

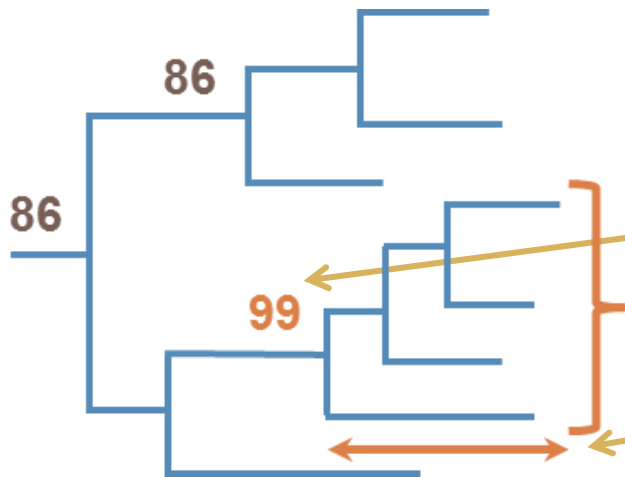
PICKING AND DESCRIBING HIV CLUSTERS IN PHYLOGENETIC TREES

Cluster Picker and Cluster Matcher

HIV transmission clusters

- Phylogenetic relationships between HIV sequences isolated from different patients can be used to investigate transmission
- HIV transmission clusters are important for the study of ongoing transmission
- Clusters are identified in trees based on high support for the grouping and low within cluster genetic distance
- The **Cluster Picker** and **Cluster Matcher** automate the process of cluster identification and analysis

Cluster definition



- Clusters are identified based on
 - high **bootstrap** and
 - low within cluster **genetic distance**

Tutorial aims

- Find clusters in two phylogenetic trees
 - ▣ One built from sequences collected in Europe up until 2003
 - ▣ One built from ALL sequences collected in Europe
- Describe the clusters according to data in an annotation file
- Match the clusters between the trees to see how they have changed over time

Tutorial pre-requisites

- To follow this tutorial, you will need
 - ▣ Java 1.6
 - You can check what version of java you are running by typing in the command: `java -version` from your command prompt
 - ▣ R
 - With the ape package installed*
 - ▣ FigTree

To install ape in R:

Go to: Packages → Install package

Choose the CRAN Mirror closest to you

Select the “ape” package in the list and click “OK”

File download

- Download ClusterTutorial.zip to a folder on your computer. As an example, we will download to:
 - ▣ C:\MyDocuments\Clusters
- Once you have extracted it, the new folder address will be
 - ▣ C:\MyDocuments\Clusters\ClusterTutorial
- This directory contains all the files for the tutorial, as well as the jar files for the programs

Sequence datasets

- European sequences from 11 countries were downloaded from the [Los Alamos National Laboratories](#) database into two data sets#:
 - ▣ Sequences collected up until 2003*:Europe1866.fas
 - ▣ All sequences: Europe3031.fas

	BE	CY	CZ	DE	DK	ES	FR	GB	GR	IT	PT	total
Up to 2003*	52	1	71	290	51	156	280	539	47	351	28	1866
All	54	88	71	295	63	286	287	1375	47	437	28	3031

BE Belgium, CY Cyprus, CZ Czechoslovakia, DE Germany, DK Denmark, ES Spain, FR France, GB Great Britain, GR Greece, IT Italy, PT Portugal

duplicate sequences were removed using [ElimDupes](#). 1seq/ patient only.

* Including sequences collected in 2003

Building trees

- You can build your own trees, or use the ones you downloaded as part of the tutorial. These were built in FastTree:
 - ▣ EU1866.nwk
 - ▣ EU3031.nwk
- You can build your trees using any other software. Just remember to check whether bootstraps on the tree are displayed out of 1 or out of 100. In FastTree they are out of 1.

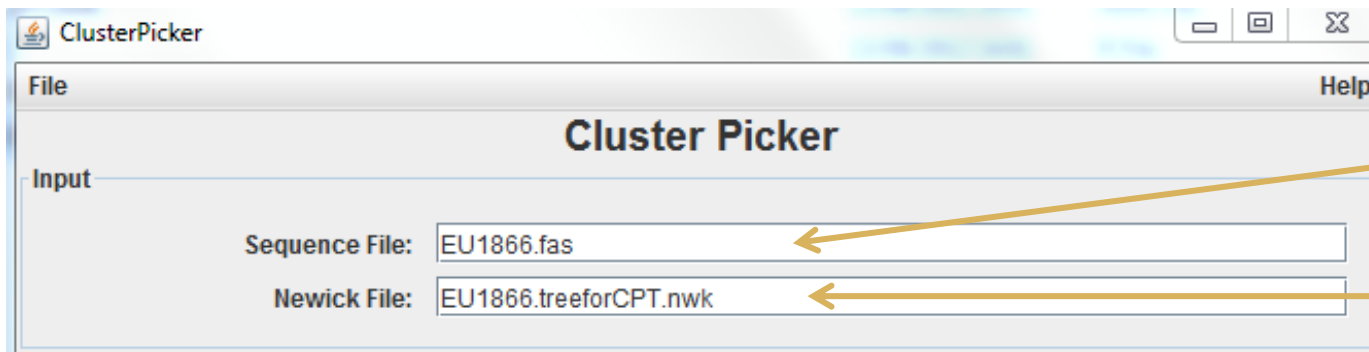
We have ensured these files can be handled by the Cluster Picker and Matcher. If yours do not get processed properly, please see the manual (polytomies are a problem, so **all duplicate sequences must be removed** before building the tree).

Editing trees in R

- The trees need to be edited in R for use in the Cluster Picker using the script fixTreeNodees.R. Open the script in notepad and edit as follows:
 - ▣ First set the working directory to the folder which contains your files:
 - `setwd("C:/MyDocuments/Clusters/ClusterTutorial")`
 - Note that R requires forward slashes
 - ▣ Then edit the tree file name to the newick file you want to process.
 - `treeFileName <- "EU1866.nwk"`
 - ▣ Then paste the whole script into your R window.
 - ▣ The script will output a tree file in the same folder with the extension: `treeforCPT.nwk`
 - ▣ Check that this tree contains as many tips as the tree you put in (in FigTree for example)!

Cluster Picker Input

- Double click on `ClusterPickerGUI_1.3.jar`



- Click in each of the boxes and navigate to the ClusterTutorial folder. Select
 - ▣ the fasta file containing your aligned sequences and
 - ▣ the treeforCPT nwk file.

Don't forget to use the newick tree that was edited in R!

Cluster Picker settings (1)

- The Cluster picker then asks for:
 - ▣ An initial threshold
 - ▣ A main support threshold for clusters
 - ▣ A genetic distance threshold for clusters
- The initial support threshold is used to split the tree into subtrees to reduce the number of computations. This initial support threshold must be \leq the main support threshold for clusters.

Cluster Picker settings (2)

- Support thresholds depend on whether bootstraps are displayed out of 100 or 1 in your tree. In FastTree, they are out of 1, so we will choose:
 - 0.9
 - 0.9 (90% bootstrap support for clusters)
 - 4.5 (maximum 4.5 substitutions/site within clusters)
 - If bootstraps were displayed out of 100, we would type 90, 90, 4.5

The screenshot shows a 'Settings' window with four input fields. Yellow arrows from the list above point to each field: 0.9 to 'Initial Threshold', 0.9 to 'Main Support Threshold', 4.5 to 'Genetic Distance Threshold', and 5 to 'Large Cluster Threshold'.

Setting	Value
Initial Threshold:	0.9
Main Support Threshold:	0.9
Genetic Distance Threshold:	4.5
Large Cluster Threshold:	5

- Finally, the Cluster Picker gives an option to output lists of clusters above a certain size. If you don't need this, type 0. Here we will output clusters ≥ 5 , so we type:
 - 5

Press:

GO

Cluster Picker output (1)

- In this example, 9 files have been output, all of which have “clusterPicks” in their name. The other 4 are:
 - ▣ 5 are lists of clusters with at least 5 sequences.
 - ▣ A fasta file of clustered sequences in which the names of sequence in clusters have been annotated
 - >B.GB.79261.JN100976_2003
 - >Clust6_B.GB.79261.JN100976_2003
 - ▣ A newick tree with sequence names annotated in the same way
 - ▣ A log file
 - ▣ A FigTree file

Cluster Picker output (2)

- The log file contains:
 - The input file names and settings
 - Details of the clusters
 - You can open the file in Excel

This file is tab delimited, you can paste it into Excel.

**** Cluster Picker Results ****

Input sequences = C:\MyDocuments\Clusters\ClusterTutorial\EU1866.fas

Input tree = C:\MyDocuments\Clusters\ClusterTutorial\EU1866.nwk

Initial support threshold= 0.9

Support threshold= 0.9

Genetic distance threshold= 0.045

Large cluster threshold= 5

** Sequences with cluster assignment output with new names

** Tree modified to contain new names

** new names have form: Clust(C)_(SequenceName) where C = cluster number, e.g. Clust25_139320

Output sequences = C:\MyDocuments\Clusters\ClusterTutorial\EU1866_EU1866_clusterPicks.fas

Output tree= C:\MyDocuments\Clusters\ClusterTutorial\EU1866_clusterPicks.nwk

Output figtree= C:\MyDocuments\Clusters\ClusterTutorial\EU1866_clusterPicks.nwk.figTree

There are 1866 sequences

Tree has 1866 tips

Found 71 clusters

ClusterNumber	NumberOfTips	NumberOfTipsCheck	TipNames	Bootstrap	GD
1	2	2	[Clust1_B.GB.1182_48_8095_20030909.DQ879092_2003, Clust1_B.GB.80597.JN101915_1998]	0.998	0.01901901901901902
2	2	2	[Clust2_B.CZ.82729PL1.AY694293_2000, Clust2_B.CZ.86543PL1.AY694321_2001]	1.0	0.013013013013013013
3	2	2	[Clust3_B.GB.67444.JN101626_1999, Clust3_B.GB.73199.JN101836_2002]	1.0	0.015015015015015015
4	2	2	[Clust4_B.GB.78956.JN100840_1998, Clust4_B.GB.93301.JN101878_2002]	0.926	0.03003003003003003
5	2	2	[Clust5_B.GB.72086.JN100900_2001, Clust5_B.GB.74246.JN100899_1997]	0.984	0.04104104104104104

1866 sequences in the tree

71 clusters

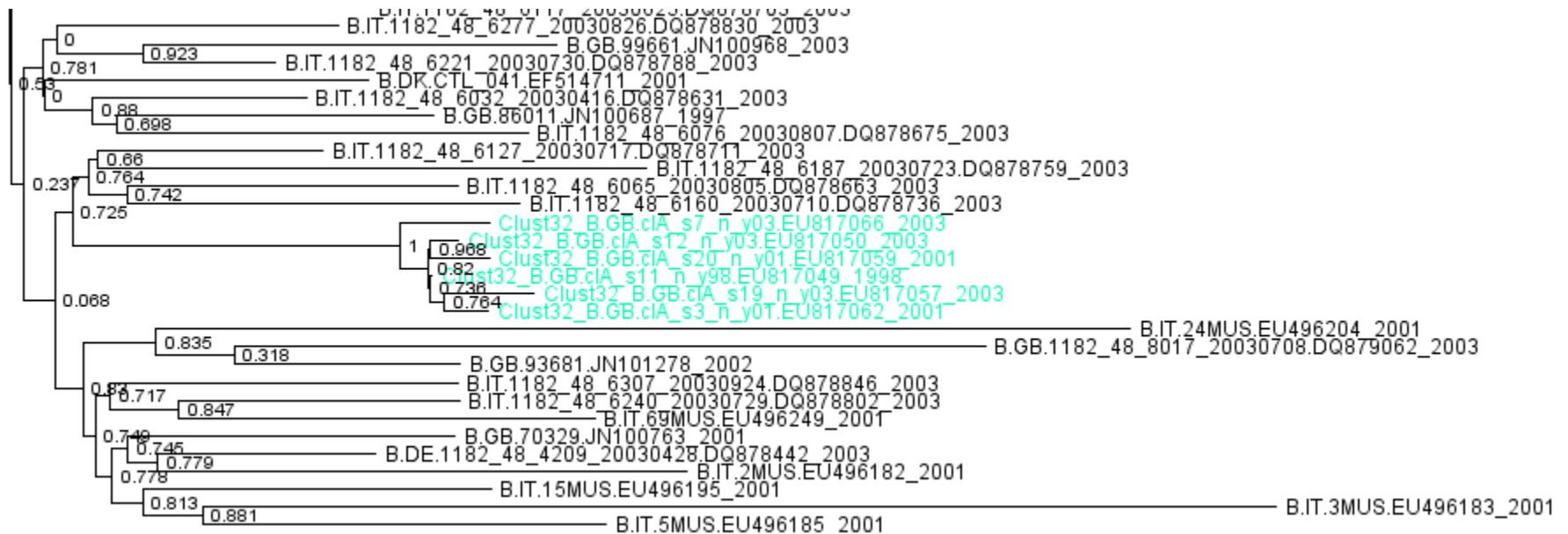
Cluster size

Bootstrap

Genetic distance

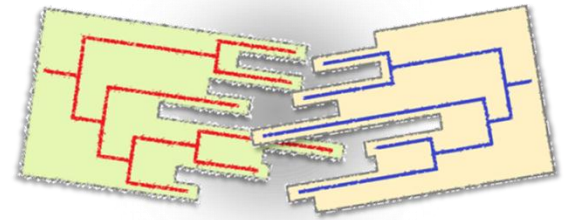
Cluster Picker output (3)

- A FigTree with annotated names and sequences coloured by cluster, which can be displayed with the program FigTree.



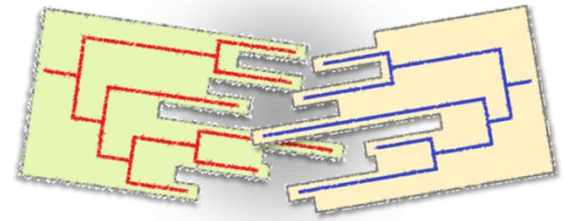
All the output files for this run and EU3031 can be found in the CPoutput folder.

Cluster Matcher



- The Cluster Matcher can describe the clusters epidemiologically based on annotations associated with each sequence.
- It also matches clusters between runs of the Cluster Picker
- The Cluster Matcher takes as input the clusterPicks.nwk file output by the Cluster Picker and an annotation file (optional).

Cluster Matcher



Cluster Matcher

Import Data

☒ 1 data set ☐ 2 data sets

Data Set 1

Newick file:

☐ Annotation file:

Data Set 2

Read Files

Output Folder

Output:

1 Data Set

Return only clusters with more than sequences

☐ and which at least % of the sequences have a value of ?

☐ Out of all sequences in the cluster, including those with no value for this field ?

Preview files will be returned ?

☐ Embed annotations in the FigTree files ?

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

Produce FigTree Files

About the Output

Created by Emma Hodcroft - 2011
Leigh Brown Group, Edinburgh University

Input

Output

Selection criteria

Navigate to the directory which contains "ClustMatch1.2.1.jar" and double click to launch it.

Cluster Matcher annotation file

- The annotation file should be in .csv format (here displayed in Excel), and contains:
 - ▣ The sequence name, followed by
 - ▣ Epidemiological data about that patient, in columns:

FastaLabel	Risk factor	Country	Sampling city	Year	Drug naive
B.IT.1182_48_6273_20030923.DQ878826_2003	Bisexual	ITALY	NA	2003	no
B.IT.1182_48_6274_20030923.DQ878827_2003	Bisexual	ITALY	NA	2003	no
B.IT.CV91_470_04.AY672456_2004	Heterosexual	ITALY	NA	2004	no
B.IT.1182_48_6275_20030923.DQ878828_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6276_20030826.DQ878829_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6277_20030826.DQ878830_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6278_20030826.DQ878831_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6279_20030826.DQ878832_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6286_20030820.DQ878834_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6287_20030828.DQ878835_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6288_20030924.DQ878836_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6292_20030910.DQ878837_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6293_20030910.DQ878838_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6294_20030930.DQ878839_2003	Heterosexual	ITALY	NA	2003	no
B.IT.1182_48_6298_20030912.DQ878840_2003	Heterosexual	ITALY	NA	2003	no

Please see the manual for instructions on how to build your own .csv file.

Cluster Matcher input: 1 file

Select “1 data set”

Click in the “Newick file” box, and locate the clusterPicks.nwk file

In the same folder, you will find your annotation file: TestDataset_3031_epiData.csv*

Click on “Read files”

Select a folder in which to write your output

The screenshot shows the 'Import Data' window of the Cluster Matcher software. It features two radio buttons at the top: '1 data set' (selected) and '2 data sets'. Below, there are two sections: 'Data Set 1' and 'Data Set 2'. In 'Data Set 1', the 'Newick file' is set to 'U1866.treeforCPT_clusterPicks.nwk' and the 'Annotation file' checkbox is checked with the path 'ces\TestDataset_3031_epiData.csv'. A 'Read Files' button is located to the right of these fields. At the bottom, there is an 'Output Folder' section with an 'Output:' label and an empty text box. Four yellow arrows point from the text instructions on the left to the corresponding UI elements: the '1 data set' radio button, the 'Newick file' text box, the 'Annotation file' checkbox, and the 'Read Files' button.

*Note that the “risk factor” and “drug-naïve” columns of this table have been edited at random because so much data was missing in the original table downloaded from LANL.

Also, this file contains epi data on all 3031 sequences – but this doesn’t matter to the Cluster Matcher.

Cluster matcher settings: 1 file

1 Data Set

Return only clusters with more than sequences

☒ and which at least % of the sequences have a value of

☐ Out of all sequences in the cluster, including those with no value for this field

files will be returned

☒ Embed annotations in the FigTree files

☒ Print a .csv file with extended information about the clusters

Created by Emma Hodcroft - 2011
Leigh Brown Group, Edinburgh University

- In this example, we are asking for the Cluster Matcher to return FigTree files for all of the clusters
 - With a least three sequences
 - Where at least 1% of the sequences are from the UK.
- Annotations from the .csv file will be embedded in the FigTree files,
- The .csv file will contain information on these clusters

You can read more about the Cluster Matcher settings in the manual.

Cluster matcher output: 1 file

- With our settings, the Cluster Matcher will output 4 FigTree files, a log file and a .csv file.
- The log file reminds us of our settings and summarises our results.

Input files used:

Newick file: C:\Tutorial\EU1866.treeforCPT_clusterPicks.nwk

Annotation file: C:\\Tutorial\Europe_3031_epiData.csv

Data Set 2:

Newick file:

The data set had 1866 sequences and 71 clusters (containing 171 sequences (9.16%))

***FigTree Files Written*:**

Output Location: C:\Tutorial\CMresults

5 clusters in data set 1 (7.04%) have more than 3 sequences.

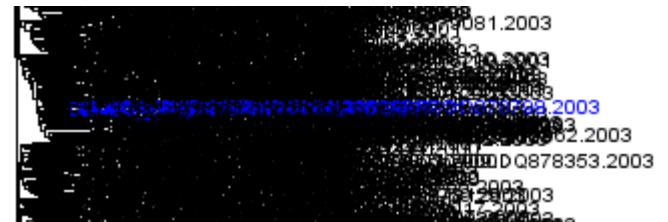
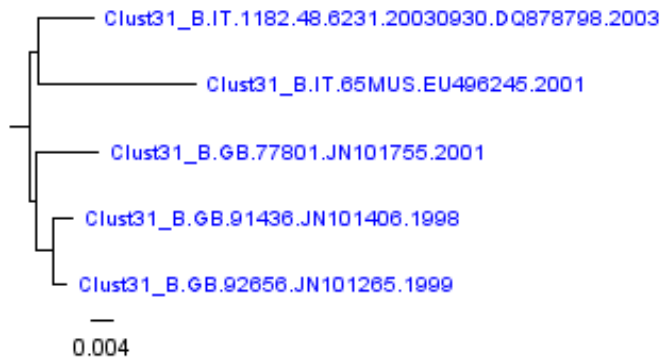
Of these, 4 clusters (5.63%) have at least 1% sequences with a Country value of UNITED KINGDOM

Of the 22 sequences in these clusters:

- 0 (0.0%) are SPAIN
- 0 (0.0%) are BELGIUM
- 0 (0.0%) are GERMANY
- 0 (0.0%) are FRANCE
- 20 (90.91%) are UNITED KINGDOM
- 0 (0.0%) are GREECE
- 2 (9.09%) are ITALY

Cluster matcher output: 1 file

- Each of the 3 FigTree files contain two trees: the cluster on its own and the cluster highlighted within the whole tree.



You can click from one tree to the next using these buttons in FigTree:



Cluster matcher output: 1 file

- The .csv file contains epidemiological data from the annotation file for each of the clusters. You can open it in Excel.

```
Clust_ID,Num_Seqs,Antwerp,BIRMINGHAM,Madrid,London,Sampling city_NA,no,yes,Drug  
naive_NA,UNITED KINGDOM,BELGIUM,SPAIN,CZECH  
REPUBLIC,ITALY,GERMANY,FRANCE,GREECE,Country_NA,Male Sex with Male,Heterosexual,IV Drug  
User,Risk factor_NA  
43,5,0,0,0,0,5,3,2,0,5,0,0,0,0,0,0,0,0,4,0,1,0  
40,6,0,0,0,0,6,1,5,0,6,0,0,0,0,0,0,0,0,5,1,0,0  
32,6,0,0,0,0,6,0,6,0,6,0,0,0,0,0,0,0,0,6,0,0,0  
31,5,0,0,0,0,5,4,1,0,3,0,0,0,2,0,0,0,0,5,0,0,0
```

Clust_ID	Num_Seqs	Madrid	Antwerp	London	BIRMINGHAM
43	5	0	0	0	0
40	6	0	0	0	0
32	6	0	0	0	0
31	5	0	0	0	0

Note that the “no” and “yes” columns refer to the “Drug-naïve” column in the annotation file but this is not stated

Cluster Matcher input: 2 files

- If you run both EU1866 and EU3031 through the Cluster Picker (or use the output files provided), we can match the clusters between those two runs and see how existing clusters changed after 2003

Import Data

☐ 1 data set ☒ 2 data sets

Data Set 1

Newick file:

☒ Annotation file:

Data Set 2

Newick file:

☒ Annotation file:

Matches File:

☒ Sequence names are the same in both data sets. ?

Read Files

In this case, the sequence names are the same in the two data sets. If working with data where they are not, just input a matches file.

Cluster Matcher settings: 2 files

- We are looking at clusters that existed in both data sets (at least one match), where at least 1% of the sequences are from the UK.

The screenshot shows the '2 Data Sets' configuration window. It includes a 'Preview' button and a status indicator showing '40 files will be returned'. Below this, there are two checked options: 'Embed annotations in the FigTree files' and 'Print a .csv file with extended information about the clusters'. At the bottom, there are two buttons: 'Produce FigTree Files' and 'About the Output'.

2 Data Sets

Return only clusters with more than sequences that match between datasets

☒ and which at least % of the sequences have a

value of in

☐ Out of all sequences in the cluster, including those with no value for this field

40 files will be returned

☒ Embed annotations in the FigTree files

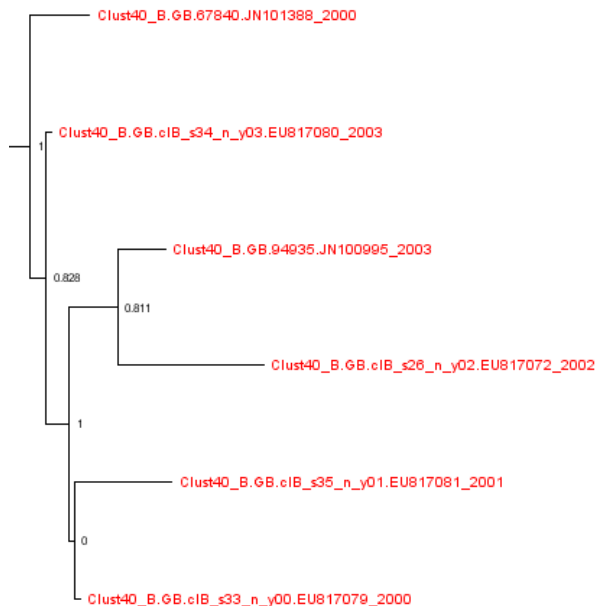
☒ Print a .csv file with extended information about the clusters

You can read more about the settings in the Cluster Matcher manual.

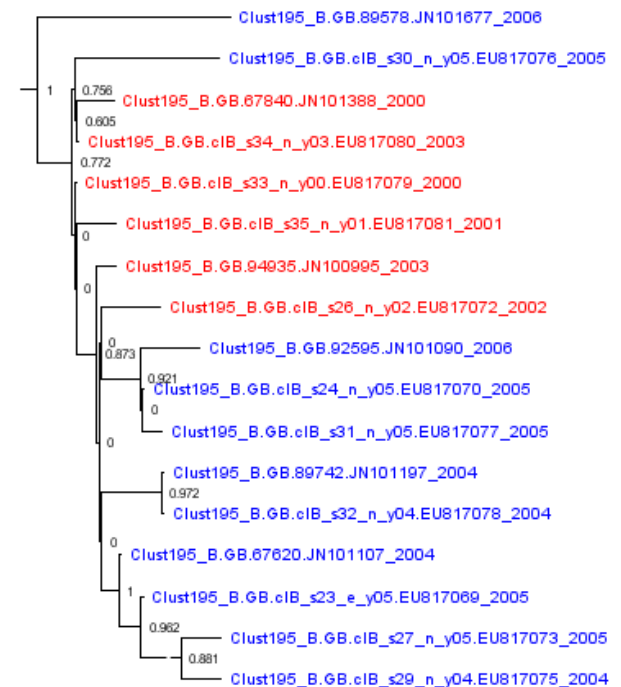
Cluster Matcher output: 2 files

- The FigTree files shows the same cluster at two time points: sequences in blue are the ones added to the cluster since 2003.

□ 2003



□ 2011



Cluster Matcher output: 2 files

- The .csv file gives extended information on the clusters at each time point:
 - ▣ Number of sequences
 - ▣ Number of matches in the other data set
 - ▣ Composition (according to annotation file)

DataSet	Clust_ID	Matching_Clust_ID	Num_Seqs	Num_Seq_wMatch	Madrid
2	202	64	5	3	0
1	64	202	3	3	0
2	23	18	3	2	0
1	18	23	2	2	0
2	208	49	2	2	0
1	49	208	2	2	0
2	5	5	2	2	0
1	5	5	2	2	0
2	219	9	4	3	0

Author contact details

Please let us know if you have any comments or questions!

Samantha Lycett
Emma Hodcroft
Manon Ragonnet
Andrew Leigh-Brown

samantha.lycett@ed.ac.uk
emmahodcroft@gmail.com
manon.ragonnet@ed.ac.uk
A.Leigh-Brown@ed.ac.uk

HIV & Flu Research Group,
Institute of Evolutionary Biology,
University of Edinburgh,
Edinburgh, UK.