Using the IMfig program

Jody Hey

Director, Center for Computational Genetics and Genomics
Department of Biology
Temple University
Philadelphia, PA 19122
hey@temple.edu

Table of Contents

Introduction

IMfig is a program for generating a figure of an isolation-with-migration model from the output file generated by running either the IMa2 or the IMa3 program.   IMfig generates an eps (encapsulated post script) file by default and will also generate files in other formats (jpeg, png, pdf).   Examples of the figures generated by IMfig are given in (Hey 2010; Hey, et al. 2018) and later in this document.

This kind of figure represents an estimate of history as a fat phylogenetic tree made up of boxes (for sampled and ancestral populations), horizontal lines (for splitting times) and curved arrows (for migration). Time is represented as depth on the vertical axis, with the sampled populations/species at the top of each figure at the most recent time. Population size is represented as width along the horizontal axis. Each population is represented by a box, the height of which refers to how long it has lasted and the width of which refers to its effective size. The confidence interval for a populations size is given both by a double headed arrow extending from the right margin of that population's box and by

faint boxes representing the lower and higher 95% Highest Posterior Density (HPD) intervals. Similarly 95% HPD intervals for splitting times are given by dashed lines and doubled-headed arrows.

Migration arrows refer to migration forward in time (i.e. an arrow from population 1 to population 2 means refers to genes moving from 1 to 2 each generation).   These arrows are generated with corresponding text of *2Nm* values. When printing migration arrows with only statistically significant (i.e. non-zero) migration rates (options –ms and –mS),) the meaning of asterisks are as follows: "*" for p<0.05; "**" for p < 0.01, and " ***" for p < 0.001.   In these cases the actual statistical test was conducted, not on *2Nm* but rather on the migration rate parameter (i.e. *m/u*) using the method of Nielsen and Wakeley (2001) .

The source code for IMfig is a Python script, and execution requires that Python be installed (http://www.python.org/). The latest version has been tested with Python 3.6 but will probably run ok with any version later than Python 2.7.   Users are also recommended to use an eps file viewer. Getting the settings right for a particular figure can take some playing around and may ultimately require several runs of IMfig. The program runs quickly, but the process of getting the best settings may require multiple runs while trying out various options.

Generating an IMa2/IMa3 output file suitable for IMfig

The file that is read by IMfig should be an output file generated using a run of IMa2 or IMa3 that generates estimates of population sizes and migration rates.   These runs will use a fixed topology that is included in the data file, and they will not use hyperpriors (an option in IMa3). IMfig extracts information from the phylogenetic tree in the original data file and from the following tables in the IMa2 or IMa3 output file:

- `Marginal Peak Locations and Probabilities`
- `MARGINAL DISTRIBUTION VALUES AND HISTOGRAMS OF POPULATION SIZE AND MIGRATION PARAMETERS`
- `MARGINAL DISTRIBUTION VALUES AND HISTOGRAMS OF PARAMETERS IN MCMC`
- `MARGINAL DISTRIBUTION VALUES IN DEMOGRAPHIC UNITS`

The last table (for demographic units) will be skipped by IMfig if the `-d` option is used on the command line. Doing an IMa2 or IMa3 analysis in demographic units requires that the program be run using demographic units, which requires in turn some information on mutation rates and generation times (see IMa2 or IMa3 documentation).

Command line options:

Running the program without any flags should cause the following text to appear on the screen:

```
IMfig command line terms  (-i is required):
-a : include ancestral population #'s in plot
-b : adjust width spacing of population boxes, values > 0, default = 1
-c : output format, default is eps, see also -w
      -c j : make a jpeg file
      -c p : make a pdf file
      -c n : make a png file
-d : do not use demographic scale information even if in input file
-e : space split times evenly  (not proportional to time,  no confidence intervals shown)
-f : font.  Default=Arial. Use postscript fonts available on the computer
     e.g. Arial, Helvetica, Times-roman, Courier
-g : global plot scale sets the size of the plot, max = 1, default = 1
-h : print only this help menu
-i : input file name
-j : arrow width, default = 1
-k : print population names on an angle
-m : options for printing of arrows and 2Nm values for migration :
      -m x :  do not print migration arrows
      -m a : 2Nm migration arrows for all 2NM > 0
      -m s : 2Nm migration arrows only if m is statistically significant p <= 0.05 (default)
      -m S : 2Nm migration arrows only if m is statistically significant p <= 0.01
      -m # : '#' is a number, migration arrows appear when 2NM >= # (e.g. -m0.1)
-n : file with alternative species names
-o : output file name, default is imfig_output
-p : fontsize (default is 14 for full scale, default follows global scale)
-q : no confidence interval boxes for population boxes printed
-r : no confidence interval arrows for population boxes printed
-s : print square, rather than landscape
-t : relative height of oldest time point, values between 0 and 1
     default value = 1/(# sampled populations+1)
-u : simple colors, blue for population boxes, red arrows for migration (default grayscale)
-v : multiple colors for population boxes, red arrows for migration (default grayscale)
-w : file image width,  integer multiple of 720 pixels (only if using -c)
-x : adjust width of plot,  >1 means wider, <1 means narrower
-y : adjust height of splittimes, relative to bottom of figure, max = 1.
-z : exclude the ghost population from the figure
```

Only the $-i$ (input file) flat is required in order to generate a figure. All other flags are optional. Several flags affect the size/shape of the plot. If you need to generate multiple figures in which the height and width of the figures are on the same scale (e.g. so that ½ inch of height corresponds to a certain length of time) then you can use the $-t$ and/or the $-y$ options. Similarly to have a unit distance on the horizontal axis mean the same thing (e.g. ½ inch corresponds to a certain effective population size) for different figures, you can play around with the $–x$ option. If you use $–g$ to make the plot bigger or smaller you may also need to adjust the font size ($-f$) and the arrow head size ($-j$).

Running IMfig

IMfig.py was tested on python version 3.6 and may run ok on earlier versions, but has not been tested thoroughly in this way. Running IMfig.py in a shell window (e.g. cmd, terminal, bash shell etc) should be fairly straightforward. For example if your IMa2 or IMa3 output file is called `my_results.out` and you want the eps file to be named `my_results.eps` then enter the following at the command prompt:

```
python IMfig.py -imy_results.out -omy_results
```

(Note that this example assumes the operating system knows where python is (i.e. it is in the current directory or in the system path), and that IMfig.py and my_results.out are in the current directory.)

Functionality of the **-c** option (writing files in formats other than eps) requires that the python pillow module (https://python-pillow.org/) be installed. The pillow module requires in turn that Ghostscript be installed (https://www.ghostscript.com/). IMfig.py will run without these but the corresponding command line options will not work. IMfig.py also uses the Colormath module, if it is installed (https://pypi.python.org/pypi/colormath/), for better color picking when using -v , though it does not make a big difference.

Output Files

By default, IMfig generates a vector-based encapsulated postscript (eps) file. There are many programs for viewing and editing eps files such as GIMP, Inkscape and Illustrator. If the -c option is used then IMfig can generate an image in one of several other formats, including jpeg, png and pdf. The -w option can be used in these cases to control resolution.

IMfig also generates a text file with name ending in '_caption.txt' that describes the figure and that can be used as a starting point for preparing a caption in a publication.

Citing IMfig

The first version of the program was written to generate figures for (Hey 2010), however it was not actually mentioned in that paper. Nevertheless based on the actual history and precedent that is probably the best paper to cite. Alternatively one could cite the more recent (Hey, et al. 2018).

References:

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.

Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff SA, Soudsa VC, Wang Y. 2018. Phylogeny Estimation by Integration over Isolation with Migration Models. submitted.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Examples

On the next several pages are shown some results using IMa3 output files from analyses of a 3

population model and 5 population model with a ghost. On each page is given a command line, a brief

description, the resulting figure and the text from the caption file.

Example 1: Default settings
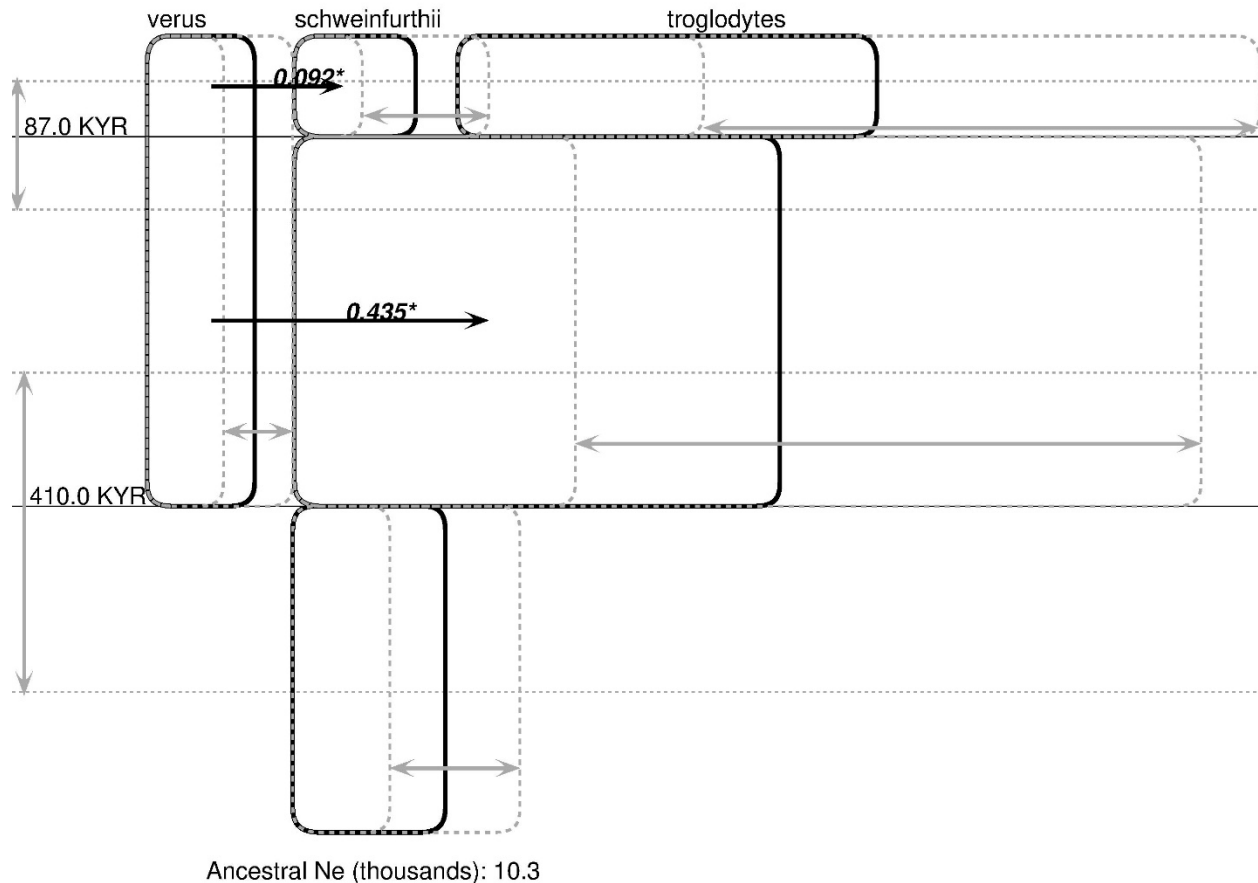
```
python IMfig.py -iIMa3_3pop_example.out
```


Figure 1. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 2. Demonstrating the use of the –d option to use mutation rate scales rather than demographic scales.

```
python IMfig.py -iIMa3_3pop_example.out –d
```
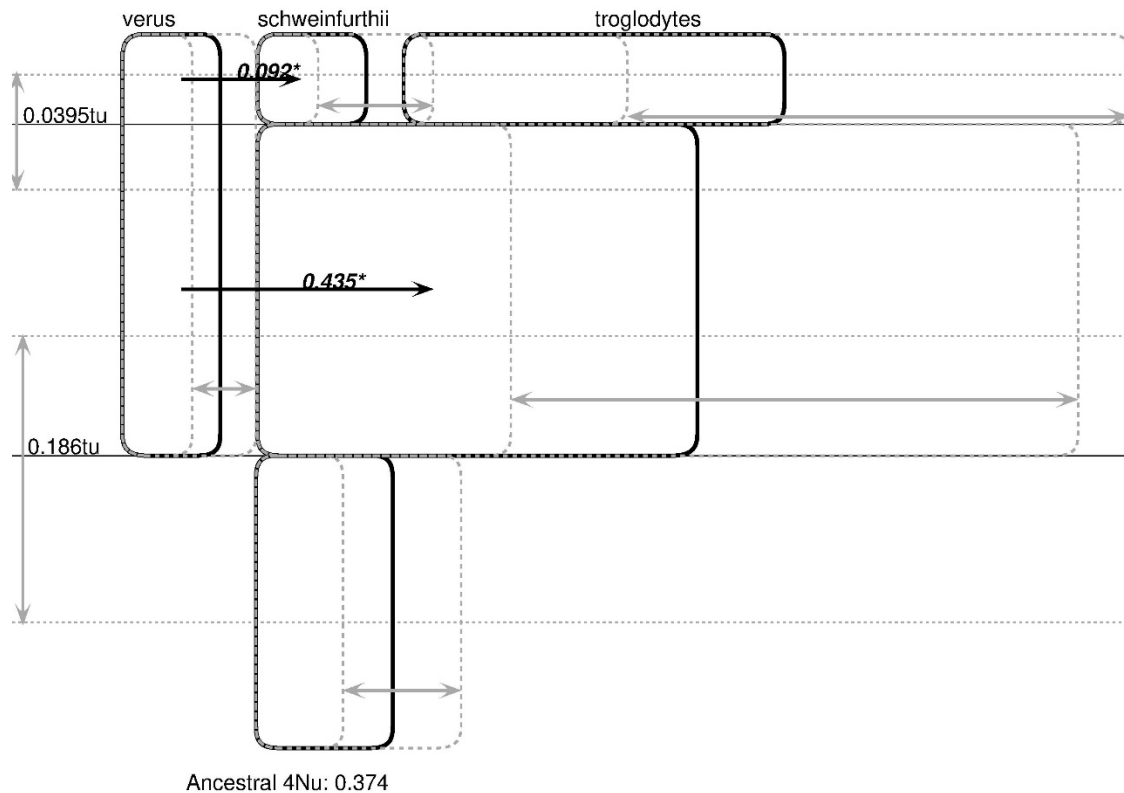

Figure 2. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001). Population size (Ne) and splitting times values are scaled by the geometric mean of the mutation rates of the loci used for the analysis.

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 3. Demonstrating the use of the –a option to have the population label number for ancestral populations included in the figure

```
python IMfig.py -iIMa3_3pop_example.out -a
```
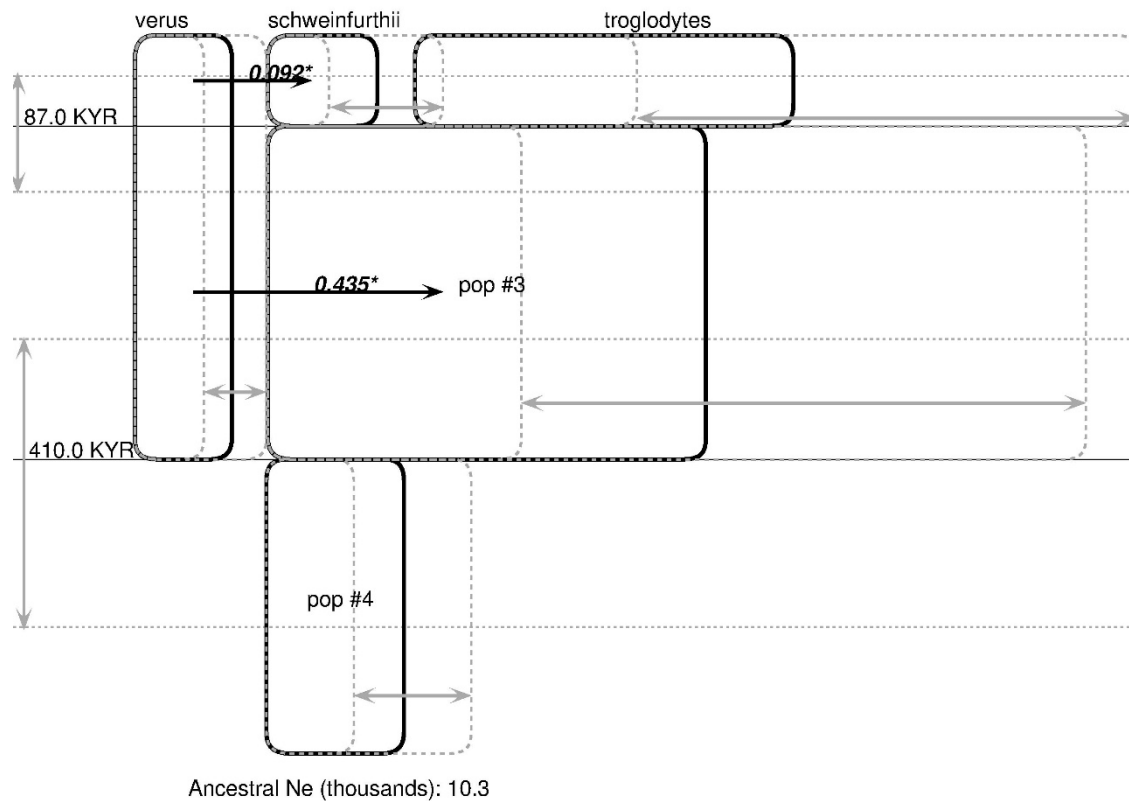

Ancestral Ne (thousands): 10.3

Figure 3. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001). Ancestral population numbers are shown in ancestral boxes.

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 4. Demonstrating the use of the –g option to rescale the entire plot, in this case a much smaller figure.

```
python IMfig.py -iIMa3_3pop_example.out -g0.3
```
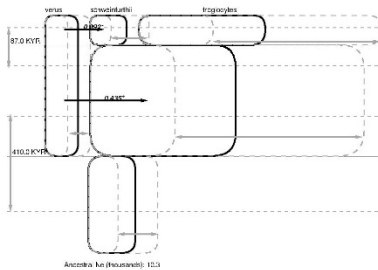

Figure 4. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 5. Demonstrating the use of the –ma option to print all arrows for all migration rates where 2NM>0

```
python IMfig.py -iIMa3_3pop_example.out –m a
```
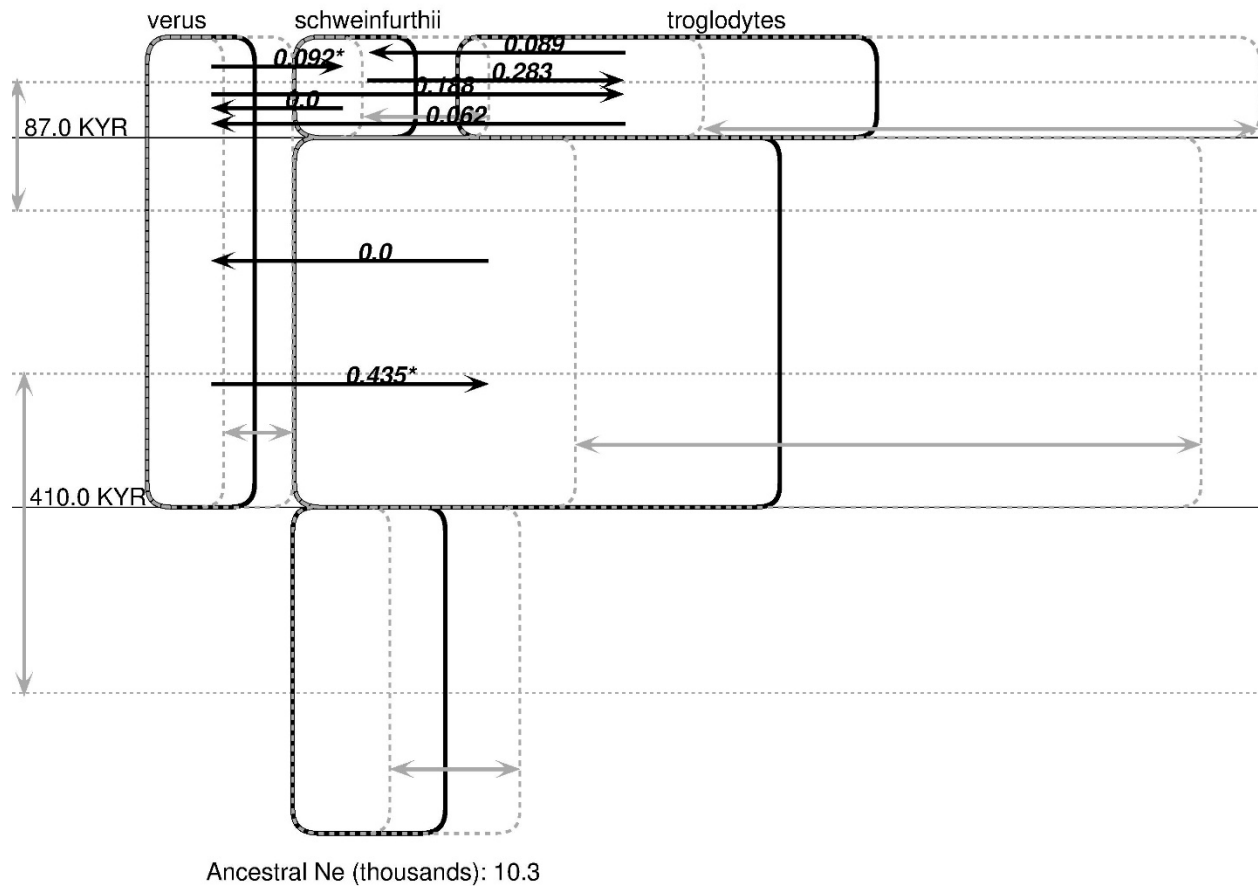

Figure 5. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.

Example 6. Demonstrating the use of the –b option to alter the width of population boxes to avoid overlapping confidence intervals

```
python IMfig.py -iIMa3_3pop_example.out –b3.0
```


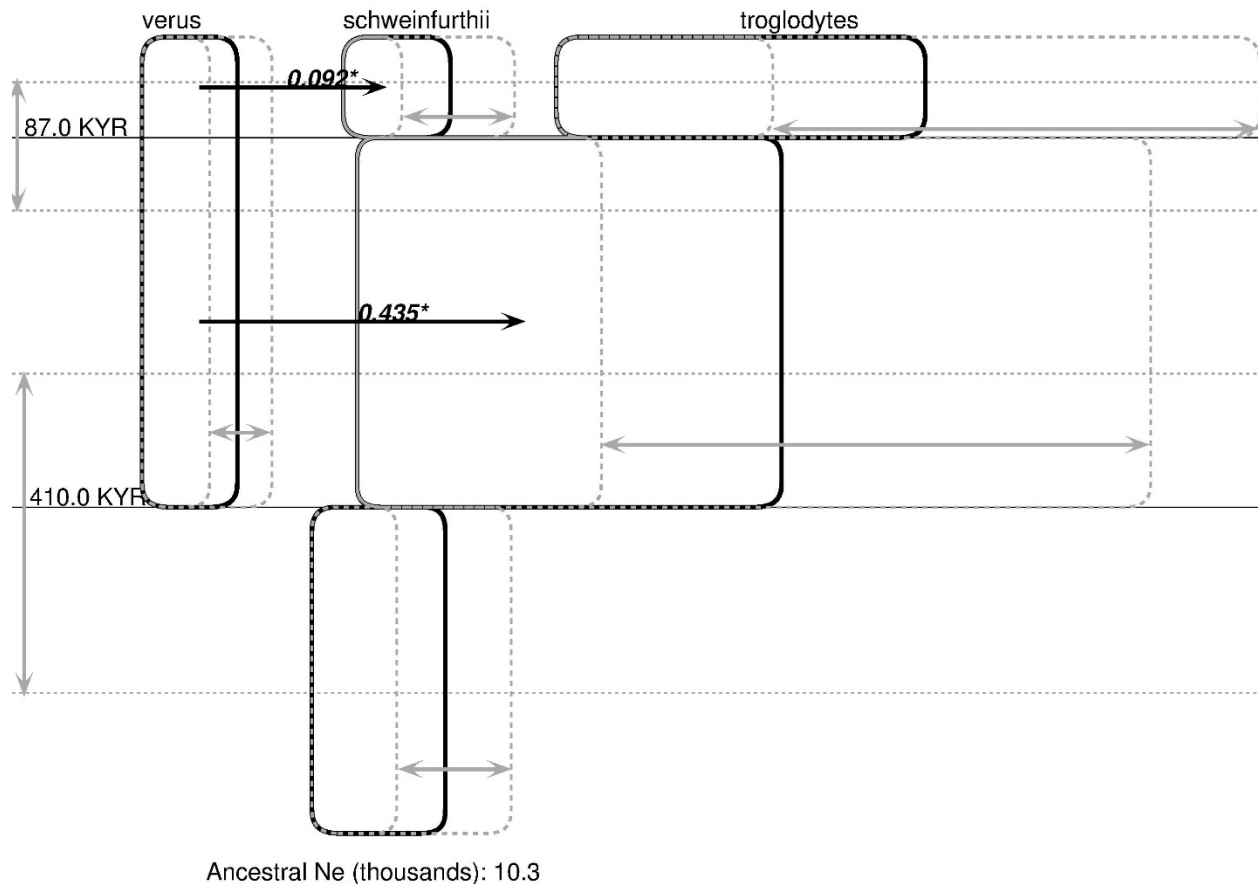Ancestral Ne (thousands): 10.3

Figure 6. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 7. Demonstrating the use of the –c option to print in color
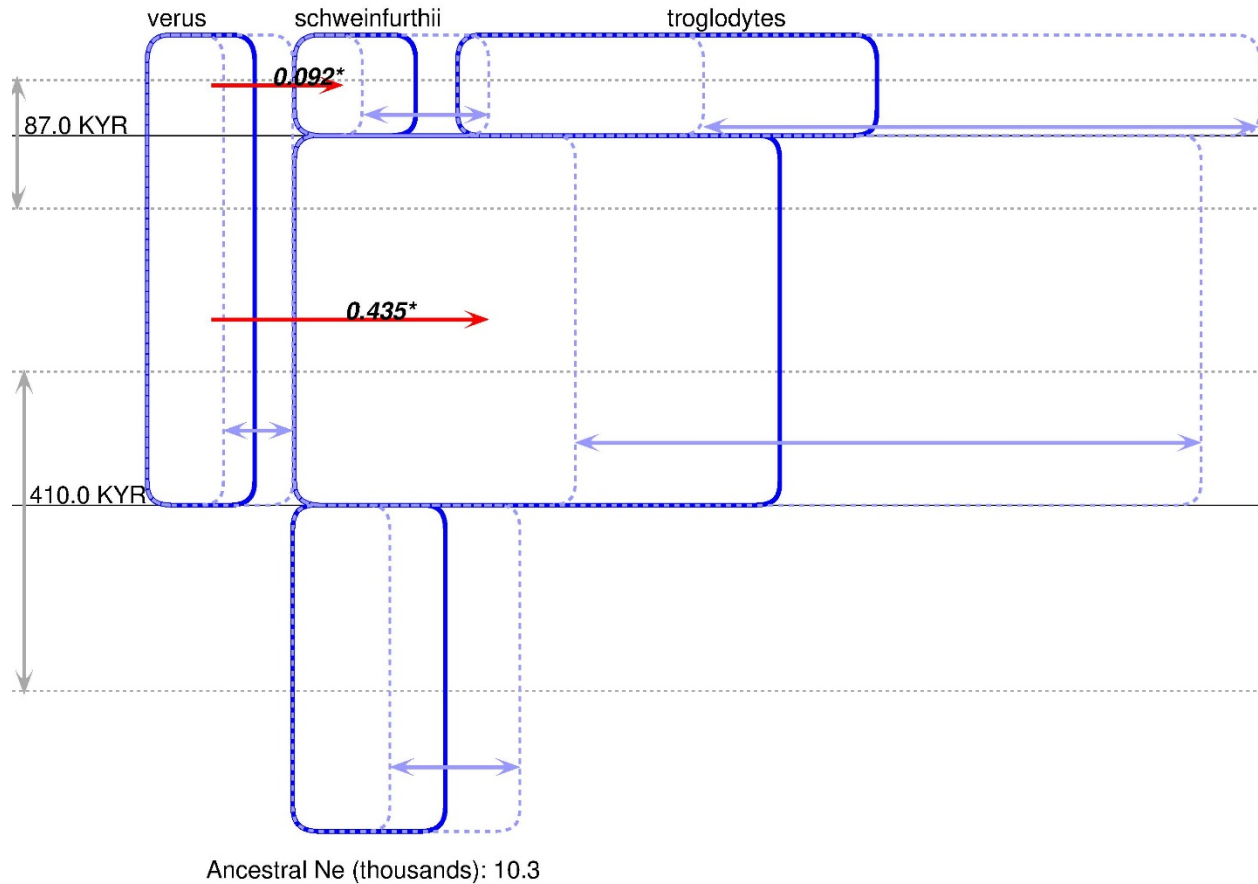
```
python IMfig.py -iIMa3_3pop_example.out -u
```


Figure 7. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 8. Demonstrating the use of the –v option to print in color

```
python IMfig.py -iIMa3_3pop_example.out -v
```
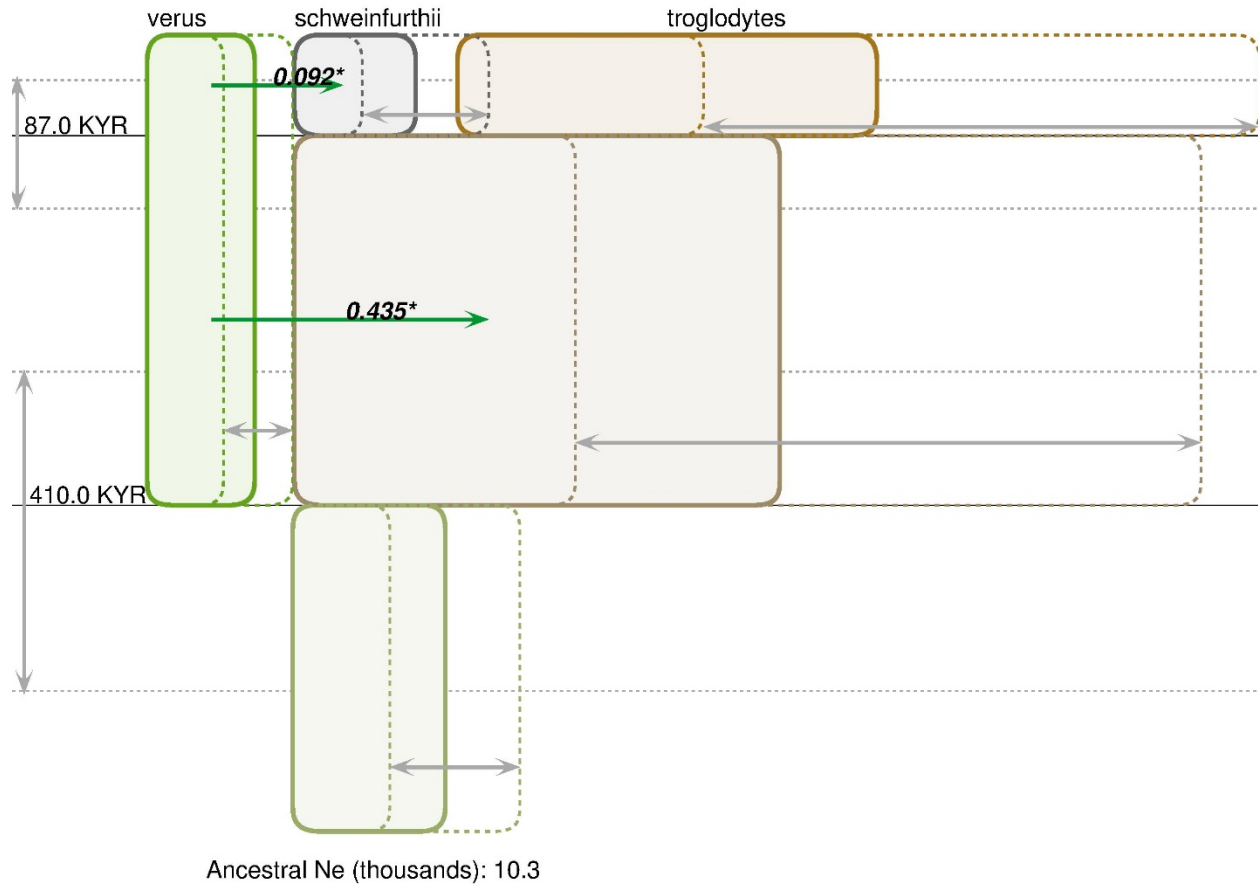

Figure 8. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 9. Demonstrating the use of the -k –n and –v options to print a figure with alternative population names, on an angle, and with color . For an input file with m populations, the file with alternative names should have m lines with line j having the alternative name for population j in the input file.

```
python IMfig.py -iIMa3_3pop_example.out -k -n popnames.txt -v
```
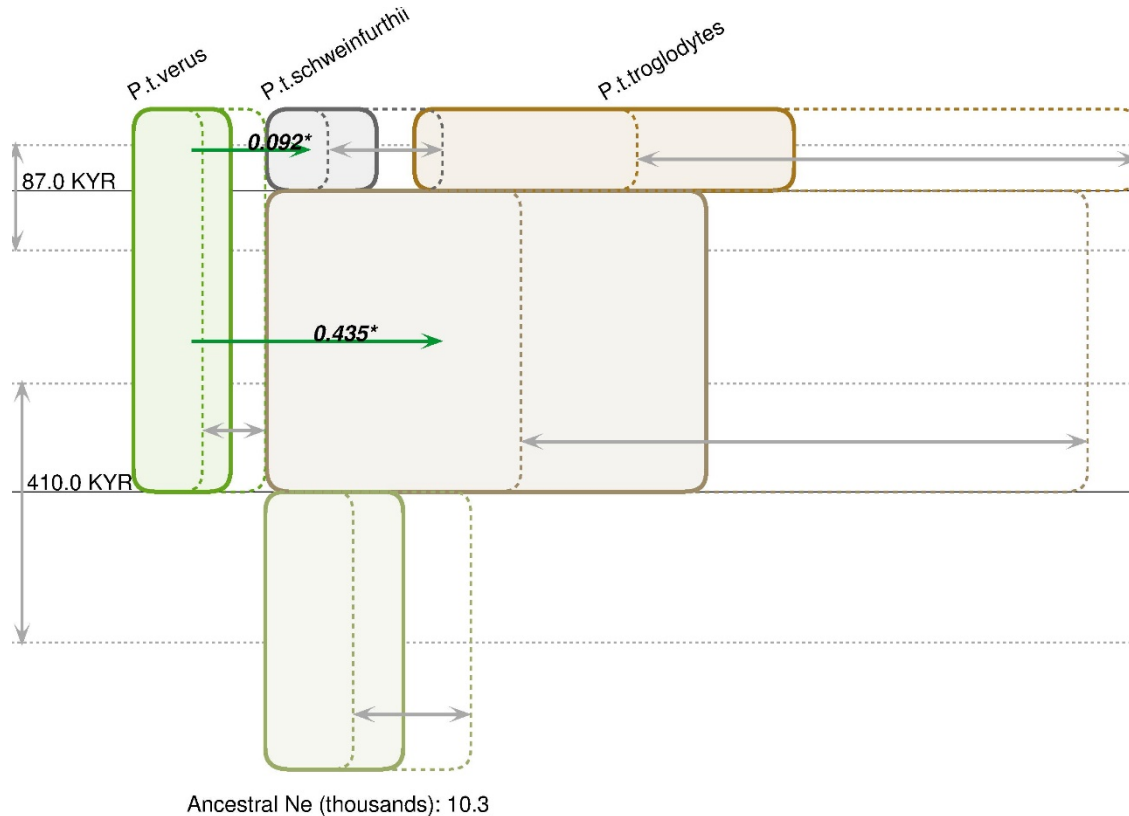

Figure 9. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* $p < 0.05$,** $p < 0.01$,*** $p < 0.001$).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 10. Demonstrating the use of the –q –r and –e options.

```
python IMfig.py -iIMa3_3pop_example.out -q -r -e
```
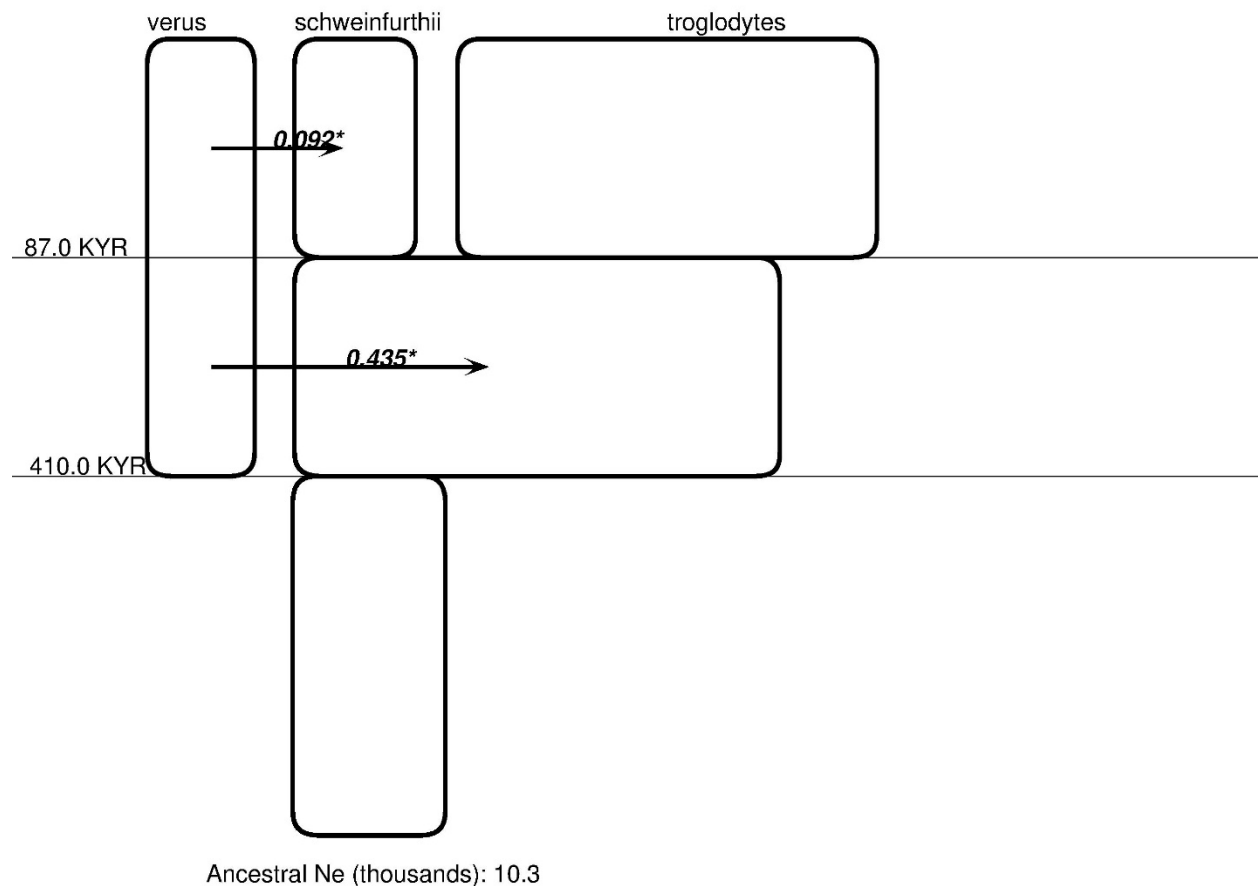

Figure 10. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne.Splitting times, positioned at even intervals, are depicted as solid horizontal lines, with text values on the left. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 11. Demonstrating inclusion of a ghost population

```
python IMfig.py -i Pan_ghost.out -k -n pan5pop_names.txt  -v
```
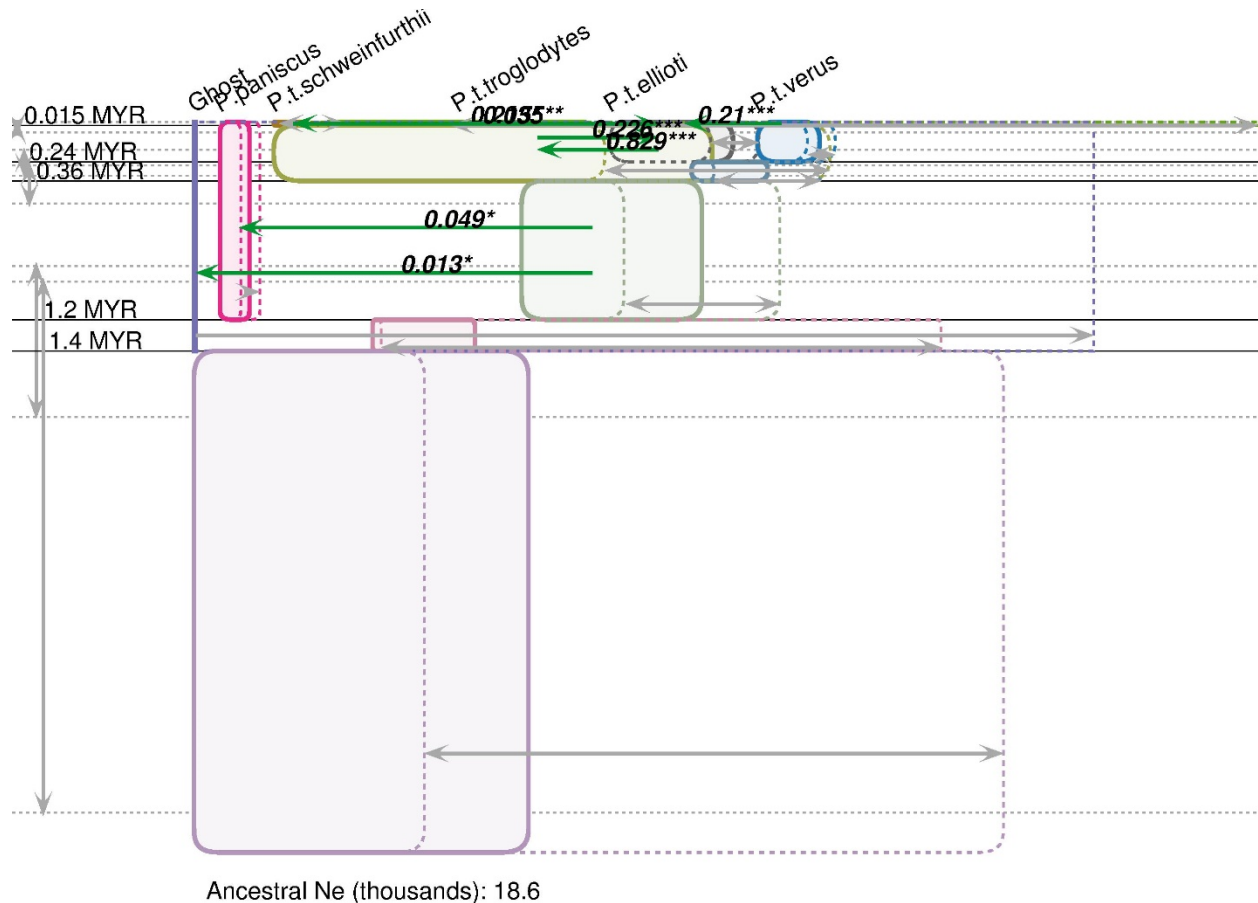

Ancestral Ne (thousands): 18.6

Figure 11. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for each Ne value are shown as dashed lines to the right of the left side of the corresponding population box. Gray arrows to the 95% Ne intervals are also shown extending to the left and right of the right boundary of each population box. Splitting times are depicted as solid horizontal lines, with text values on the left. Confidence intervals for splitting times are shown as vertical gray arrows on the left, and parallel dashed lines. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001).

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.

Example 12. Demonstrating exclusion of a ghost in the figure (when a ghost was included in the analyses);   also other formatting options. Compare with Example 11 run on the same out file.

```
python IMfig.py -i Pan_ghost.out -k -n pan5pop_names.txt -e -v -z -q
```
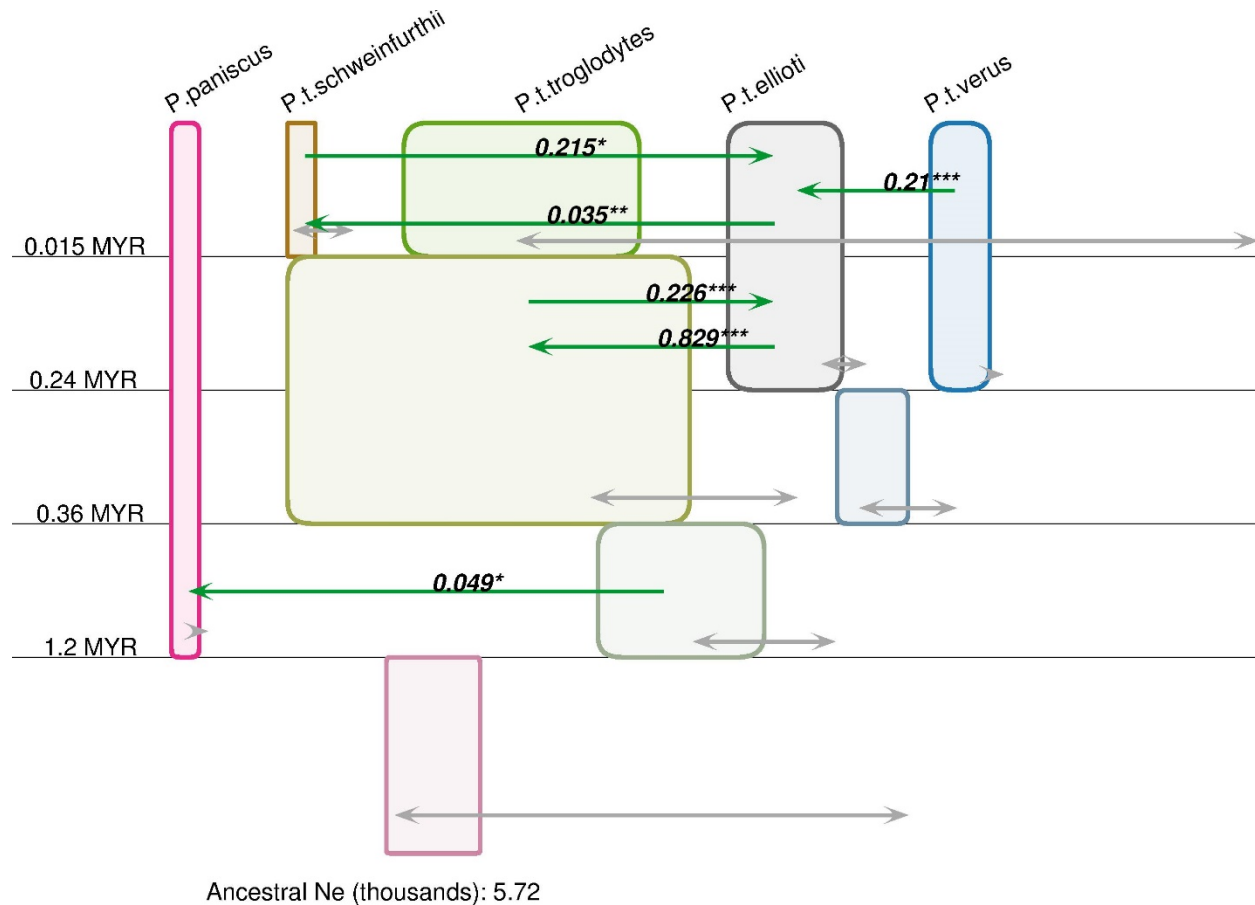

Figure 12. A representation of an estimated Isolation with Migration model generated by the IMfig program (Hey 2010). The phylogeny is depicted as a series of boxes organized hierarchically, with ancestor boxes positioned in between the corresponding descendants, and the width of boxes proportional to estimated Ne. 95% confidence intervals for Ne values are shown as gray arrows extending to the left and right of the right boundary of each population box. Splitting times, positioned at even intervals, are depicted as solid horizontal lines, with text values on the left. Migration arrows (if shown) indicate estimated 2Nm values from one population to another over the time interval when both populations exist. Arrows are shown only for estimated migration rates that are statistically significant (Nielsen and Wakeley, 2001) at or above the 0.05 level (* p < 0.05,** p< 0.01,*** p < 0.001). The ghost population is not shown in this figure.

Hey J. 2010. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. Mol Biol Evol 27:921-933.
Nielsen R, Wakeley J. (21297244 co-authors). 2001. Distinguishing migration from isolation. A Markov chain Monte Carlo approach. Genetics 158:885-896.