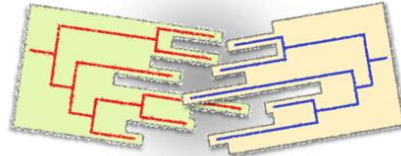# CLUSTER MATCHER 1.2 MANUAL

© 2012 EMMA HODCROFT

ANDREW LEIGH BROWN GROUP

UNIVERSITY OF EDINBURGH

INTRODUCTION

Cluster Matcher is a Java program that allows the user to explore previously-defined clusters (exported from Sam Lycett's Cluster Picker program) in a more intuitive way, select clusters with characteristics of interest, and export summary data about the chosen clusters.

You can use Cluster Matcher with one phylogeny or with two phylogenies. If used with one phylogeny, the user can choose clusters that fit a specific definition to be output to FigTree files (annotated if the user wishes) and also output a file with summary information about the chosen clusters.

If used with two phylogenies, Cluster Matcher will identify sequences that match between the two trees (either because they have the same name, or using a file that details the names of sequences that match), and thus identify 'matching clusters.' These are clusters that contain sequences that match across the two phylogenies. The user can then chose to export clusters (and their matches) to FigTree files (again, annotated if the user wishes) and output a file with summary information about the chosen clusters.

This guide will help you through the process of using Cluster Matcher, going through each panel of the user interface in turn. There is one set of instructions for using one phylogeny (one data set) and one set of instructions for using two phylogenies (two data sets). The program is also accompanied by a manual and test files.

LICENSE AND DISCLAIMER

The Cluster Matcher 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 v 3.0 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.

SYSTEM REQUIREMENTS

JAVA

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

LIBRARIES

The openCSV library is required for the Cluster Matcher to work; however it is included in the Cluster Matcher jar and does not need to be installed separately.

FIGTREE

The Cluster Matcher outputs phylogenetic trees in FigTree format.

HARDWARE

The Cluster Matcher is platform-independent and should run on any personal computer.

INSTALLING THE CLUSTER MATCHER

The Cluster Matcher, along with its sister program the Cluster Picker can be downloaded as an executable jar file from http://hiv.bio.ed.ac.uk/software.html, along with a tutorial and test data sets.

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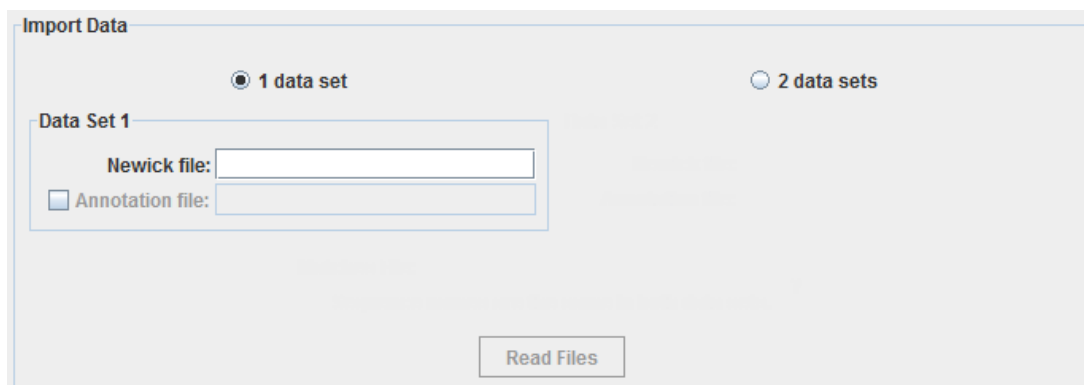
USING THE CLUSTER MATCHER

Double click on ClustMatch1.2.3.jar to launch.

INPUT

The Cluster Matcher takes as input the annotated newick file(s) output by the Cluster Picker. As well as this, an annotation file in .csv format can be input containing epidemiological data associated with each sequence.

ONE DATA FILE:

Ensure that you have run your original phylogeny through Cluster Picker with the bootstrap and genetic distance values that will define clusters. This will output a FigTree file and a Newick file – you will need to

import **only** the Newick file into Cluster Matcher. (Sequences in clusters will be labelled 'Clust##_seqname', where 'seqname' is the sequence name.)

If you want to select clusters by annotation value (for example, clusters with over 20% sequences annotated as female), you will have to supply an annotation file as well. See the section 'Preparing an Annotation File'.

Ensure that '1 data set' is selected, that the 'Annotation' check box is selected or not selected as appropriate, and click on the text areas to select the Newick file and Annotation file (if applicable).

After the files are selected, you will be able to press the 'Read Files' button.

TWO DATA FILES:

Ensure that you have run both of your original phylogenies through Cluster Picker with the bootstrap and genetic distance values that will define clusters. Each run will output a FigTree file and a Newick file – you will need to import **only** the Newick files into Cluster Matcher. (Sequences in clusters should be labelled 'Clust##_seqname', where 'seqname' is the sequence name.)

If you want to select clusters by annotation value (for example, clusters with over 20% sequences annotated as female), you will have to supply annotation files as well. See the section 'Preparing an Annotation File' – be sure to read the section that applies to two data sets.

Ensure that '2 data sets' is selected, that the 'Annotation' check boxes are selected or not selected as appropriate, and click on the text areas to select the Newick files and Annotation files (if applicable).

MATCHES FILE

If matching sequences have different sequence names in the two newick files, prepare a two-column .csv file with a one-line 'header' containing matched sequence names for DataSet1 in column one, DataSet2 in column two.

```
DataSet1,DataSet2
83033,133433
93433,126933
88333,134833
63037,123333
36836,103383
```

If sequences that match between phylogenies have the same sequence name (they must be an exact match), you can select the box next to 'Sequence names are the same in both data sets.' Note that if you use this option, the annotation file supplied under 'Data Set 2' will be applied to the sequences in Data Set 1 and 2 (a

separate Data Set 1 annotation file cannot be provided). For more information see the section 'Preparing an Annotation File' – subsection 'If matching sequences have the same name'.

After the files are selected you will be able to press the 'Read Files' button.

SETTINGS

OUTPUT FOLDER

Output Folder

Output: [   ]

Click in the text area next to 'Output' to select the folder where all generated files will be written. This will be where the log file, cluster FigTree files, and extended information file (if selected) will be output. If the user tries to output a second time to the same folder, they will be asked to either change the output folder or delete the files from the previous run. This is to prevent confusion over which clusters were output from separate runs.

CLUSTER SELECTION

Cluster Matcher allows you to select clusters that match attributes of particular interest. To preview how many clusters will be returned with the currently selected attributes, press the 'Preview' button. This will also output a summary of the returned clusters to the log file. You can explore your data in real-time by refreshing the log file while using the 'Preview' button to try different cluster selection criteria.

You can output the FigTree files for the clusters that match the criteria you've selected by pressing the 'Produce FigTree Files' button.

BY NUMBER OF SEQUENCES

ONE DATA FILE:

1 Data Set

Return only clusters with more than [0] sequences

You can select to output only clusters which contain more than a user-specified number of sequences by typing in a number other than 0. If you don't want to select clusters based on size, leave the value at 0.

TWO DATA FILES:

2 Data Sets

Return only clusters with more than [0] sequences that match between datasets

You can select clusters which contain more than a user-specified number of sequences matching between data sets by typing in a number other than 0. If you don't want to select clusters base on the number of matching sequences, leave the value at 0.

BY ANNOTATION VALUE

Clusters can only be selected by annotation value if annotation files have been supplied! A 'field' is the general category of an annotation, and the 'values' are the possible values that field can hold. For example, in the "sex" field the values are 'male' and 'female'. Remember that **only fields with fewer than 11 values will be displayed**.

ONE DATA FILE:

Select by annotation value by first checking the box. Select the field and value that you'd like to select clusters by, then type in the minimum percentage of sequences in the cluster that should have that value. By default, the second check box is un-selected.

This determines how the percentage is calculated. If the box is un-selected, it will only consider sequences in the cluster that have a value for the selected field. So in a cluster of 3 sequences with one sequence that has no value for 'sex' and the other holding values of 'male' and 'female,' only the sequences with the values 'male' and 'female' would be considered. In this case, the cluster would be considered 50% male and 50% female. (It would be selected if the user wanted clusters that were 50% female.)

If the box is selected, it will consider ALL sequences in the cluster, including those with no value for the field. So in the same cluster as before, it will now consider all the sequences and consider the cluster 30% male, 30% female, and 30% no value. (It would not be selected if the user wanted clusters that were 50% female.)

TWO DATA FILES:

Select by annotation value by first checking the box. Select the field and value that you'd like to select clusters by, then type in the minimum percentage of sequences in the cluster that should have that value. You can choose whether clusters in Data Set 1 will be selected by these criteria (but the matching clusters in Data Set 2 may or may not fit the criteria), whether clusters in Data Set 2 will be selected by these criteria (but the matching clusters in Data Set 1 may or may not fit the criteria), or whether both clusters in Data Set 1 **and** their matching cluster in Data Set 2 should fit the criteria.

By default, the second check box is un-selected.

This determines how the percentage is calculated. If the box is un-selected, it will only consider sequences in the cluster that have a value in the selected field. So in a cluster of 3 sequences with one sequence that has no value for 'sex' and the other holding values of 'male' and 'female,' only the sequences with the values 'male' and 'female' would be considered. In this case, the cluster would be considered 50% male and 50% female. (It would be selected if the user wanted clusters that were 50% female.)

If the box is selected, it will consider ALL sequences in the cluster, including those with no value for the field. So in the same cluster as before, it will now consider all the sequences and consider the cluster 30% male, 30% female, and 30% no value. (It would not be selected if the user wanted clusters that were 50% female.)

OUTPUTTING FILES

☑ Embed annotations in the FigTree files ?

You can select whether to embed annotations in the FigTree files that are output. However, if you have selected clusters by annotation value, you must embed annotations in the FigTree files so that you can check the clusters have been selected appropriately. All annotations provided (even those not chosen for cluster selection) will be embedded.

☐ Print a .csv file with extended information about the clusters ?

You can also select whether to output a .csv file with extended information about the clusters. This will contain the cluster number, the matching cluster number (if applicable), the number of sequences in the cluster, the number of sequences in the cluster that match to the other data set (if applicable), and information on the proportion of sequences in the cluster with each annotation value. (Only fields with fewer than 11 values will be output here.) This file can be linked to the Cluster Picker log file using the merge function in R.

Print just .csv file

Finally, you can output only the csv file without the FigTree files.

OUTPUT

The Cluster Matcher outputs three types of files:

- A log file that contains general information about the tree read in and records all 'Preview' and 'Produce FigTree Files' results in detail is written in real time.
- A .csv file with extended information about the clusters can be output if the user wishes. See the second paragraph of the section "Outputting Files".
- A FigTree file for each cluster.

FIGTREE FILE INFORMATION:

The sequences in the output FigTree files are coloured:

- Red = The sequence in cluster X in Data Set 1 has a match in a cluster in Data Set 2
- Blue = This sequence has no match, or no match that is in a cluster
- Green = the sequence is part of the 'matching' cluster in data set 2 and has a match in Data Set 1, but the match is not in the cluster being examined in this file – it is in another cluster.

ONE DATA FILE:

Each FigTree file is named after the cluster it contains, and contains two trees: one of the whole input tree, with the cluster coloured, and one of just the cluster.

TWO DATA FILES:

Each FigTree file is named after the Data Set 1 cluster it contains, and contains at least 4 trees: one of the whole Data Set 1 input tree, with the cluster coloured; one of just the Data Set 1 cluster; one or more of the matching cluster(s) from Data Set 2; and one of the whole Data Set 2 input tree, with the matching cluster(s) coloured.

MANIPULATION IN FIGTREE

FigTree files output by Cluster Matcher with embedded annotations can be labelled according to those annotations. To modify tip labels, open the tree in FigTree, and click on the arrow next to "Tip Labels" to display options. You can then select to display the annotations on the tree instead of the sequences names, or you can colour the names according to the annotations. If you want to colour branches according to annotations, follow the same steps under "Appearance".

CAVEATS AND WARNINGS

PREPARING AN ANNOTATION FILE

The annotation file should be a comma-delimited (.csv) file that contains clinical or other data about each sequence on separate rows. The sequence name MUST BE in the first column.

To make the following modifications to your .csv annotation file, using R is recommended, as this should make replacing and standardizing column names and field values easier.

It is recommended that discrete rather than quantitative traits be used. Only fields with fewer than 11 possible values will be displayed for cluster selection. However, annotations that are not displayed will still be embedded in the FigTree files if this option is selected.

The match for all values must be exact. If sequences in your data set have 'male,' 'female,' 'm,' and 'f' for 'sex' values, all four options will be considered independent ('m' will not be equal to 'male', and also 'male' would not be equal to 'Male').

Similarly, 'Unknown', 'Not Known' will be classed as different by the program (they will be listed as two different trait values), so all values that represent missing information should be changed to '**NA**', which will be read by the program as missing information.

TWO DATA FILES:

Only column names that match between the two data sets will be displayed. If one data set has a column called 'sex' that contains 'male' and 'female' and the other has a column called 'gender' that contains 'male' and 'female,' neither column will be displayed for cluster selection. Similarly, 'Sex' and 'sex' are not the same.

Values for matching fields should also be exactly the same between data sets. If for a column called 'sex' one data set has the values 'male' and 'female' and the other data set has the values 'woman' and 'man' or 'm' and

'f', you will be able to select all four options to select clusters by – but no clusters in Data Set 2 will have sequences with the value 'male' and no clusters in Data Set 1 will have sequences with the value 'man'. Ensure that the same values are used in both data sets.

IF MATCHING SEQUENCES HAVE THE SAME NAME

If matching sequences have the same name in the two data sets, the annotation file supplied under 'Data Set 2' will be applied to the sequences in Data Set 1 and 2 (a separate Data Set 1 annotation file cannot be provided). This is due to restrictions in how FigTree works and dealing with one sequence name that has two different annotations attached. For optimal performance in this situation, submit an annotation file that contains information for ALL the sequences in both Data Set 1 and Data Set 2.

FAQ

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

CONTACT

Emma Hodcroft (Cluster Matcher): emmahodcroft@gmail.com

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