Contents

PopART (Population Analysis with Reticulate Trees) is free, open source population genetics software that was developed as part of the Allan Wilson Centre Imaging Evolution Initiative. This is a collaborative project involving mathematicians and biologists from five universities and research institutes across New Zealand to develop better software to understand evolutionary relationships among populations.

When you launch PopART, the main window will contain a data view area on the left, and a network view area on the right. You can switch between the network and map views to visualise your data in different ways.

You can interact with PopART using the menu bar and the toolbar.

Menus

File

The file menu lets you get data into and out of PopART. The standard "open", "close", and "save as" items do what you would expect them to. PopART's normal file format for reading and saving files is Nexus. When you save your work, PopART will prompt you for a new file name. Normally you can save to the same name as your input file. However, if your input Nexus file contains blocks that PopART doesn't use, you should save to a new file, because these won't be saved. Any colours or fonts that you change in your network will be saved along with your network.

You can also import alignments in Phylip formats and other data in CSV format, or space- or tab-delimited tables. You can export networks and maps as graphics in PNG, SVG, and PDF formats. You can also export your network as a table if you'd like to be able to open it with other programs that can import networks as text files.

◎ Read more about getting data into PopART

◎ Read more about importing tables
◎ Read more about Nexus format

Edit

The edit menu lets you modify various attributes of the network or map view. You can change colours, fonts, and node sizes. You can also redraw the network. This can be useful if you think your network is especially hard for PopART to draw: in this case, you should probably increase the number of iterations used for drawing.

Network

The network menu allows you to infer haplotype networks from your sequence data using a number of different algorithms.

◎ Read more about network inference methods

View

The view menu lets you switch between network and map views. It also lets you change certain aspects of how your data are displayed in these views. If you have both geotag and other trait information loaded, you can toggle between the two to colour your network, or to place your sequences on the map. You can also change the way mutations are displayed on networks, and the map's theme. Finally, you can choose whether you want to display a legend on the map or in a separate window.

Statistics

The statistics menu lets you see which sequences PopART sees as identical, as well as look at various statistics about your data. You can also perform all statistical tests (and optionally save the results to a file). Note that for Tajima's D statistic and AMOVA, significance is assessed using a parametric distribution (a beta distribution for Tajima's D, and an F distribution for AMOVA).

Segretating sites: The number of sites that differ among sequences
Parsimony-informative sites: The number of sites containing at least two states that occur in at least two sequences each
Tajima's D: A measure of selection (see references).
AMOVA: A measure of molecular variation among demes (see references). This uses the groups defined by geotags or traits.

Toolbar

The toolbar includes buttons to open and save Nexus files, and some other useful tools for interacting with networks and maps. You can change the colours using several pre-defined themes, zoom in and out of networks, and rotate networks, and you can show or hide viewers that let you see the names of identical sequences at network nodes, or view trait distributions as bar charts, rather than pie charts.

Data View

The data view lets you look at your alignment or traits data. The traits viewer displays the currently "active" traits. If you have traits associated with both geotag and regular traits data, you can toggle between the two using the "view" menu. This will be reflected in the traits viewer, as well as in the map or network view. Currently, PopART does not allow data editing, but this feature will be added in an upcoming release. You can undock this window or dock it in at the left or bottom of the main window.

Map View

The map view area lets you place your sequences on a map. You can do this by including at least one type of data in your input nexus file. You can include geotags for your sequences in the `GeoTags` block in your Nexus file. You can also include geographical coordinates for the traits described in your `Traits` block. Finally, if your traits represent geographical locations, you can give them sensible names that will be used to search for geographical coordinates. You can also import geotag or trait data.

When you switch to the map view, PopART will search for geographical coordinates for your traits, unless you have provided them already. You'll be asked to confirm the results. You should then see a pie chart representing the frequencies of the different sequences at each location. Don't worry if you made a mistake (or if PopART made a mistake) identifying your locations, you can right-click on the pie charts and tell PopART to move them. You can change the colours associated with individual sequences by right-clicking on the keys in the legend, or if you'd like to change to a new pre-defined colour theme, you can use the "Change colour theme" button on the toolbar.

You can move the map around and zoom in or out until you're happy with the way things look. If you have a lot of sequences, you might find that the legend obscures too much of the map: you can use the "view" menu to show the legend in a separate window.

Network View

The network view window lets you view and interact with your network. You can move nodes around, or move labels independently. You can move the legend around, or right-click on keys to change the colours associated with different traits. You can grab the edges of the network plot area (indicated by a dotted line) to change the size of the area that will be saved if you export your network as a graphics file. The toolbar includes tools that let you do various other things to change the way your network looks, as do the "edit" and "view" menus.

Data Input

PopART reads Nexus-format alignments. An example (including the optional trees and traits blocks) can be found here. Currently binary (`Standard`) and nucleotide data types are supported.

Your data file must have an alignment, either in `Taxa + Characters` blocks, or a `Data` block. If you want the nodes to be coloured according to phylogeographic or phenotype information (or in some other sensible way), you need to include a `Traits` block as well. You can also include geotagged sequences in the `GeoTags` block. These will be used to cluster sequences into groups that can be used to colour networks. See the example for the syntax for `Traits` and `GeoTags` blocks. In both cases, coordinates may be given as negative numbers to indicate southern or western hemispheres, or hemisphere may be specified with a letter (e.g. 25.61S or 76W).

You can also include a `Trees` block, if you want to be able to infer Ancestral Parsimony networks. This method is supposed to use all most parsimonious trees, but if you'd like to use it with some other set of trees (e.g., some trees produced by another inference method, or trees within 1 or 2 steps of the MP tree), the method should still produce a network.

At the moment you can't use PopART to edit either the alignment or the traits data, but this feature will be added at some point.

PopART can also import alignments in Phylip formats and traits or geotags in tabular formats (tab, space, or comma-delimited).

Nexus Format

This is an example of a Nexus file, modified from a sample file distributed with DNASP. I've separated the individual blocks with text describing the important points, but they should all be included in a single text file, as shown in here. The example should run in PopART if you copy it into a text file. Note that any text appearing between square braces is a comment and will be ignored by the program.

Nexus files begin with the #NEXUS command. All "blocks" begin with the keyword Begin, followed by the title of the block (and a semi-colon), and end with the keyword End;.This example then contains a Taxa block, followed by a Characters block. Nexus allows sequence alignments to be stored in either these two blocks, or else in a single Data block. Either option can be read with PopART.

```
#NEXUS

BEGIN TAXA;
DIMENSIONS NTAX=7;

TAXLABELS
seq_1
seq_2
seq_3
seq_4
seq_5
seq_6
seq_7
;

END;
```

In this example, the Characters block includes the MatchChar argument (indicating that the '.' character means that a sequence contains the same state as the first sequence at the position in question). This is allowed, but not necessary: you can explicitly include all characters for all sequences.

```
BEGIN CHARACTERS;
DIMENSIONS NCHAR=56;
FORMAT DATATYPE=DNA MISSING=? GAP=- MATCHCHAR=. ;
```

```
MATRIX

seq_1 ATATACGGGGTTA---TTAGA----AAAATGTGTGTGTGTTTTTTTTTTTCATGTGG

seq_2 ......---.A..---...C.-----.G...C.A...C..C...C............

seq_3 ..........A..---...T.-----.G............................

seq_4 ..........A..---G...T----................................A

seq_5 ..........A..---G...G----................................C

seq_6 ..........A..---G...C----................................T

seq_7 ..........A..---G....----................................A

;

END;
```

The `Traits` block is specific to PopART. The numbers in the matrix are numbers of samples associated with each trait. The order of the columns should match the order of `TraitLabels`. `Separator` can be `Comma`, `Space`, or `Tab`.

You can optionally include `TraitLatitude` and `TraitLongitude` if you have geographical locations associated with your trait labels (these should again be in the same order as `TraitLabels`). If you do not include these keywords, PopART will try to search for your `TraitLabels` as though they name geographical locations when you switch to the map view.

```
BEGIN TRAITS;
  Dimensions NTRAITS=5;
  Format labels=yes missing=? separator=Comma;
  TraitLatitude 53 43.6811 5.4 -25.61 -0;
  TraitLongitude 16.75 87.3311 26.5 134.355 -76;
  TraitLabels Europe Asia Africa Australia America;
  Matrix
seq_2 10,5,0,6,0
seq_7 0,0,5,0,0
seq_5 4,0,10,0,0
seq_4 0,0,0,4,2
seq_3 0,0,0,3,5
seq_1 0,0,0,3,3
seq_6 0,0,0,7,3
;

END;
```

The `GeoTags` block is also specific to PopART. It lets you specify geographical coordinates for each sequence. These will be

clustered into the number of groups specified by NClusts. Clustering is done using the k-means algorithm.

There are a few keywords you can optionally include in the GeoTags block. If you specify ClustLatitude and ClustLongitude, PopART will use these as centres for sequence clustering: sequences will be assigned to the nearest centre. If you include ClustLatitude and ClustLongitude, you can also include ClustLabels, which will be used as names for the sequence clusters.

The lines in the Matrix each contain sequence name, latitude, longitude, number of samples. Sequences may appear more than once. Number of samples is optional. You can also give a number indicating which cluster a sequence belongs to (between 1 and NClusts).

```
Begin GeoTags;
Dimensions NClusts=5;
Format labels=yes separator=Spaces;
ClustLatitude 53 43.6811 5.4 -25.61 -0;
ClustLongitude 16.75 87.3311 26.5 134.355 -76;
ClustLabels Europe Asia Africa Australia America;
Matrix
seq_5 48.3621687364 -10.6803478956 4
seq_5 20.9011934738 41.0061010737 10
seq_4 -34.5248765744 156.980424598 4
seq_4 6.47997227196 -126.310890191 2
seq_7 12.3724817276 32.9959970947 5
seq_6 -25.3125032726 132.069876577 7
seq_6 -17.5420017453 -81.8875028076 3
seq_1 -28.1278668152 133.927583142 3
seq_1 1.90619338755 -100.675193346 3
seq_3 -9.2031496606 156.495614547 3
seq_3 2.76141711712 -95.3738206052 5
seq_2 43.4068001627 40.4459241442 10
seq_2 36.0290665027 68.1753092664 5
seq_2 -21.8525898158 120.606615453 6
;
End;
```

This is a Trees block (this was saved by PAUP*, but any Nexus-format Trees block will do). You only need this block if you want to build an Ancestral Parsimony network. These were all MP trees, but feel free to use ML or Bayesian trees, or sub-optimal parsimony trees, or any other trees you feel you'd like to use with the Ancestral Parsimony algorithm.

```
Begin trees;  [Treefile saved Sun Jun 10 12:51:09 2012]
[!
>Data file = /home/jleigh/workspace/Hapnet/examples/dnasp.nex
>Heuristic search settings:
```

```
>   Optimality criterion = parsimony
>     Character-status summary:
>       Of 56 total characters:
>         All characters are of type 'unord'
>         All characters have equal weight
>         45 characters are constant
>         8 variable characters are parsimony-uninformative
>         Number of parsimony-informative characters = 3
>     Gaps are treated as "missing"
> Starting tree(s) obtained via stepwise addition
> Addition sequence: simple (reference taxon = seq 1)
> Number of trees held at each step during stepwise addition = 1
> Branch-swapping algorithm: tree-bisection-reconnection (TBR)
> Steepest descent option not in effect
> Initial 'MaxTrees' setting = 100 (will be auto-increased by 100)
> Branches collapsed (creating polytomies) if maximum branch
   length is zero
> 'MulTrees' option in effect
> Topological constraints not enforced
> Trees are unrooted
>
>Heuristic search completed
>   Total number of rearrangements tried = 976
>   Score of best tree(s) found = 16
>   Number of trees retained = 14
>   Time used = <1 sec (CPU time = 0.00 sec)
]
  Translate
    1 seq_1,
    2 seq_2,
    3 seq_3,
    4 seq_4,
    5 seq_5,
    6 seq_6,
    7 seq_7
    ;
tree PAUP_1 = [&U] (1,((2,3),((4,6,7),5)));
tree PAUP_2 = [&U] (1,((2,3),(((4,7),6),5)));
tree PAUP_3 = [&U] (1,((2,3),(((4,7),5),6)));
tree PAUP_4 = [&U] (1,((2,3),(((4,6),7),5)));
tree PAUP_5 = [&U] (1,((2,3),((4,7),(5,6))));
tree PAUP_6 = [&U] (1,((2,3),((4,(5,6)),7)));
tree PAUP_7 = [&U] (1,((2,3),(4,(5,6),7)));
tree PAUP_8 = [&U] (1,((2,3),((4,5),6,7)));
tree PAUP_9 = [&U] (1,((2,3),(4,5,6,7)));
tree PAUP_10 = [&U] (1,((2,3),((4,6),5,7)));
tree PAUP_11 = [&U] (1,((2,3),(((4,6),5),7)));
tree PAUP_12 = [&U] (1,((2,3),((4,5,7),6)));
tree PAUP_13 = [&U] (1,((2,3),(((4,5),7),6)));
tree PAUP_14 = [&U] (1,((2,3),(((4,5),6),7)));
End;
```

The Networks

PopART includes several network inference methods. Minimum Spanning, Median Joining, and TCS are all popular methods that have been described elsewhere. Ancestral Parsimony, Integer Neighbor-Joining, and Tight Span Walker are new methods that I'll describe briefly here.

Ancestral Parsimony

This method shows different ways the haplotype sequences might be connected through ancestral sequences inferred by the Maximum Parsimony criterion. You'll find that these networks tend to have a lot of edges, because there's often a lot of uncertainty in the inference of ancestral sequences. You can reduce this somewhat by increasing the "Minimum Ancestral Frequency" value.

Integer Neighbor-Joining

This method begins with a Neighbor-Joining tree, but sets the branch lengths to integer values so that they represent the number of mutations between sequences. Then, for pairs of haplotypes whose distances on the tree are longer than the distances between the sequences, edges are added to shorten the distance. New edges won't be added if they create any distances on the network that are shorter than the distances between sequences. If a new edge is longer than the improvement produced by its addition (i.e., the difference between the distance on the network and the sequence distance), it won't be added. You can increase the "Reticulation tolerance" to allow the addition of these edges if they aren't much too long.

Tight Span Walker

This method uses some properties of the tight span for the distance matrix for your sequences to build a haplotype network. We'll be describing this method in detail soon, but it's beyond the scope of this documentation. The distances on the network should match your sequence distances exactly.

References

1. French N, Yu S, Biggs P, Holland B, Fearnhead P, Binney B, Fox A, Grove-White D, Leigh JW, Miller W, Muellner P, Carter P (2014). Evolution of Campylobacter species in New Zealand. In: Sheppard SK & Méric G. (Eds), *Campylobacter Ecology and Evolution* (221–240). Norfolk, England: Caister Academic Press.
2. Bandelt H, Forster, P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*, 16(1):37–48.
3. Clement M, Snell Q, Walker P, Posada D, Crandall K (2002). TCS: Estimating gene genealogies. *IPDPS* 2:184.

Importing Tables

You can import two kinds of data that will be used to colour the nodes in your haplotype networks. Geotags are latitude/longitude pairs that are associated with each sequence. Traits are arbitrary labels associated with one or more sequences (and each sequence can be associated with more than one trait). Both are imported from text files containing tables (either tab- space-, or comma-delimited).

Geotags Tables

Geotag tables are fairly strict. There should be a row for each sequence, containing the latitude, longitude, and (optionally) number of samples. Sequences can appear more than once (i.e., if a sequence was found in more than one location). If the number of samples is left out, PopART will assume that a single sequence was found at the given latitude and longitude. You can optionally include a header row, but it will be ignored, so make sure you give latitude and longitude in the correct order.

For southern hemisphere latitudes or western hemisphere longitudes, negative numbers can be used. Alternatively, you can specify a hemisphere (e.g., 25.61S or 76W, equivalent to -25.61 or -76, respectively). You can specify hemisphere for northern and eastern hemispheres as well (e.g., 15N or 62E).

```
seq_2 54.756365N 20.559500E 10
seq_2 43.362073N 85.087465E 5
seq_2 27.743932N 134.852156E 6
seq_7 6.903639N 28.416291E 5
seq_5 53.102734N 14.241575E 4
seq_5 5.671317N 26.812297E 10
seq_4 26.568643N 137.810478E 4
seq_4 10.557455N 69.352084E 2
seq_3 26.903287N 136.965627E 3
seq_3 6.620725N 72.943461E 5
seq_1 26.216272N 135.401545E 3
seq_1 13.450485N 69.562928E 3
seq_6 22.851512N 134.605860E 7
seq_6 8.989081N 74.288348E 3
```

Traits Tables

A traits table should include a column with sequence names, and columns for each trait. Table entries should be the number of samples of a particular sequence that were associated with a trait (for the table below, 10 samples of seq_2 were associated with Europe). In this example, the trait labels are clearly geographical regions, but they don't need to be: you could have a phenotypic label, or any other trait that might be associated with a sequence. For traits, the header row is still optional, but it's probably a good idea to include one, or your traits will get names like trait1, trait2, etc.

```
,Europe,Asia,Africa,Australia,America
seq_2,10,5,0,6,0
seq_7,0,0,5,0,0
seq_5,4,0,10,0,0
seq_4,0,0,0,4,2
seq_3,0,0,0,3,5
seq_1,0,0,0,3,3
seq_6,0,0,0,7,3
```

References

1. Tajima F (1999). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
2. Excoffier L, Smouse P, Quattro J (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
3. French N, Yu S, Biggs P, Holland B, Fearnhead P, Binney B, Fox A, Grove-White D, Leigh JW, Miller W, Muellner P, Carter P (2014). Evolution of Campylobacter species in New Zealand. In: Sheppard SK & Méric G. (Eds), *Campylobacter Ecology and Evolution* (221–240). Norfolk, England: Caister Academic Press.
4. Bandelt H, Forster, P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*, 16(1):37–48.
5. Clement M, Snell Q, Walker P, Posada D, Crandall K (2002). TCS: Estimating gene genealogies. *IPDPS* 2:184.