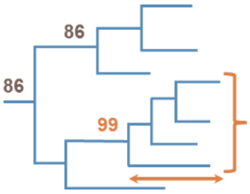
Cluster Picker 1.2 Manual

©2012 Samantha Lycett ©2015 Emma Hodcroft

Andrew Leigh Brown Group, University of Edinburgh



NEW!

The Cluster Picker 1.2 is now able to process rooted and unrooted trees and trees with identical sequences!

Introduction

In the literature, groups of related HIV infections, or HIV "clusters", are widely defined based on support for the phylogenetic grouping (bootstrap or posterior probability) and/or within cluster genetic distance. However, there is no widely available tool which is able to identify clusters in phylogenetic trees using these criteria.

The Cluster Picker is a Java program which addresses this problem.

This instruction manual will take you through the required steps to set up an analysis. The program is also accompanied by a tutorial and test files.

License and disclaimer

The Cluster Picker is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation v3.0 and as long as the contribution of previous workers is recognised.

This program is distributed in the hope that it will be useful, but without any warranty; without even the implied warranty of merchantability or fitness for a particular purpose. See the GNU General Public License for more details.

Source code is available here: https://github.com/emmahodcroft/cluster-picker-and-cluster-matcher

System requirements

Java

The Cluster Picker requires Java Running Environment (JRE) 1.6.0 or higher. Java can be downloaded here.

FigTree

The Cluster Picker outputs a phylogenetic tree in FigTree format; FigTree is thus required to visualise this tree.

Hardware

The Cluster Picker should run on any personal computer.

Installing the Cluster Picker

The Cluster Picker, along with its sister program the Cluster Matcher, can be downloaded as an executable jar files from https://github.com/emmahodcroft/cluster-picker-and-cluster-matcher and http://hiv.bio.ed.ac.uk/software.html

A tutorial and test data sets are available at those same addresses.

Cluster Picker input and output

Input

The Cluster Picker takes as input

- a fasta file of aligned sequences and
- a phylogenetic tree in newick format with support values on nodes, built from those same sequences.

Trees can be constructed in any software. We suggest FastTree or RaxML for maximum likelihood trees. Although the program will appear to work if the sequence names do not match between the alignment and tree, we do not recommend doing this as it can lead to errors.

Trees must have support values on nodes.

Trees must be fully bifurcating (see Identical sequences below).

Settings

- The Cluster Picker then asks for:
  - An initial threshold
  - A main support threshold for clusters
  - A genetic distance threshold for clusters
  - A large cluster threshold
- The initial support threshold is used to split the tree into subtrees to reduce the number of computations. This initial support threshold must be ≤ the main support threshold for clusters. We suggest using the same number.
- The Cluster Picker gives an option to output lists of clusters above a user-specified size. If you don't need this, type 0.

Output

The Cluster Picker outputs at least 4 files, all of which contain "clusterPicks" in their name:

- A fasta alignment file of clustered sequences in which sequence names have been replaced by 'Clust##_seqname'
- A newick tree file (identical to the one input) in which sequence names have been replaced by 'Clust##_seqname'
- A FigTree file in which sequence names have been replaced by 'Clust##_seqname' and where sequences are coloured by cluster.
- A log file detailing the user-input settings and data on each of the clusters (number of tips, tip names, bootstrap and maximum genetic distance)
- A file for each of the large clusters containing sequence names if this option has been selected.

The log file is in .csv format and can be opened in Microsoft Excel for viewing.

Running the Cluster Picker

GUI

- Double click on ClusterPickerGUI_1.3.jar to launch.
- Click in each of the boxes and navigate to the folder containing the fasta file and nwk file. Select these as input.

Command line version

From the command prompt, navigate to the folder containing the Cluster Picker. Type:

```
java -jar ClusterPicker_1.2.jar
```

At this point you have two options. If you press enter, you can interact with the command line and it will ask for each necessary input one at a time. The first time you use the command line version, we suggest you process in this way. Otherwise you can include your input in the initial command line like this:

```
java -jar ClusterPicker_1.2.jar input-fasta.fas input-tree.nwk
bootstrap bootstrap genetic-distance maxClusterSize
```

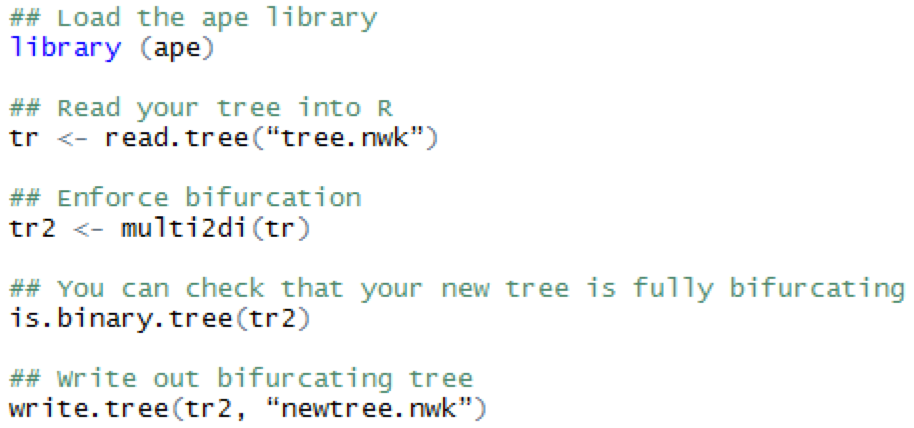
Caveats and warnings

Identical sequences

Previously, the Cluster Picker was unable to deal with identical sequences, but this new version of the Cluster Picker will process trees with identical sequences as long as they are not stored as polytomies in the tree. RaxML produces fully bifurcating trees but FastTree and other programs allow polytomies. If your tree contains a polytomy (other than the root polytomy of unrooted trees), the Cluster Picker will return an error.

You can enforce bifurcations using the “ape” package in R.

Enforcing bifurcation



```
## Load the ape library
library (ape)

## Read your tree into R
tr <- read.tree(“tree.nwk”)

## Enforce bifurcation
tr2 <- multi2di(tr)

## You can check that your new tree is fully bifurcating
is.binary.tree(tr2)

## Write out bifurcating tree
write.tree(tr2, “newtree.nwk”)
```

Missing bootstraps

If you enforce bifurcation though the method explained above, you will end up with an unannotated node. In this case, the branch lengths to both tips will be 0, so the Cluster Picker will automatically set support to 100.

Some programs, like RaxML, produce fully bifurcating trees, but the split between identical sequences is not annotated with bootstrap support. Similarly, the Cluster Picker will annotate the node with a support of 100.

In all other cases of unannotated nodes, when branch lengths ≠ 0, support will be set to 0.

Rooted/ unrooted trees

The Cluster Picker can process both rooted trees and unrooted (with a polytomy at the root) trees.

Genetic distance measures

Within cluster genetic distance can be calculated in a number of ways: the mean of the pairwise genetic distances [1], their median [2] and their maximum [3]. Another alternative is "single linkage", where a sequence is included in a cluster if its distance to just one other sequence in the cluster is below the threshold [4]. The Cluster Picker uses MAXIMUM genetic distance because this is the best approximation of time to most recent common ancestor used in time-stamped trees [3].

Genetic distance options

In the GUI Cluster Picker, pairwise genetic distance is calculated as the p-distance for all sites except gaps ("gap").

The command line version of the Cluster Picker can calculate genetic distance in four ways:

- gap: as above
- abs: p-distance based on all sites including gaps
- valid: p-distance for a, c, t, g sites only
- ambiguity: p-distance but returns matches for ambiguity codes as summarised in the table below

IUPAC code	matches
A	A
C	C
G	G
T/U	T
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	A or G or C or T

The distance algorithm can be modified by adding the preferred algorithm to the end of the command:

```
java -jar ClusterPicker_1.2.jar input-fasta.fas input-tree.nwk
bootstrap bootstrap genetic-distance maxClusterSize dist
```

Suggested settings

Genetic distance

In the literature, bootstraps of 70% [5], 80% [6], 90% [7, 8], and up to 99% [1, 9] have been used.

Bootstrap

In the literature, within-cluster genetic distances of 1.5% [7, 9, 10], 3% [8], and up to 4.5% [11, 12] substitutions per site have been used.

Available scripts

Merging output with Cluster Matcher

Script available to combine output form Cluster Picker and Cluster Matcher into one data frame and output as a csv file: combine_CPCM.R

Launching Cluster Picker in a loop

The command line version can be launched in a loop on lots of files. Our GitHub has a python script to do this: launchCPloop.py

Contact

Emma Hodcroft: emma.hodcroft@ed.ac.uk

Manon Ragonnet: manon.ragonnet@ed.ac.uk

Reference List

1. Hue S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS **2004**;18:719-28.
2. Prosperi MC, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, et al. A novel methodology for large-scale phylogeny partition. Nat Commun **2011**;2:321.
3. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis **2011**;204:1463-9.
4. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The Global Transmission Network of HIV-1. J Infect Dis **2014**;209:304-13.
5. Cuevas M, Fernandez-Garcia A, Sanchez-Garcia A, Gonzalez-Galeano M, Pinilla M, Sanchez-Martinez M, et al. Incidence of non-B subtypes of HIV-1 in Galicia, Spain: high frequency and diversity of HIV-1 among men who have sex with men. Euro Surveill **2009**;14.
6. Pilon R, Leonard L, Kim J, Vallee D, De RE, Jolly AM, et al. Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs. PLoS ONE **2011**;6:e22245.
7. Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis **2010**;10:262.
8. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. J Acquir Immune Defic Syndr **2008**;49:9-16.
9. Bezemer D, van SA, Lukashov VV, van der Hoek L, Back N, Schuurman R, et al. Transmission networks of HIV-1 among men having sex with men in the Netherlands. AIDS **2010**;24:271-82.
10. Mehta SR, Kosakovsky Pond SL, Young JA, Richman D, Little S, Smith DM. Associations Between Phylogenetic Clustering and HLA Profile Among HIV-Infected Individuals in San Diego, California. J Infect Dis **2012**;205:1529-33.
11. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog **2009**;5:e1000590.
12. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med **2008**;5:e50.