Has iniciat la sessió

En enviar aquest formulari, es registraran el teu nom d'usuari (**jan.izquierdo@alum.esci.upf.edu**) i les respostes.

CANVIA DE COMPTE

Comparative and Functional Genomics. Practical session: Gene Order

IMPORTANT_1: Submit your answers with your ESCI-UPF gmail account. You can only submit your answers once.
IMPORTANT_2: Friday 14th of June 2024, is the last day to submit your answers.
IMPORTANT_3: When you submit your answers, a confirmation message will appear on your screen "Your response has been recorded"

jan.izquierdo@alum.esci.upf.edu Canvia de compte

* Indica que la pregunta és obligatòria

Adreça electrònica *

Registra **jan.izquierdo@alum.esci.upf.edu** com el correu electrònic que s'inclourà a la meva resposta

1) Genome browsers

Genome browsers are a graphical interface displaying information on genomic data from biological databases. They are usually specific to model organisms with dedicated websites and specific databases. It allows to navigate assembled genomic sequences (ideally at chromosome level, although can be used for scaffolds and supercontigs) and explore various annotations in different tracks, such gene orientation, gene structure (exon and introns), the way in which a series of genes are distributed in the genome (gene order), and so on. In this exercise you will practice how to navigate and compare gene order in a genus of ascomycotan fungus.

Go to this link (if it does not work, try to refresh or go to https://www.ncbi.nlm.nih.gov/gdv and look for **Aspergillus fumigatus** and click Browse genome) This is the NCBI genome viewer and it allows to visualize the assembled genomes of all the species present in NCBI with its genes (green tracks) and other annotations that can be configured ad hoc. For example, you can visualize the GC content along the chromosome by clicking on **Tracks and User Data**, **options, Configure tracks, Sequence** and **G+C content**.

1.1) Search for the *gliP* gene (Gliotoxin biosynthesis protein P, **AFUA_6G09660**) in the search bar. Ask the browser to show you 50 kb around the gene of interest. List the names of the six **PROTEINS** that are on either side of *gliP*. *

AFUA_6G09670, AFUA_6G09710, AFUA_6G09

1.2) *Aspergillus fumigatus* is a primary and opportunistic pathogen, and its virulence may be augmented by the production of mycotoxins. *gliP* has an important role in the production of gliotoxin, an immunosuppressive mycotoxin. Do you think any of the surrounding proteins are also involved?

- (●) Yes
- ( ) No

Esborra la selecció

2) The GFF file

Not every species genome is available in online genome browsers. In such circumstances, we have to go to download and work with raw information. General Feature Format (GFF) files are text files containing genomic features with tab separated fields including sequence name, start and end positions of the feature, strand direction, frame, etc. You can find information about it on the ensembl website. Keep in mind that there are other file formats with similar/equivalent information. A not minor feature of each of these formats is whether they start at the base position 0 or 1 (1-based formats: VCF, MAF, SAM, GFF and Wiggle; 0-based formats: Bed, BAM, BCFv2 and PSL). Here you will navigate a GFF and compare it with the genome browser tool.

2.1) Go to EnsemblFungi and find the GFF3 file corresponding to Aspergillus flavus NRRL3357 (Assembly: JCVI-afl1-v2.0). Uncompress the file and search the **GliG** protein. Find a gene with this putative function in *Aspergillus flavus* genome Which is the location of this gene?

97373-98280

2.2) How many exons does this gene have in this genome?

- ○ 1
- ○ 2
- ◉ 4
- ○ 5

Esborra la selecció

2.3) Are there genes potentially related to gliotoxin both in the 10Kb upstream and downstream to **GliG** as in *A. fumigatus*? If so, how many there are?

- ○ None
- ○ 3, only upstream
- ○ 3, upstream and downstream
- ○ 5, only upstream
- ◉ 5, upstream and downstream

Esborra la selecció

2.4) What may be the advantages of using GFF files to explore genome features over online Genome Browsers?

The information is more condensed and its easier to view the information of genes at the same time, its more user friendly, as more people know how to work with regular files that a specific website. You can also use it in scripts and it's accessible offline.

3. Genome Alignment (Mauve)

Mauve is a software package that allows aligning regions among two or more genome sequences that have undergone local and/or large-scale changes.

IMPORTANT: As Mauve can be difficult to install you can complete the exercise just by analyzing the provided screenshots. Otherwise, Mauve is available for Linux, Mac and Windows systems (http://darlinglab.org/mauve/download.html). You can find installation instructions here: https://darlinglab.org/mauve/user-guide/installing.html. For Linux, after downloading it, open the terminal and go to the folder where the software has been downloaded. Then:

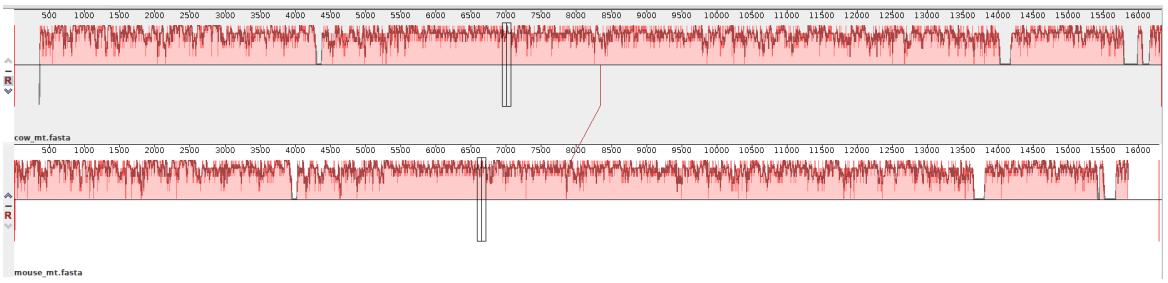
- Extract the file: e.g., *tar xvzf mauve_linux_snapshot_2015-02-13.tar.gz*
- The run it: *./mauve_snapshot_2015-02-13/Mauve*

Alternatively, you can install mauve from conda: *conda install -c bioconda mauve*.

3.1) Go to NCBI and download in fasta format the mitochondrial genomes of mouse (**GenBank: V00711.1; https://www.ncbi.nlm.nih.gov/nuccore/V00711**) and cow (**GenBank: V00654.1; https://www.ncbi.nlm.nih.gov/nuccore/V00654**).

Open mauve and select **File -> Align with progressiveMauve**. It will open a window, where you have to insert the two genomes by clicking in **Add Sequence**. Once you add both sequences, **select an Output name** and **press Align…**

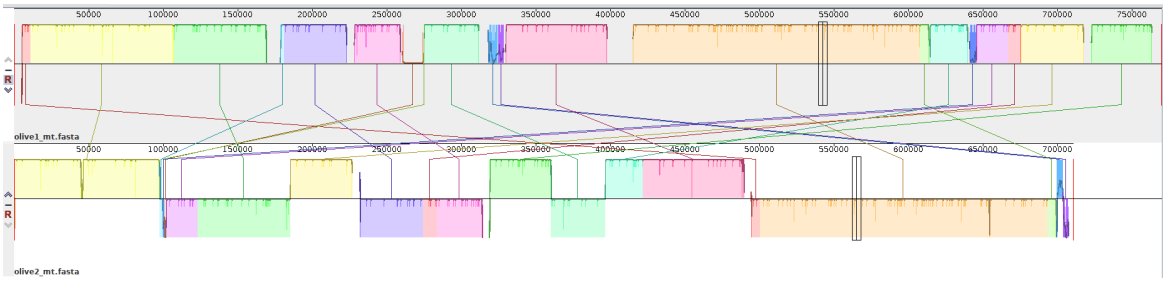
Cow versus mouse mitochondrial genome (output from Mauve in case you have not installed it)



Are the mitochondrial genomes conserved? Why? What do you think the empty spaces represent?

Yes, the line marks the regions as homologous. The empty spaces represent parts of the sequence that aren't homologous, non-conserved regions.

Now download the mitochondrial genomes of two olive individuals: **GenBank: MG372119.1 (https://www.ncbi.nlm.nih.gov/nuccore/MG372119)** and **GenBank: MG372117.1 (https://www.ncbi.nlm.nih.gov/nuccore/MG372117)**. Align both genomes as before.



3.2) Which of the following statements are true, if any?

- [x] There seems to be a substantial amount of homologous regions between the two mitochondrial genomes
- [x] The two mitochondrial genomes have many rearrengements
- [x] There are some regions that are not homologous
- [x] These mitochondrial genomes present inverted regions respect to each other

3.3) In this exercise we have explored the mitochondrial genomes of two individuals that belong to the same species (olive tree) and in the previous exercise two different species (mouse and cow, less related). Is this represented in the results? If not, why do you think this could be happening?

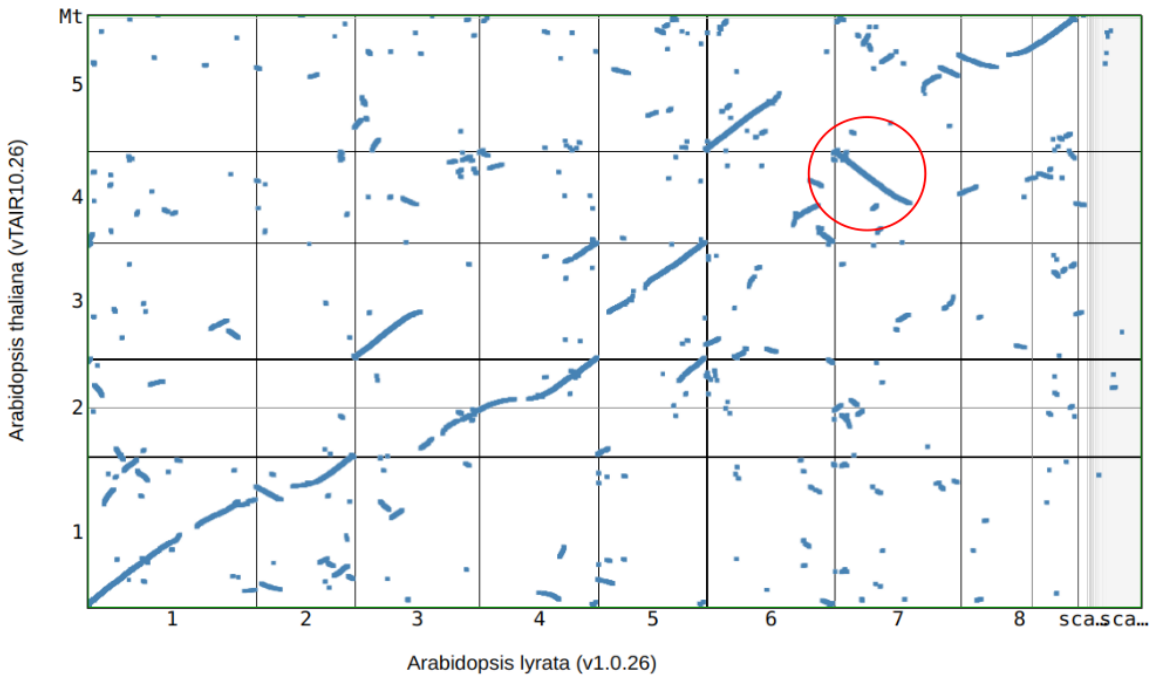
Not represented. Plants can survive with more changes in the order of the genome than animals do, therefore 2 olive plants can have a very different genome order and still be functional where animals wouldn't. Animal genomes will always be more similar between species.

4) Synteny browsers (CoGe)

Synteny is the conservation of blocks of order within genes when comparing chromosomes. When evaluating synteny between different species, we will explore whether the order of a set of orthologous genes is conserved between the compared genomes.

CoGe is a platform that can be used for synteny conservation analyses. Among its key features it includes several programs that allow you to easily compare genomes based on gene order, or see regions in the genome that have conserved gene order.

Go to **Tools; SynMap**. SynMap will compare two genomes, it is better when the genomes have CDS predicted so select the two species: ***Arabidopsis lyrata* (id6589)** and ***Arabidopsis thaliana* (id39871)** and press on **Generate SynMap.**



4.1) How do you interpret the graph? What do the blue lines mean? Which genomic variation does the line inside the red circle represent?

The blue lines are base-base correspondences form the 2 species. The variant means an inversion.

4.2) A new alternative way to explore syntenic map is Comparative Genome Viewer from NCBI. Go there and select *Rattus norvegicus* against *Mus musculus*.

Go to dotplot view. Is the gene order more conserved than in the previous example?

- ○ Yes
- ◉ No
- ○ More or less the same

Esborra la selecció

4.3) Take a look at chromosome 4 from *Rattus norvegicus* compared to *Mus musculus*. What genetic phenomenon explains this display, and which *M. musculus* chromosomes are involved?

HINT: You can switch between view modes and click on specific chromosomes to highlight

There is a translocation of chromosome 4 of Rattus norvegicus to chromosome 6 of Mus musculus. A fragment also transloactes to chromosome 5.

5. Protein association networks (String)

STRING is a database of known and predicted protein-protein direct (physical) and indirect (functional) interactions. http://version10.string-db.org/help/getting_started/

Go to the string database (https://string-db.org/) and search the protein ***GliG*** in the organism **Aspergillus flavus NRRL3357** and click continue.

The first image you see is a network. This network shows genes that are known to interact with your gene of interest.

5.1) Do all the proteins in this network connect directly with your protein of interest? What does it mean that so many proteins interact with your protein of interest?

Yes, all these proteins interact with my protein of interest. The broadness of the interactions implies that my protein is essential to the cellular function of the organism, thus its probably involved in various processes doing different roles.

5.2) Now press the **+more** button on the lower right page. Do all proteins still directly interact with your protein of interest?

- ○ Yes
- ◉ No

Esborra la selecció

5.3) As we saw a few sessions ago, proteins can be grouped into groups using clustering strategies. In this case the proteins will be clustered not by sequence similarity but by interactivity. Expand the network again (**press 3 times +more**). Go to **Clusters** and apply the two available clustering methods: mcl and k-means with default parameters. Do the two methods agree? Which method do you think makes more sense?

Both methodologies produce similar results, but k-means creates 1 cluster with 1 protein which means that MCL is a better choice.

5.4) Go back to the base network. Now go to settings and have a look at the Basic settings. As you can see the lines indicate the different kinds of evidence the researchers used to find the interactions between proteins. If you select only experimental and co-expression evidence, how many proteins form the network?

- ○ 9
- ◉ 5
- ○ 8
- ○ 3

Esborra la selecció

5.5) Now switch the lines from evidence to confidence (keeping only experimental and coexpression evidence as before). Write just the name of the protein in this graph which is more confidently connected to **GliG** (**AFLA_064530**).

HINT: Browse and explore the information available in the different tabs.

AFLA_049300

Envia

Esborra el formulari