

Clustering on NGS data' learning

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

June 2015

To know about clustering

- **There are two main methods:**

- **Classification = supervised method:**

Bring together elements into categories you defined before 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, you don't know how many clusters you have to do)



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, you don't know how many clusters you have to use)

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

Be careful :

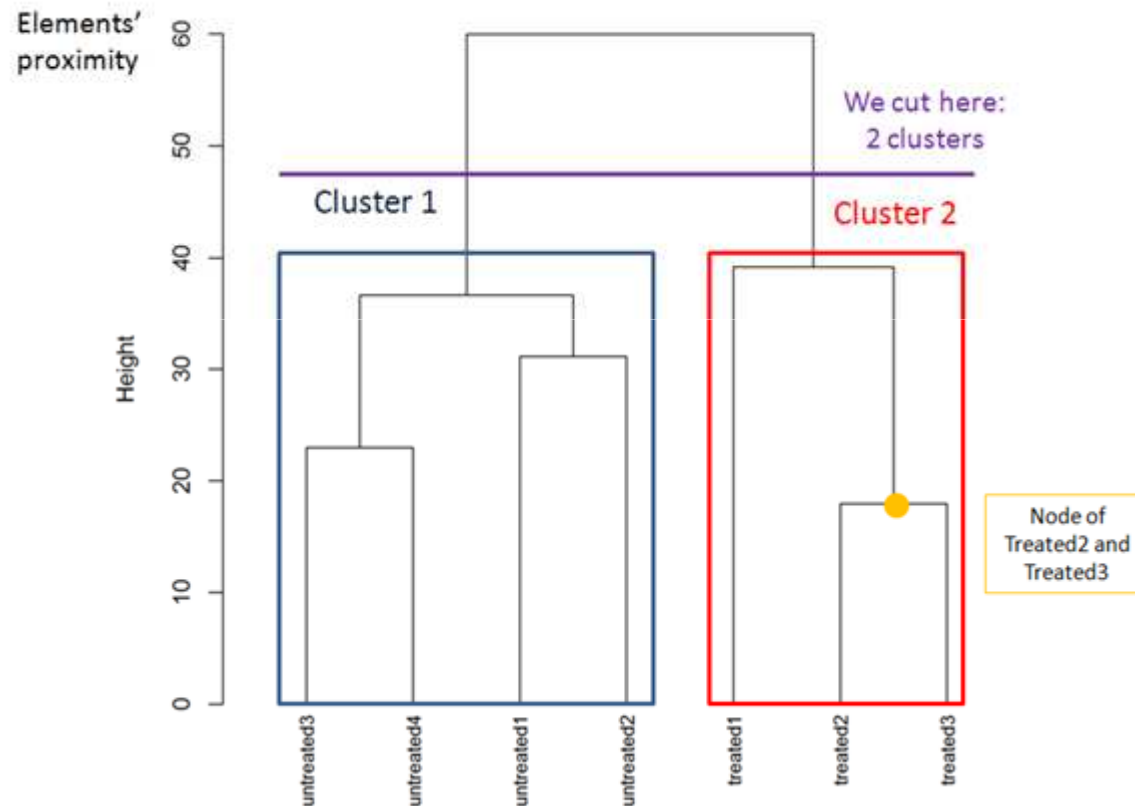
Clustering \neq Classification

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 a 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 depends on these choices. Remember, Hierarchical clustering is 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 count data **according to different parameters**.
 - *Input data file* : contains counts for each gene (csv ou txt with tabular as separator)
 - *Samples group member file*: optional, allows to color labels of samples in the graphic (tabular file with only one column, no title for column)
 - *Tags group member file*: optional, allows to color labels for tags (ie reads, genes) (tabular file with only one column, no title for column)

To know about clustering

- You can transform or not the datas:
 - *None*: the data are not transformed
 - *Rld*: the data are transformed into a specific log2 transformed. It allows to minimizes differences between samples for genes / contigs with small counts
 - *Vsd*: the data are transformed by an algorithm which calculates a variance stabilizing transformation

Best
method
for
RNAseq
data

To know about clustering

- Advices:

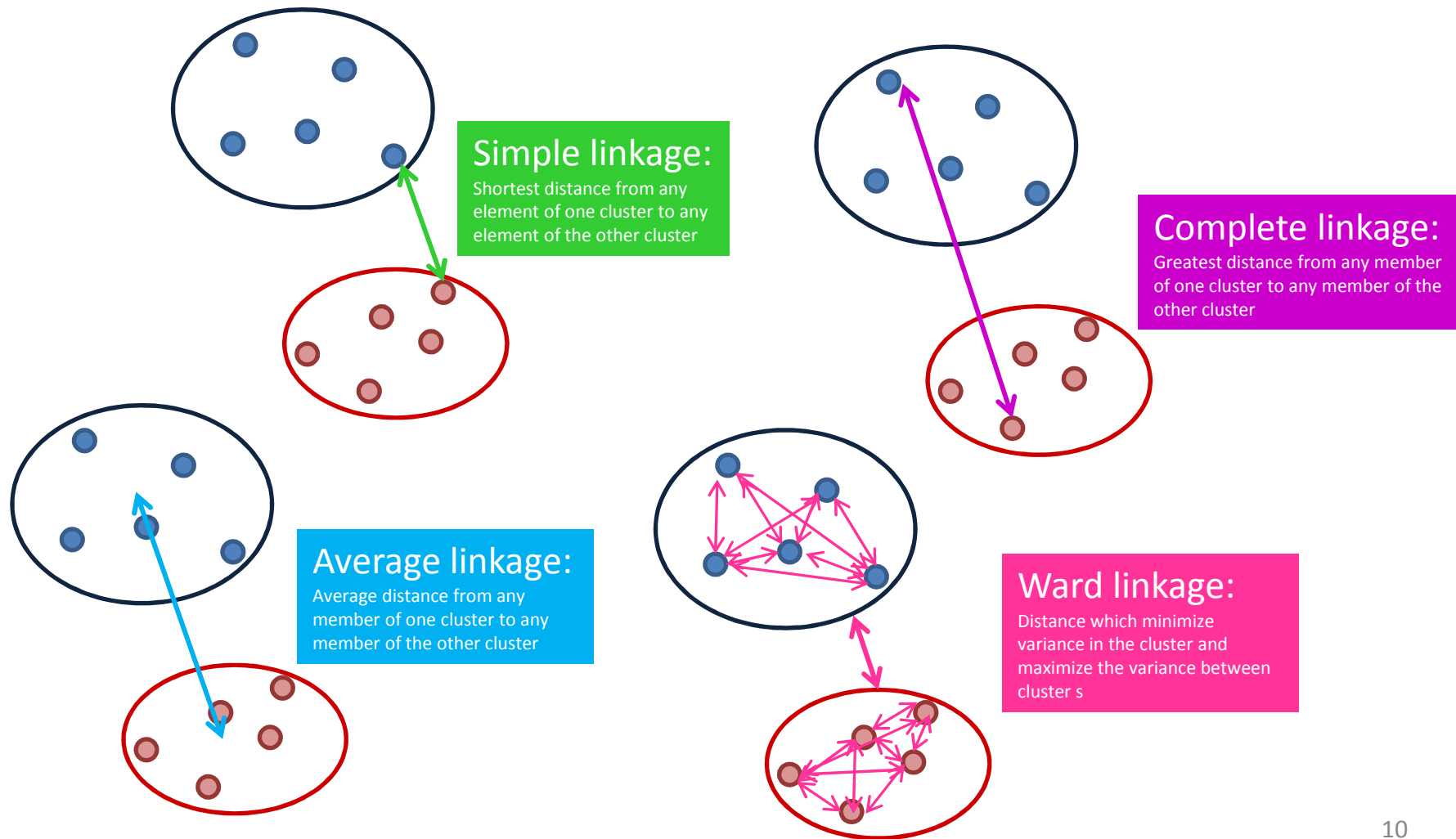
Normally when we do an hierarchical clustering, **we should have homoscedastic data** which means that the variance of an observable quantity (i.e., the expression strength of a gene) does not depend on the mean. **In RNA-Seq data, however, variance grows with the mean.** So different methods are implemented **to stabilized the variance** such as rld (regularized-logarithm transformation) and vsd (varianceStabilizingTransformation). **Currently, the regularized-logarithm transformation is the best method** in order to have homoscedastic data (This method is more robust in the case when the size factors of samples vary widely, that the varianceStabilizingTransformation method).

To know about clustering

- The distances measures to be used (one choice mandatory):
 - The distance between elements, must be one of "euclidean", "correlation" or "maximum". The most distance measure used is "euclidean" and "correlation".
 - The agglomeration method to be used (one choice mandatory): should be one of "ward", "single", "complete" or "average". The most distance measure used 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 genes.
 - *Number of top genes to use for clustering, selected by highest row variance*. If NULL all the genes 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 high resolution which can take times to open)

What you should have to begin

- You can have 3 files:
 - File contains the count data (**mandatory**) and looks like this:

1

| | A | B | C | D | E | F | G | H | |
|----|-------------|------------|------------|------------|------------|----------|----------|----------|--|
| 1 | gene_id | untreated1 | untreated2 | untreated3 | untreated4 | treated1 | treated2 | treated3 | |
| 2 | FBgn0000003 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| 3 | FBgn0000008 | 92 | 161 | 76 | 70 | 140 | 88 | 70 | |
| 4 | FBgn0000014 | 5 | 1 | 0 | 0 | 4 | 0 | 0 | |
| 5 | FBgn0000015 | 0 | 2 | 1 | 2 | 1 | 0 | 0 | |
| 6 | FBgn0000017 | 4664 | 8714 | 3564 | 3150 | 6205 | 3072 | 3334 | |
| 7 | FBgn0000018 | 583 | 761 | 245 | 310 | 722 | 299 | 308 | |
| 8 | FBgn0000022 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 9 | FBgn0000024 | 10 | 11 | 3 | 3 | 10 | 7 | 5 | |
| 10 | FBgn0000028 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | |
| 11 | FBgn0000032 | 1446 | 1713 | 615 | 672 | 1698 | 696 | 757 | |
| 12 | FBgn0000036 | 2 | 1 | 0 | 0 | 1 | 0 | 1 | |
| 13 | FBgn0000037 | 15 | 25 | 9 | 5 | 20 | 14 | 17 | |
| 14 | FBgn0000038 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 15 | FBgn0000039 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | |
| 16 | FBgn0000042 | 101664 | 120163 | 45880 | 53201 | 127363 | 76099 | 83164 | |
| 17 | FBgn0000043 | 33402 | 41118 | 16007 | 18360 | 56048 | 31421 | 34344 | |
| 18 | FBgn0000044 | 21 | 63 | 15 | 13 | 64 | 28 | 16 | |
| 19 | FBgn0000045 | 9 | 15 | 1 | 5 | 14 | 4 | 6 | |
| 20 | FBgn0000046 | 12 | 50 | 13 | 14 | 53 | 31 | 13 | |

Be careful,
NA are
not allowed

What you should have to begin

- You can have 3 files:
 - File contains the tag member group file (optional) and looks like this :

2

```
1
1
1
1
2
2
2
```

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

3

| | A |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |
| 6 | 2 |
| 7 | 2 |
| - | |

Upload data

- Example of upload file:

1

The screenshot displays the Siganae web application interface. The main window is titled "Upload File (version 1.1.3)". The "File Format" dropdown is set to "Auto-detect". The "File" field contains the text "Aucun fichier sélectionné." and a "Parcourir..." button, which is highlighted with a yellow circle. Below the "File" field, there is a section for "URL/Text:" and a "Convert spaces to tabs:" checkbox. The "Genome:" field is set to "unspecified (?)". An "Execute" button is located at the bottom of the main window. A file explorer window is open over the main window, showing the "Data" folder selected, with the file "gene_counts" highlighted. The file explorer window shows the "Nom" column with entries like ".Rhistory", "gene_counts", and "member". The "Nom du fichier:" field at the bottom of the file explorer is set to "gene_counts". The background of the Siganae interface shows a sidebar with "Tools" and "YOUR DATA" sections, and a "History" panel on the right showing a list of files including "member.csv" and "smaman member.txt".

Upload data

- Example of upload file:

2

Sigenae - Welcome slamarre Analyze Data Workflow Shared Data Visualization Help User

Tools

search tools

YOUR DATA

Upload Data

- Upload File
- Upload File from genotoul
- EBI SRA ENA SRA
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- Get Microbial Data
- BioMart Central server
- Compress zip or tar file

Download Data

FILES MANIPULATION

Text Manipulation (e-learning)

- Add column to an existing dataset
- Compute an expression on every row

Upload File (version 1.1.3)

File Format:

Auto-detect

Which format? See help below

File:

Parcourir... gene_counts.txt

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

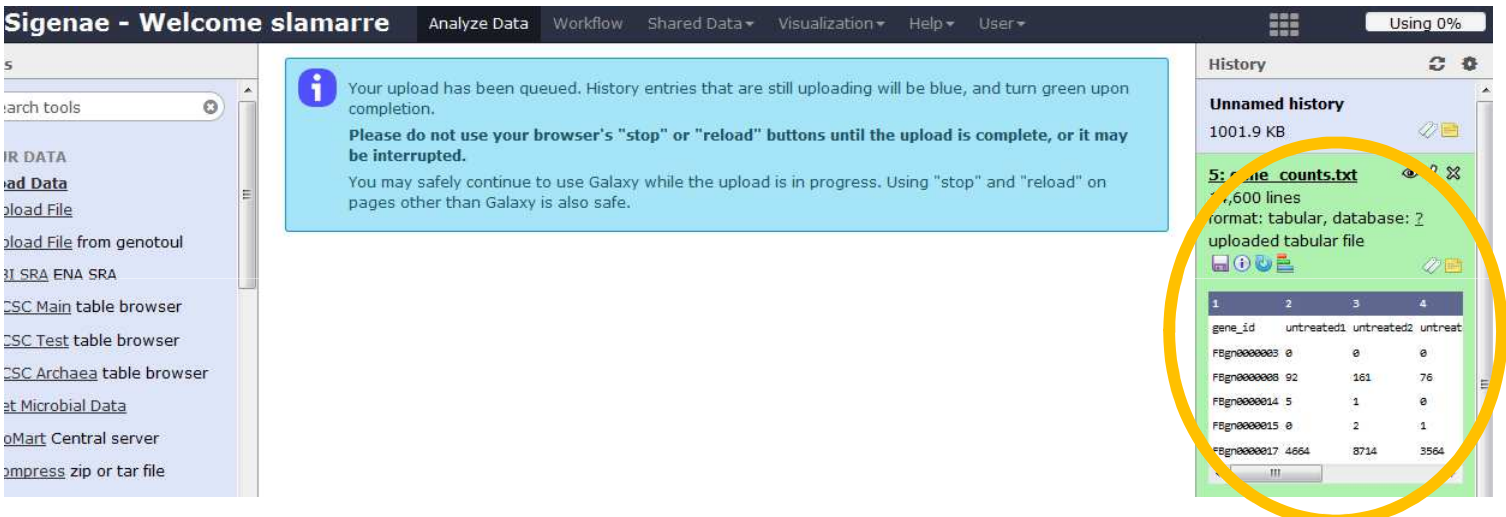
unspecified (?)

Execute

Upload data

- Example of upload file:

3



Signae - Welcome slamarre Analyze Data Workflow Shared Data Visualization Help User Using 0%

i Your upload has been queued. History entries that are still uploading will be blue, and turn green upon completion.
Please do not use your browser's "stop" or "reload" buttons until the upload is complete, or it may be interrupted.
You may safely continue to use Galaxy while the upload is in progress. Using "stop" and "reload" on pages other than Galaxy is also safe.

History Unnamed history 1001.9 KB

5: gene_counts.txt 1,600 lines
format: tabular, database: 2
uploaded tabular file

| 1 | 2 | 3 | 4 |
|-------------|------------|------------|---------|
| gene_id | untreated1 | untreated2 | untreat |
| FBgn0000003 | 0 | 0 | 0 |
| FBgn0000008 | 92 | 161 | 76 |
| FBgn0000014 | 5 | 1 | 0 |
| FBgn0000015 | 0 | 2 | 1 |
| FBgn0000017 | 4664 | 8714 | 3564 |

Successful upload

Ready for clustering

The screenshot shows the Sigenae web interface. The top navigation bar includes "Sigenae - Welcome slamarre" and menu items for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". On the left, a "Tools" sidebar lists various data processing options. The "Graph/Display Data" option is highlighted with a red rectangle, and the "Hierarchical clustering on NGS data" option is highlighted with a yellow circle. A central notification box contains an information icon and text regarding data uploads.

Sigenae - Welcome slamarre Analyze Data Workflow Shared Data Visualization Help User

Tools

- dataset
 - [Trim](#) leading or trailing characters
 - [Line/Word/Character count](#) of a dataset
 - [Secure Hash / Message Digest](#) on a dataset
 - [Filter on ambiguities](#) in polymorphism datasets
 - [Arithmetic Operations](#) on tables
- Filter and Sort**
- Join, Subtract and Group**
- Convert Formats**
- BED Tools**
- Graph/Display Data**
- [Histogram](#) of a numeric column
- [Scatterplot](#) of two numeric columns
- [Graphics class](#) (beta version)
- Hierarchical clustering on NGS data**
- [Cluster](#)

i Your upload has been queued. History entries that are still uploading will be blue, and turn green upon completion.
Please do not use your browser's "stop" or "reload" buttons until the upload is complete, or it may be interrupted.
You may safely continue to use Galaxy while the upload is in progress. Using "stop" and "reload" on pages other than Galaxy is also safe.

Ex 1

Examples of clustering

- Example on the samples: you only have count data file

1

Tools

- dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset
- Secure Hash / Message

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt

Clustering is performed on samples or tags ?:
Clustering performed on samples

Do you have an input sample / tag group member file ?:
No

Count data transformation for graphical display (one choice mandatory):
rld

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

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

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. MAXIMUM GENES NUMBER = 300:
100

We use the rld data transformation method (the most used)

Count data file

Clustering on samples

We don't have members files

The most used measure for calculate distances between elements

The most used measure for calculate distances between group of elements

We realize the clustering based on 100 genes which have the highest row variance

Ex 1

Examples of clustering

- Example: you only have count data file

2

The screenshot shows a web interface for clustering analysis. On the left is a navigation menu with the following items: 'columns', 'Graphics_desc (beta version)', 'Hierarchical clustering on NGS data', 'Cluster', 'DE_Seq Run Differential Expression analysis from SAM To Count data', 'SAM/BAM To Counts Produce count data from SAM or BAM files', 'DESeq2 Differential gene expression analysis based on the negative binomial distribution', and 'edgeR - Estimates differential gene expression for short read sequence count using methods appropriate for'. The main content area contains several input fields: 'An overall title for the plot (without white space):' with the value 'My first clustering on s'; 'A title for the x axis (without white space):' with the value 'x axis'; 'A title for the y axis (without white space):' with the value 'y axis'; 'The width of the graphics region in inches:' with the value '7'; 'The height of the graphics region in inches:' with the value '7'; and 'The nominal resolution in ppi:' with the value '300'. A yellow box highlights these last three fields. At the bottom is an 'Execute' button, which is circled in orange. To the right of the interface are two callout boxes: a green one with the text 'You can give your own title for the plot, for axis' and a yellow one with the text 'More graphics options (optional)'.

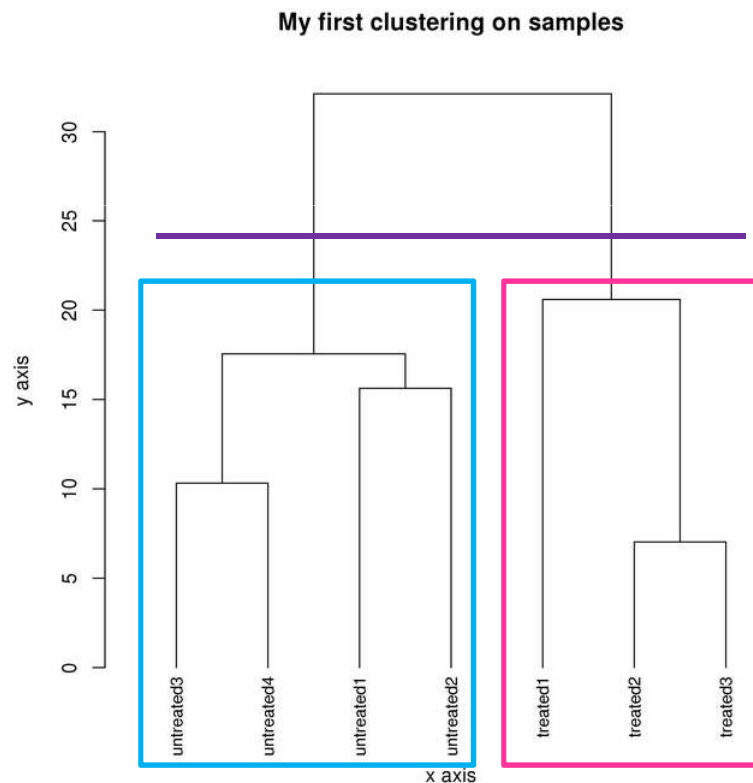
Ex 1

Examples of clustering

- Example: you only have count data file

Hierarchical classification :

3



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

| History | |
|---------------------------------------|---------------------------|
| Unnamed history | 1002.3 KB |
| 6: Hierarchical classification report | View data |
| 5: gene_counts.txt | |
| 4: /work/smaman/member.csv | |
| 3: member.csv | |
| 2: /work/smaman/gene_member.txt | |
| 1: /work/smaman/gene_counts.txt | |

Click here for view the clustering

We can cut the dendrogram in two clusters, on one hand (left) you have the untreated samples and on the other hand (right) you have the treated samples. The conditions are distinct, that is we expect.

Ex 2

Examples of clustering

- Example on tags (reads): you only have count data file

1

We use the rld data transformation method (the most used)

The most used measure for calculate distances between group of elements

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt

Clustering is performed on samples or tags ?:
Clustering performed on tags

Clustering is performed on samples or tags. Warning : clustering is performed on samples you need to give a member file with samples. If clustering is performed on tags you need to give a member file with tags. Instead your galaxy data

Do you have an input sample / tag group member file ?:
No

Count data transformation for graphical display (one choice mandatory):
rld

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

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

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. MAXIMUM GENES NUMBER = 300:
100

Count data file

Clustering on tags (reads)

We don't have members files

The most used measure for calculate distances between elements

We realize the clustering based on 100 genes which have the highest row variance

Ex 2

Examples of clustering

- Example on tags (reads): you only have count data file

2

An overall title for the plot (without white space):

My first clustering on t

A title for the x axis (without white space):

x axis

A title for the y axis (without white space):

y axis

The width of the graphics region in inches:

7

The height of the graphics region in inches:

7

The nominal resolution in ppi:

300

You can give your own title for the plot, for axis

More graphics options (optional)

Execute

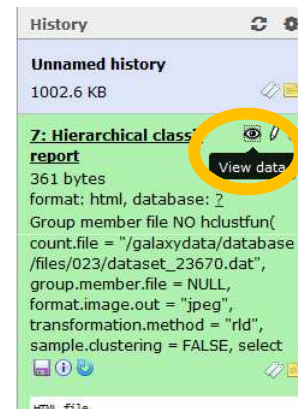
Ex 2

Examples of clustering

- Example on tags (reads): you only have count data file

3

Hierarchical classification :



Click here for
view the
clustering

We can cut the dendrogram in two clusters (blue and pink). In each cluster, you have the tags (reads) which look alike.

But you can also, cut in more number of clusters (for example, 4 groups), here in green.

Ex 3

Examples of clustering

- Example on samples: you have count data file + samples member group file

1

We use the rld data transformation method (the most used)

The most used measure for calculate distances between group of elements

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt — Count data file

Clustering is performed on samples or tags ?:
Clustering performed on samples — Clustering on tags (reads)

Clustering is performed on samples or tags. Warning: clustering is performed on samples you need to give a member file with samples. If clustering is performed on tags you need to give a member file with tags. Instead your galaxy dataset will be in error (red).

Do you have an input sample / tag group member file ?:
Yes — Yes — Samples member group file

Input sample/tag group member file:
4: /work/smaman/member.csv

Count data transformation for graphical display (one choice mandatory):
rld — The most used measure for calculate distances between elements

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

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

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. MAXIMUM GENES NUMBER = 300:
100 — We realize the clustering based on 100 genes which have the highest row variance

Ex 3

Examples of clustering

- Example on samples: you have count data file + samples member group file

2

An overall title for the plot (without white space):

A title for the x axis (without white space):

A title for the y axis (without white space):

The width of the graphics region in inches:

The height of the graphics region in inches:

The nominal resolution in ppi:

Execute

You can give your own title for the plot, for axis

More graphics options (optional)

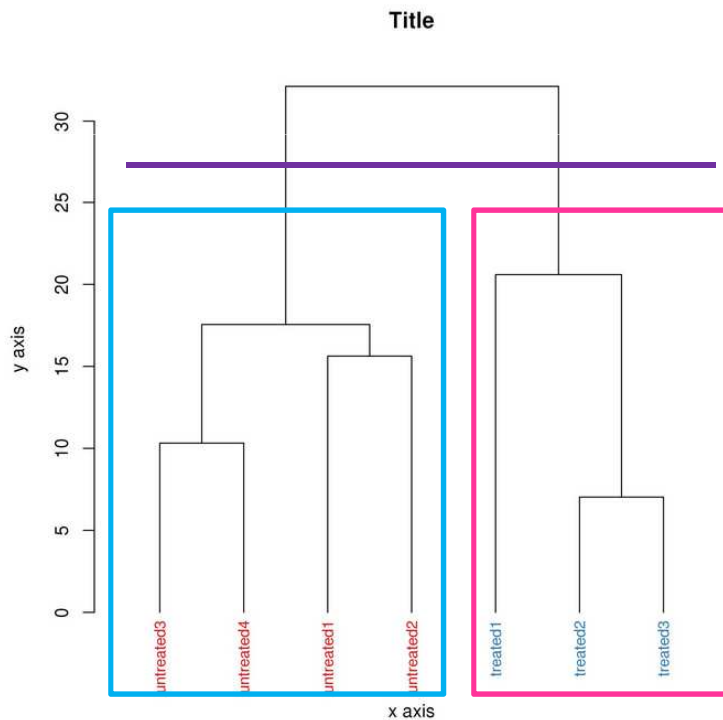
Ex 3

Examples of clustering

- Example on samples: you have count data file + samples member group file

3

Hierarchical classification :



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



Click here for view the clustering

We can cut the dendrogram in two clusters, on one hand (left) you have the untreated samples and on the other hand (right) you have the treated samples. The conditions are distinct, that is we expect.

Moreover, you see the labels of samples with their colors, so it is easy for you to see if the conditions are distinct or not.

Ex 4

Examples of clustering

- Example on tags: you have count data file + tags member group file

1

We use the rld data transformation method (the most used)

The most used measure for calculate distances between group of elements

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt — **Count data file**

Clustering is performed on samples or tags ?:
Clustering performed on tags — **Clustering on tags (reads)**

Do you have an input sample / tag group member file ?: Yes — **Genes member group file**

Input sample/tag group member file:
2: /work/smaman/gene_member.txt

Count data transformation for graphical display (one choice mandatory):
rld — **The most used measure for calculate distances between elements**

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

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

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. MAXIMUM GENES NUMBER = 300
100 — **We realize the clustering based on 100 genes which have the highest row variance**

Ex 4

Examples of clustering

- Example on tags: you have count data file + tags member group file

2

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. **MAXIMUM GENES NUMBER = 300:**

An overall title for the plot (without white space):

A title for the x axis (without white space):

A title for the y axis (without white space):

The width of the graphics region in inches:

The height of the graphics region in inches:

The nominal resolution in ppi:

You can give your own title for the plot, for axis

More graphics options (optional)

Ex 4

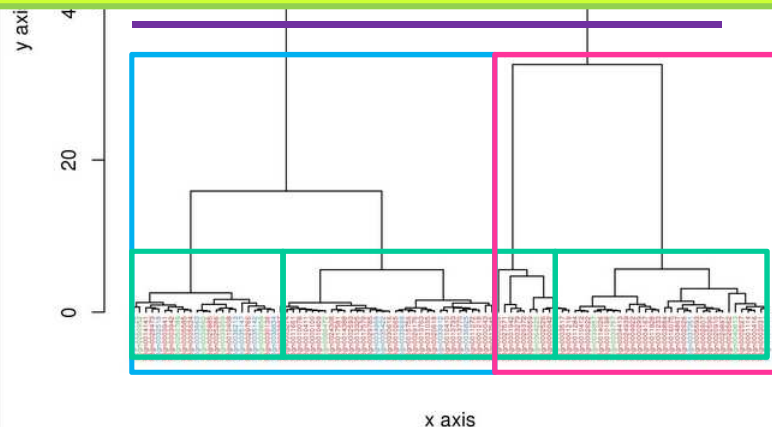
Examples of clustering

- Example on tags: you have count data file + tags member group file

3

Hierarchical classification :

We can see here, that the colors aren't equivalent to the clusters (maybe here it is a statistician which don't know the characteristics of these genes 😊 and he gives randomly a group member for each gene)



History

Unnamed history
1003.3 KB

9: Hierarchical classification report
361 bytes
format: html, database:
group member file YES hdistfun(
count.file = "/galaxydata
/database/file
/023/dataset_
group.memb
/database/file
/016/dataset_
format.image.
transformation

View data

HTML file

8: Hierarchical classification report
361 bytes
format: html, database:
group member file YES hdistfun(
count.file = "/galaxydata
/database/file
/023/dataset_
group.memb
/database/file
/016/dataset_
format.image.
transformation

HTML file

7: Hierarchical classification report
361 bytes
format: html, database:
group member file YES hdistfun(
count.file = "/galaxydata
/database/file
/023/dataset_
group.memb
/database/file
/016/dataset_
format.image.
transformation

HTML file

Click here for view the clustering

We can cut the dendrogram in two clusters (blue and pink). In each clusters, you have the tags (reads) which look alike. But you can also, cut in more number of clusters (for example, 4 groups), here in green.

Ex 5

Examples of clustering

- Example with other distances (in order to see the difference, if the results are more suitable for our data): you have count data file + samples member group file

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt

Clustering is performed on samples or tags ?:
Clustering performed on samples

Clustering is performed on samples or tags. Warning : Give a sample or tag member file is not mandatory. If clustering is performed on samples you need to give a member file with samples. If clustering is performed on tags you need to give a member file with tags. Instead your galaxy dataset will be in error (red).

Do you have an input sample / tag group member file ?:
Yes

Input sample/tag group member file:
4: /work/smaman/member.csv

Count data transformation for graphical display (one choice mandatory):
rld

The distance measure to use:
correlation

The agglomeration method to use:
average

Number of top genes to use for clustering, selected by highest row variance. If NULL all the genes are selected. MAXIMUM GENES NUMBER = 300:
100

Number of top genes to use for clustering, selected by hi selected. MAXIMUM GENES NUMBER = 300:
100

An overall title for the plot (without white space):
Title

A title for the x axis (without white space):
x axis

A title for the y axis (without white space):
y axis

The width of the graphics region in inches:
7

The height of the graphics region in inches:
7

The nominal resolution in ppi:
300

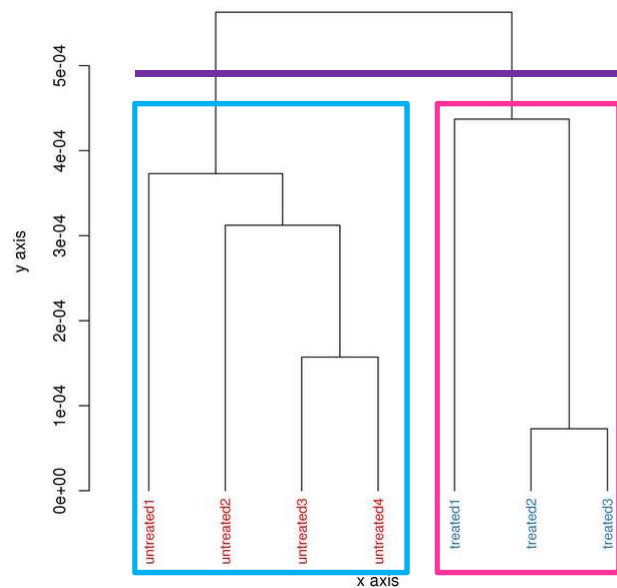
Execute

We try these distances

Ex 5

Examples of clustering

- Example with other distances (in order to see the difference, if the results are more suitable for our data): you have count data file + samples member group file



We see conditions are distinct. But in details, the aggregation of samples (noticeable especially for untreated condition) is particular (stairs shape). It is due to average distance (to convince, you can do a clustering with euclidean distance and average distance).

Ex 6

Examples of clustering

- Example with other transformation: you have count data file + samples member group file

Hierarchical clustering (version 1.0.0)

Input count file:
5: gene_counts.txt

Clustering is performed on samples or tags ?:
Clustering performed on samples

Clustering is performed on samples or tags. Warning : Give a sample or tag member file is not mandatory. If clustering is performed on samples you need to give a member file with samples. If clustering is performed on tags you need to give a member file with tags. Instead your galaxy dataset will be in error (red).

Do you have an input sample / tag group member file ?:
Yes

Input sample/tag group member file:
4: /work/smaman/member.csv

Count data transformation for graphical
none

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

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

Number of top genes to use for clustering, selected selected. MAXIMUM GENES NUMBER = 300:
100

An overall title for the plot (without white space):
Title

A title for the x axis (without white space):
x axis

A title for the y axis (without white space):
y axis

The width of the graphics region in inches:
7

The height of the graphics region in inches:
7

The nominal resolution in ppi:
300

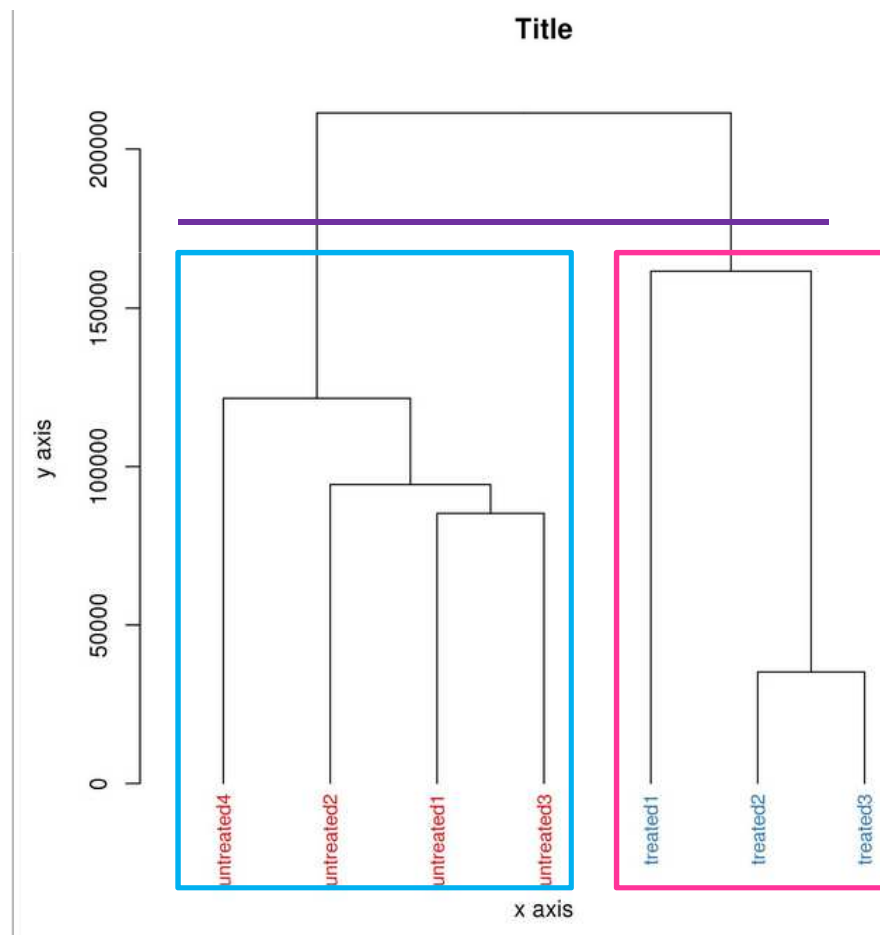
Execute

We don't want to perform a transformation

Ex 6

Examples of clustering

- Example with other transformation: you have count data file + samples member group file



*We see conditions are distinct.
The samples of untreated condition aren't cluster as the same that when we make a rld transformation.*

To conclude

- **The clustering is a exploratory analysis:**
 - The results depends on the parameters you choose
 - There is not a bad clustering, is the biologist which know if the clustering is suitable for its datas
 - It is important to do others exploratories analysis like Principal Component Analysis in order to confirm the results of clustering