

Hierarchical Clustering Tutorial

Ignacio Gonzalez, Sophie Lamarre, Sarah Maman, Luc Jouneau
CATI Bios4Biol - Statistical group

March 2017

To know about clustering

- There are two main methods:

- **Classification = supervised method:**

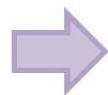
Bring together elements into categories you defined previously to launch the classification



For prediction

- **Clustering = unsupervised method:**

Bring together elements which are similar into the same cluster (you don't know the clusters, nor how many clusters you have)



For exploratory analysis

To know about clustering

- There are two main methods:
 - **Classification = supervised method:**
[...]

- **Clustering = unsupervised method:**

Bring together elements which are similar into the same cluster (you don't know the clusters, nor how many clusters you have to use)

Hierarchical clustering analysis = HCA, is an unsupervised, exploratory method

Be careful:

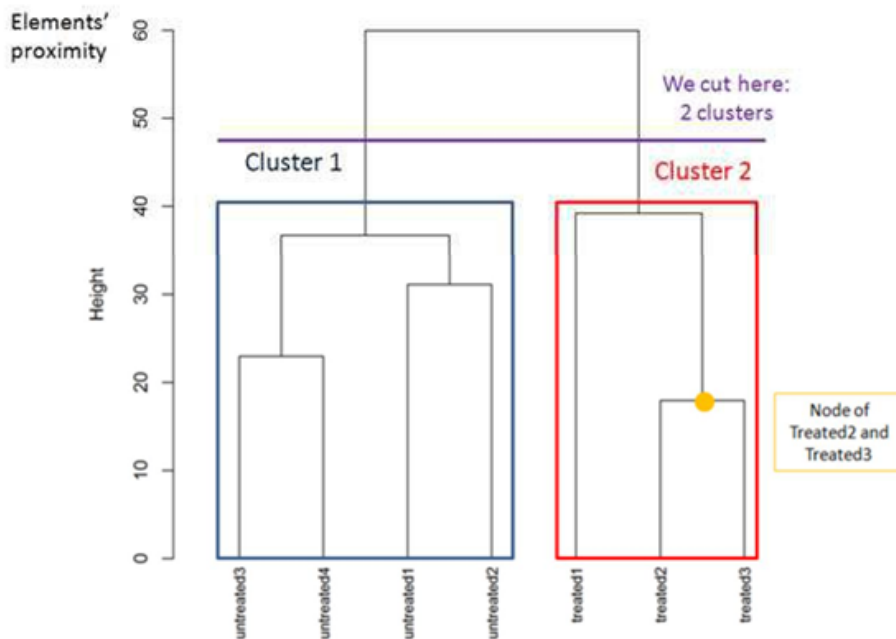
Clustering \neq Classification

To know about clustering

- **Hierarchical clustering analysis** of n objects is defined by a stepwise algorithm which merges two objects at each step, the two which are the most similar. In order to group together the two objects, we have to choose a distance measure (Euclidean, maximum, correlation). Then we bring together the clusters of objects by choosing an agglomeration method (ward, single, complete, average).
- Either rows or columns of a matrix can be clustered, in each case we have to choose the appropriate distance measure and agglomeration method that we prefer, the results depend on these choices.
Remember, Hierarchical clustering is an exploratory analysis method.

To know about clustering

- Example of clustering



Interpretation:

We expect to find replicates of the same condition in the same cluster.

Here, the dendrogram highlights there are 2 clusters, one for “untreated” condition and one for “treated” condition. The replicates are classified as we expect.

The closest objects are Treated2 and Treated3 (the small height node).

To know about clustering

- The Clustering Galaxy module allows to **generate hierarchical clustering analysis** on a table of numeric data **according to different parameters**.
 - *Input data file*: contains counts for each gene (txt with tabular as separator)
 - *group member file*: optional, allows to color labels of samples in the graphic (file with only one column, no header for column)

To know about clustering

- Data should be normalized before applying hierarchical clustering algorithm:

Normally when we do a hierarchical clustering, **we should have homoscedastic data**, which means that the variance of an observable quantity (i.e., the expression strength of a gene) does not depend on the mean.

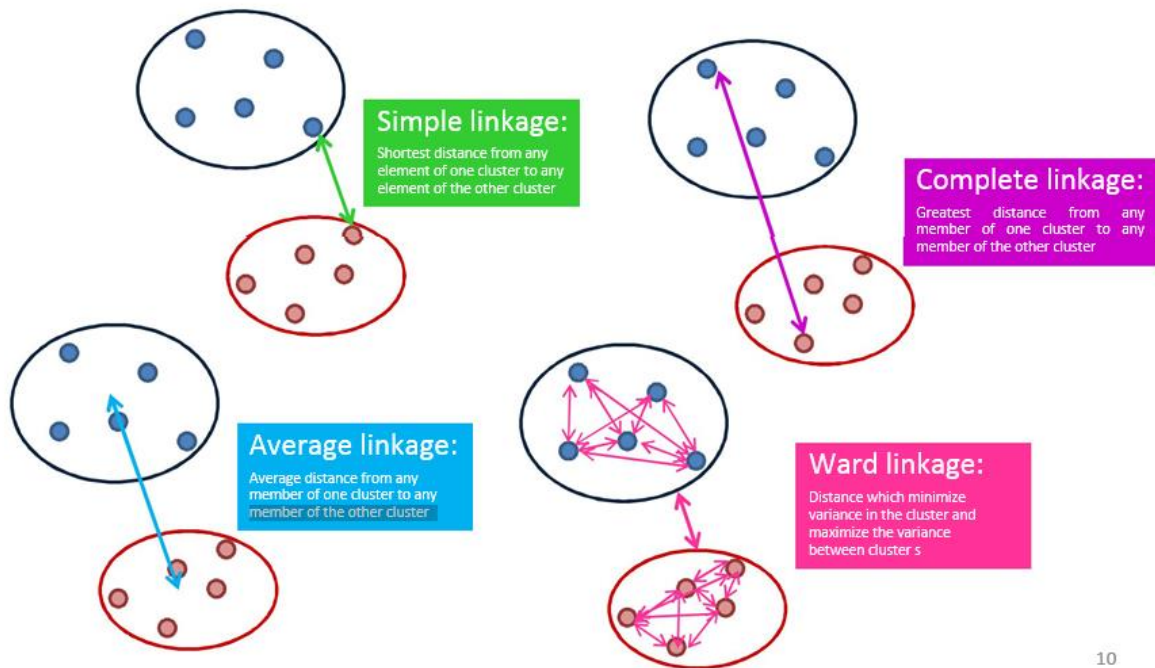
Please refer to Normalization galaxy module.

To know about clustering

- The measures to be used:
 - The distance between elements, must be one of "euclidean", "correlation" or "maximum". The most commonly used distance measures are "euclidean" and "correlation".
 - The agglomeration method to be: should be one of "ward", "single", "complete" or "average". The most commonly used measure is "ward".

To know about clustering

– The agglomeration method



To know about clustering

- You can also choose:
 - *Clustering is performed on the samples:* if YES clustering is performed on the samples. if NO clustering is performed on the variables.
 - *Number of top elements to use for clustering, selected by highest row variance. If NULL all the elements are selected: enter a number (maximum is 300).*
 - *An overall title for the plot:* enter a title for the plot
 - *A title for the x axis:* enter a title for the x axis
 - *A title for the y axis:* enter a title for the y axis
 - *The width of the graphics region in inches:* enter a number
 - *The height of the graphics region in inches:* enter a number
 - *The nominal resolution in ppi:* enter a number (a higher number means a higher resolution which can take times to open)

What you should have to begin

You can have 2 files:

- The **mandatory** file contains the numeric data to cluster and looks like this:

	A	B	C	D	E	F	G	H	I	J	K
1	Name	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m
2	SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7
3	CLAY	10.76	7.4	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5
4	KARPOV	11.02	7.3	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2
5	BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1
6	YURKOV	11.34	7.09	15.19	2.1	50.42	15.31	46.26	4.72	63.44	276.4
7	WARNERS	11.11	7.6	14.31	1.98	48.68	14.23	41.1	4.92	51.77	278.1
8	ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268
9	McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.1
10	MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.6	4.92	52.33	262.1
11	HERNU	11.37	7.56	14.41	1.86	51.1	15.06	44.99	4.82	57.19	285.1
12	BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.1	4.72	55.4	282
13	NOOL	11.33	7.27	12.68	1.98	49.2	15.29	37.92	4.62	57.44	266.6
14	BOURGUIGN	11.36	6.8	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.7
15	Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5	70.52	280.01
16	Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.9	69.71	282
17	Karpov	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11
18	Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.4	58.46	265.42
19	Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.9	55.39	278.05
20	Zsivoczky	10.91	7.14	15.31	2.12	49.4	14.95	45.62	4.7	63.45	269.54
21	Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.8	57.76	264.35

Missing values will be replaced with mean value of the considered variable, across all samples.

What you should have to begin

You can also have an optional file for coloring your cluster by groups:

- File contains the sample/variable member group file (**optional**) and looks like this:

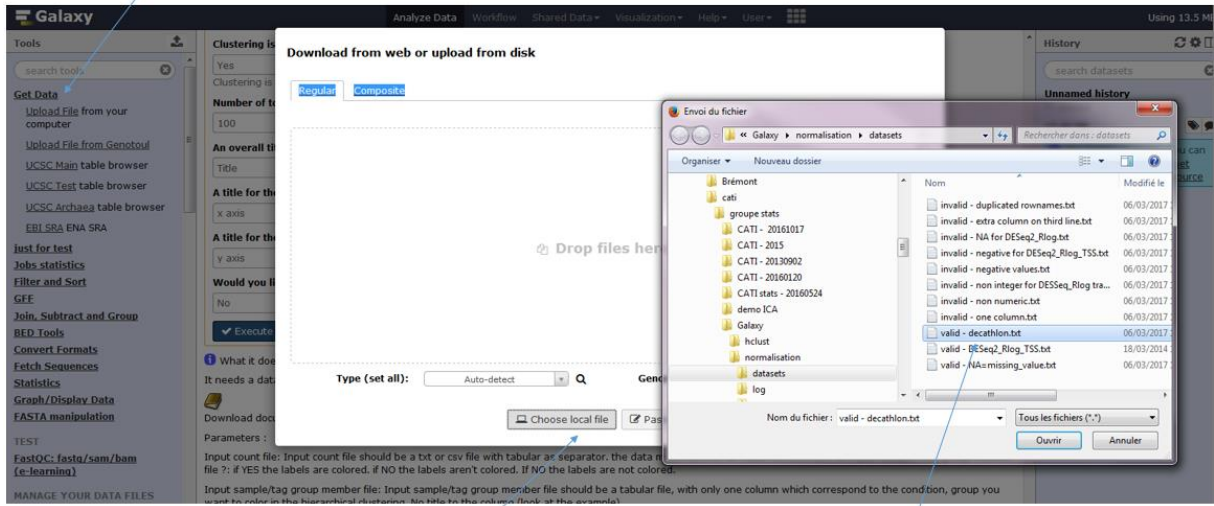
1
1
3
2
2
1
...

Each number corresponds to a group and will have a specific color in the clustering. If clustering is done on individuals (decathlon competitors in our example) the group number is assign to each individual as it is ordered in the numeric data table:

SEBRLE	group 1
CLAY	group 1
KARPOV	group 3
...	

Upload data

① Click on Get data






② Click on Choose local file

③ Select a file on your computer

Download from web or upload from disk

Regular Composite

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 valid - decathlon.txt	2.7 KB	Auto-det...	unspecified (?)		

Type (set all): Auto-detect Q Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Pause Reset **Start** Close

④ Click on Start, then Close

Upload data



Your data is now imported into your Galaxy history

Ready for clustering

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 13.5 MB

WELCOME TO GALAXY WORKBENCH

Galaxy is a workbench available for biologists from Sigeneae Platform. Galaxy objectives are:

- Make bioinfo Linux tools accessible to biologists.
- Hide the complexity of the infrastructure.
- Allow creation, execution and sharing of workflows.

Warnings :

- All jobs running on galaxy are sent to BioInfo Genotoul cluster.
- Your data are stored in work/ directory. Consequently, BioInfo Genotoul platform reserves the right to purge all files not accessed since 120 days on work/ disk space.
- Contact your support : sigeneae-support@listes.inra.fr
- La liste de diffusion user-galaxy-toulouse@listes.inra.fr permet de diffuser des informations à

Tools search tools

- Get Data
- just for test
- Jobs statistics
- Filter and Sort
- GFF
- Join, Subtract and Group
- BED Tools
- Convert Formats
- Fetch Sequences
- Statistics
- Graph/Display Data**
 - Bar chart for multiple columns
 - Boxplot of quality statistics
 - Histogram of a numeric column
 - Scatterplot of two numeric columns
 - Summary statistics (beta version)
 - Normalization Normalize your data with some well known methods
 - Hierarchical clustering**

History search datasets

Unnamed history
1 shown, 49 filtered
13.46 MB

S1: valid - decathlon.txt

Choose:

Graph/Display data > Hierarchical clustering
sub menus

Galaxy

Tools search tools

- Get Data
- just for test
- test
- Jobs statistics
- Filter and Sort
- GFF
- Join, Subtract and Group
- BED Tools
- Convert Formats
- Fetch Sequences
- Statistics
- Graph/Display Data**
 - Bar chart for multiple columns
 - Boxplot of quality statistics
 - Histogram of a numeric column
 - Scatterplot of two numeric columns
 - Summary statistics (beta version)
 - Hierarchical clustering (documentation available)**
 - PCAFactoMineR PCA using FactoMineR package
 - Normalization Normalize your data with some well known methods

Example of clustering #1

All options used with default value:

Hierarchical clustering (documentation available) (Galaxy Version 1.0.0) Options

Data file on which clustering will be performed
S1: valid - decathlon.txt

Do you have an input variable/individual group member file ?
No

The distance measure to be used (one choice mandatory)
euclidean

The agglomeration method to be used (one choice mandatory)
ward

Clustering is performed on the columns
Yes

Number of top elements to use for clustering, selected by highest row variance. If NULL all the elements will be used
NULL

Label used for Missing values
NA

An overall title for the plot
Title overall title for the plot

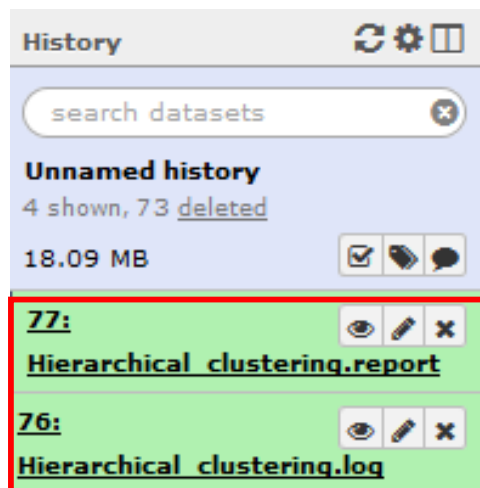
A title for the x axis
x axis for the x axis

A title for the y axis
y axis for the y axis

Would you like to parameter more graphic option
No

Execute

Click on **Execute** button and you will get two new boxes in your history:



Click on the **Eye** icons to look at the content.

Hierarchical_clustering.log:

In case of success (the box is green), you should see following message:

✓ Your clustering process is successfull !

In case of error (the box is red), you should see a message with the reason of the error:

△ An error occurred while trying to read your table.

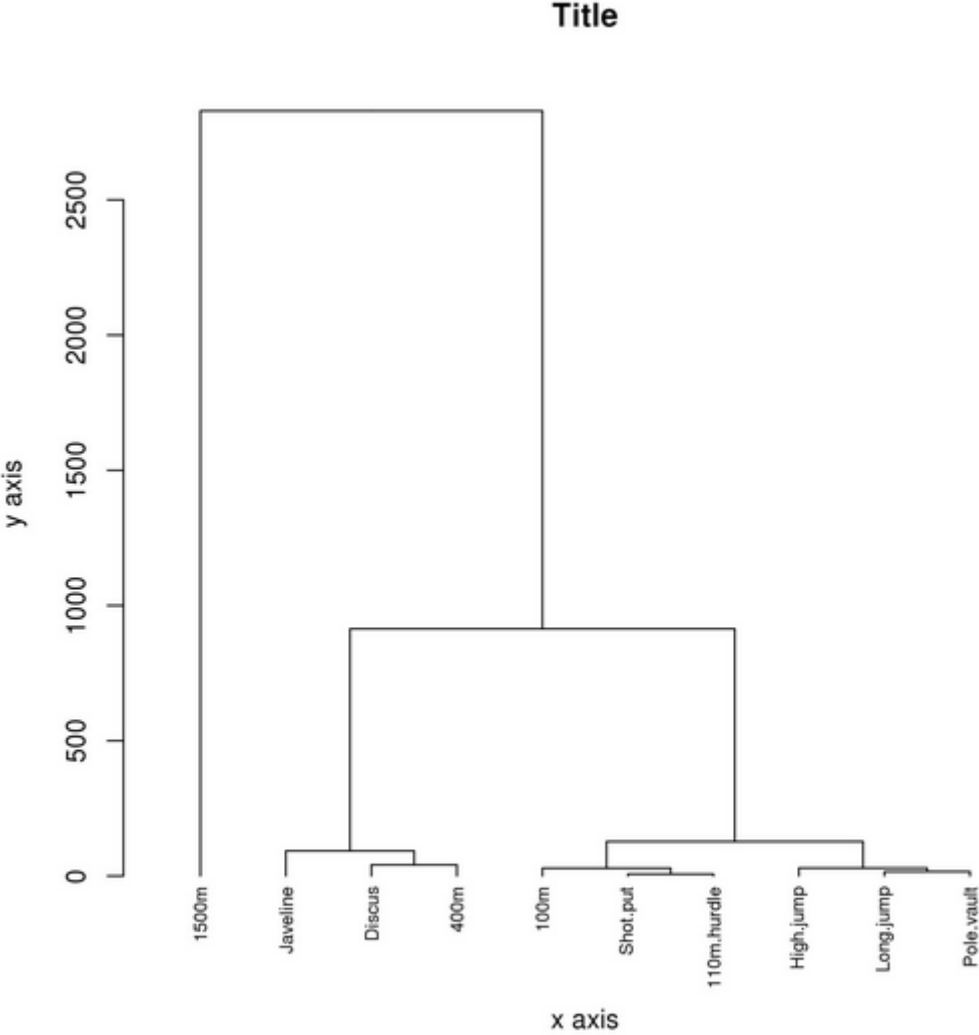
Please check that:

- the table you want to process contains the same number of columns for each line
- the first line of your table is a header line (specifying the name of each individual)
- the first column of your table specifies the name of each variable
- both individual and variable names should be unique
- each value is separated from the other by a **TAB** character
- except for first line and first column, table should contain a numeric value
- this value may contain character '.' as decimal separator or 'NA' for missing values

Error messages recieved :

The table on which you want to do a clustering must be a data table with at least 2 rows and 2 columns

Hierarchical_clustering.report:



[Download here your hierarchical classification map.](#)

Right click with the mouse on the link at the end of the page and use "Save link as ..." menu to save your graph on your disk.

Example of clustering #2

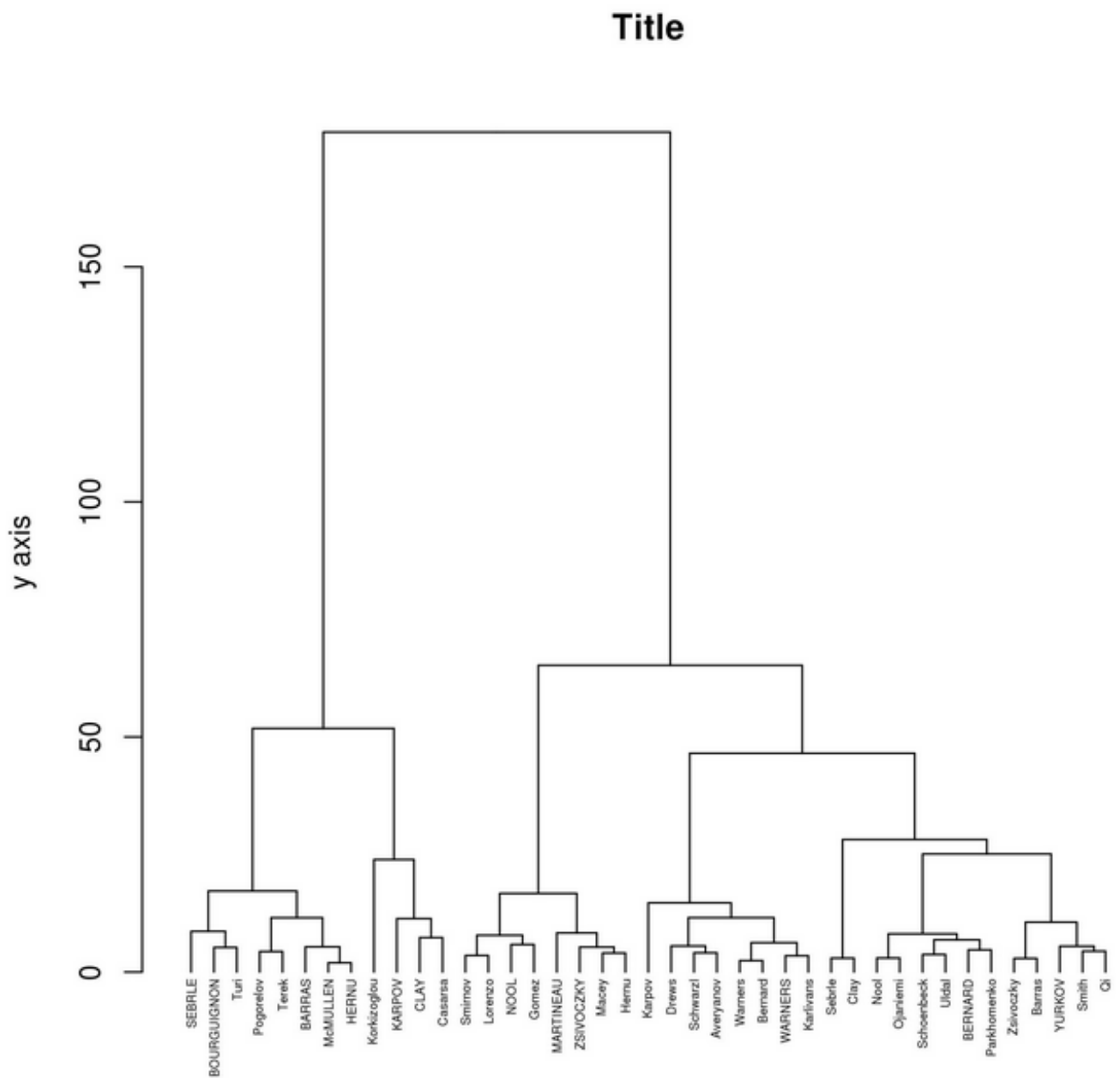
Clustering on lines (i.e. sportmen names):

Clustering is performed on the columns

Yes

Yes

No



Example of clustering #3

Clustering on lines with a group definition file:

Hierarchical clustering (documentation available) (Galaxy Version 1.0.0) Options

Data file on which clustering will be performed
51: valid - decathlon.txt

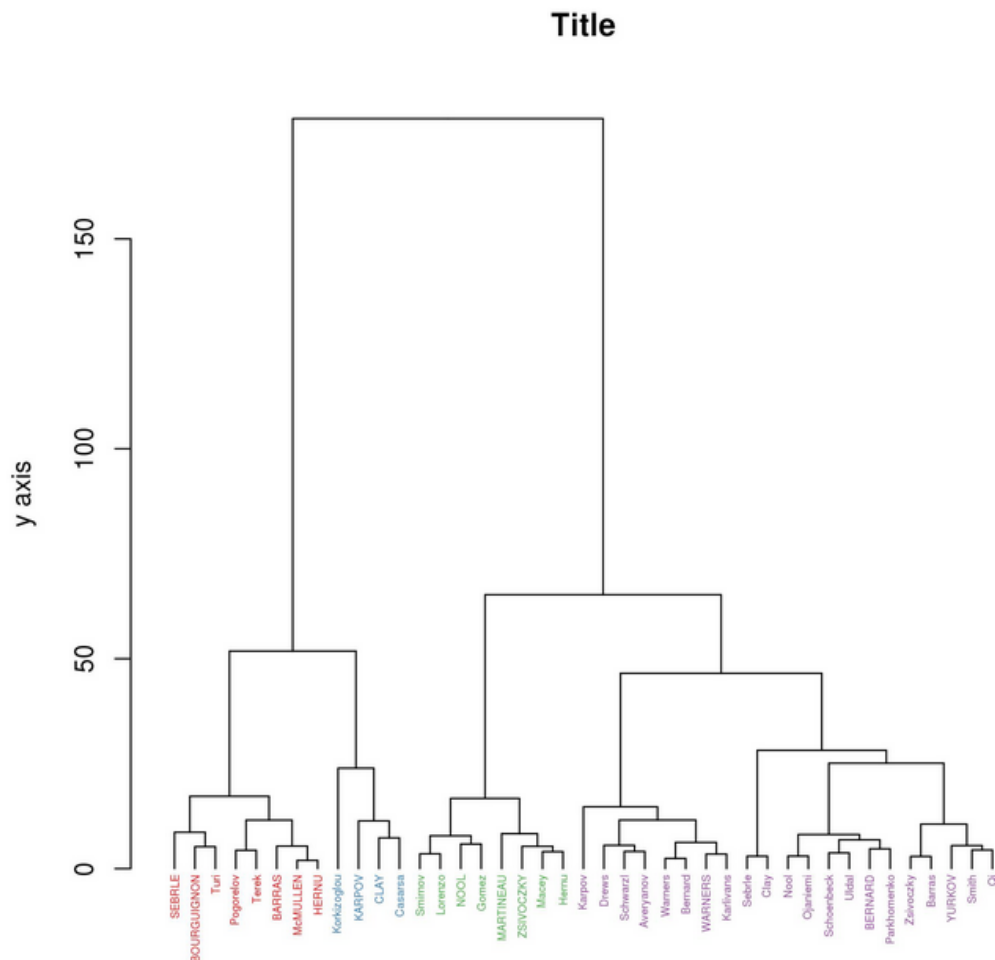
Do you have an input variable/individual group member file ?
Yes

Input variable/individual group member file
56: competitors_groups - 1 column.txt

The distance measure to be used (one choice mandatory)
euclidean

The agglomeration method to be used (one choice mandatory)
ward

Clustering is performed on the columns
No



Example of clustering #4

Configure title, axis labels, image size:

An overall title for the plot

Decathlon competitors clustering

A title for the x axis

Competitors

A title for the y axis

Euclidean distance (ward linkage method)

Would you like to parameter more graphic option

Yes

The width of the graphics region in inches

10

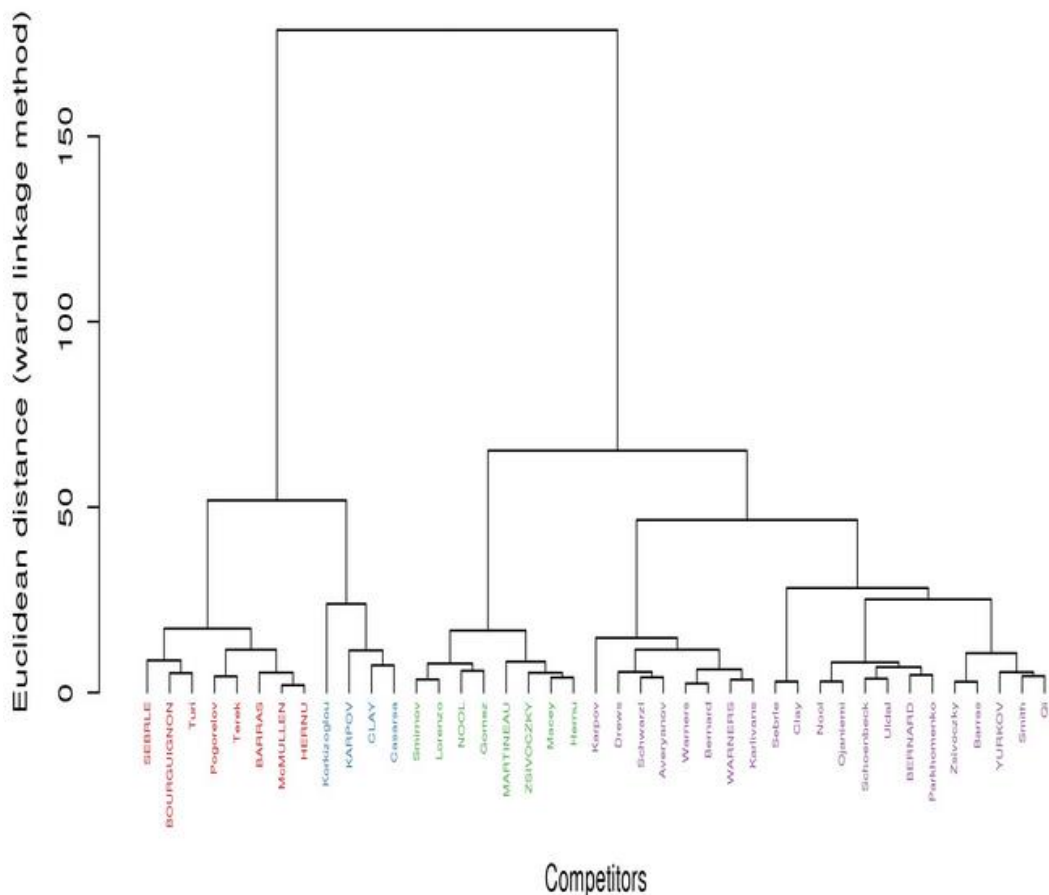
The height of the graphics region in inches

5

The nominal resolution in ppi

300

Decathlon competitors clustering



Special note for users having NGS counts or microarray data

In the example shown above, the table has the variables (sports) in columns and individual (sportmen) in lines. If you want to cluster a table containing NGS data or microarray data, you will usually have:

- Sample names in columns
- Variables (genes, probes, ...) in lines

If you want to cluster samples, beware to have “Yes” value in the parameter “[Clustering is performed on the columns](#)”.

If you want to cluster the genes, probes, ... you should have this parameter set to “No” but beware to put a non NULL value in the parameter :

“[Number of top elements to use for clustering](#)” otherwise the software will try to cluster ALL variables (genes, probes) and if you have more than one hundred variable, either your clustering will fail because it would need too much memory than available, or the gene/probe names will not be readable in the output image.

A last warning for NGS counts data: Do not use hierarchical clustering on counts data. This would lead to a wrong clustering, due to the fact that few genes are counted a lot. Variation of counts for these genes will decide of the clustering instead of taking into account all genes. You should first apply a RLog normalization process as proposed in the “Normalization” Galaxy module.