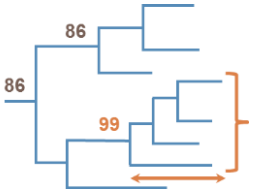
CLUSTER PICKER 1.3 MANUAL

©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. The Cluster Picker can of course be used for other infectious diseases, and the test files include influenza and hepatitis C virus sequences as well.

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

BEAST 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 is platform-independent and 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 executable jar files from http://homepages.ed.ac.uk/eang09/software.html, along with a tutorial and test data sets.

The Cluster Picker can be used with a GUI or as a command line version.

Source code for the program is available on Google code (http://code.google.com/p/cluster-picker-and-cluster-matcher/) under GNU GPLv3.

USING THE CLUSTER MATCHER

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. The Cluster Picker can also analyse BEAST maximum clade credibility trees (MCC), but this requires some processing. 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.

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 main support threshold is the minimum support you want to define a cluster.
- The genetic distance threshold is the maximum genetic distance you want within defined 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 tab delimited and can be opened in Microsoft Excel for viewing.

GUI VERSION OF THE CLUSTER PICKER

- Double click on ClusterPickerGUI.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.
- Select thresholds.
- Ideally, there should be no "Sequences with no tips" or "Tips with any sequences".
- Press GO.

COMMAND LINE VERSION OF THE CLUSTER PICKER

- From your command prompt, navigate to the folder containing ClusterPicker_command.jar
- Type into the command prompt (changing the input file names and thresholds):

  java -jar ClusterPicker_command.jar inputSeq.fas inputTree.nwk 0.9 0.9 0.045 10
- Used in this way, the Cluster Picker can easily be launched in a loop (for example from a python script).

USING THE CLUSTER PICKER ON MCC TREES

The Cluster Picker can be used on BEAST [1] maximum clade credibility (MCC) trees generated by Tree Annotator, but this requires some processing of the MCC tree. The MCC tree can be processed in R using the script "MCC_to_NWK.R" included as part of this tutorial.

MCC_TO_NWK.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 and scroll to the end.
- 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 MCC file you want to process (keep the quote marks):
  - mccName <- "mytreefile.figTree"
- Paste the entire scrip into your R window and press Enter.
- The script will output a newick tree file in the same folder that can be used in the Cluster Picker.

CAVEATS AND WARNINGS

SEQUENCE NAMES

Sequence names cannot contain spaces or characters that have a special meaning in newick or nexus files. (ie: &,();) as these are used as input and output for the CP. We have tested sequence names up to 84 characters with no problems.

GENETIC DISTANCE MEASURES

Within cluster genetic distance can be calculated in a number of ways: the mean of the pairwise genetic distances [2], their median [3] and their maximum [4]. 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 [5, 6]. 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 [4]. 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% [7-9], 3% [10], and up to 4.5% [11, 12] 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 [13]. You can do this in the Cluster Picker by setting the genetic distance threshold to the maximum pairwise genetic distance in your data. Clusters will then be identified solely based on the bootstrap support. If you want to define clusters based solely on genetic distance, choose your genetic distance cut-off and then set bootstrap to 0.

BOOTSTRAP THRESHOLDS

In the literature, bootstraps of 70% [14], 80% [15], 90% [8, 10], and up to 99% [2, 7] have been used.

Note that bootstraps on trees can be displayed out of 1 (for example in FastTree) or out of 100 (in RaxML). You must change the threshold you set accordingly to 0.9 or 90.

FAQ

None yet! But if you do have questions, don't hesitate to contact us at the email addresses below.

CONTACT

Samantha Lycett (Cluster Picker): s.lycett@ed.ac.uk

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

Reference List

1. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees**. *BMC Evol Biol* 2007, **7:**214.
2. 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:**719-728.
3. Prosperi MC, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, Di GS, Bruzzone B, Capetti A, Vivarelli A, Rusconi S, Re MC, Gismondo MR, Sighinolfi L, Gray RR, Salemi M, Zazzi M, De LA: **A novel methodology for large-scale phylogeny partition**. *Nat Commun* 2011, **2:**321.
4. 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:**1463-1469.
5. 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:**123-130.
6. Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS, Kitahata M, Rodriguez B, Dennis AM, Boswell SL, Haubrich R, Smith DM: **Characterizing HIV Transmission Networks Across the United States**. *Clin Infect Dis* 2012.
7. Bezemer D, van SA, Lukashov VV, van der Hoek L, Back N, Schuurman R, Boucher CA, Claas EC, Boerlijst MC, Coutinho RA, de WF: **Transmission networks of HIV-1 among men having sex with men in the Netherlands**. *AIDS* 2010, **24:**271-282.
8. Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, Vogelaers D, Vandekerckhove L, Verhofstede C: **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.
9. 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:**1529-1533.
10. 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.
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.
13. Stadler T, Kouyos R, von W, V, Yerly S, Boni J, Burgisser P, Klimkait T, Joos B, Rieder P, Xie D, Gunthard HF, Drummond AJ, Bonhoeffer S: **Estimating the basic reproductive number from viral sequence data**. *Mol Biol Evol* 2012, **29:**347-357.
14. Cuevas M, Fernandez-Garcia A, Sanchez-Garcia A, Gonzalez-Galeano M, Pinilla M, Sanchez-Martinez M, Garcia V, Perez-Alvarez L: **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**.
15. Pilon R, Leonard L, Kim J, Vallee D, De RE, Jolly AM, Wylie J, Pelude L, Sandstrom P: **Transmission patterns of HIV and hepatitis C virus among networks of people who inject drugs**. *PLoS ONE* 2011, **6:**e22245.