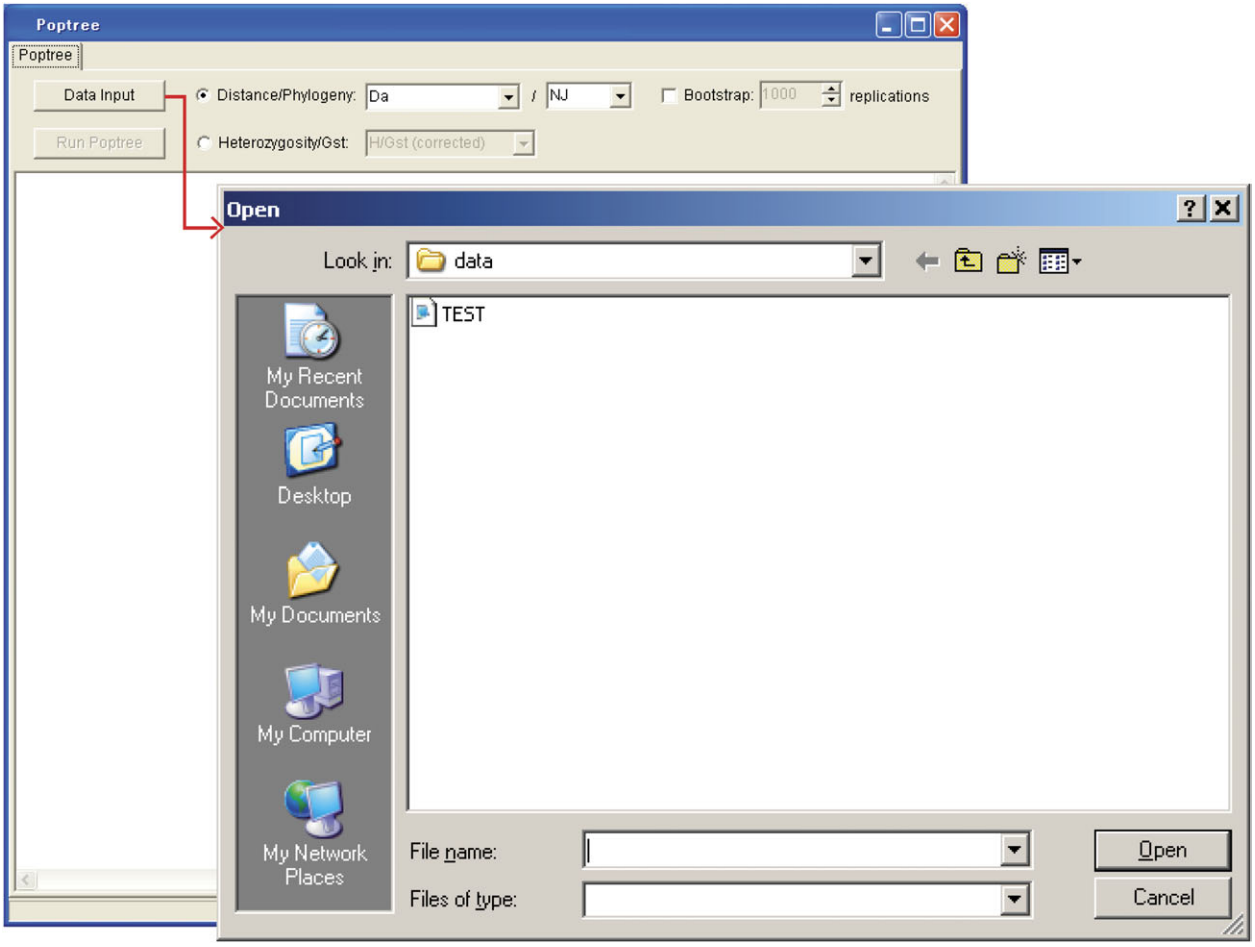

FIG. 1. The Poptree window that appears right after starting POPTREE2 and the dialog box for specifying an input data file. When users start POPTREE2, the Poptree window appears. A click on the Data input button on the upper-left corner of the Poptree window will show the dialog box. An input data file can be specified in the dialog box.

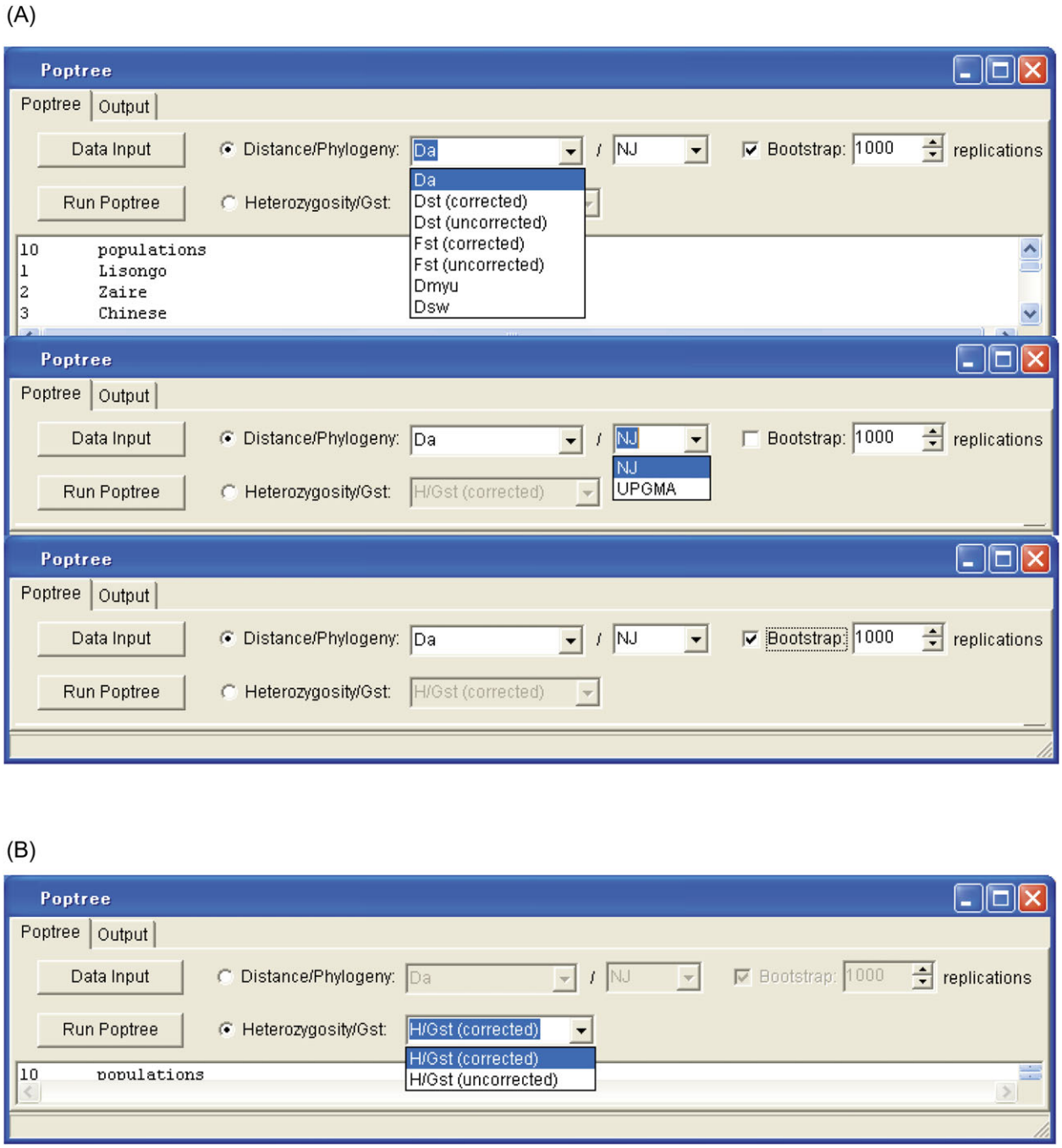

FIG. 2. The computational methods and their options. The computational methods, either computation of distance values and construction of phylogenetic tree ("Distance/Phylogeny") or computation of heterozygosities and G_{ST} ("Heterozygosity/Gst"), can be specified by checking the radio button at the left corner of the upper section of the Poptree window. (A) Computation of distance values and construction of phylogenetic tree. The distance options "Da," "Dst," "Dst (u)," "Fst," "Fst (u)," "Dmyu," and "Dsw" correspond to D_A, D_{ST} (sample size bias corrected) (Nei 1978), D_{ST} (bias uncorrected) (Nei 1972), F^*_{ST} (bias corrected) (Nei 1987), F^*_{ST} (bias uncorrected) (Latter 1972), $(\delta\mu)^2$ (Goldstein et al. 1995), and D_{SW} (Shriver et al. 1995), respectively. Users need to choose "NJ" or "UPGMA" as a tree construction method. Bootstrap test can be done for a phylogenetic tree constructed by checking the box of "Bootstrap." The number of bootstrap replication can be specified in the edit box on the right side. (B) Computation of heterozygosities and G_{ST}. H and G_{ST} will be computed with unbiased estimator (Nei and Roychoudhury 1974; Nei 1987) with the option "H/Gst (corrected)" and by the original method without bias correction (Nei 1973) with the option "H/Gst (uncorrected)."

more options for tree presentation are available. The graphical presentation of a phylogenetic tree copied to the clipboard can be pasted to other graphic software for publication.

For computation of heterozygosities and G_{ST}, users should choose "H/Gst (corrected)" or "H/Gst (uncorrected)" in the pull-down menu on the right side of the radio button (fig. 2*B*). H and G_{ST} will be computed with

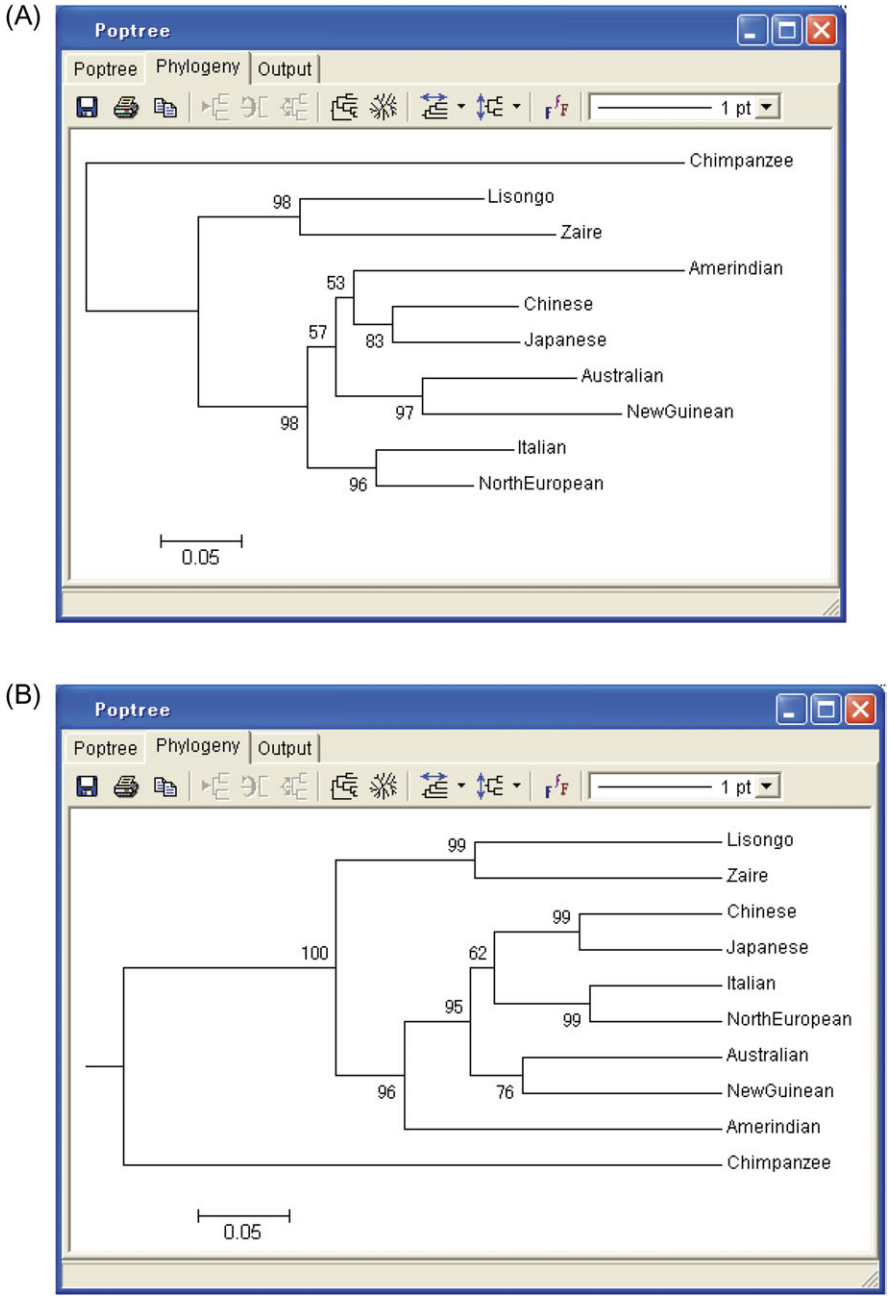

FIG. 3. Presentation of phylogenetic trees. (A) A NJ tree. (B) A UPGMA tree. The root of a default NJ tree is given by the midpoint rooting method. The root position can be changed by placing the cursor on the branch where the root should be located and clicking icon 4 (icons are numbered from left to right) on the tool bar. The root position of the UPGMA tree cannot be changed because it is determined by the method. The other functions of the icons on the upper section of the window are as follows: to save the tree in a Newick format (icon 1), print the tree (icon 2), copy the tree to the clipboard of the system (icon 3), and change the vertical orders of two descendant clusters of a branch (icons 5 and 6), tree style (rectangular and radial presentations) (icons 7 and 8), size of the tree in the horizontal and the vertical directions (icons 9 and 10), font of the population names (icons 11), and line width (icon 12).

unbiased estimator (Nei and Roychoudhury 1974; Nei 1987) for H/Gst (corrected) and by the original method without bias correction (Nei 1973) for the option H/Gst (uncorrected). Clicking the Run Poptree button, one can obtain the values of H for each population, G_{ST}, and the number of alleles in the Output page (fig. 4B). The merits of using G_{ST} instead of classical F_{ST} (Wright 1951) have been discussed by Nei and Kumar (2000, p. 238–244) and Crow (2004).

More details of the computational methods and the options of tree presentation are available in the user guide of POPTREE2.

Computer Platforms and Future Direction

Although POPTREE2 can run only on Windows, we continue to distribute the command-line version. It can be used in Linux, the Terminal window of MacOS, and the Command Prompt of Windows. Furthermore, we are

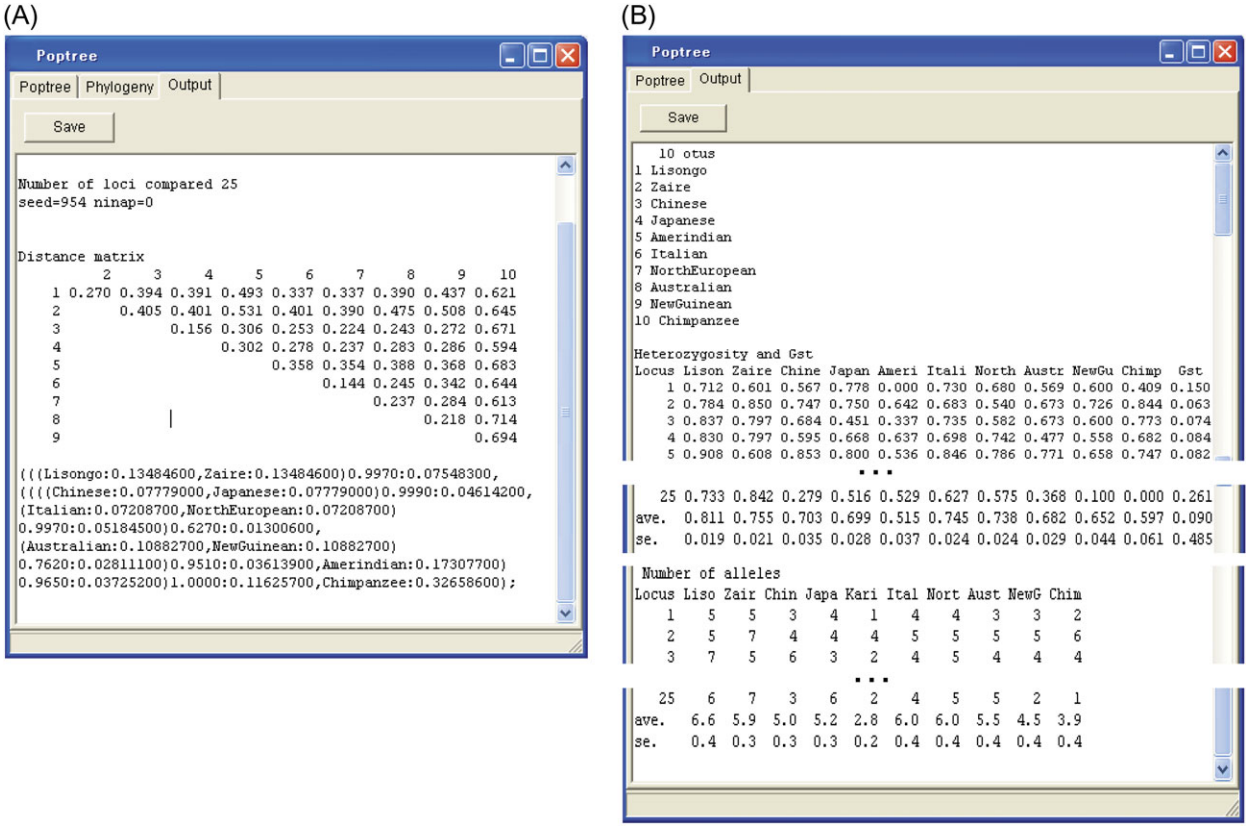

FIG. 4. (A) The distance values. By choosing "Distance/Phylogeny" as a computational method and running Poptree, the distance values used for construction of the phylogenetic tree will be shown in the Output page. Below the distance matrix, the phylogenetic tree constructed is shown in the Newick format. With a click on the Save button, the content of the Output page will be saved in a text file. (B) Results of computation of G_{ST} and *H*. By choosing "Heterozygosity/Gst" as a computational method (fig. 2), estimates of heterozygosity and G_{ST} will appear in the Output page. The heterozygosities for each population and G_{ST} are shown for each locus, and the averages of all loci and their standard errors are shown below. Furthermore, the number of alleles for each population is shown. With a click on the Save button, the content of the Output page can be saved in a text file.

preparing for development of Web version, which is accessible online from the different computer platforms. In the future, we also plan to expand the analyses of POPTREE2 to include population tree analysis of mitochondrial DNA and Y chromosome data.

Acknowledgments

We thank Masafumi Nozawa for his comments on an earlier version of the manuscript. National Institutes of Health (grant GM020293 to M.N.).

References

Brito PH, Edwards SV. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* 135:439–455.

Crow JF. 2004. Assessing population subdivision. In: Wasser SP, editor. Evolutionary theory and processes: modern horizons. Nordrecht (The Netherlands): Kluwer Academic Publishers. p. 35–42.

DeSalle R, Amato G. 2004. The expansion of conservation genetics. *Nat Rev Genet*. 5:702–712.

Estoup A, Jarne P, Cornuet JM. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol Ecol*. 11:1591–1604.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA*. 92:6723–6727.

Latter BDH. 1972. Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics* 70:475–490.

Nei M. 1972. Genetic distance between populations. *Am Nat*. 106:283–291.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci USA*. 70:3321–3323.

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nei M, Roychoudhury AK. 1974. Sampling variances of heterozygosity and genetic distance. *Genetics* 76:379–390.

Nei M, Tajima F, Tateno Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. *J Mol Evol*. 19:153–170.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4: 406–425.

Shriver MD, Jin L, Boerwinkle E, Deka R, Ferrell RE, Chakraborty R. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol Biol Evol*. 12:914–920.

Sneath PHA, Sokal RR. 1973. Numerical taxonomy. San Francisco (CA): W.H. Freeman.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol*. 24:1596–1599.

Wright S. 1951. The genetical structure of populations. *Ann Eugen*. 15:323–354.