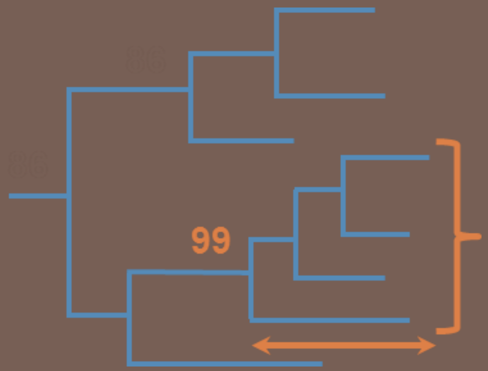
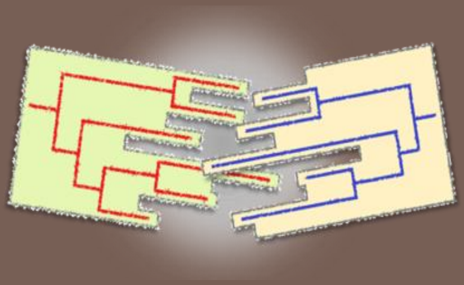




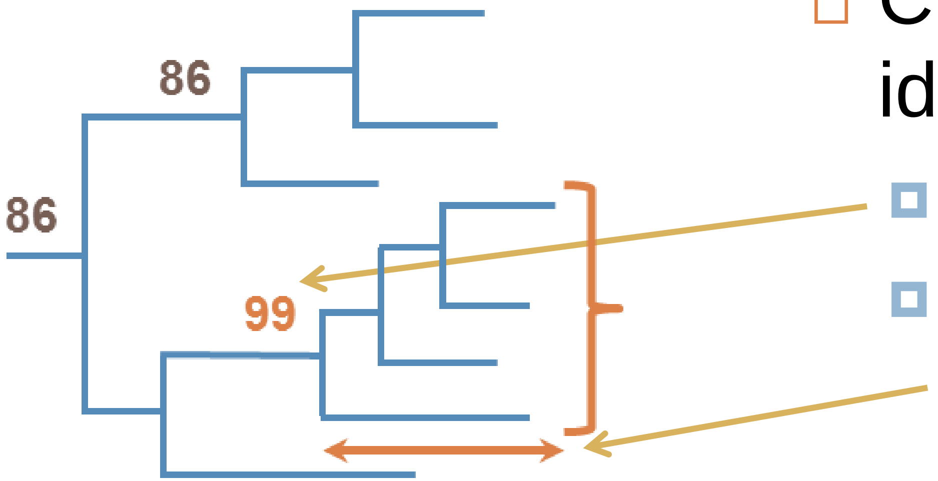
PICKING AND DESCRIBING HIV CLUSTERS IN PHYLOGENETIC TREES

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 a phylogenetic tree
  - Built from all 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

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
  - FigTree

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

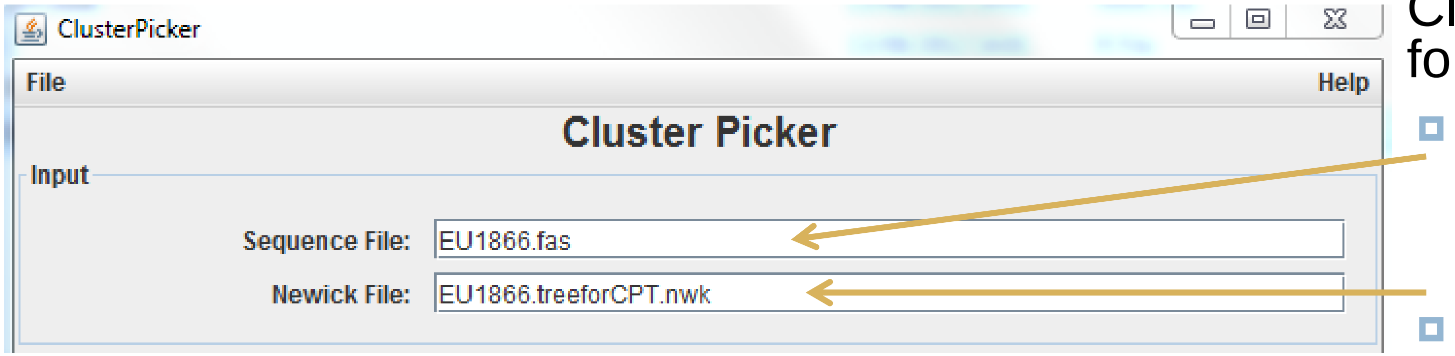
Building trees

- You can build your own trees, or use the ones you downloaded as part of the tutorial. These were built in FastTree:
  - EU1866.treeforCPT.nwk
  - EU3031. treeforCPT.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 must be resolved, see “Enforcing bifurcation”.).

Cluster Picker Input

- Double click on `ClusterPicker_GUI.jar`



Sequence File: EU1866.fas

Newick File: EU1866.treeforCPT.nwk

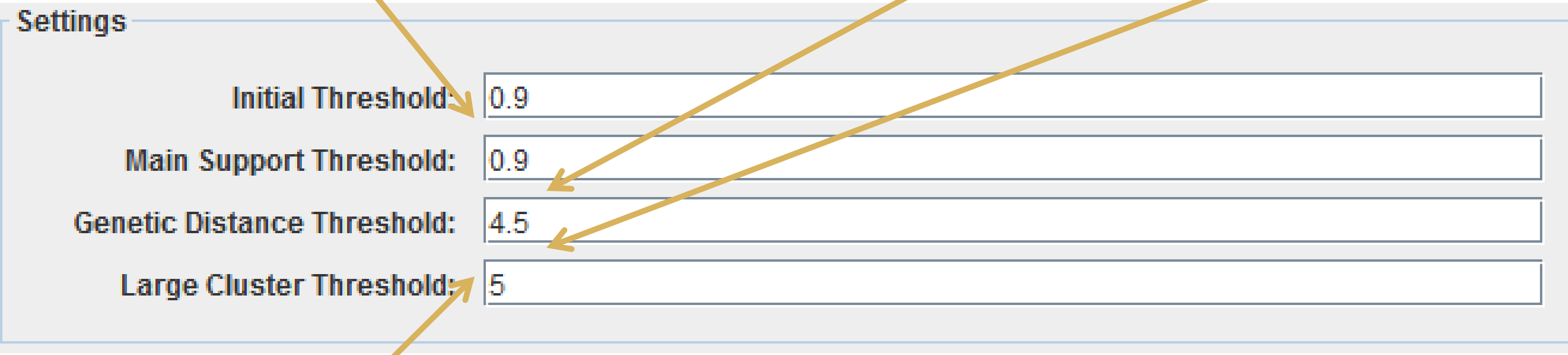
- Click in each of the boxes and navigate to the ClusterTutorial folder. Select
  - the fasta file containing your aligned sequences and
  - the treeforCPT nwk file.

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 ≤ 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`

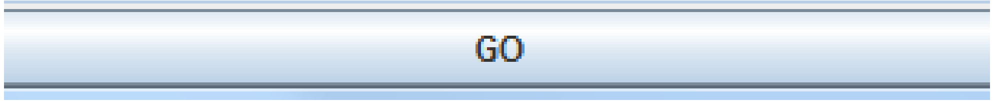


Settings

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:



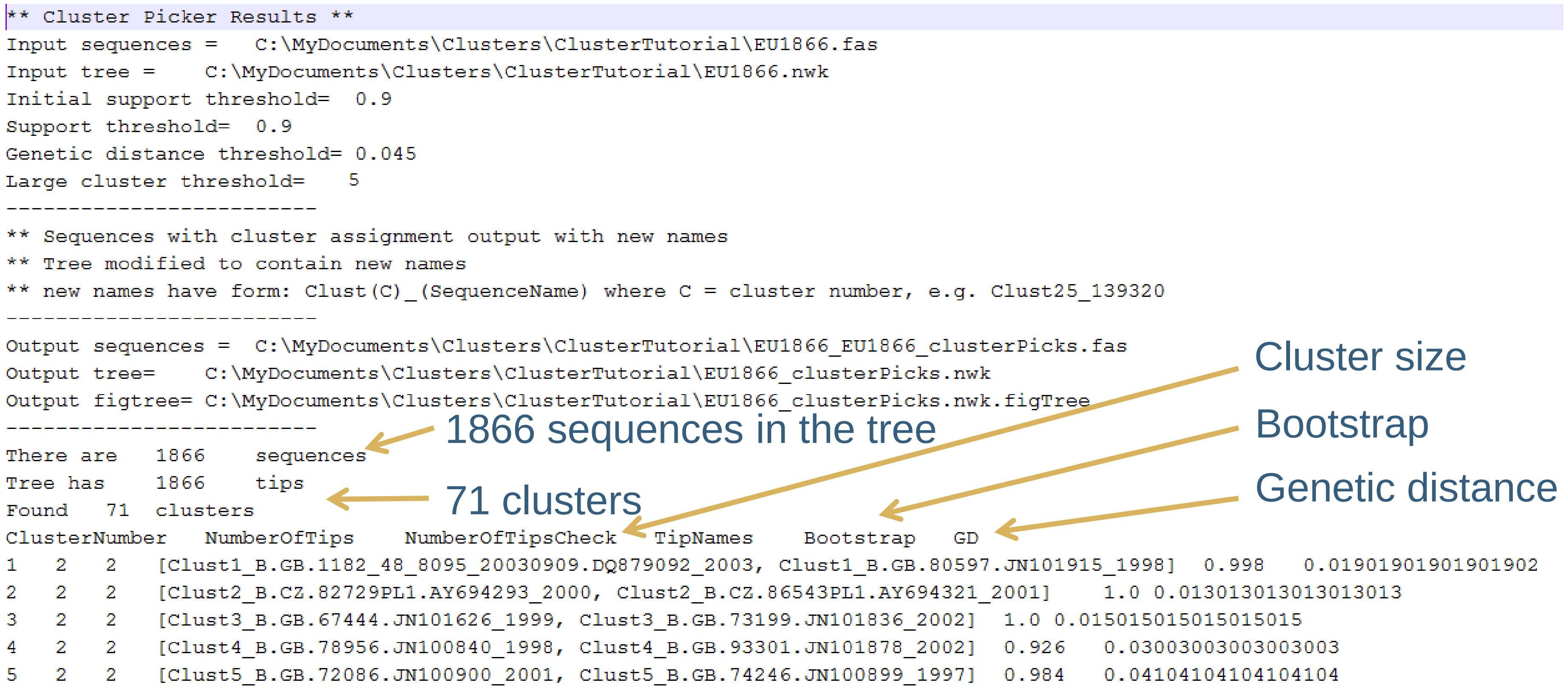
Cluster Picker output (1)

- 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

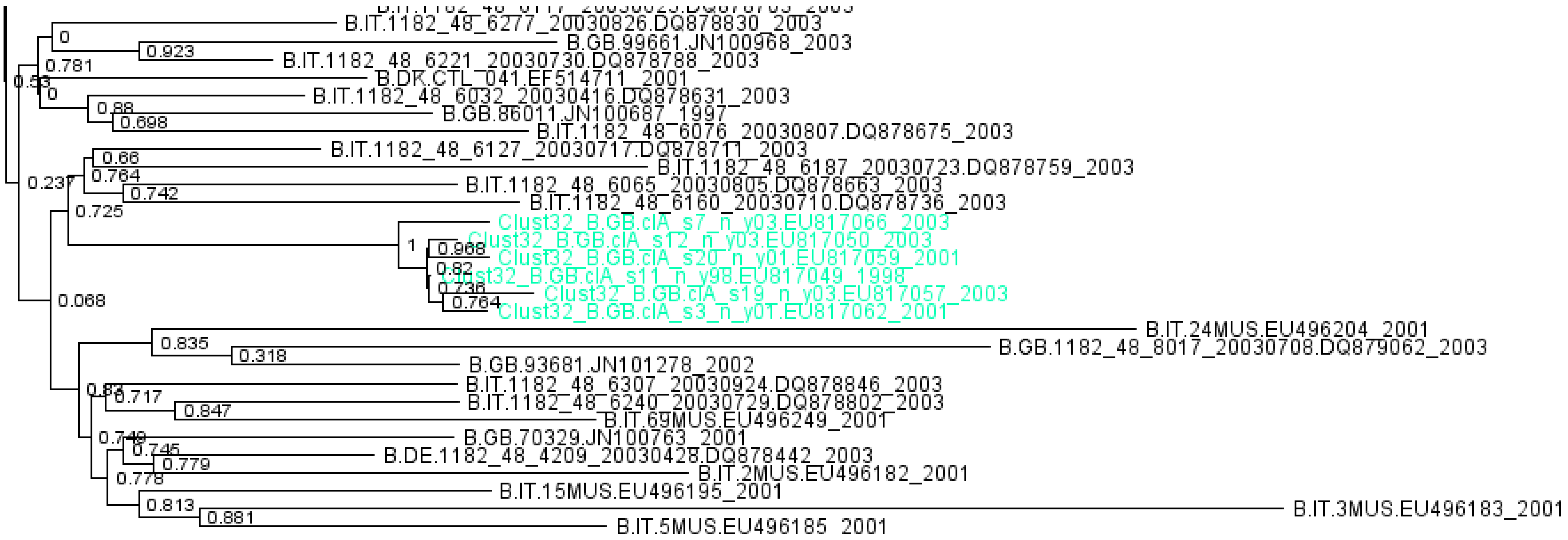
Cluster size

Bootstrap

Genetic distance

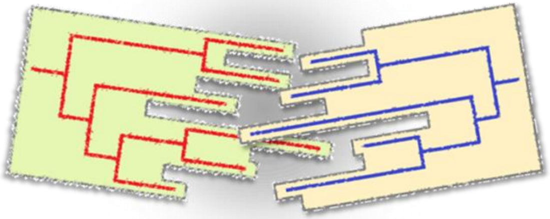
Cluster Picker output (3)

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


B.IT.1182_48_6277_20030826.DQ878830_2003
B.GB.99661.JN100968_2003
B.IT.1182_48_6221_20030730.DQ878788_2003
B.DK.CTL_041.EF514711_2001
B.IT.1182_48_6032_20030416.DQ878631_2003
B.GB.86011.JN100687_1997
B.IT.1182_48_6076_20030807.DQ878675_2003
B.IT.1182_48_6127_20030717.DQ878711_2003
B.IT.1182_48_6187_20030723.DQ878759_2003
B.IT.1182_48_6065_20030805.DQ878663_2003
B.IT.1182_48_6160_20030710.DQ878736_2003
Clust32_B.GB.clA_s7_n_y03.EU817066_2003
Clust32_B.GB.clA_s12_n_y03.EU817050_2003
Clust32_B.GB.clA_s20_n_y01.EU817059_2001
Clust32_B.GB.clA_s11_n_y98.EU817049_1998
Clust32_B.GB.clA_s19_n_y03.EU817057_2003
Clust32_B.GB.clA_s3_n_y01.EU817062_2001
B.IT.24MUS.EU496204_2001
B.GB.1182_48_8017_20030708.DQ879062_2003
B.GB.93681.JN101278_2002
B.IT.1182_48_6307_20030924.DQ878846_2003
B.IT.1182_48_6240_20030729.DQ878802_2003
B.IT.69MUS.EU496249_2001
B.GB.70329.JN100763_2001
B.DE.1182_48_4209_20030428.DQ878442_2003
B.IT.2MUS.EU496182_2001
B.IT.15MUS.EU496195_2001
B.IT.3MUS.EU496183_2001
B.IT.5MUS.EU496185_2001
0
0.923
0.781
0.53
0
0.88
0.698
0.66
0.764
0.237
0.742
0.725
1
0.968
0.82
0.736
0.764
0.068
0.835
0.318
0.83
0.717
0.847
0.749
0.745
0.779
0.778
0.813
0.881

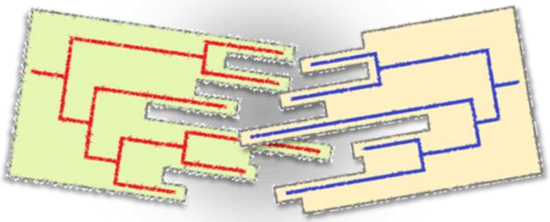
All the output files for this run and EU3031 can be found in the CPoutput folder.

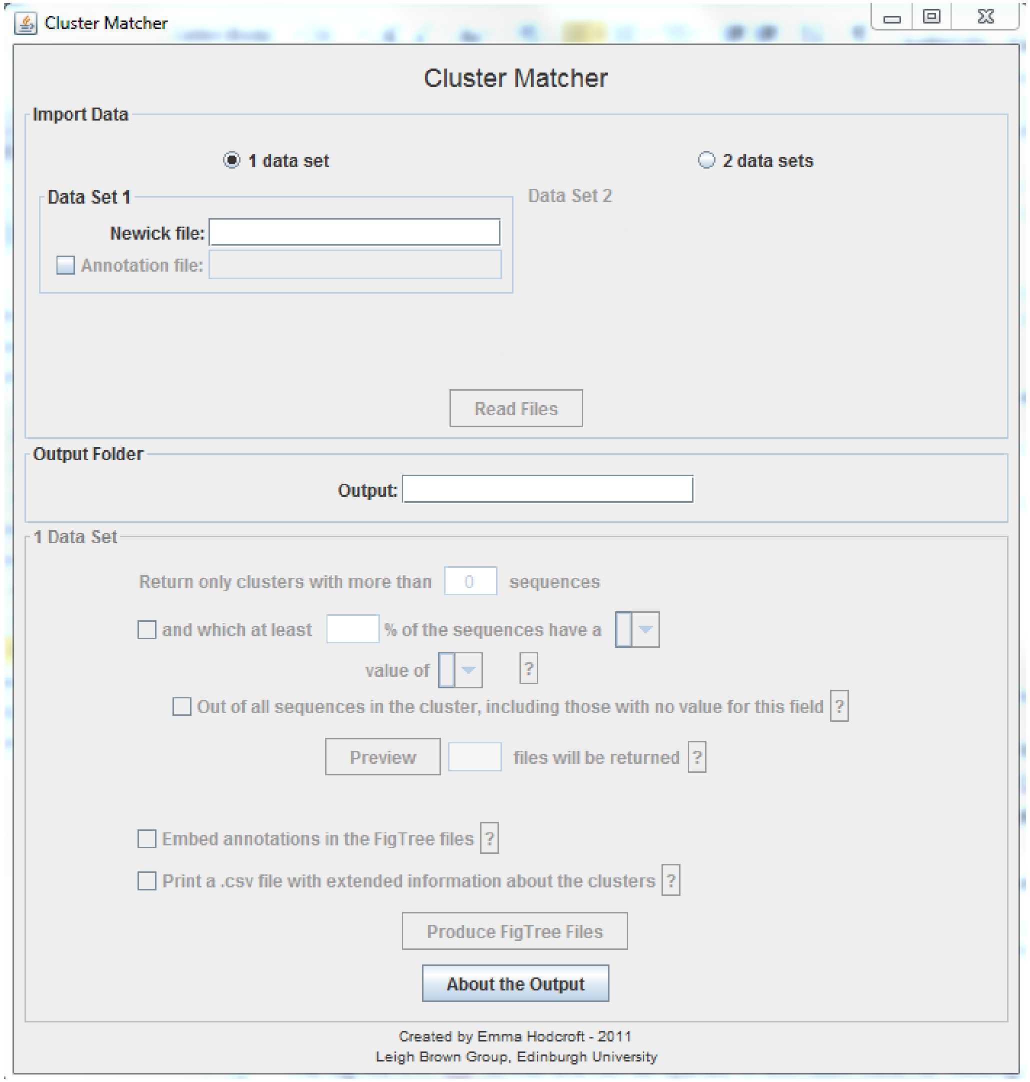
Cluster Matcher



- 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

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 0 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

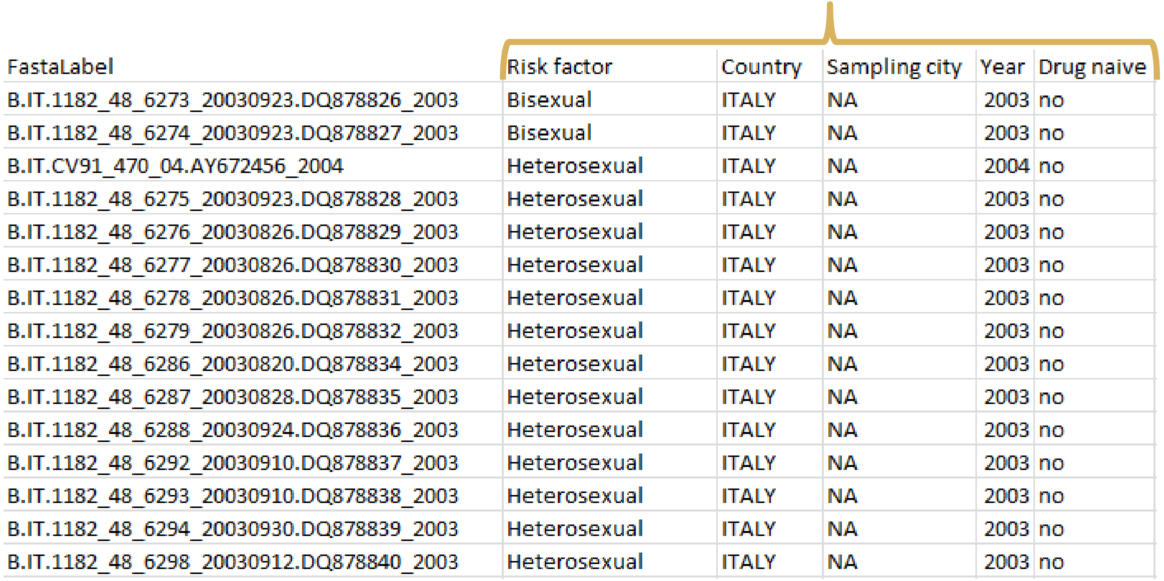
Navigate to the directory which contains “ClustMatch1.2.1.jar” and double click to launch it.

Output

Selection criteria

Cluster Matcher annotation file

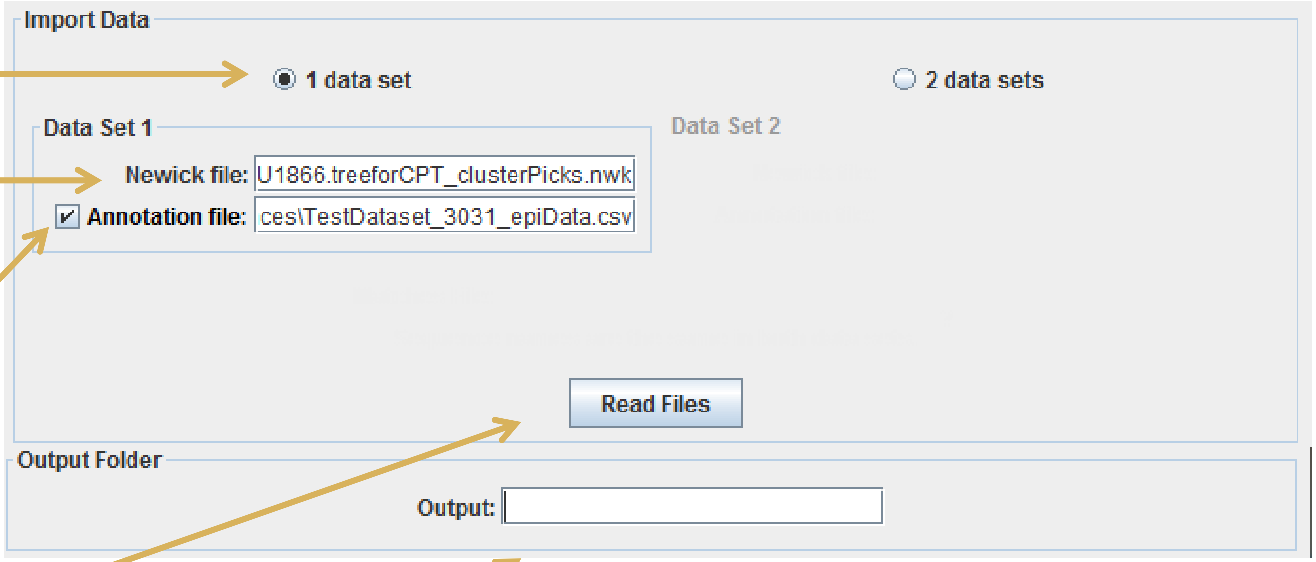
- The annotation file should be in .csv format (here displayed in Excel), and contains:
  - The sequence name, followed by
  - Epidemiological data about that patient, in columns:



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



Import Data

● 1 data set ○ 2 data sets

Data Set 1

Data Set 2

Newick file: U1866.treeforCPT_clusterPicks.nwk

☑ Annotation file: ces\TestDataset_3031_epiData.csv

Output Folder

Output:

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*

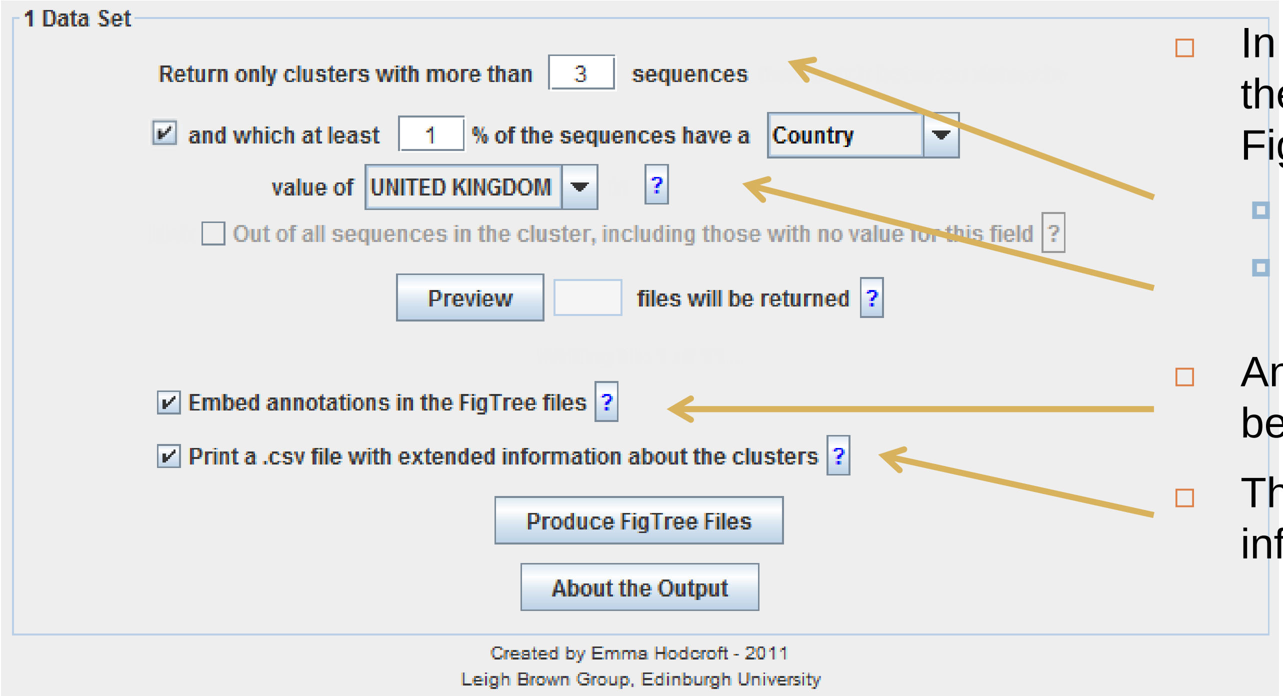
Click on “Read files”

Select a folder in which to write your output

*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 3 sequences

☑ and which at least 1 % of the sequences have a Country

value of UNITED KINGDOM ?

□ 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 ?

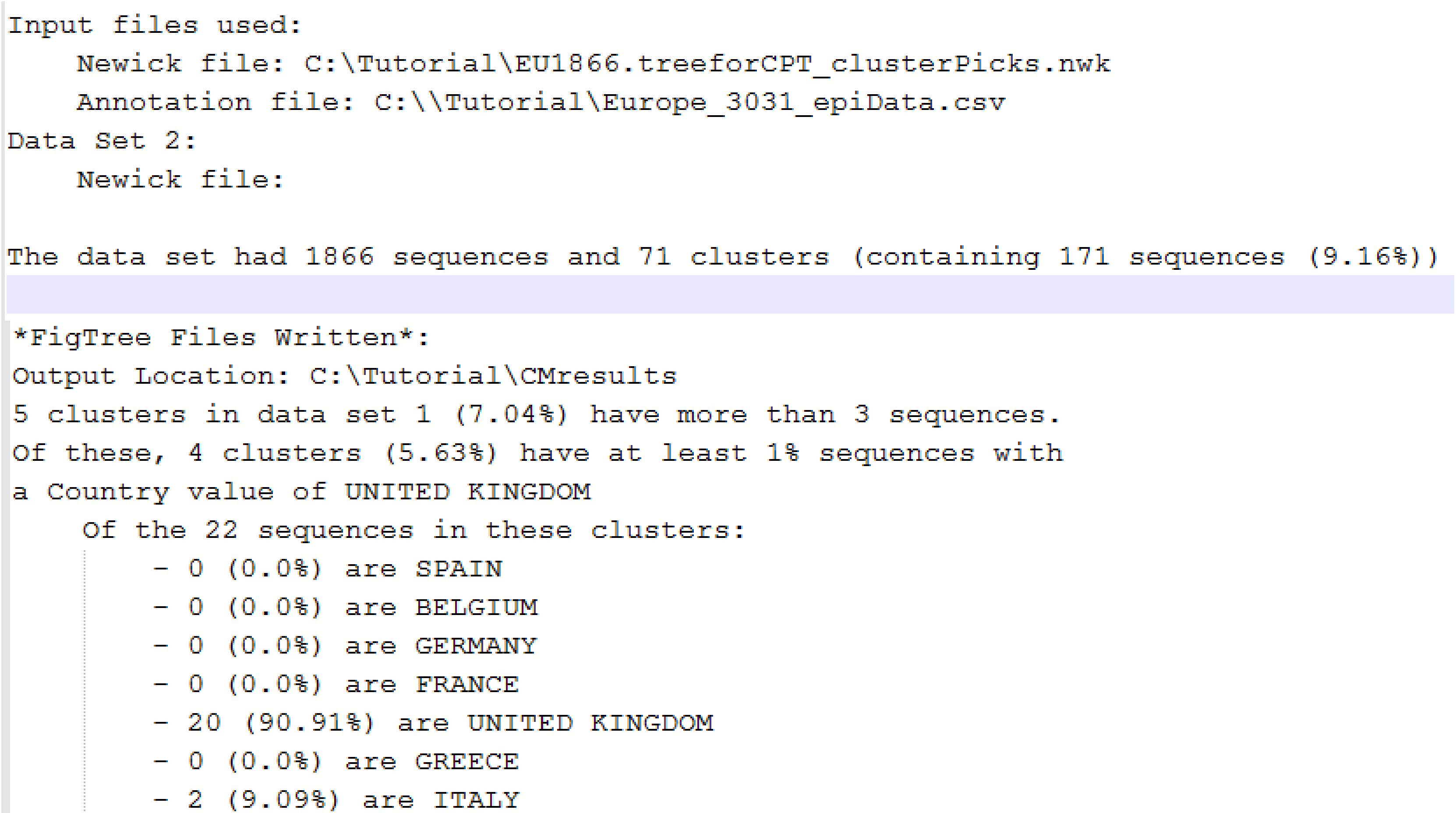
Created by Emma Hodcroft - 2011
Leigh Brown Group, Edinburgh University

- 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

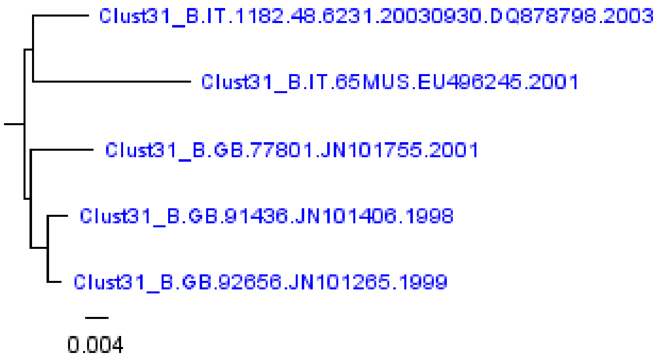
- With our settings, the Cluster Matcher will output 4 FigTree files, a log file and a .csv file.
- The log file reminds us of our settings and summarises our results.

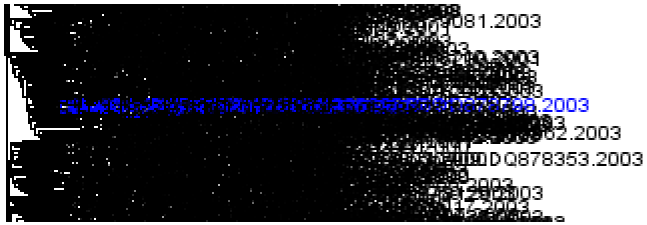


```
Input files used:
    Newick file: C:\Tutorial\EU1866.treeforCPT_clusterPicks.nwk
    Annotation file: C:\\Tutorial\Europe_3031_epiData.csv
Data Set 2:
    Newick file:

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
```

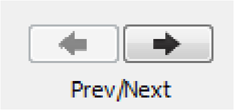
Cluster matcher output: 1 file

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


Clust31_B.IT.1182.48.6231.20030930.DQ878798.2003
Clust31_B.IT.65MUS.EU496245.2001
Clust31_B.GB.77801.JN101755.2001
Clust31_B.GB.91436.JN101406.1998
Clust31_B.GB.92656.JN101265.1999
0.004

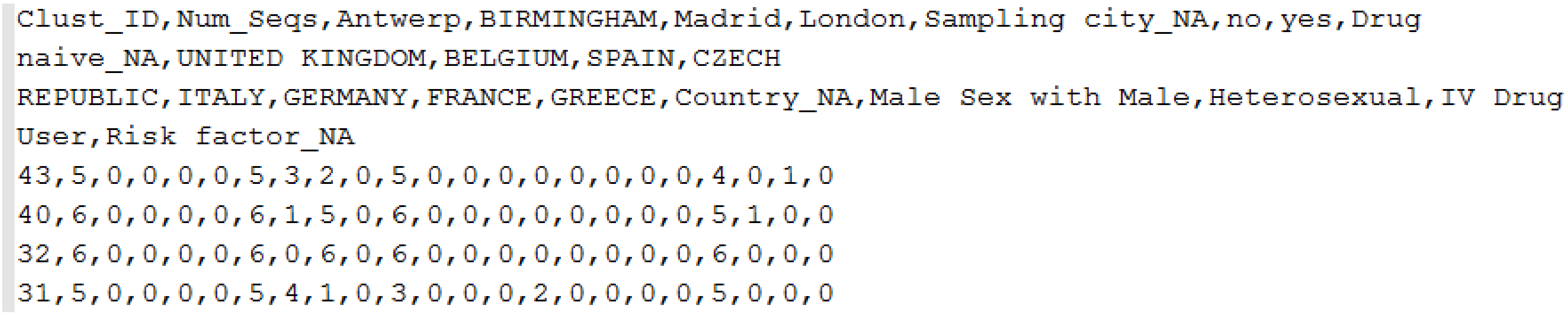


You can click from one tree to the next using these buttons in FigTree:

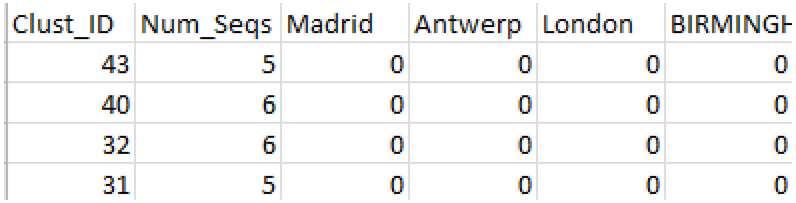


Cluster Matcher output: 1 file

- 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
```

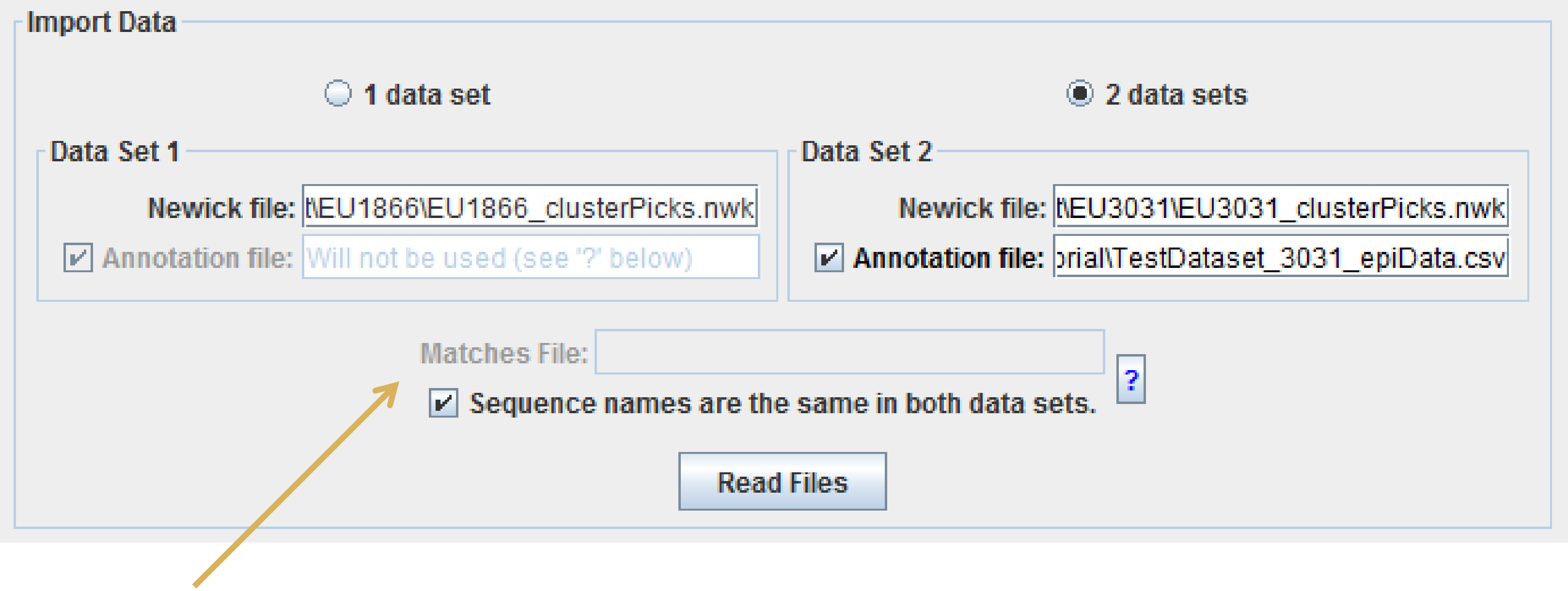


Clust_ID	Num_Seqs	Madrid	Antwerp	London	BIRMINGH
43	5	0	0	0	0
40	6	0	0	0	0
32	6	0	0	0	0
31	5	0	0	0	0

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: t\EU1866\EU1866_clusterPicks.nwk

☑ Annotation file: Will not be used (see '?' below)

Data Set 2

Newick file: t\EU3031\EU3031_clusterPicks.nwk

☑ Annotation file: ɔrial\TestDataset_3031_epiData.csv

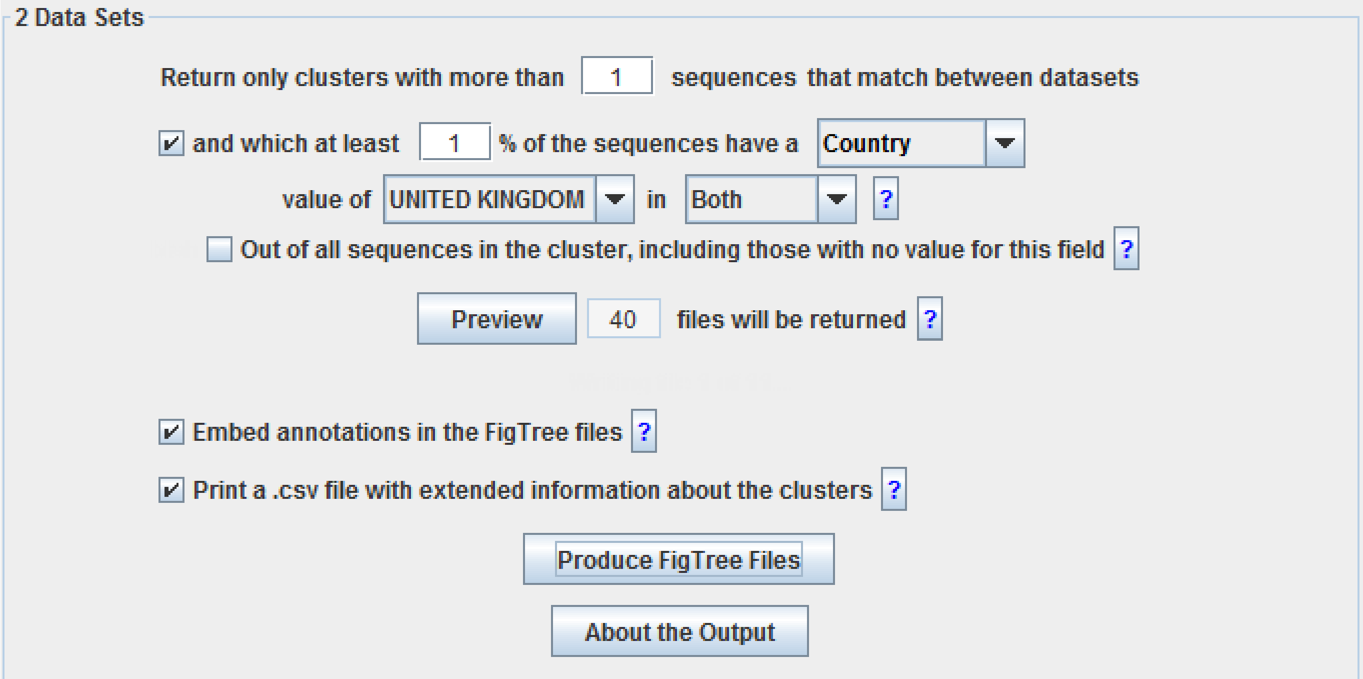
Matches File:

☑ Sequence names are the same in both data sets.

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.



2 Data Sets

Return only clusters with more than 1 sequences that match between datasets

☑ and which at least 1 % of the sequences have a Country

value of UNITED KINGDOM in Both

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

40 files will be returned

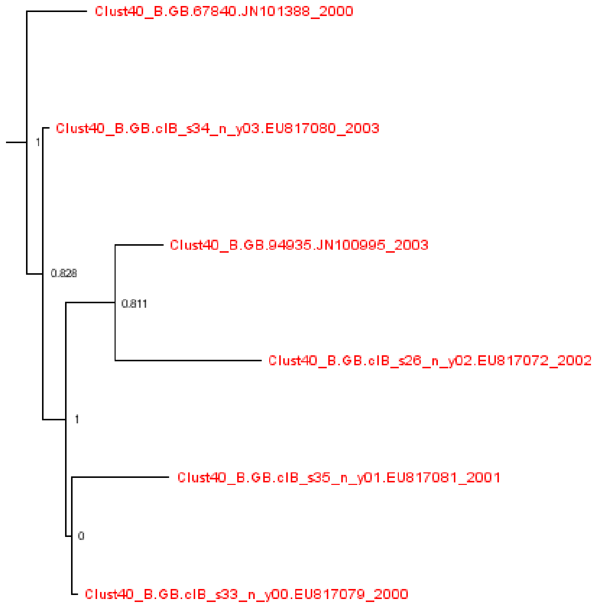
☑ Embed annotations in the FigTree files

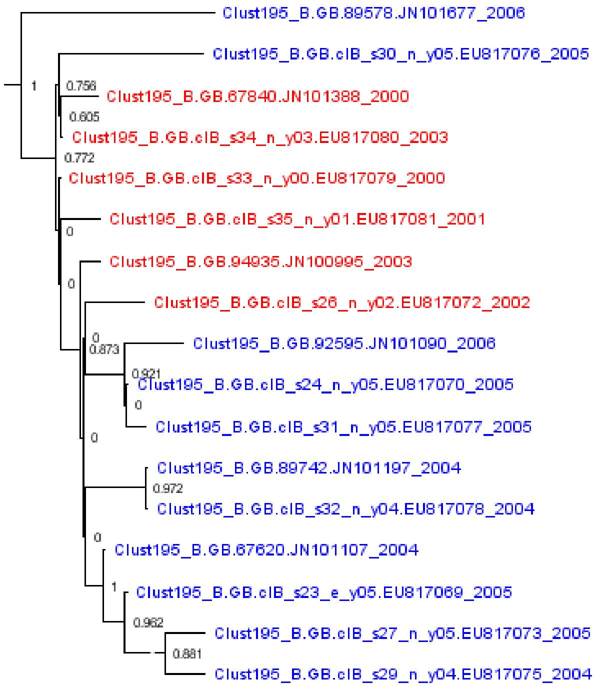
☑ Print a .csv file with extended information about the clusters

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

Cluster Matcher output: 2 files

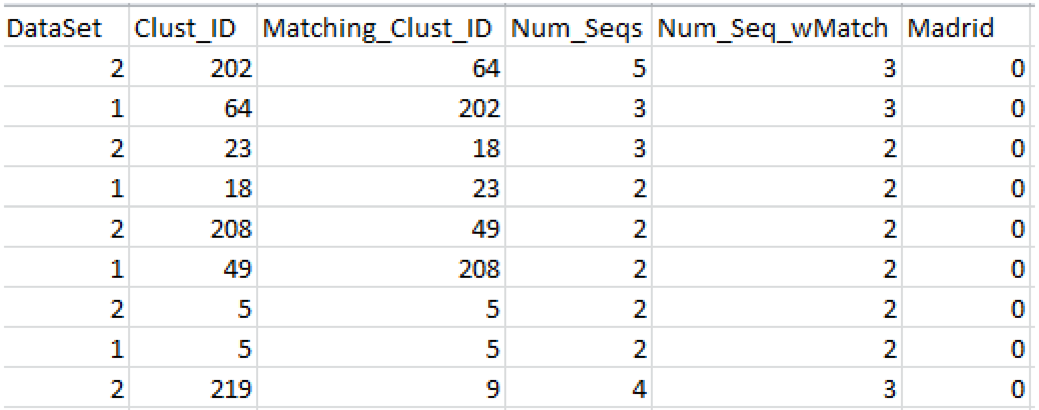
- The FigTree files shows the same cluster at two time points: sequences in blue are the ones added to the cluster since 2003.
    - 2003
    - 2011


Clust40_B.GB.67840.JN101388_2000
1
Clust40_B.GB.clB_s34_n_y03.EU817080_2003
Clust40_B.GB.94935.JN100995_2003
0.828
0.811
Clust40_B.GB.clB_s26_n_y02.EU817072_2002
1
Clust40_B.GB.clB_s35_n_y01.EU817081_2001
0
Clust40_B.GB.clB_s33_n_y00.EU817079_2000


Clust195_B.GB.89578.JN101677_2006
Clust195_B.GB.clB_s30_n_y05.EU817076_2005
1
0.756
Clust195_B.GB.67840.JN101388_2000
0.605
Clust195_B.GB.clB_s34_n_y03.EU817080_2003
0.772
Clust195_B.GB.clB_s33_n_y00.EU817079_2000
Clust195_B.GB.clB_s35_n_y01.EU817081_2001
0
Clust195_B.GB.94935.JN100995_2003
0
Clust195_B.GB.clB_s26_n_y02.EU817072_2002
0
0.873
Clust195_B.GB.92595.JN101090_2006
0.921
Clust195_B.GB.clB_s24_n_y05.EU817070_2005
0
Clust195_B.GB.clB_s31_n_y05.EU817077_2005
0
Clust195_B.GB.89742.JN101197_2004
0.972
Clust195_B.GB.clB_s32_n_y04.EU817078_2004
0
Clust195_B.GB.67620.JN101107_2004
1
Clust195_B.GB.clB_s23_e_y05.EU817069_2005
0.962
Clust195_B.GB.clB_s27_n_y05.EU817073_2005
0.881
Clust195_B.GB.clB_s29_n_y04.EU817075_2004

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

Advanced

- If you know how to use R, you might find the following scripts useful:

Merging Cluster Picker and Cluster Matcher output

- Script available to combine output into one data frame and creata a csv file: combine_CPCM.R

Launching Cluster Picker in a loop

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

- Both available in this tutorial and on our GitHub https://github.com/emmahodcroft/cluster-picker-and-cluster-matcher

Author contact details

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

Emma Hodcroft	emmahodcroft@gmail.com
Manon Ragonnet	manon.ragonnet@ed.ac.uk
Andrew Leigh-Brown	A.Leigh-Brown@ed.ac.uk

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