POPTREE2

Software for constructing population trees from allele frequency data and computing other population statistics with Windows-interface

User guide

Naoko Takezaki, Masatoshi Nei, and Koichiro Tamura

Table of contents

1. Introduction

POPTREE2 constructs phylogenetic trees from allele frequency data by 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. Some population statistics are also computed.

Distance measures that can be used for the phylogeny construction are as follows:

(i) D_A distance (Nei et al. 1983)

(ii) Nei's standard genetic distance (D_{ST}) (Nei 1972) with or without sample size bias correction

(iii) F_{ST}^* distance (Latter 1972) with or without sample size bias correction

(iv) $(\delta\mu)^2$ distance (Goldstein et al. 1995)

(v) 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 repeats. By contrast, D_A, D_{ST}, and F_{ST}^* can be used for any kind of allele frequency data.

In addition to construction of the phylogenetic trees, POPTREE2 can compute the following quantities:

(1) Average heterozygosity and its standard error for each population with and without sample size bias correction

(2) Number of alleles per locus for each population

(3) G_{ST}, a measure of genetic differentiation among subdivided populations for multiple alleles (Nei 1973), standardized G_{ST} (Hedrick 2005), Jost's (2008) D with and without sample size bias correction

(4) Values of the distance measures (i) – (v)

In POPTREE2 all the computations can be done through Windows-interface and the display of phylogenetic trees can easily be changed, copied to other applications, and printed by using icons.

2. Getting started

2.1 Installing POPTREE2 on your computer

Download POPTREE2.zip on your PC and uncompress it. Under the POPTREE2 folder, data folder, program folder, and POPTREE2 icon will appear.

2.2 Example file

There is an example file in the data folder. The allele frequency data and the genotype data in GENEPOP format can be used as an input file.

Allele frequency data:

test.dat: example input file [modified from Nei and Takezaki (1996)]

test1.dat: example input file [modified from Deka et al. (1995)]

Genotype data:

test_genepop1.dat

test_genepop2.dat

3. Running POPTREE2

First, prepare your allele frequency data or genotype data in the format as described in the section **4**. Open the example file test.dat or test_genepop1.dat with Notepad or Word. Looking at the example files will help the preparation of your input file.

To run POPTREE2, click the POPTREE2 icon. The “Poptree” window will appear.

Specify the input file, and choose the computational method in the “Poptree” window, as described in the following.

3.1 Starting POPTREE2

Double click the POPTREE2 icon in the POPTREE2 folder. “Poptree” window will appear.

3.2 Specifying input file

Click the "Data input" box at the upper-left corner of the “Poptree” window. An input file can be specified here. Please see section **4** for the format of the input file. If the input file is specified, the content will appear on the lower box of the “Poptree” window.

3.3 Choosing computational methods

First, you should choose either computation of distance values and construction of phylogenetic trees (“Distance/Phylogeny”) or computation of heterozygosities and G_{ST} (“Heterozygosity/Gst”) by checking the radio button in the upper section of the “Poptree” window.

3.4 Computation of distance values and construction of phylogenetic trees

If you choose “Distance/Phylogeny”, you should choose the distance measure to be calculated and used for tree construction and the method of tree construction in the pull-down menus on the right-hand side of the radio button.

3.4.1 Choosing distance measures

The following seven options are available [see the details of distance measures in

section **6**].

(1) “Da”: D_A distance

(2) “Dst”: Nei's standard genetic distance ($\hat{D}_{ST}$) (bias corrected)

(3) “Dst (u)”: Nei's standard genetic distance (D_{ST}) (bias uncorrected)

(4) “Fst”: F_{ST}^{*} distance (F_{ST}^{*}) (bias corrected)

(5) “Fst (u)”: F_{ST}^{*} distance (F_{ST}^{*}) (bias uncorrected)

(6) “Dmyu”: $(\delta\mu)^2$ distance

(7) “Dsw”: D_{SW} distance

“Dmyu” [$(\delta\mu)^2$] and “Dsw” (D_{SW}) are applicable only for microsatellite data. The number of repeats [it can be a relative value (see section **6**)] for each allele has to be specified in the input file for the computation of “Dmyu” [$(\delta\mu)^2$] and “Dsw” (D_{SW}). To compute “Da” (D_A), “Dst” and “Dst (u)” (D_{ST}), and “Fst” and “Fst (u)” (F_{ST}^{*}), the repeat number of each allele is not necessary.

3.4.2 Choosing tree construction methods

The following two options are available.

(1) “NJ”: construction of phylogenetic tree by a neighbor-joining method

(2) “UPGMA”: construction of phylogenetic tree by UPGMA

Please see section **5** for the NJ method and the UPGMA.

3.4.3 Bootstrap tests

If you would like to do a bootstrap test for the NJ or UPGMA tree, check the "Bootstrap test" box at the upper section of the “Poptree” window. Then, specify the number of bootstrap replications to be done in a box on the right side of the "Bootstrap test” box.

3.5 Computation of heterozygosities (*H*) and G_{ST}

If you choose “Heterozygosities/Gst”, you should choose the following options in the pull-down menu on the right-hand side of the radio button.

(1) “H/Gst (corrected)”: computation of heterozygosity, G_{ST}, standardized G_{ST} and Jost’s *D* (bias corrected)

(2) “H/Gst (uncorrected)”: computation of heterozygosity and G_{ST}, standardized G_{ST} and Jost’s *D* (bias uncorrected)

Please see section **7** for heterozygosity, G_{ST}, standardized G_{ST} and Jost’s *D*.

3.6 Carrying out computation

Click the "Run Poptree" box at the upper section of the “Poptree” window and the computation will be done.

3.7 Output

3.7.1 "Distance/Phylogeny"

3.7.1.1 Phylogenetic tree

If you choose "Distance/Phylogeny" in 3.3, the phylogenetic tree will be displayed in the "Phylogeny" window after clicking the "Run Poptree" box. You can change the shape of the tree by using the icons at the second row of the "Phylogeny" window [see section **8** for the

details].

3.7.1.2 Distance matrix

If you choose "Distance/Phylogeny" in 3.3, the distance values used for tree construction will be shown in the "Output" window. To see the distance matrix, click the "Output" box at the top row of the "Poptree" window.

Below the distance matrix in the "Output" window, the phylogenetic tree constructed is shown in Newick format. If you put this part in a separate file, this file can be used as an input file of MEGA (Tamura et al. 2013) and other software. MEGA has more options for displaying the tree. The Newick format of the phylogenetic tree can also be saved in a file in the "Phylogeny" window (see section **8**).

The content of the "Output" window can be saved in a file by clicking the "Save" box in the upper section of the "Output" window.

3.7.2 "Heterozygosities/Gst"

If you choose "Heterozygosities/Gst" in 3.3, the heterozygosities and G_{ST}, standardized G_{ST}, Jost*'s D* [with bias correction for "H/Gst (corrected)" and without bias correction for "H/Gst (uncorrected)"] computed for all populations will appear in the "Output" window. The number of alleles will also be shown below the heterozygosity and G_{ST} values [see details in section **7**].

The results of the computation can be saved in a file by clicking the "Save" box in the upper section of the "Output" window.

3.8 Terminating POPTREE2

If you click the "×" box at the upper-right corner of the "Poptree" window, POPTREE2 will be terminated.

4. Input file format

4-1. Allele frequency data

Please look at the example file test.dat. The first line indicates the number of populations. In the following lines the population names are shown. Each population name is shown in one line. Then, allele frequency data are shown.

n populations

1 population1

2 population2

3 population3

.

.

.

@locus 1 locusname

allele1

allele2

.

.

.

#the number of chromosomes examined in each population

@locus 2 locusname

.

.

The line for different alleles consists of the number of nucleotide repeats for microsatellite DNA loci (the name of allele for other data) and allele frequencies of populations separated by "*".

XX * frq1 frq2 frq3 ...

XX is the number of repeats for microsatellite DNA (not the fragment size) or for the name of an allele for other kinds of data. frq1, frq2, and frq3 are frequencies of the allele for populations 1, 2, and 3. After all allele frequencies for one locus are shown, the number of chromosomes (not the number of individuals) examined is shown. The line for the number of chromosomes should start with "#".

The number of repeats and the number of chromosomes do not have to be integer. If you use only D_A, D_{ST}, and F_{ST}^* distances, XX does not have to be the number of repeats for an allele and can be anything. The numbers of repeats for alleles are necessary for the computation of $(\delta\mu)^2$ and D_{SW} distances (see section **6**).

n populations

1 population1

2 population2

3 population3

.

.

.

@locus 1 locusname

1	*	.2600	.0942	.0000	..
2	*	.0000	.0000	.0054	..
2.5	*	.0000	.0000	.0000	..

.

.

.

#	100	138	186	...

@locus 2 locusname

.

.

4-2. Genotype data.

Genotype data in GENEPOP (Rousset 2008) format can be used as an input file of

POPTREE2. The input data has the following format.

Title line:
Locus name 1
Locus name 2
…

Pop population name 1

Sample 1
Sample 2
…
POP population name 2
…
pop
…

It has three sections: 1) title line, 2) locus names, and 3) genotype data of samples in different populations.

1) Title line

The first line is a title line. Any comment can be written in it.

2) Locus names

Following the title line are names of loci. They can be shown in different lines or separated by a comma.

3) Genotype data of each sample in populations. The data of different populations are separated by “Pop” tags. Following the “Pop” tag, the name of population can be specified. If a population name is not specified, Data of each sample are shown in a line. The sample name can be shown before a comma. Genotypes of loci for a sample are separated by a space or a tab.

```
Pop
Grange des Peres , 0201 003003 0102 0302 1011 01
Grange des Peres , 0202 003001 0102 0303 1111 02
Grange des Peres , 0102 004001 0202 0102 1010 01
Grange des Peres , 0103 002002 0101 0202 1011 02
Grange des Peres , 0203 002004 0101 0102 1010 01

POP

…
```

Allele name: Missing data is represented by the allele name that only consists of letter zero (e.g., 00, 000), as in Genepop. Otherwise, allele names do not have to consist of

only numbers 0-9. However, if a user wishes to use $(\delta\mu)^2$ or D_{SW}, allele names should represent the number of repeats (not a fragment size in bp) at a microsatellite locus. (The repeat number does not have to be an integer.)

Genotype data : For a genotype data at a cell, if the number of letters is smaller than or equal to 3, the data is considered to be a haploid data. If the number of letters is larger than 3, the data is considered to be a diploid data, and so it should be an even number. Otherwise, it is considered as an error. The first half and the second half of the data are respectively regarded as an allele name.

Exclusion of locus data due to missing data: If the data for a locus are all missing data (e.g., 00, 000), then this locus is excluded from the computation. After exclusion of these loci, if a population has only missing data for a locus, it is considered as an error.

Number of populations: The number of populations should be larger than or equal to 3.

5. Neighbor-joining (NJ) and UPGMA trees

In the NJ method (Saitou and Nei 1987), starting from a star-tree (all branches are connected to one node), a pair of taxa (populations) which gives the smallest sum of branch lengths are combined into a cluster and form a composite taxa. This process is repeated until an unrooted tree is produced. The branch lengths are computed by the least-squares method in each step.

In the UPGMA (Sneath and Sokal 1973), a pair of taxa with the smallest distance are combined into one cluster and form a composite taxa. This process is repeated until a rooted tree is made. The branch lengths are calculated so that the sum of the branch lengths from the taxa to the node connecting the two taxa is half the distance of the two taxa. In the UPGMA,

the molecular clock (rate constancy) is implicitly assumed (Chakraborty 1977). If the rate constancy approximately holds, the UPGMA can be efficient in constructing the correct tree topology (Takezaki and Nei 1996).

In the bootstrap test (Felsenstein 1985), the loci are resampled with replacement in POPTREE2. The phylogenetic tree is constructed with the distance values calculated from the same number of resampled loci as that of the original input dataset in each replication. The number of replications in which the branch (the grouping of the taxa separated by the branch) appeared is counted and the proportion of this number in the total replications is shown in percent on the branch of the tree in the "Phylogeny" window. In the case of the UPGMA tree, the bootstrap numbers are counted by removing the root of the tree.

NJ method and UPGMA produce bifurcating trees. However, in the phylogenetic tree displayed in the "Phylogeny" window, the branches with length zero or negative values are treated as though they do not exist. Because of this treatment a multifurcating node sometimes appears.

See Nei and Kumar (2000) for the NJ and UPGMA methods.

6. Measures of genetic distance between populations

6.1 D_A distance

D_A distance (Nei et al. 1983) is defined by

$$D_A = 1 - \frac{1}{r}\sum_j^r \sum_i^{m_j} \sqrt{x_{ij} y_{ij}}$$

where x_{ij} and y_{ij} are the frequencies of the i-th allele at the j-th locus in populations X and Y, respectively, m_j is the number of alleles at the j-th locus, and r is the number of loci used. Note that Cavalli-Sforza and Edwards' (1967) chord distance (D_C) is

$$D_C = (2/\pi r)\sum_j^r \sqrt{2(1-\sum_i^{m_j}\sqrt{x_{ij}y_{ij}})}\,.$$

D_C for the *j*-th locus measures the chord distance of populations X and Y represented on the multidimensional hypersphere with coordinates of allele frequencies of this locus. The angle (q_j) of the two populations is given by $\cos\theta_j = \sum_i^{m_j}\sqrt{x_{ij}y_{ij}}$.

6.2 Nei's standard genetic distance without sample size bias correction (D_{ST})

Nei's standard genetic distance without sample size bias correction (D_{ST}) (Nei 1972) is defined by

$$D_{ST} = -\ln\frac{J_{XY}}{\sqrt{J_X J_Y}}$$

where $J_X = \sum_j^r\sum_i^{m_j} x_{ij}^2/r$ and $J_Y = \sum_j^r\sum_i^{m_j} y_{ij}^2/r$ are average homozygosities over loci in populations X and Y, respectively, and $J_{XY} = \sum_j^r\sum_i^{m_j} x_{ij}y_{ij}/r$. In the distance option "Dst (u)", the distance is calculated with this formula.

6.3 Nei's standard genetic distance with sample size bias correction ($\hat{D}_{ST}$)

Unbiased estimators of J_X and J_Y are $\hat{J}_X = \frac{1}{r}\sum_j^r(n_{Xj}\sum_i^{m_j} x_{ij}^2 - 1)/(n_{Xj}-1)$, and $\hat{J}_Y = \frac{1}{r}\sum_j^r(n_{Yj}\sum_i^{m_j} y_{ij}^2 - 1)/(n_{Yj}-1)$, where n_{Xj} and n_{Yj} are the number of chromosomes examined at the *j*-th locus for populations X and Y, respectively. An unbiased estimate of D_{ST} can be obtained by replacing J_X and J_Y with $\hat{J}_X$ and $\hat{J}_Y$, as shown below (Nei 1978).

$$\hat{D}_{ST} = -\ln\frac{\hat{J}_{XY}}{\sqrt{\hat{J}_X\hat{J}_Y}}\,.$$

where $\hat{J}_{XY} = J_{XY}$. In POPTREE, D_{ST} is computed with this formula.

In the infinite allele model (Kimura and Crow 1964), the expectation of D_{ST} increases

linearly with time for populations under the mutation-drift balance. That is, E(D_{ST}) = E($\hat{D}_{ST}$) = 2vt, where v is the mutation rate per locus per generation and t is the time in generations after the divergence of the two populations. In the infinite allele model a new allele is always created by a new mutation, and it can apply to classical markers such as blood groups and isozymes and other markers such as single nucleotide polymorphism (SNP).

6.4 F_{ST}^{*} distance

F_{ST}^{*} distance (Latter 1972) is given by

$$F_{ST}^{*} = \frac{(\hat{J}_X + \hat{J}_Y)/2 - \hat{J}_{XY}}{1 - \hat{J}_{XY}}$$

where $\hat{J}_X, \hat{J}_Y$, and $\hat{J}_{XY}$ are unbiased estimators of J_X, J_Y, and J_{XY} and computed by the formulas shown in 6.3 (Nei 1987). In the distance option "Fst" (with bias correction) the distance is calculated with the above formula. The distance option "Fst (u)" (without bias correction), the distance value is calculated by replacing $\hat{J}_X$ and $\hat{J}_Y$ with J_X and J_Y.

The expectation of F_{ST}^{*} is given by

$$E(F_{ST}^{*}) = 1 - e^{-t/(2N)}$$

if populations with the effective size N diverged t generations ago (Nei 1987, p359).

6.5 Distance measures developed for microsatellite DNA data

In the case of microsatellite DNA data, most of the changes by mutation are changes of the number of nucleotide repeats, and the majority of the repeat number changes occur by one. The mutational pattern of microsatellite loci roughly follows the stepwise mutation model (Ohta and Kimura 1973), in which the state of allele increases or decreases by one with an equal probability (Estoup, Jarne, and Cornuet 2002; Ellegren 2004). By taking into account the mutational pattern of microsatellite loci, $(\delta\mu)^2$ distance (Goldstein et al. 1995) and D_{SW}

(Shriver et al. 1995) were developed.

6.6 $(\delta\mu)^2$ distance

$(\delta\mu)^2$ distance (Goldstein et al. 1995) is given by

$$(\delta\mu)^2 = \sum_j^r (\mu_{X_j} - \mu_{Y_j})^2 / r,$$

where $\mu_{X_j} (= \sum_i i x_{ij})$ and $\mu_{Y_j} (= \sum_i i y_{ij})$ are average number of repeats of allele at the *j*-th locus. and x_{ij} and y_{ij} are the frequencies of the allele with *i* repeats at the *j*-th locus in populations X and Y. Under the stepwise mutation model (Ohta and Kimura 1973), $(\delta\mu)^2$ increases linearly with time for populations under the mutation-drift balance. $E[(\delta\mu)^2] = 2vt$, where *v* is a mutation rate per locus per generation and *t* is the number of generations after the two populations diverged.

6.7 D_{SW} distance

D_{SW} distance (Shriver et al. 1995) is given by

$$D_{SW} = W_{XY} - (W_X + W_Y)/2,$$

where $W_X = \sum_k^r \sum_{i \neq j} |i-j|\, x_{ik} x_{jk} / r$, $W_Y = \sum_k^r \sum_{i \neq j} |i-j|\, y_{ik} y_{jk} / r$, $W_{XY} = \sum_k^r \sum_{i \neq j} |i-j|\, x_{ik} y_{jk} / r$.

x_{ij} and y_{ij} are the frequencies of the allele with *i* repeats at the *j*-th locus in populations X and Y.

6.8 Number of repeats for allelic specification in the input file

In microsatellite DNA data the actual number of repeats of an allele is often unknown, but the fragment size (bp) of an allele is presented. If the repeat unit size (usually 2-5 bp) is known for the locus, (fragment size)/(repeat unit size) can be specified as the allele size in the input file for the computation of $(\delta\mu)^2$ and D_{SW}. In the computation of $(\delta\mu)^2$ and

D_{SW}, the number of nucleotides in the flanking regions outside the repeat region (divided by the repeat unit size) is subtracted.

6.9 Efficiencies of distance measures for constructing phylogenetic trees

Although $(\delta\mu)^2$ and D_{SW} were developed for microsatellite DNA data by taking into account the mutational pattern, the efficiency of these distance measures for constructing phylogenetic trees is low particularly for the data with a small number of loci. The probabilities of obtaining the correct tree topology were much higher for distance measures such as D_A and D_{ST} developed for classical markers than $(\delta\mu)^2$ and D_{SW} developed for microsatellite data in computer simulation (Takezaki and Nei 1996) and in the analysis of actual data (Takezaki and Nei 2008).

It should be noted that D_A distance appears to be more efficient in obtaining the correct tree topology for microsatellite data as well as for classical markers than other distance measures including D_{ST} (Takezaki and Nei 1996, 2008).

Please refer to Nei (1987) and Nei and Kumar (2000) for more details of the genetic distance measures.

7. Heterozygosity and G_{ST}

7.1 Heterozygosity

The heterozygosity of a locus (h) for a population is defined as

$$h = 1 - \sum_i^m x_i^2,$$

where m is the number of alleles for this locus, and x_i is the i-th allele of this locus. In the

heterozygosity is estimated by

$$\hat{h} = \frac{n}{n-1}(1 - \sum \hat{x}_i^2),$$

where n is the number of chromosomes examined. This is an unbiased estimator of h (Nei and Roychoudhury 1974). The average heterozygosity (H) over loci is estimated by

$$\hat{H} = \sum_j^r \hat{h}_j / r,$$

where r is the number of loci examined and h_j is the estimate of heterozygosity at the j-th locus. The sampling error of H is estimated as

$$S(\hat{H}) = \sqrt{V(\hat{h})/r}$$

where $V(\hat{h})$ is the variance of $\hat{h}$ and is given by

$$V(\hat{h}) = \sum_j^r (\hat{h}_j - \hat{H})^2 / (r-1).$$

The heterozygosity is also calculated without bias correction by replacing $\hat{h}$ with h in above formulae.

Please refer to chapter 8 [equations (8.3) – (8.8)] of Nei (1987).

7.2 G_{ST}, standardized G_{ST}, and Jost's D

G_{ST} is a measure of gene differentiation among subdivided populations and called the coefficient of gene differentiation (Nei 1973). It is defined for multiple alleles and is equivalent to Wright's fixation index (F_{ST}) for two alleles.

Let us consider the case where a population is subdivided into s subpopulations. The expected heterozygosity within populations for a locus (h_S) is given by

$$h_S = 1 - \sum_k^s \sum_i x_{ki}^2 / s$$

where x_{ki} is the frequency of the i-th allele in the k-th subpopulation.

The expected heterozygosity (h_T) for the total population is

$$h_T = 1 - \sum_i \bar{x}_i^2$$

where $\bar{x}_i = \sum_k^s x_{ki} / s$.

G_{ST} can be defined as

$$G_{ST} = (h_T - h_S) / h_T .$$

The estimates of h_S and h_T with sample size bias correction are given by

$$\hat{h}_S = n_m (1 - \sum \bar{\hat{x}}_i^2) / (n_m - 1), \text{ and}$$

$$\hat{h}_T = 1 - \sum \bar{x}_i^2 + \hat{h}_S / (n_m s),$$

respectively, where n_m is the harmonic mean of the number of chromosomes of subpopulations. These are equations (8.31) and (8.32) in Nei (1987). G_{ST} for all loci is computed by the average of $G_{ST} = (h_T - h_S) / h_T$ for each locus. The unbiased estimate ($\hat{G}_{ST}$) of G_{ST} is also computed by replacing h_S and h_T by $\hat{h}_S$ and $\hat{h}_T$, respectively. When sample size is small, $\hat{G}_{ST}$ sometimes become negative. However, the value of G_{ST} never becomes negative (Nei 1973, 1987).

The variance of G_{ST} is computed by the jackknife method.

For the theory of genetic variation in subdivided populations, see chapter 8 in Nei (1987) and chapter 12 in Nei and Kumar (2000). The theoretical advantages of G_{ST} over Wright's (1951) classical F_{ST} are discussed in Nei and Kumar (2000) and Crow (2004).

The value of G_{ST} may become considerably smaller than 1 when the mutation rate is high even if alleles are not shared by different populations (Nei and Kumar 2000). Because of this problem, Hedrick (2005) proposed standardized G_{ST} (G_{ST}') which is G_{ST} divided by the maximum value ($1 - h_S$).

$G_{ST}' = G_{ST} (s - 1 + h_S) / [(s - 1)(1 - h_S)]$

Jost (2008) developed a measure of genetic differentiation based on the effective number of alleles, which takes value of 0 to 1.

$D = k/(k-1)(h_T - h_S)/(1 - h_S)$.

Jost (2008) suggested use of unbiased estimators derived by Nei and Chesser (1983). These estimators are shown as equations (7.38) and (7.39) in Nei (1987) and equations (12.26) and (12.27) in Nei and Kumar (2000).

For computation of Hedrick's G_{ST}' and Jost's D, the estimates of h_T ($\hat{h}_T$) and h_S ($\hat{h}_S$) are also used because they generally give smaller sampling variances than those by Nei and Chesser (1983) under the assumption that each population is in Hardy-Weinberg equilibrium (Nei 1987). The variances of Hedrick's G_{ST}' and Jost's D are computed by the jackknife method.

8. Displays of phylogenetic trees

The shape of the phylogenetic tree can be changed by clicking the icons at the second row of the "Phylogeny" window.

Icon 1: Newick format of the phylogenetic tree (text file) is saved. This file can be used for further change of the appearance of the tree by other programs such as MEGA 6 (Tamura et al. 2013) (http://www.megasoftware.net/). MEGA 6 has more options for the appearance of the tree.

Icon 2: Print the tree. The tree in the "Phylogeny" window is printed.

Icon 3: Copy the tree to the Clipboard. The tree can be copied to other applications such as PowerPoint and Word, after clicking this icon and pressing Ctl+V in the other

application.

Icon 4: Root. After placing the cursor at a certain branch, clicking this icon will give the root of the tree on the branch. This function is available only for the NJ tree. The position of the root is not given by the NJ method. By default the root of the NJ method in the "Phylogeny window" is calculated by the mid-point rooting method, in which the root is placed in the mid-point of the longest path of two taxa. This function is not available for the UPGMA tree because the position of the root of the UPGMA tree is automatically given by the method.

Icon 5: Flip. After placing the cursor at a certain internal branch, clicking this icon will flip the two descendant clusters of the branch (like a mirror image).

Icon 6: Swap. After placing the cursor at a certain internal branch, clicking this icon will swap the two descendant clusters of the branch (vertical positions of the taxa within the cluster will remain the same).

Icon 7: Tree style (Traditional). Change the style of the tree to rectangular presentation.

Icon 8: Tree style (Radiation). Change the style of the tree to radial presentation.

Icon 9: Expand/Compress the tree in the horizontal direction.

Icon 10: Expand/Compress in the tree vertical direction.

Icon 11: Font of the taxon names can be changed.

Icon 12: Change the line width by point size.

9. References

Cavalli-Sforza LL, Edwards AWF. 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19: 233-257.

Chakraborty, R. 1977. Estimation of time of divergence from phylogenetic studies. Can J Genet Cytol 19:217-223.

Crow JF. 2004. Assessing population subdivision. In: Wasser SP, editor. Evolutionary Theory and Processes. Modern Horizons. Netherlans: Kluwer Academic Publishers. p 35-42.

Deka R, Shriver MD, Yu LM, Ferrell RE, Chakraborty R. 1995. Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis 16:1659-1664.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5: 435–445.

Estoup A, Jarne P, Cornuet J. 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.

Hedrick PW. 2005. A standardized genetic differentiation measure. Evolution 59: 1633-1638.

Jost L. 2008. G_{ST} and its relatives do not measure differentiation. Mol Ecol 17: 4015-4026.

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725-738.

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, Chesser RK. 1983. Estimation of fixation indices and gene diversities. Ann Hum Genet 47: 253-259.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

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.

Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201-204.

Rousset F. 2008. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol Ecol Resour 8: 103-106.

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. W. H. Freeman, San Francisco.

Takezaki N, Nei M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389-399.

Takezaki N, Nei M. 2008. Empirical tests of the reliability of phylogenetic trees constructed with microsatellite DNA. Genetics 178:385-392.

Tamura K, Stecher G, Peterson D, Filipsli A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol. 30: 2725-2729.

Wright S. 1951. The general structure of populations. Ann Eugen 15:323-352.

10. Author contact

Windows-interface of POPTREE2 is developed by Koichiro Tamura, Professor of Tokyo Metropolitan University, and a primary author of MEGA 6 (Molecular Evolutionary Genetics Analysis version 6), software for evolutionary genetics analyses of DNA and protein sequences.

Masatoshi Nei is Evan Pugh Professor of the Department of Biology at Pennsylvania State University and Director of the Institute of Molecular Evolutionary Genetics. He has developed a number of statistical methods in the study of population genetics, molecular evolution and phylogenetics including D_A and D_{ST} distances, the neighbor-joining method, and G_{ST} computed in POPTREE2. He is also an author of MEGA and started development of MEGA under his auspice. His webpage is http://www.bio.psu.edu/People/Faculty/Nei/Lab/.

If you have questions or problems, please let me know.

Naoko Takezaki
Life Science Research Center
Kagawa University
Ikenobe 1750-1, Mikicho, Kitagun
Kagawa 760-0793
Japan
takezaki at med.kagawa-u.ac.jp