GECKO-MGV: Web-based layered analysis of multiple genome comparisons including external post-processing services: Guided exercises

Version v1: 28th March 2019.

Developed by:
Sergio Diaz-Del-Pino
Samuel Benavides
Jose Arjona
Oscar Torreño
Oswaldo Trelles

Report incidences to:
ots@ac.uma.es

Contents:

Exercise 1: Diving into GECKO-MGV: Pairwise genome analysis mode

1.1 Introduction

In this exercise we will use the multiple genome comparison mode, in which we first perform simple tasks such as filtering and zooming; second we identify and extract a collection of repeated fragments containing the "transposase" annotation; third, the sequences will be retrieved (forward or reverse complementary sequences are retrieved -depending on the "strand" field in the fragment data- from their corresponding forward or reverse-complementary full genome sequence. These sequences are stored in a file. Both services, sequence retrieval and obtaining the reverse-complementary sequence are internal. All these sequences will be used in a fourth step to be aligned with the external service ClustalW at the EBI, and finally the multiple sequence alignment will be displayed.

1.1.a Load comparison results and perform simple actions.

We can start loading all the fragments files as local files previously calculated, but files stored in the server could be used as well. After the loading process, the application displays the different views and layers. An Horizontal and vertical view per input file is generated, and a map with the active layers is shown. At this point, the data analyst can interact with the comparisons by zooming, filtering, searching for annotations, etc.

1.1.b. Searching for functional annotations.

Next step in this exercise is to search for "Transposase" annotation associated with genome rearrangements. We proceed simply by using the annotation-search engine. Results are displayed in the main canvas in the positions (both genomes) where the annotation occurs. Noteworthy observe, the searching task is performed over all active datasets.

1.1.c Fragments selection and sequences retrieval

Just in this exercise we can observe the transposase functional annotation appears in several positions in sequence J (MH-J) that could belong to repetitions. We select some of the fragments containing the annotation. In this case we simply use the zoom and selection functionality. The coordinates of such fragments can be stored and used as

input to an internal service to retrieve the sequences defined by the fragments coordinates from the full genome sequence. Resulting multi-fasta file is stored in the local machine. This retrieval sequence is prepared to identify the coding direction of the sequence (forward or reverse DNA strain) and call another internal service to obtain the reverse complementary from the forward sequence.

1.1.d Invoking external services

Next, we want to launch a multiple sequence alignment to be performed over the set of retrieved sequences. ClustaW is the service we will use. This service is available at EBI as a Web-Service and can be invoked from our framework, including the sequences file and parameter. Results from such alignment and displayed in a new tab of the browser

1.2 Exercise development

1. - Enter to http://pistacho.ac.uma.es

2. - Once the application is loaded, we proceed to login as registered user. Clicking the 'Sign in' button a dropdown menu will be showed to introduce our login information. In this case:

User: guest
Pass: guest

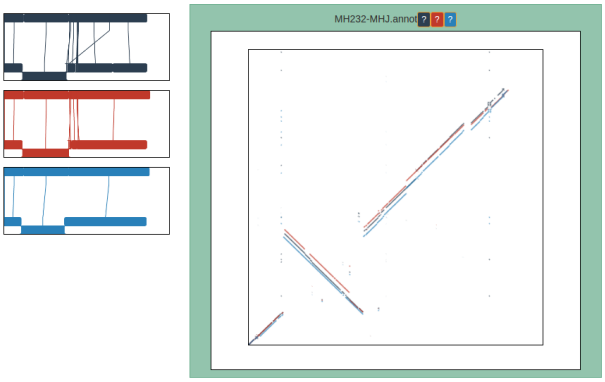
2. - As registered user we proceed to load our comparisons obtained by the Gecko Workflow:

- Mycoplasma hyopneumoniae 232
- Mycoplasma hyopneumoniae 7422
- Mycoplasma hyopneumoniae 7448
- Mycoplasma hyopneumoniae J

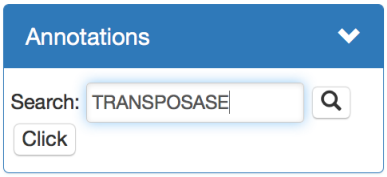
So we click in the 'Load frags from local' icon, go through our file system to the folder where the comparison files are in CSV format and click in 'Open'.

Alternatively we can directly open them from server using the 'Load frags from server' icon, those files have been placed in the user's file system in advance.

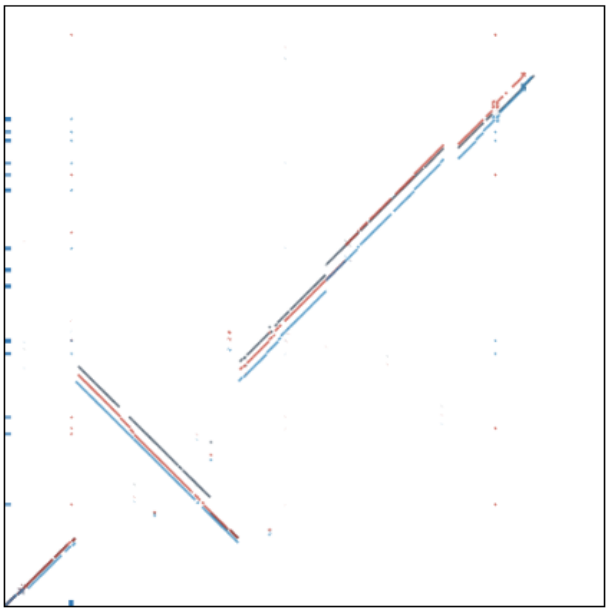
3. - After loading the comparisons, this should be our main views:



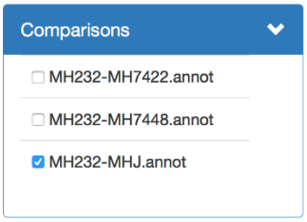
4. - Now, in the ‘Annotations’ menu, we are going to search for ‘TRANSPOSASE’ in the annotations.



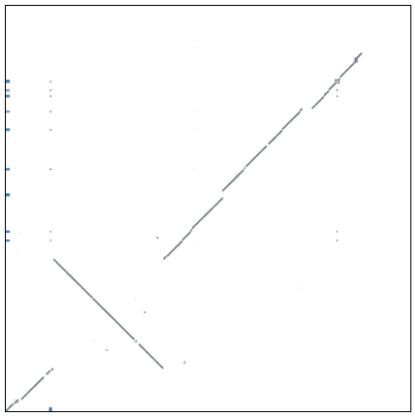
5. - Gene-annotations related to ‘TRANSPOSASE’ are painted in the main canvas.



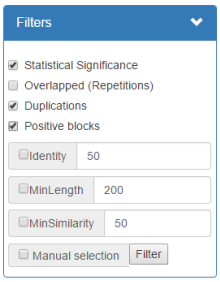
6. - Those annotations can be consulted in the modal menu with the gene-annotation information after clicking in the ‘Annotations’ button (A) of the top menu.
7. - For this first exercise we are going to work only with the MH232-MHJ in order to make the exercise easier to follow but it can be used with any other. So, in the right panel, in the ‘Comparisons’ menu, we keep active only that comparison.



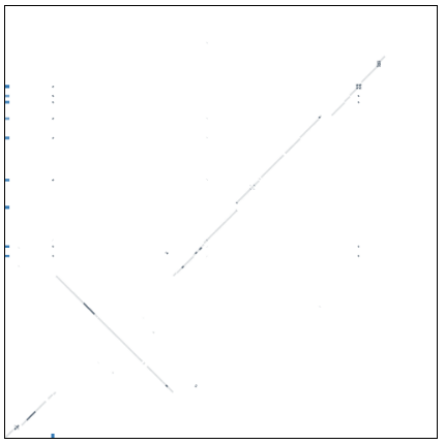
8. - Our main canvas should look this way now:



9. - We are going to filter some data. We are goin to apply the ‘‘Statistical significance’, ‘Duplications’ y ‘Positive blocks’ filters. We should have now a cleaner data representation, so we can focus in what we are looking for.

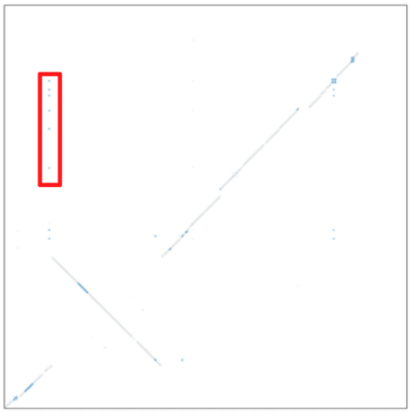


Our main canvas should look like this:

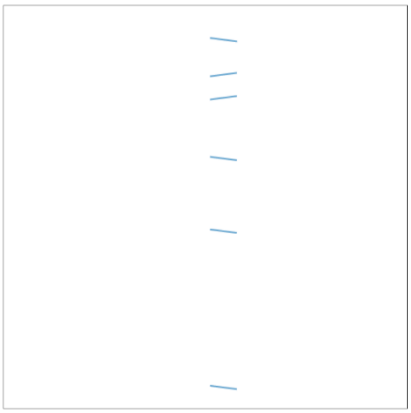


10. - Once filtered our comparison and seen where we can find the transposase, we are going to zoom the upleft corner, in which we detected some repetitions and we located some transposase related genes before.

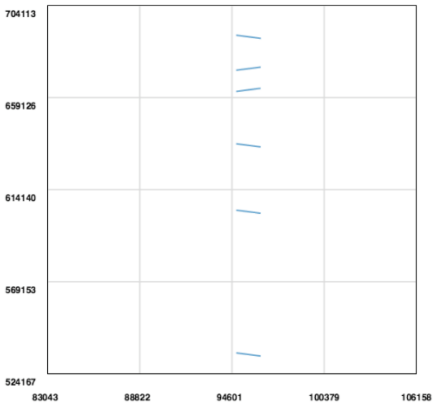
For zooming we proceed to draw a small cuadrangle, by clicking with the left click of the mouse and dragging in until all the repetitions were inside the drawed area.



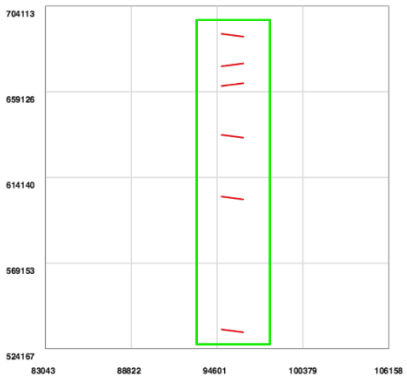
This will take us to:



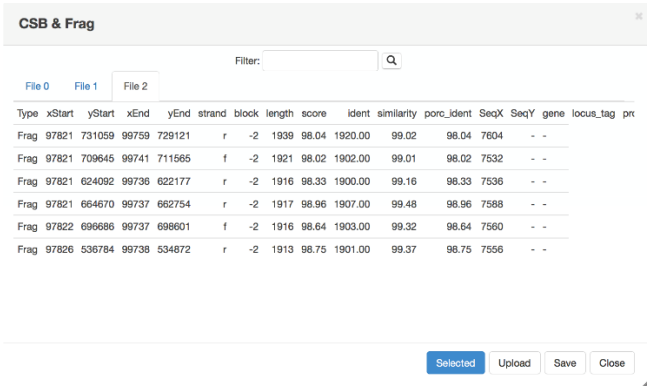
11. -Now, we are going to activate the grid, just to have a better perception of where we are in the genome. We click in the 'Grid' ( ) button in the top menu. As result, the grid is shown and we can now see more accurately the position of our repetitions.



12. - We are going to select theses fragments represented in the screen holding the ‘Shift’ key in our keyboard and repeating the same process that we followed when we zoom before. This time the fragments will be painted in red, as a signal of they are selected.



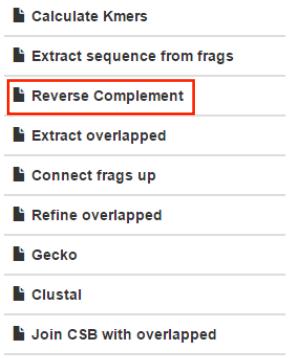
13. - To store the selection we should click in the ‘CSB & Frag info’ () button. This will show a modal menu with the information of all the fragments in the comparison. After clicking ‘Selected’ this information will be only related to the selected fragments.



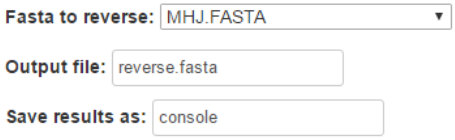
14. - To store that selection in server side we just click in ‘Upload’.

15. – Now it is important to have our FASTA files in server side. So first we are going to the ‘File Manager’ view and we are going to upload them by clicking in the ‘upload file’ button and selecting it.

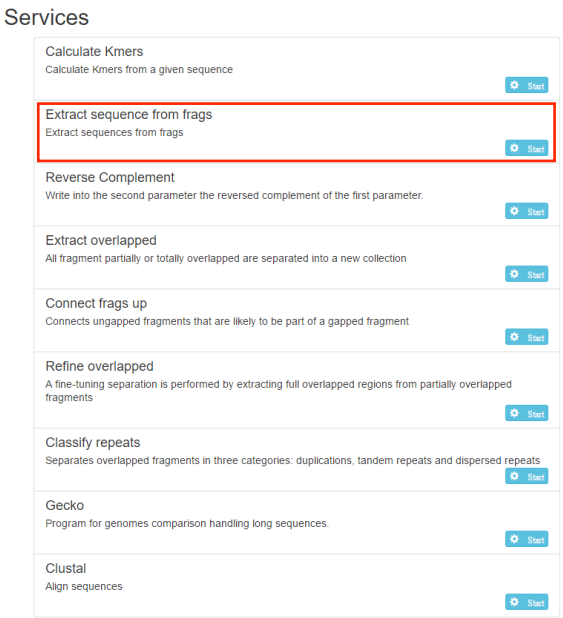
16. -The first step should be to obtain the reverse complementary from the Y sequence, in this case ‘MHJ.FASTA’, for that we will use the reverse complement service placed on the  button althouhgt it can be also found at the ‘Services’ tab. and we click on it.



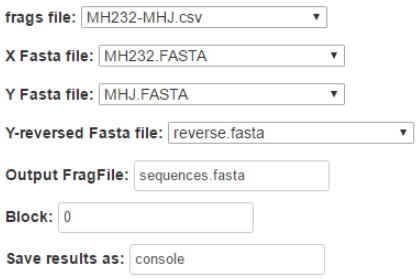
17. - We must fill the required fields. In this exercise we will use the following parameters, and then we click on the Submit button.



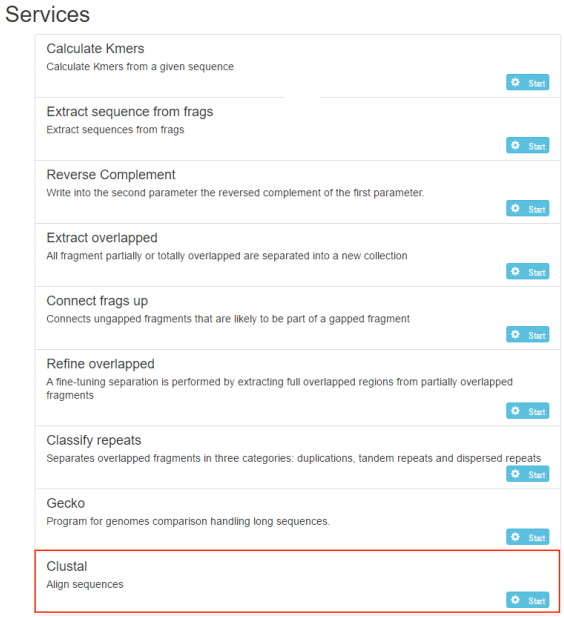
18. - Now we should open a new tab in order to keep saved our workspace, there go to pistachio.ac.uma.es again and then access to the 'Services' tab in our main menu. There we will find the list of avaliable services. We are going to use the 'Extract sequence from frags', which, from a given group of frags, retrieves its sequences.



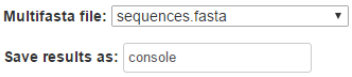
19. - We must fill the required fields. In this exercise we will use the following parameters, and then we click on the Submit button.



20. - Now we are aligning this sequences using a third party service from EBI. Again in the service list we will run 'Clustal Omega' with our generated file.



21. - In this exercise we will use the following parameters, and then we click on the Submit button.



22. - The results will be shown automatically in a different tab using the MSA-Viewer from BioJS.

Exercise 2

2.1 Introduction

This exercise illustrates the execution of two external services aimed at automatically detecting repeated regions between two genomes. A second aspect of this exercise is to illustrate the concept of layers. Layers might be helpful to compare results from different executions. In this exercise, we will use the layer concept and different services to compare specific fragments which could be interesting for the user.

The execution result will be showed in a new layer, becoming part of the session data. Both layers will be displayed then in the main canvas. Using the first layer as a guide, repetitions in the second layer located at the same place where “transposase” was found will be selected. Then, the sequences will be extracted and aligned using annotated regions that were detected in the exercise 1. In this exercise, annotations of one genome will not be available.

2.1.a Using the web interface to calculate the comparison

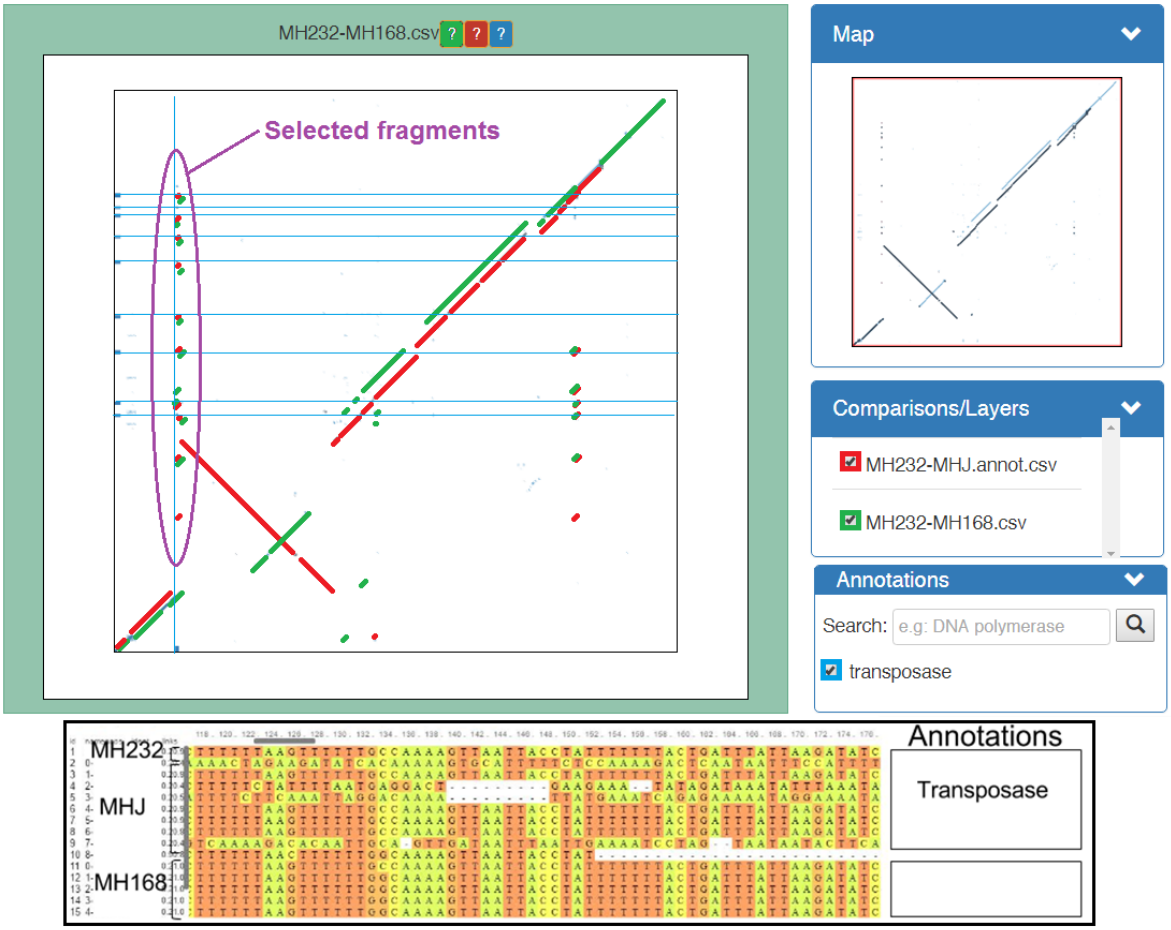
This second exercise starts with the execution of GECKO to compare Mycoplasmas MH-232 (NC_006360.1) and MH-168 (NC_021831.1) through the web interface developed in GECKO-MGV. The result produced by executing GECKO is added as a new layer. As a second step, we run the GECKO-CSB service to calculate repeated regions and the Computational Synteny Blocks (CSBs), which will be also loaded in new layers. Note that both repeated regions and CSBs do not have annotated information.

2.1.b Using layers to compare results from different executions.

In the next step, the exercise focuses on the usage of layers. In this case, we want to compare repetitions detected in the first exercise, located in layer 1 and repetitions detected previously, located in layer 2.

Repetitions in the second layer do not have any annotations because they were not provided. However, we can observe that both comparisons (layer 1 and 2) have repetitions at the same place where “transposase” was found (see Figure below). Using the custom service “extract sequences from Frags”, sequences from repetitions in both layers are extracted. Then, using an external service “ClustalW” these sequences are aligned, using the sequence X as a reference.

The result of such alignment reveals high similarities at the annotated region (see Figure 4 at bottom). This would imply that the unannotated region (extracted from results in layer 2) could share the same functionality than the annotated region (from layer 1) and therefore, the same annotation.



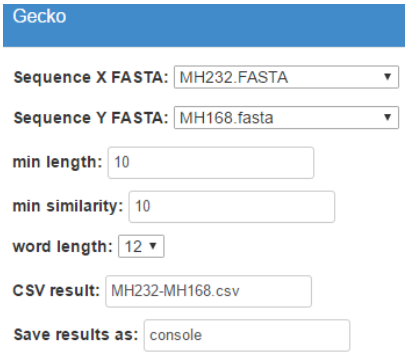
Representation of two comparisons. First layer (red) obtained in exercise 3.1. Second layer (green) obtained in this exercise. Blue lines represent the place where transposase was found in the first layer. Purple circle shows the fragments which have been selected in the exercise to be aligned. At bottom, representation of the obtained multiple sequence alignment. It is worth noting that although for the new compared sequence (i.e. MH168) there is no annotation, the good alignment suggests that this region of MH168 could potentially share the same annotation.

2.2 Exercise development

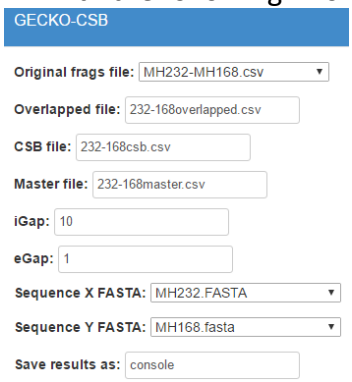
1. - If we followed correctly the instructions described on the previous exercise we only have to uncheck the other two layers, but if instructions wasn't followed or just want to do the second exercise then the first steep will be to load the comparison MH232-MHJ, as mentioned in the previous exercise, in order to do that we open them from server using icon, the file has been placed in the user's file system in advance.

2. - Directly from the main page we will click on the button and in the GECKO service

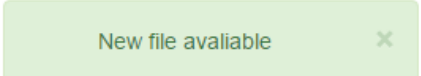
3. - We must fill the required fields using the following parameters and clicking on the Submit button.



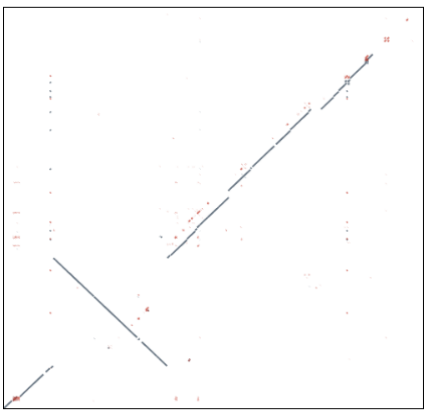
4. – We will repeat the process but this time clicking on the GECKO-CSB service instead of Gecko and we will fill the form with the following info:



5. - After the notification on the right up corner has appeared we proceed to load the result file named previously as '232-168overlapped.csv' file from the server by button and then clicking on the file.

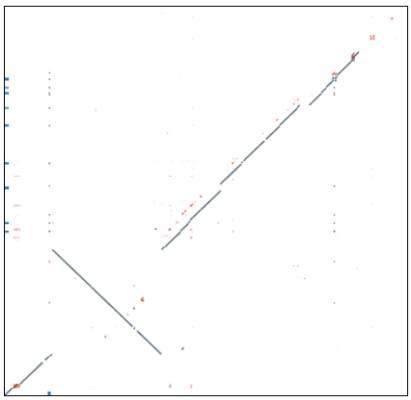


Then we should see something like this:

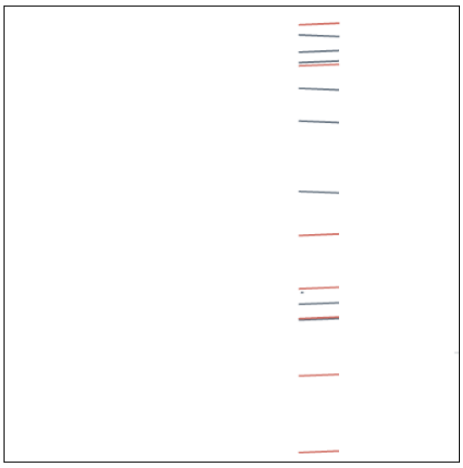


6. - We proceed to search into the annotations the transposase enzyme, to do that we will click the A button, meaning annotations info where we will write into the text box next to the search label the word transposase. After pressing the 'enter' button or clicking on the 🔍 button. Alternatively we can do the same from the general info panel or the 'Annotations' panel placed on the right down corner.

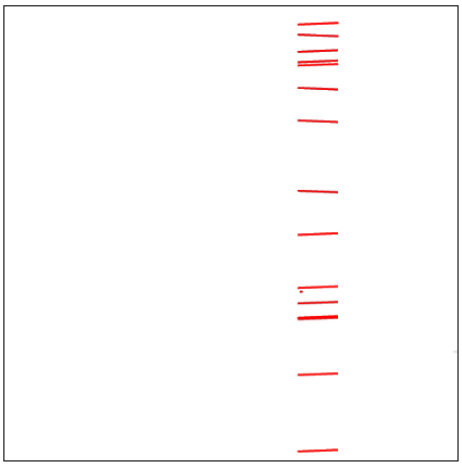
A few marks will appear on the main canvas indicating where on the x and Y axis is located that annotation. Also a checkbox will appear on the 'Annotations' panel to activate one annotation.
Please note that if few words are used they will appear few checkboxes and it will work with 'and' logic.
After this is done we should see something like this:



7. - We proceed to zoom the marked area by clicking and dragging the mouse over it so we should see something like this:



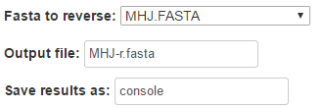
8. - We are going to save both selections in order to align them later, in order to do that we must firstly select all the fragments shown on the main view by clicking, dragging and holding 'Shift' button, the result should be similar to this:



9. - Next we will go to the info menu by clicking on the  button and then we will click on the  button so there will be displayed only those we have selected.

10. - Then we will click on the  button, and it will create automatically the files on the user's file system, displaying a notification on the right up corner.

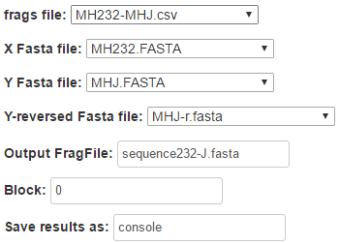
11. - The next step is to get the reverse complementary of our two Y sequences (MHJ and MH168) to get them we recommend to open a new tab to keep our workspace saved, there go to pistachio.ac.uma.es again and there click on the 'Services' tab and then clicking on the  button next to the 'Reverse complement' label and fill the form as shown on the image.



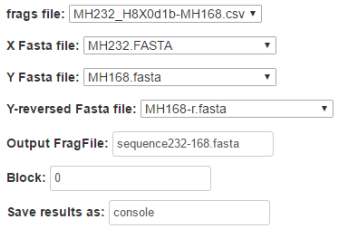
Then we click on the 'submit' button.

12. - We repeat the previous step but this time filling the form with the data from MH168.

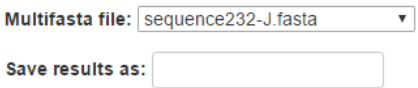
13. - We proceed to extract the sequences. We will do this by clicking on the 'Services' tab and then clicking on the Start button next to the 'Extract sequence from frags' label and fill the form as shown on the image. Then we will click on the 'Submit' button.



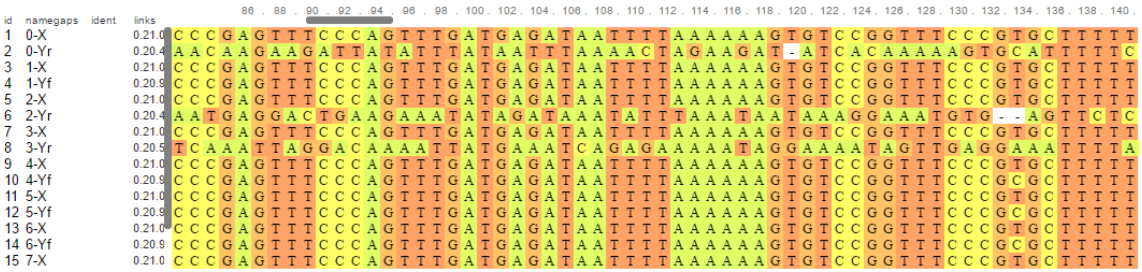
14. – We repeat the previous step but this time filling the form with the data from MH232-MH168.



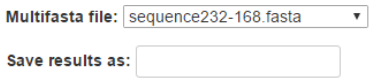
15. - To visualize the alignment of the sequences we must firstly go to the 'Services' tab and then click on the Start button next to the 'Clustal' label and fill the form as shown on the image. Then we click on the 'Submit' button.



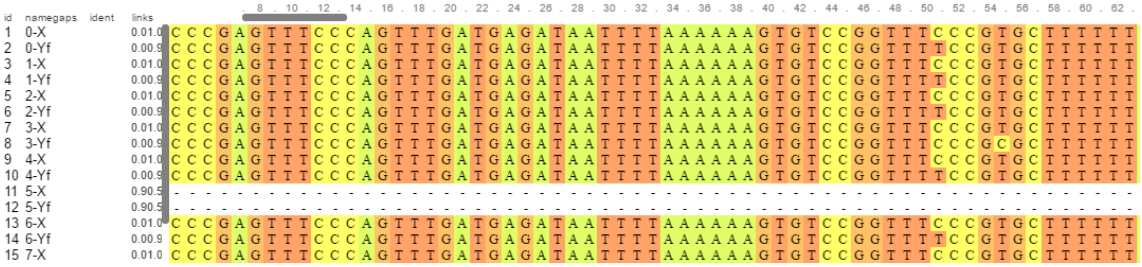
After a while the visualization of the alignment will be loaded, this could take several seconds as it is an external service, it normally takes around 20 seconds.



16. – We repeat the previous step but this time filling the form with the data from MH232-MH168.



This time the result obtained is the following.



The result of such alignment reveals high similarities at the annotated region. This would imply that the unannotated region (extracted from results in layer 2) could share the same functionality than the annotated region (from layer 1) and therefore, the same annotation.

Exercise 3

3.1 Introduction

This exercise is aimed to understand the general usage of the services area on the GECKO-MGV web tool.

During this exercise we will start by the creation of the comparison using the GECKO service and we will follow a workflow of services allocated in our web tool in order to join the fragments obtained into Computational Synteny Blocks (CSBs) integrating in them the overlapped fragments.

The workflow followed will be the following:

As a result we will obtain a set of CSB that will have included into them all overlapped fragments making the visualization more user friendly while keeping on it most of the information from the original comparison.

Please, submit any recommendation or suggestions directly to ots@ac.uma.es

3.2 Exercise development

1. - Enter to http://pistacho.ac.uma.es

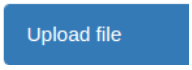
2. - Once the application is loaded, we proceed to login as registered user. Clicking the 'Sign in' button a dropdown menu will be showed to introduce our login information. In this case:

User: guest
Pass: guest

3. - As registered user we proceed to load our sequences on the file manager tab:

- NC_014760.1.fasta
- NC_013948.1.fasta

The sequences should be already on the user's file system so we can avoid the next step.

So we click in the 'Upload file' 

icon placed on the top part of the 'File manager tab' , click on choose file, go through our file system to the folder where the sequence files are in FASTA format and click in 'Upload' on each of them.

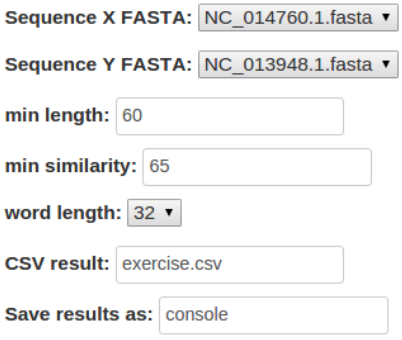
Alternatively we can copy and paste our sequence on the 'Content' field clicking on the 

button on the top right corner of the 'File manager' tab and then giving some name to the file and clicking on 'Create'.

4. - After having the sequences on the user's file system we click on the 'Home' tab and then on the  button although we can do it also by clicking on the 'Services' tab on the top of the webpage.

5. - One of the last services on the list must be GECKO (a tool to compare genomes), there we will click on the 'Gecko' label.

6. - We must fill the required fields using the following parameters.

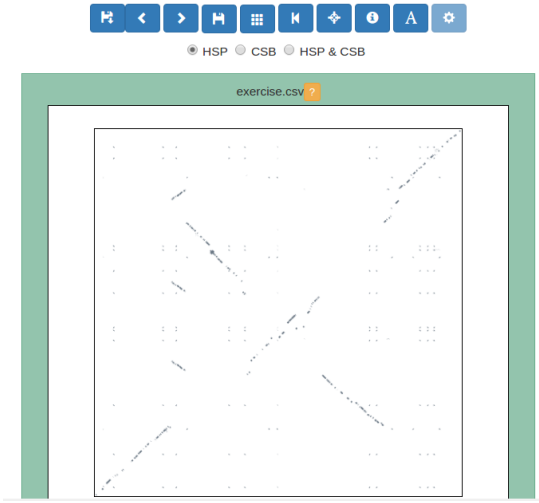


7. - Once we have all the fields filled we click on the Submit button and after a few seconds we will obtain the results.

8. - Now we will click again on ‘File manager’ to see the result of the comparison.

9. - To see the file we click on the Open button next to the ‘exercise.csv’ label.

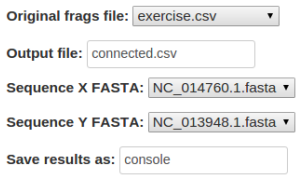
10. - Then we click on the ‘Home’ tab and then on the button on the upper part and then we click on the ‘exercise.csv’ file displayed on the floating window.
Then we should see something like this:



11. - We click again on the button, there we will go this time to ‘Connect frags up’ service.

12. - We cilck on the ‘Connect frags up’ label.

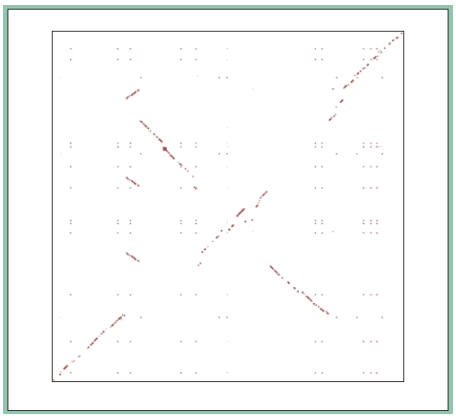
13. - We must fill the required fields, in this exercise we will use the following parameters.



14. - Once we have all the fields filled we click on the Submit button and after a while we will obtain the results and the console log will be displayed.

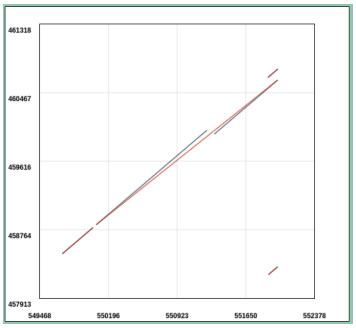
15. - Then we click on the ‘Home’ tab and then on the button on the upper part and then we click on the ‘exercise.csv’ file displayed on the floating window.

16. - Then we click again on the button and this time we click on the ‘connected.csv’ file displayed on the floating window.
Then we should see something like this.



17. - To appreciate better what we have done we will active the grid by clicking on the  button.

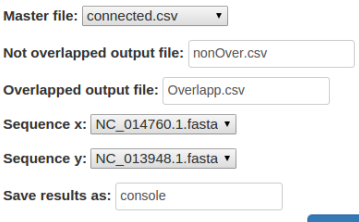
18. - Then we will zoom by clicking and dragging and then releasing in several steps untill visualize to the coordenates on order of (550920, 459615). There we will see somethig like this



19. - We click once more on the  button, there we will go this time to ‘Extract overlapped’ service.

20. - We cilck on the ‘Extract overlapped’ label.

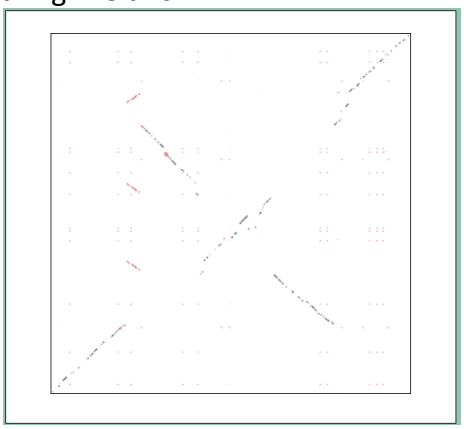
21. - We must fill the required fields, in this exercise we will use the following parameters.



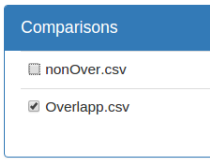
22. - Once we have all the fields filled we click on the  button and after a while we will obtain the results and the console log will be displayed.

23. - Then we click on the ‘Home’ tab and then on the  button on the upper part and then we click on the ‘nonOver.csv.csv’ file displayed on the floating window.

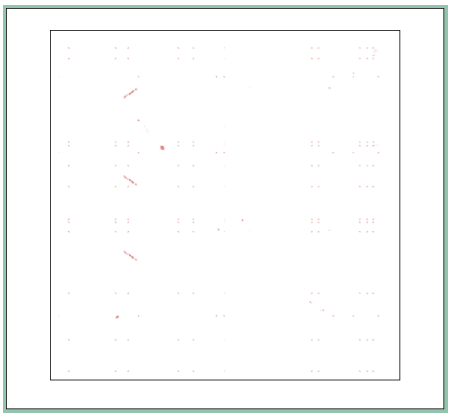
24. - Then we click again on the  button and this time we click on the ‘Overlapp.csv’ file displayed on the floating window.
Then we should see something like this.



25. - To see only the overlapped fragments we will uncheck the ‘nonOver.csv’ layer on the comparisons panel placed on the right side.



Then we should see something like this.

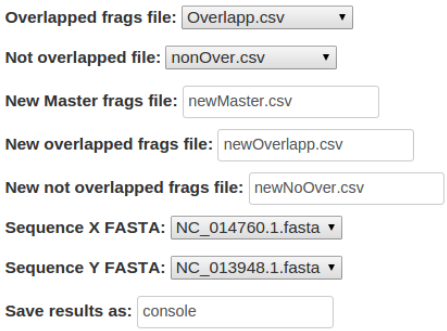


26. - We click once more on the ‘Services’ tab, there we will go this time to ‘Refine overlapped’ service.

27. - We cilck on the 

button next to the ‘Refine overlapped’ label.

28. - We must fill the required fields, in this exercise we will use the following parameters.

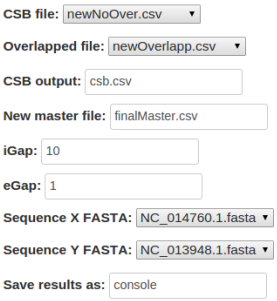


29. - Once we have all the fields filled we click on the Submit button and after a while we will obtain the results and the console log will be displayed.

30. - We click once more on the 'Services' tab, there we will go this time to 'Join CSB with overlapped' service.

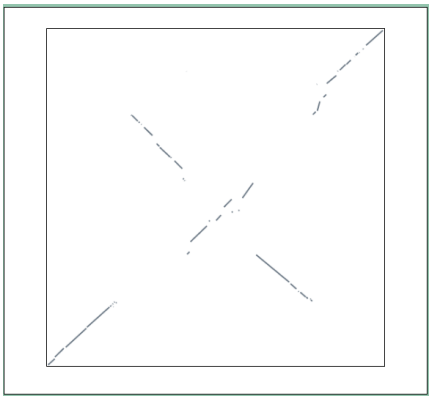
31. - We cilck on the Start button next to the 'Join CSB with overlapped' label.

32. - We must fill the required fields, in this exercise we will use the following parameters.



33. - Once we have all the fields filled we click on the Submit button and after a while we will obtain the results and the console log will be displayed.

34. - Then we click on the 'Home' tab and then on the button on the upper part and then we click on the 'csb.csv.csv' file displayed on the floating window.
Then we should see something like this.



We can change between HSP, CSB or both by clicking on 

Exercise 4

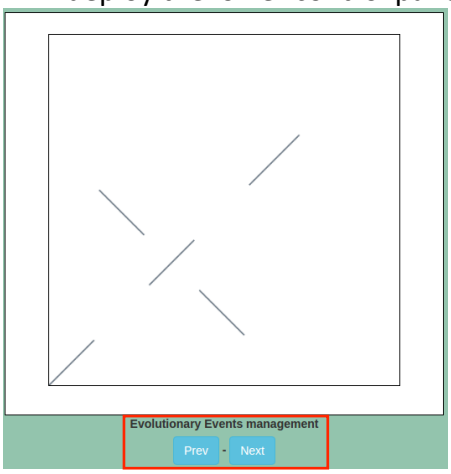
4.1 Introduction

In this exercise we are going to learn how to use the evolutionary events panel, allowing us to navigate through all the evolutionary events occurred on a genome.

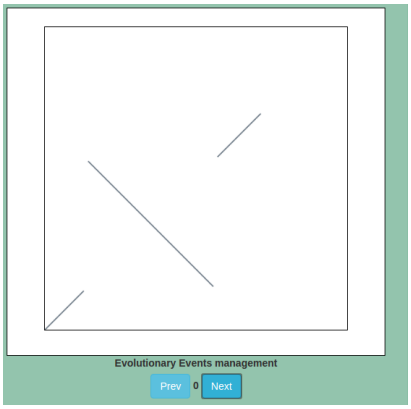
4.2 Exercise development

1. - Being on the 'Home' tab we will proceed to load our Evolutionary Events file, it can be found on the user's file system and can be loaded directly from the server by clicking on the  button and then clicking on the file 'ExampleEE.csv'

2. - Automatically after loading the file the software will detect by its own that it is an Evolutionary Event file and will deploy the lower control panel.



3. - We click on the 'next' button to go back on the time line showing the result after a short animation.



4. - Between the 'previous' and 'next' button we can find the stage where we can find the genome starting by '–' continuing by '0'

Annex 1:

Data Structure

The main data structure is **Frags**, which contains the information relative to a given fragments. Here we do not mention the additional associated fields that need to be declared to facilitate data management; i.e. “number of fragments”
The following are files used during experiments:

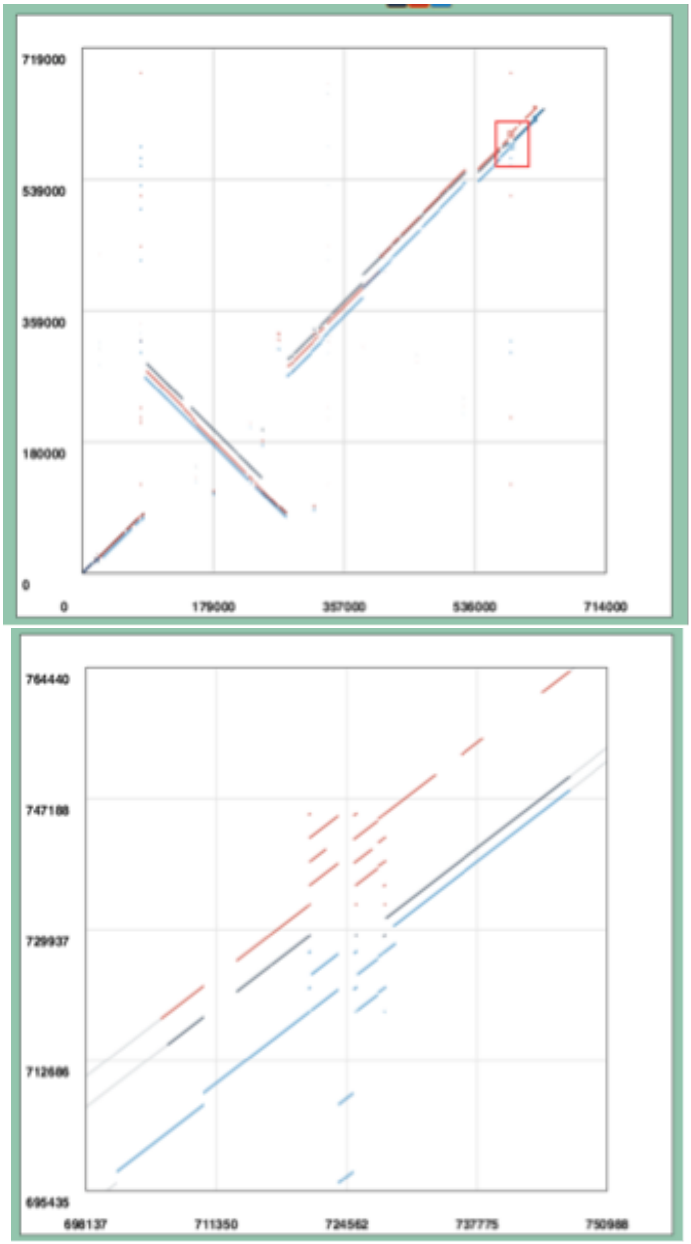
- Input files:
  - Sequence files: in FASTA format
  - Annotations file: in GFF format (optional)
  - Score matrix (inter residues scores) --optional
- Intermediate files:
  - Dictionaries: contains k=32-mers and its repetitions. From these files is possible to obtain repeats, tandem repeats, word frequencies, words appearing over and below the expected values, etc.
  - Hits: occurrences of the same word in two sequences
  - FrequencyKmer: K-mer frequency (k=1) to obtain karlin parameters.
  - Karpar: Karlin parameters (to calculate p-value)
- Output files:
  - Fragments file (binary) contains the description and coordinates for each fragment. Usually combines forward and reverse-complementary frags.
  - *.INF file: metadata information about the procedure to obtain the fragments files (name of the genomes, sequences, parameters, etc.)
  - Fragments (CSV). Combines both fragments and INF file into a readable CVS file. This file is much easier to manage, since it can be post-processed using other edition tools (i.e. a simple spreadsheet editor); and could be extended with other information for each fragment
  - Extended frags: CSB-Master format

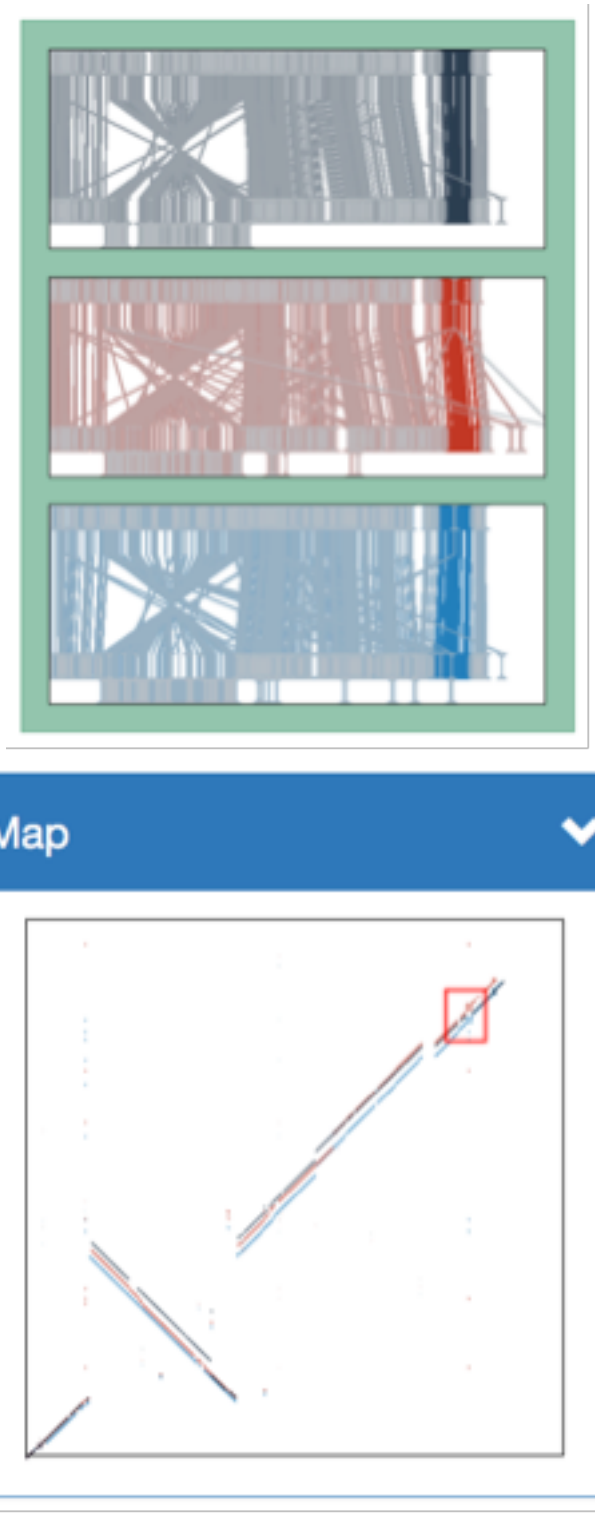
Annex 2:

1 Interacting with the comparison

1.1 Zoom-in and zoom-out

Zooming interaction is done in the main canvas. At first a quadrangle must be drawn while clicking with the left button of the mouse over the zone we want to zoom. After releasing the left button zoom is processed and all the views are updated. Two buttons to go back and go forward have been implemented to track every step in the analysis.





Once we have drawn the square (1) zoom is reflected in the main canvas (2) and the auxiliary views are updated (3) to track the information we are viewing.

1.2 Selecting and post-processing

Once the comparison files have been loaded into the system frags and CSB can be selected to retrieve information or to use them as input data for services. This post-processing services are customizable, but some of them, like the one to get repetitions, retrieve sequences for those frags, align frags… are provided by default.

The selection of the frags is done by using ‘shift’ key at the same time that a quadrangle is drawled, like in the zoom case. This time, frags inside the quadrangle are represented in red in both views and its information can be seen in a modal menu in the interface.
It is possible not to select areas but select just one frag by clicking on it while pressing ‘Shift’. This will add the fragment to the selection in case the user has already selected some of them.
Once a group of fragments have been selected the user can invoke services with those fragments and the results will be sent back to the application and it will be showed in a new layer.