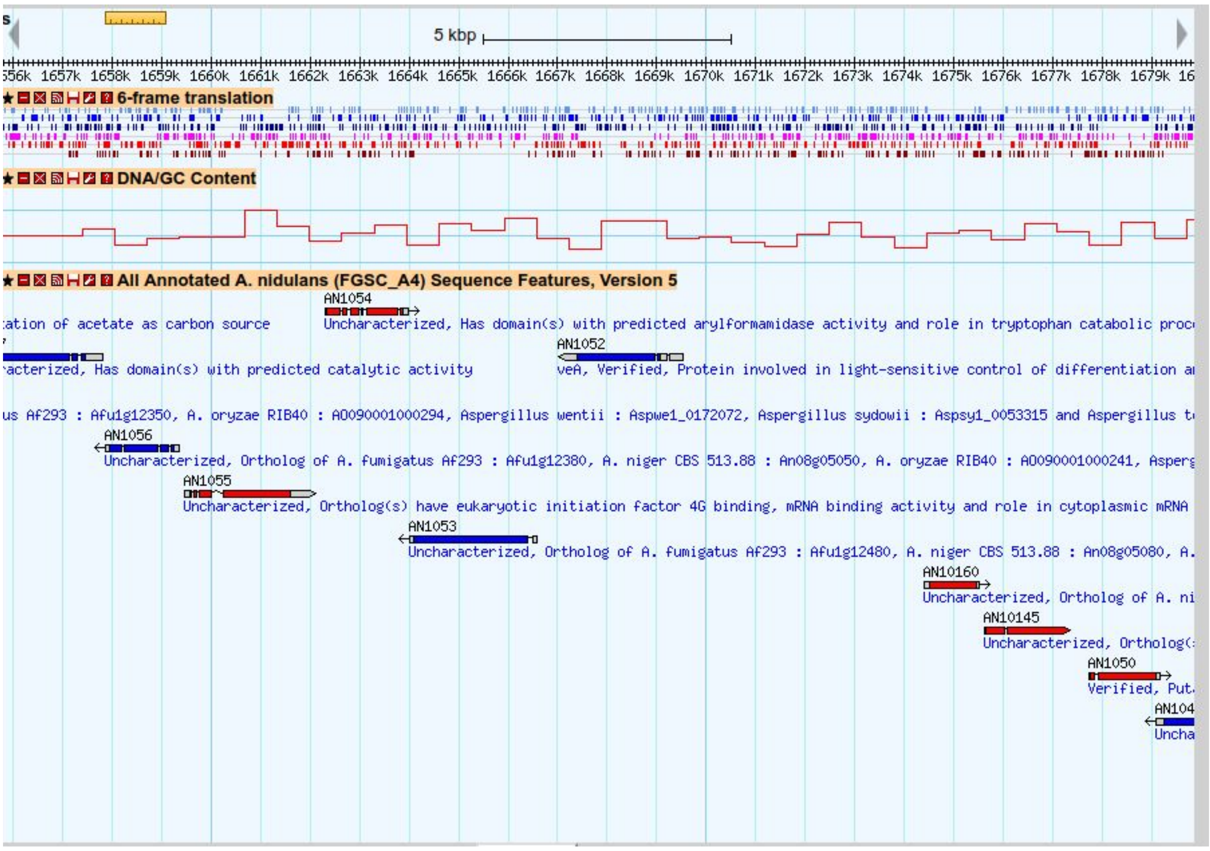
Session 6 - Gene order

A.- How to find where you 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.

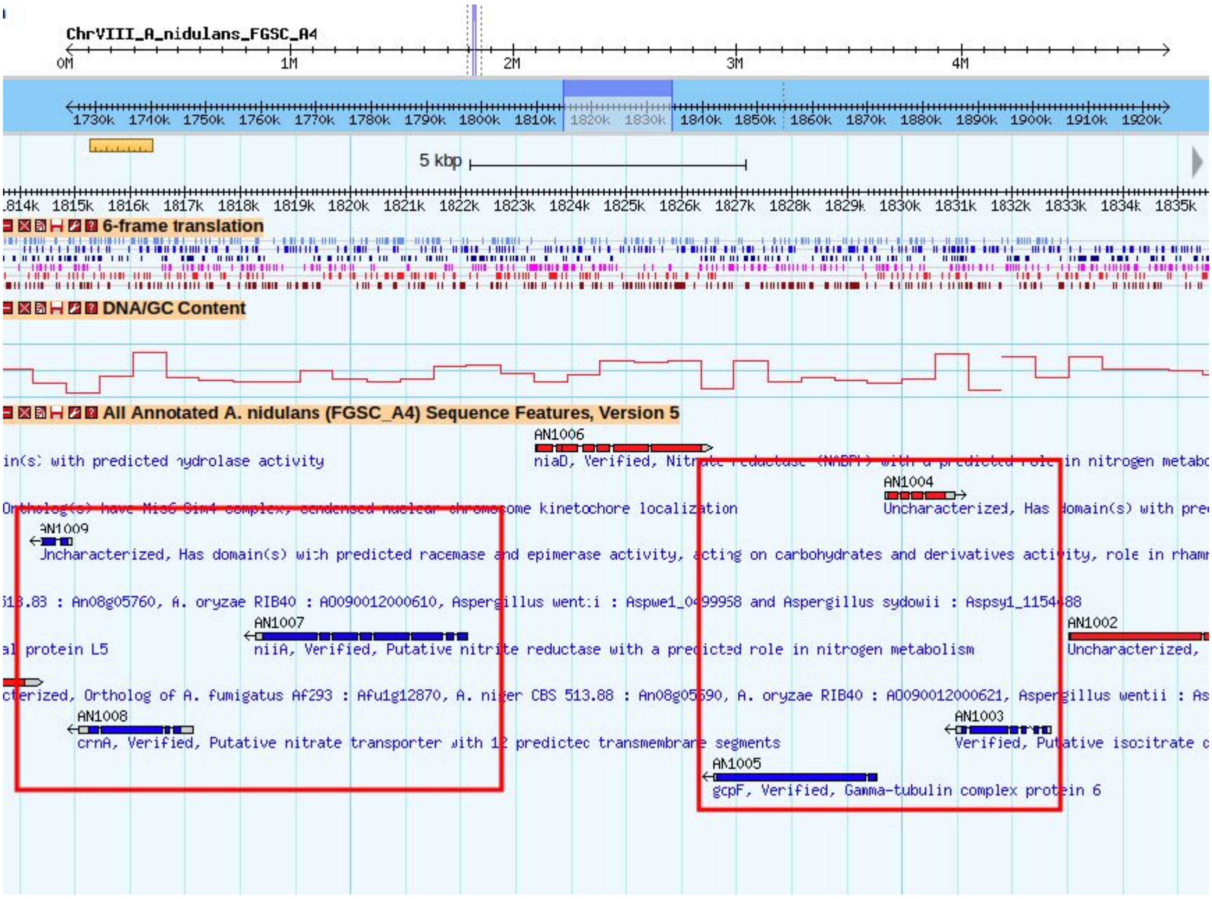
A tool that us often used to explore gene order are the genome browsers. They usually are only available for model species but their presence has been increasing over time. Go to the Aspergillus genome database (http://www.aspergillusgenome.org/). This database has been designed to hold all the information found for a few specific Aspergillus species. Once there go to the GBrowse of *A. nidulans*. The browser will open and you will see a region of the *A. nidulans* genome. The squares at the bottom of the image in blue and red indicate the exons:


5 kbp
1657k 1658k 1659k 1660k 1661k 1662k 1663k 1664k 1665k 1666k 1667k 1668k 1669k 1670k 1671k 1672k 1673k 1674k 1675k 1676k 1677k 1678k 1679k
6-frame translation
DNA/GC Content
All Annotated A. nidulans (FGSC_A4) Sequence Features, Version 5
AN1054
ation of acetate as carbon source
Uncharacterized, Has domain(s) with predicted arylformamidase activity and role in tryptophan catabolic proc
AN1052
acterized, Has domain(s) with predicted catalytic activity
veA, Verified, Protein involved in light-sensitive control of differentiation a
us Af293 : Afu1g12350, A. oryzae RIB40 : AO090001000294, Aspergillus wentii : Aspwe1_0172072, Aspergillus sydowii : Aspsy1_0053315 and Aspergillus t
AN1056
Uncharacterized, Ortholog of A. fumigatus Af293 : Afu1g12380, A. niger CBS 513.88 : An08g05050, A. oryzae RIB40 : AO090001000241, Asperg
AN1055
Uncharacterized, Ortholog(s) have eukaryotic initiation factor 4G binding, mRNA binding activity and role in cytoplasmic mRNA
AN1053
Uncharacterized, Ortholog of A. fumigatus Af293 : Afu1g12480, A. niger CBS 513.88 : An08g05080, A.
AN10160
Uncharacterized, Ortholog of A. ni
AN10145
Uncharacterized, Ortholog(
AN1050
Verified, Put
AN104
Uncha

You can search for a specific protein encoded in the genome by going to Reference position and inserting the name of the gene of interest.

1.- Search the browser for the niaD protein (Nitrate reductase). 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.

2.- Ask the browser to show you 20 kb around the gene of interest. Can you list the three proteins that are on either side of niaD?


ChrVIII_A_nidulans_FGSC_A4
0M
1M
2M
3M
4M
1730k 1740k 1750k 1760k 1770k 1780k 1790k 1800k 1810k 1820k 1830k 1840k 1850k 1860k 1870k 1880k 1890k 1900k 1910k 1920k
5 kbp
.814k 1815k 1816k 1817k 1818k 1819k 1820k 1821k 1822k 1823k 1824k 1825k 1826k 1827k 1828k 1829k 1830k 1831k 1832k 1833k 1834k 1835k
6-frame translation
DNA/GC Content
All Annotated A. nidulans (FGSC_A4) Sequence Features, Version 5
AN1006
in(s) with predicted hydrolase activity
niaD, Verified, Nitrate reductase (NADPH) with a predicted role in nitrogen metabo
AN1004
Ortholog(s) have Mis6-Sim4 complex, condensed nuclear chromosome kinetochore localization
Uncharacterized, Has domain(s) with pre
AN1009
Uncharacterized, Has domain(s) with predicted racemase and epimerase activity, acting on carbohydrates and derivatives activity, role in rhamn
513.83 : An08g05760, A. oryzae RIB40 : AO090012000610, Aspergillus wentii : Aspwe1_0499958 and Aspergillus sydowii : Aspsy1_1154488
AN1007
AN1002
al protein L5
niiA, Verified, Putative nitrite reductase with a predicted role in nitrogen metabolism
Uncharacterized,
cterized, Ortholog of A. fumigatus Af293 : Afu1g12870, A. niger CBS 513.88 : An08g05590, A. oryzae RIB40 : AO090012000621, Aspergillus wentii : As
AN1008
AN1003
crnA, Verified, Putative nitrate transporter with 12 predicted transmembrane segments
Verified, Putative isocitrate c
AN1005
gcpF, Verified, Gamma-tubulin complex protein 6

3.- niaD has an important role in nitrogen metabolism in *Aspergillus nidulans*. Do you think any of the surrounding proteins is also involved?

crnA → Nitrate transporter
niiA → Nitrite reductase

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 *Penicillium digitatum* and download the GFF file.
2.- Uncompress the file and search the niaD protein (you will have to search for Nitrate reductase).
3.- Which is the location of this gene in the *Penicillium digitatum* genome? (Provide the contig and start and end position of the gene).

NW_014574610.1 509941:512878

4.- Are the genes we found in exercise B3 also close in the genome based on the annotation?

Nitrite reductase: NW_014574610.1 505625:509320

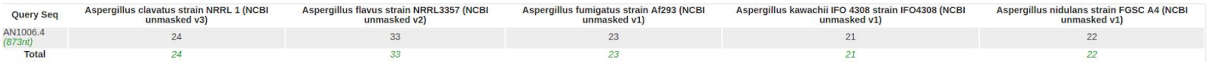
Now you know where your gene is located and the surrounding proteins. You can compare gene order across species either manually as we have done in the exercise or with a script. An alternate solution is to use tools that already exist, such as CoGe.

B.- 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 niaD. The first result of the search should belong to Aspergillus nidulans. 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. Select the following:
Aspergillus nidulans strain FGSC A4
Aspergillus clavatus strain NRRL 1
Aspergillus flavus strain NRRL3357
Aspergillus fumigatus strain Af293
Aspergillus kawachii IFO 4308 strain IFO4308
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?



Query Seq	Aspergillus clavatus strain NRRL 1 (NCBI unmasked v3)	Aspergillus flavus strain NRRL3357 (NCBI unmasked v2)	Aspergillus fumigatus strain Af293 (NCBI unmasked v1)	Aspergillus kawachii IFO 4308 strain IFO4308 (NCBI unmasked v1)	Aspergillus nidulans strain FGSC A4 (NCBI unmasked v1)
AN1006.4 *(873nt)*	24	33	23	21	22
Total	*24*	*33*	*23*	*21*	*22*

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

A. clavatus: 7
A. flavus: 11
A. fumigatus: 7
A. kawachii: 14
A. nidulans: 6

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 *A. kawachii* is different, why do you think that is?

Likely because it does not have a gene prediction therefore it's finding the homologous region thanks to the tblastn search even though proteins are not predicted.

The advantage of CoGe over other viewers is that you can upload your own data and use their tools to analyse your genome. To do that you only need to create a session, obtain the genome sequence and the gff file from NCBI and create an entry for your own use.

CoGe also allows us to compare genomes as a whole, not just focussing on one single gene.

4.- Go to Tools; SynMap. SynMap will compare two genomes, it is better when the genomes have CDS predicted so select the two species: *Aspergillus fumigatus strain A1163* and *Neosartorya fischeri strain NRRL 181* and press on Generate SynMap.

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

Yes, it is decently conserved

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

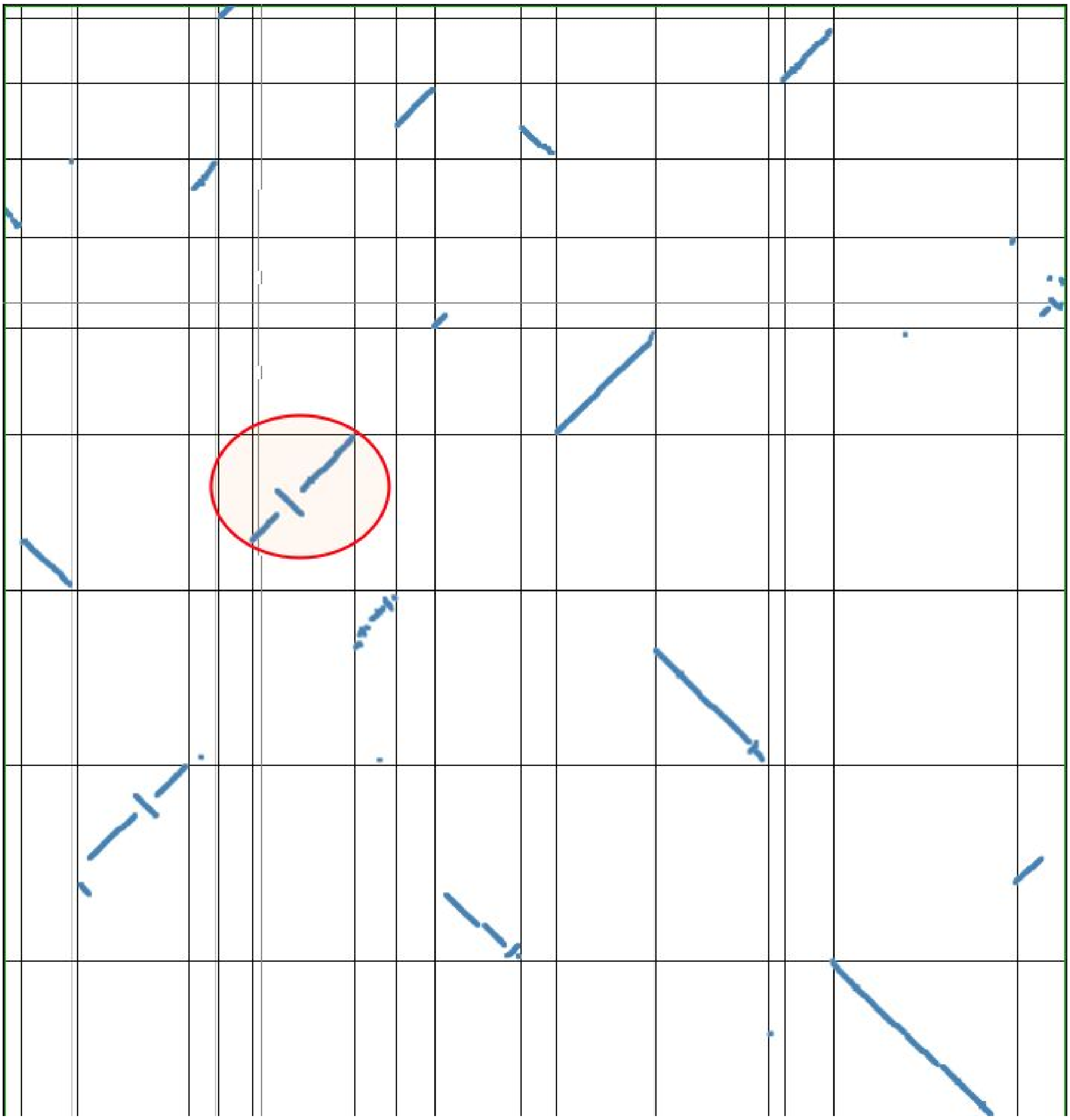
Each blue line represents homologous regions between the two genomes. A perfect match between two contigs is indicated by a diagonal in the square that represents the two homologous chromosomes. Any break in the diagonal can represent losses, duplications, inversions, translocations, etc...

5.- Make a synMap between *Aspergillus fumigatus strain A1163* and *Aspergillus clavatus strain NRRL 1*. Is the gene order more or less conserved than before? Why?

It is less well conserved. You can see that there are more breaks between these two genomes.

What do you think the broken line means? (see the diagonal circled in red)
It represents an inversion in one of the two genomes when compared to the other.

C, A, B



6.- Finally, compare *Aspergillus fumigatus strain A1163* to the following species and rank them in order from more conserved to less conserved in terms of gene order:

A. *Aspergillus nidulans*
B. *Penicillium marneffei strain ATCC18224*
C. *Aspergillus fumigatus strain Af293*

C, A, B

C.- String

String is an interesting webpage to know. Among other features it also contains information on gene order.
Go to the string database (https://string-db.org/) and search the protein pcbAB. As you can see it gives you five results in five different species, the two fungal species *Penicillium chrysogenum* and *Aspergillus flavus* and the three prokaryotes *Stigmatella aurantiaca*, *Streptomyces cattleya* and *Streptomyces clavuligerus*. First we will look at the prokaryotic sequences. Select the protein in *Streptomyces clavuligerus*.

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?

Yes. The more proteins that interact with your protein of interest, the more central its function is to the functioning of the organism.

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

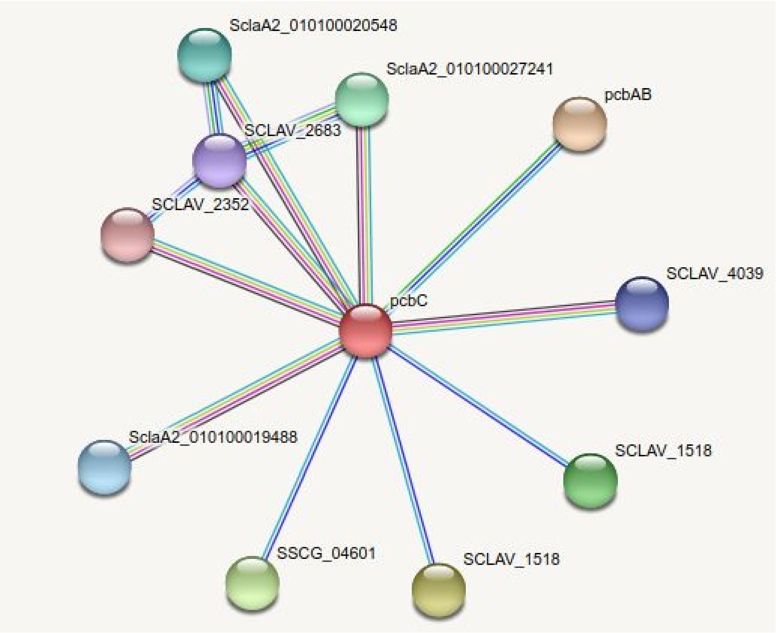
1.2.- Do all proteins still interact with your protein of interest? What happened?

No, the network was expanded to include proteins that were only indirectly interacting with our protein of interest.

1.3.- Now press on the pcbC 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 pcbC does not interact with so many proteins means that this protein is more or less central for the metabolism of this species?

It means it's less central

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.



Use the two different clustering approaches on this dataset. Do you obtain the same results?

Expand the network again (press 5 times +more) now apply the two clustering methods again. Which method do you think makes more sense and why.

The clustering approaches give different result. Kmeans forces a specific number of clusters to be present whereas mcl is not forced to such constraints. MCL is therefore the more suited algorithm of the two.

1.5.- Go back to the small 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.

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

1.5.2.- Now add neighborhood evidence. How many proteins form the network now? What kind of evidence is neighborhood?
Three. It indicates that the three proteins are found in different genomes together

1.5.3.- Now switch the lines from evidence to confidence. Which pair of proteins in this graph are more reliably joined?
pcbAB and pcbC have a thicker line therefore they are better connected.

1.6.- Search again for protein pcbAB, 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 pcbAB and pcbC. Can you zoom into the tree and say which is the species that has the strongest interaction?
Trichophyton rubrum

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?

Most proteins indicate that the orthology relationships are complex and therefore it is difficult to assess whether gene order is conserved or not. In general we do not see a big gene order conservation either in eukaryotes nor in prokaryotes.