Aina Montalbán
NIA: 103635

Session 6 - Gene order

A.- How to find where your Gene is in your genome.

In order to compare gene order across species we first need to be able to locate genes in our genome. There are mainly two ways of doing this:

Genome browsers.

1.- Search the browser for the gliP gene (Gliotoxin biosynthesis protein P, AFUA_3G12920). Now move back to the browser view by pushing on the position link. As you can see the browser is now focused on the gene you searched for. You'll have to zoom out to have a better view of the surrounding genes.

AnnotatedSequence Feature Standard Name:gliP Systematic Name:Afu6g09660 Alias:AFUA_6G09660, AFUB_036270, AFUB_075710, CADAFUAG00001584, nrps10, pesK Coordinates:Chr6_A_fumigatus_Af293:2352620..2359124 Description:Non-ribosomal peptide synthetase encoded in the gliotoxin biosynthetic gene cluster; catalyzes the first step in gliotoxin biosynthesis; regulated by the transcription factor StuA; expression increases in vivo

2.- Ask the browser to show you 50 kb around the gene of interest. Can you list the three proteins that are on either side of gliP ?

Left:

- Afu6g09650 (gliJ)
-Afu6g09640 (gliI)
-Afu6g09630 (gliZ)

Right:

- Afu6g09670 (gliC)
- Afu6g09680 (gliM)
-Afu6g09690 (gliG)

3.- 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 gliotoxin production, which is an immunosuppressive mycotoxin. Do you think any of the surrounding proteins is also involved?

Yes, all the surrounding proteins are involved. If we search the functions of the surrounding genes, we can see that all of them are involved in gliotoxin biosynthesis.

The GFF file

For most genomes there is no gene order information set in a database therefore we have to go to the original information.

1.- Go to NCBI and search for the genome page of Aspergillus flavus and download the GFF file.

Aspergillus flavus
Representative genome: Aspergillus flavus NRRL3357 (assembly JCVI-afl1-v2.0)
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Aspergillus flavus **genome**
All 56 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

2.- Uncompress the file and search the gliP protein (you will have to search for GliP).

3.- Which is the location of this gene in the Aspergillus flavus genome? (Provide the scaffold and start and end position of the gene).

AFLA_064560A;end_range=106889,.;gbkey=mRNA;locus_tag=AFLA_064560;partial=true;product=nonribosomal peptide synthase GliP-like%2C putative;start_range=.,101963;transcript_id=XM_002379975.1
NW_002477244.1 RefSeq exon 104322 106889 . - . ID=exon-XM_002379975.1-1;Parent=rna-XM_002379975.1;Dbxref=GeneID:7917588,Genbank:XM_002379975.1;Note=transcript
AFLA_064560A;end_range=106889,.;gbkey=mRNA;locus_tag=AFLA_064560;partial=true;product=nonribosomal peptide synthase GliP-like%2C putative;transcript_id=XM_002379975.1
NW_002477244.1 RefSeq exon 101963 104260 . - . ID=exon-XM_002379975.1-2;Parent=rna-XM_002379975.1;Dbxref=GeneID:7917588,Genbank:XM_002379975.1;N

Start position = 101963
End position = 106889
Scaffolds = NW_002477244.1, NW_002477244.1

4.- Are the genes we found in exercise A2 also close in the genome based on the annotation?

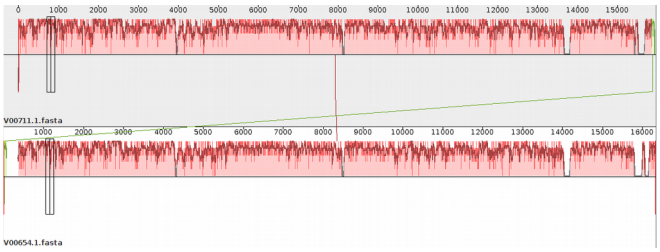
Yes, the gene found in exercise A2 are also close in the Aspergillus flavus. For example, GliC has start position: 99677 and end positions: 100034. However, gliZ is not found in the genome.

B. Genome Alignment (Mauve)

Mauve is a software package that attempts to align orthologous and xenologous regions among two or more genome sequences that have undergone both local and large-scale changes.

1.1.- Are the genomes conserved? Why?

Yes, the genomes are more or less conserved, because the structure is very similar. In the following image we can see them:



2..2.- Change the options View -> Style and add Solid LCB coloring. What can you say about the colors? Are these genomes conserved?

When we add Solid LCB coloring, a red big block is formed. At both genomes, we can find the same color and structure. Therefore, we can see that the genomes are conserved.

3.3.- What do you think the empty spaces represent?

The empty spaces represent that there is no homology found in the other genome.

2. Now download the mitochondrial genomes of two olive individuals: GenBank: MG372119.1 and GenBank: MG372117.1 . Align both genomes as before for the exercise 1.

2.1.- Are these genomes more or less conserved than the previous ones?

These genomes are less conserved than the previous ones. We can observe that the regions are not regular, some regions appear down, and also there are changes called structural variations.

2.2.- Change the options View -> Style and add Solid LCB coloring. What do you think the colors represent? Some regions appear up and down, what do you think this means?

Each color box represents a Locally Collinear Block a conserved segments that appear to be internally free from genome rearrangment. If the block appears up, it is on the forward strand of a genome, otherwise, it is on the reverse strand.

2.3.- In this exercise we have the mitochondrial genomes of two individuals that belong to the same species (very close related) and in the previous exercise two different species (mouse and cow, far related). Is this represented in the results? Can you give an explanation about why this happens.

No, this is not represented in the results. The mitochondrial genomes of olive individuals seems to be more different than the genomes of the mouse and cow that are far related. This could be because the two olive species can come from a very far away speciation that the two species evolved differently and, in consequence, they have suffred more changes.

2.4.- Which are the drawbacks of the Mauve tool?

As a genome aligner tool, it is clear that one of the main drawsbacks of MAUVE is the memory (because a genome is formed by a vast number of nucleotides) and the time consuming.

In addition, MAUVE can have difficulties to align genomes with high number of duplications. However, they had improved the MAUVE algorithm, with the Progressive Mauve Algorithm.

C.- How to check synteny (CoGe)

CoGe is a platform that allows you to compare genomes. 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.

1.- Go to the CoGe website (https://genomevolution.org/coge/) and search for the gene gliP of A. fumigatus (AFUA_3G12920). Select it and press "view details" that appears on the right side of the page. This will show you the details about the gene in this genome, but it also will show you links to the different CoGe applications.

2.- Run a CoGeBlast. Unlike other blasts, this time you will have to select the group of species you want to run the blast with.

And now run a tblastn within the CoGeBlast. The results will appear on the upper part of the page. Go there and write down:

- How many hits do we have per species?

HSP Count hide

Query Seq	Aspergillus flavus strain NRRL3357 (NCBI unmasked v2)	Aspergillus fumigatus strain Af293 (NCBI unmasked v1)	Aspergillus niger CBS 513.88 (NCBI unmasked v1)	Aspergillus oryzae strain RIB 40 (NCBI unmasked v1)	Aspergillus sojae strain NBRC 4239 (NCBI unmasked v1)
AFUA_3G12920 *(2243nt)*	20	20	20	20	20
Total	*191*	*21*	*160*	*151*	*206*

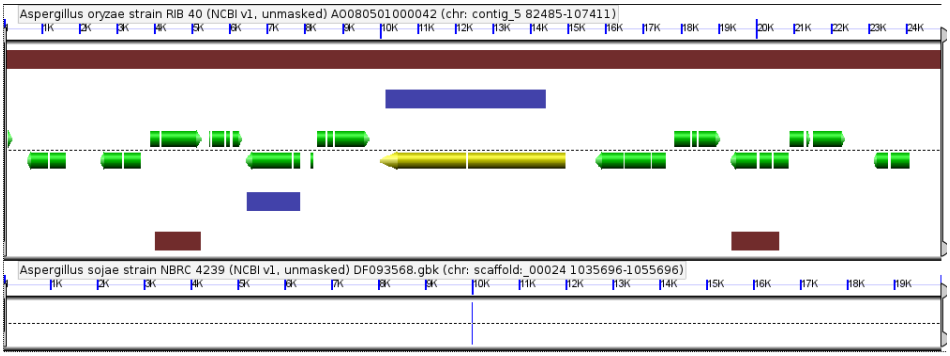
The total number of hits per species is shown in red. The hits with AFUA_3G12920 are shown in black above. The overall number of hits is 729.

- How many contigs have at least one homolog of gliP for each of the four species?

26 contigs have at least one homolog if gliP.

3.- From the list of homologs. Select the best hit for each of the five species and perform a GEvo analysis. GEvo will make an image similar to the one you found in the Gbrowser of the region where your gene of interest can be found. Your gene of interest is found in yellow. The image for Aspergillus sojae is different, why do you think that is?

It is different, because *Aspergillus sojae* doesn't seem to contain the gene of interest.



4.- 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 (v1.0.26, id25868) and Arabidopsis thaliana (vTAIR10.26, id25869) and press on Generate SynMap.

Do you think the gene order between these two genomes is conserved?

The gene order between these two genomes is more conserved in the initial part than the last region of the genome. However, we cannot say that the gene order is perfectly and totally conserved, as if they were we will find a straight linear line in the middle of the plot.

How do you interpret the graph? What do the blue lines mean?

The graph is called a syntenic dotplot, which allows us to compare the genomes of different species and identify synteny. It is useful to identify: insertions, deletions, inversions , duplications, translocations... In the x axis we can find Arabidopsis lyrata and the y axis the Arabidopsis thaliana.

The blue lines represents an homologous match between the two sequences.

What do you think the line means? (see the diagonal circled in red)

The line circled in red is an Inversion. This means that the sequence of Arabidopsis thaliana has these region of the genome inverted.

5.- Make a synMap between Mus musculus (house mouse) and Rattus norvegicus (Norway rat) Is the gene order more or less conserved than before? Why?

The gene order between the house mouse and the Norway rat is less conserved. In the dot-plot, we can observe a vast amount of evolutionary changes.

What do you think the two lines means? (see the diagonal circled in red)

These two lines shows that the chromosomes has suffered rearrengements. In this case, the chromosome 3 of *Mus musculus* is on the chromosome 3 of *Rattus norvegicus*.

6.- Finally, compare Aspergillus flavus strain NRRL3357 to the following species and rank them in order from more conserved to less conserved in terms of gene order:

1. Aspergillus oryzae strain RIB 40
2. Aspergillus niger strain CBS 513.88
3. Aspergillus nidulans strain FGSC A4
4. Aspergillus terreus strain NIH2624

D.- String

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

1.1.- Do all the proteins in this network interact directly with your protein of interest? What does it mean that so many proteins interact with your protein of interest?

All the proteins in the network interact directly with our protein. This fact can be observed by the node color, as we don’t have any white node, which ae the proteins associated with the proteins from the first shell of interactors.

If our protein of interest interact with so many proteins could mean that is involved in a relevent pathway (metabolism) or has an important role to the cell.

Now press the +more button on the lower right page.

1.2.- Do all proteins still interact with your protein of interest? What happened?

Now, not all proteins still interact with our protein of interesent. By pressing +more we have "expand" the view and we can see proteins that are a little bit far from our protein.

1.3.- Now press on the areA protein and recenter the network around this protein. Is this protein as well connected as the previous one? Keep expanding the network if you are not sure. The fact that areA interact with so many proteins means that this protein is more or less central for the metabolism of this species?

Yes, it is well connected as the previous one. It is more central for the metabolism of this species, when they interact with so many proteins.

1.4.- 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. Go back to the niaD network:
Use the two different clustering approaches on this dataset.

Do you obtain the same results?

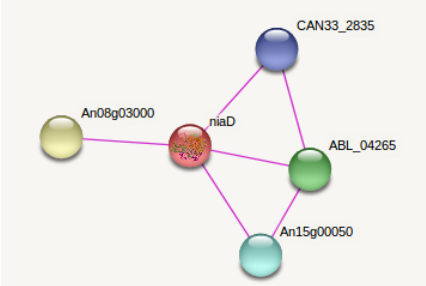
We don't obtain the same results, in K-means all the proteins have the same color.

Expand the network again (press 4 times +more) now apply the two clustering methods again. Which method do you think makes more sense and why.

It seems to make more sense the MCL cluster, as we obtain more different groups. In the kmeans cluster, we obtain three groups (red, green and blue), whereas in the MCL method we can find that some protein that are further are separeted in clusters.

1.5.1.- If you select only experimental evidence, how many proteins form the network?

By selecting only experimental evidence, there are only 5 proteins forming the network, including the protein of interest.



1.5.2.- Now add neighborhood evidence. How many proteins form the network now? What kind of evidence is neighborhood?

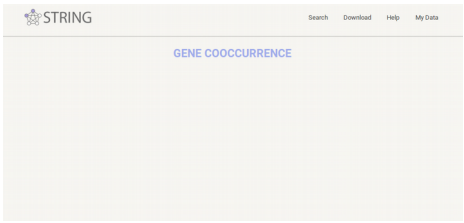
There are 6 proteins in the network now. In order to know what kind of evidence is neighborhood, I have search to the HELP webpage STRING and the neighborhood evidence is computed from the inter-gene nucleotide count.

1.5.3.- Now switch the lines from evidence to confidence. Which pair of proteins in this graph are more reliably joined to niaD ?

ABL_04265 and An08g03000 seems to be more reliably joined to niaD, for the line thickness that indicates the strength of data support.

1.6.- Search again for protein niaD , select the same species as before, go to viewers and select co-occurrence. This will show you a heatmap that indicates which species have similar interactions among the proteins shown in the network. As you can see, in Eukaryotes there is a strong correspondence between niaD and ABL_0465 and An08g03000 . Can you zoom into the tree and give an example of one species that have the strongest interaction?

The co-occurrence view doesn't seem to work:



1.7.- Now switch to the neighborhood view. This will show the gene order conservation for this gene. Is gene order conserved among prokaryotes? Is it among eukaryotes?

The gene order is conserved among prokaryotes as we can see in the image above. However in eukaryotes, the gene order is not conserved.