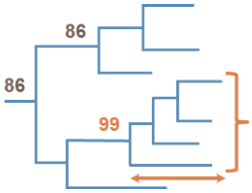
CLUSTER PICKER 1.3 MANUAL

©2012 SAMANTHA LYCETT

ANDREW LEIGH BROWN GROUP

UNIVERSITY OF EDINBURGH



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

SYSTEM REQUIREMENTS

JAVA

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

R

Phylogenetic trees should be edited in R before analysis. The "ape" package is required for this step.

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 file from http://homepages.ed.ac.uk/eang09/software.html.

A tutorial and test data sets are available at the same address.

USING THE CLUSTER PICKER

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.

Note that the sequence dataset should contain no duplicate sequences (identical name or sequence). Duplicate sequences must be removed from the alignment before the tree is constructed, for examples using the program ElimDupes. 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.

In some cases, programs will output a newick tree that is not properly read by the Cluster Picker. In order to avoid this problem, the tree can be edited using the fixtreeNodes.R script distributed with the tutorial. We recommend processing all trees before running the Cluster Picker. R will output a new newick file which should be used in the Cluster Picker along with the original fasta file.

FIXTREENODES.R

- Download R.
- Launch R and install ape:
    - Go to: Packages → Install package
    - Choose the CRAN Mirror closest to you
    - Select the “ape” package in the list and click “OK”
- Open the script in Notepad for editing
- Set the working directory to the folder that contains your tree file
    - For example: setwd(“C:/MyDocuments/Clusters/”)
    - Note that R requires forward slashes.
- Edit the tree file name to the newick file you want to process (keep the quote marks)
    - treeFileName <- “mytreefile.nwk”
- Paste the entire scrip into your R window and press enter.
- The script will output a tree file in the same folder with the extension treeforCPT.nwk

CLUSTER PICKER INPUT

- 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 treeforCPT file. Select these as input.

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.
- 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.

CAVEATS AND WARNINGS

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, 5]. 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]. We plan on adding alternative measures of genetic distance (mean and median) to future releases of the Cluster Picker.

In the literature, within-cluster genetic distances of 1.5% [6-8], 3% [9], and up to 4.5% [10, 11] substitutions per site have been used. In some studies, genetic distance thresholds are not used at all, using instead only bootstrap as a cut-off [12].

BOOTSTRAP THRESHOLDS

In the literature, bootstraps of 70% [13], 80% [14], 90% [7, 9], and up to 99% [1, 6] have been used.

FAQ

None yet! But if you do have questions, don't hesitate to contact us at the email addresses below. Do also contact us if you are interested in the source code for either program.

CONTACT

Samantha Lycett (Cluster Picker): s.lycett@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(5)**:719-728.
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(9)**:1463-1469.
4. Heimer R, Barbour R, Shaboltas AV, Hoffman IF, Kozlov AP. Spatial distribution of HIV prevalence and incidence among injection drugs users in St Petersburg: implications for HIV transmission. *AIDS* 2008; **22(1)**:123-130.
5. Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS*, et al.* Characterizing HIV Transmission Networks Across the United States. *Clin Infect Dis* 2012.
6. 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(2)**:271-282.
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. 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(10)**:1529-1533.
9. 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(1)**:9-16.
10. 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(9)**:e1000590.
11. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med* 2008; **5(3)**:e50.
12. Stadler T, Kouyos R, von W, V, Yerly S, Boni J, Burgisser P*, et al.* Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol* 2012; **29(1)**:347-357.
13. 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(47)**.
14. 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(7)**:e22245.