

GECKO-MGV

Combining strengths for multi-genome visual
analytics comparison

Guided Exercises

Version v2: 17th December 2018.



UNIVERSIDAD
DE MÁLAGA



Developed by:

Sergio Diaz-Del-Pino
Pablo Rodriguez-Brazzarola
Esteban Perez-Wohlfeil
Oswaldo Trelles

Report incidences to:

ortrelles@ac.uma.es

Or contact:

www.bitlab.es

Contents:

[Exercise 1: Diving into GECKO-MGV: Pairwise genome analysis mode](#)

[Section 1.1: Introduction](#)

[Section 1.2: Exercise development](#)

[Exercise 2: Diving into GECKO-MGV: Executing a pairwise genome comparison and extracting repetitions](#)

[Section 2.1: Introduction](#)

[Section 2.2: Exercise development](#)

[Exercise 3: Use GECKO-CSB full workflow step by step](#)

[Section 3.1: Introduction](#)

[Section 3.2: Exercise development](#)

[Exercise 4: Evolutionary events management](#)

[Section 4.1: Introduction](#)

[Section 4.2: Exercise development](#)

[Anex 1: Data Structure](#)

[Anex 2: Interacting with the comparison](#)

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

1.1 Introduction

In this exercise we will use the pairwise genome comparison mode, in which we first detect the repetitions of a comparison from Gecko using the registered service RepKiller; second, we perform simple tasks such as filtering and zooming; third, we select a group of repetitions of our interest to extract their primary sequences. Since the nucleotides sequences of our interest depend on “strand” field (forward or reverse), we first must generate the complementary sequence of the genome in the Y-axis to extract them. The retrieved sequences will be used in a fourth step to be aligned with the MUSCLE service, and lastly the multiple sequence alignment will be displayed.

1.1.a *Execute RepKiller on a comparison result.*

We start by executing a pairwise comparison of two genomes with Gecko. Afterwards, the repetitions on the fragments file are flagged and stored in the server by executing the RepKiller service on the server. When it is done, we will obtain another fragment file that identifies each fragment either as either unique, main repetition or normal repetition. It distinguishes fragments between these last two classifications, but maintains a relationship between each main repetition and the group of normal repetitions that it represents.

1.1.b *Load detected repetitions and perform simple actions.*

Now we can load the clean and repetition frag files stored in the server. After the loading process, the application displays the different views and layers. A Horizontal and Vertical view per 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.c *Repetition selection and sequence extraction*

In this exercise we will select a group of repetitions in the Y-axis, we will verify the group number that RepKiller has flagged them with. With such information, we can extract those fragments by specifying the group number in the Extract Repetitions from CSV service. With the generated frag file, we can extract the primary sequences with the Extract Sequences from CSV service, but first we must generate the reverse complementary sequence of the genome in the Y axis. The resulting multi-fasta file is stored in the server and will be used as input in the following step.

1.1.d *Align and visualize*

Next, we will launch a multiple sequence alignment to be performed over the set of retrieved sequences. Clustal Omega is the service we will use. This service is available as a Web-Service and can be invoked from our framework, including the sequences file as a parameter. Results from such alignment can be visualized with the MSA Viewer.

1.1 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:

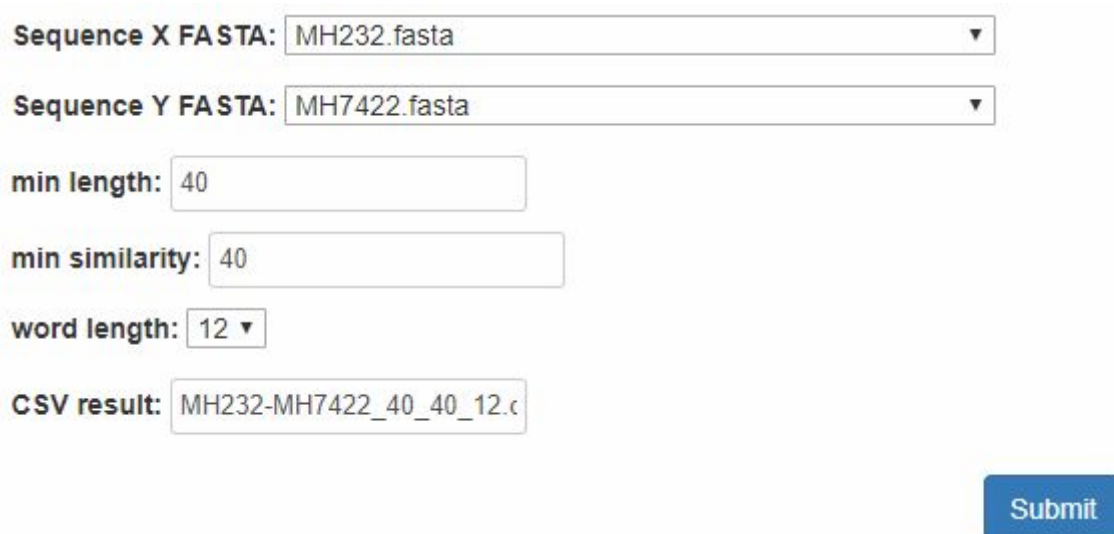
User: guest

Pass: guest



The image shows a login interface with a dark header bar. On the left is a 'Sign in' button. Next to it is a text input field containing 'guest'. To the right of that is a password input field with masked characters '.....'. On the far right is another green 'Sign in' button.

3. As registered user we proceed the Services tab, in order to execute the GECKO Workflow with the following parameters:



The image shows a configuration form for the GECKO Workflow. It includes several input fields: 'Sequence X FASTA' with a dropdown menu showing 'MH232.fasta', 'Sequence Y FASTA' with a dropdown menu showing 'MH7422.fasta', 'min length' with a text input '40', 'min similarity' with a text input '40', 'word length' with a dropdown menu showing '12', and 'CSV result' with a text input 'MH232-MH7422_40_40_12.c'. A blue 'Submit' button is located at the bottom right of the form.

(CSV result: MH232-MH7422-40-40-12.csv)

Important: CSV extension.

The comparison result will be available as soon as the process is finished in the File Manager tab. After it is finished we shall execute RepKiller service with the following parameters:



The image shows a configuration form for the RepKiller service. It includes several input fields: 'CSV frags file' with a dropdown menu showing 'MH232-MH7422_30_30_12.csv', 'Output Marked CSV' with a text input 'RK-MH232-MH7422-40-40-1', 'Similarity Length' with a text input '0.7', and 'Similarity Position' with a text input '0.7'. A blue 'Submit' button is located at the bottom right of the form.

(Output Marked CSV: RK-MH232-MH7422-40-40-12.csv)

Important: CSV extension

The output will be stored in the server, however we can download them by going to the File Manager tab in the navigation bar.



4. Now we shall split the output from RepKiller into a clean file that contains all the sequences that were not flagged as a repetition and another one that keeps the flagged fragments. To achieve this we must execute the Remove Repetitions from CSV service with the following parameters:

CSV file:

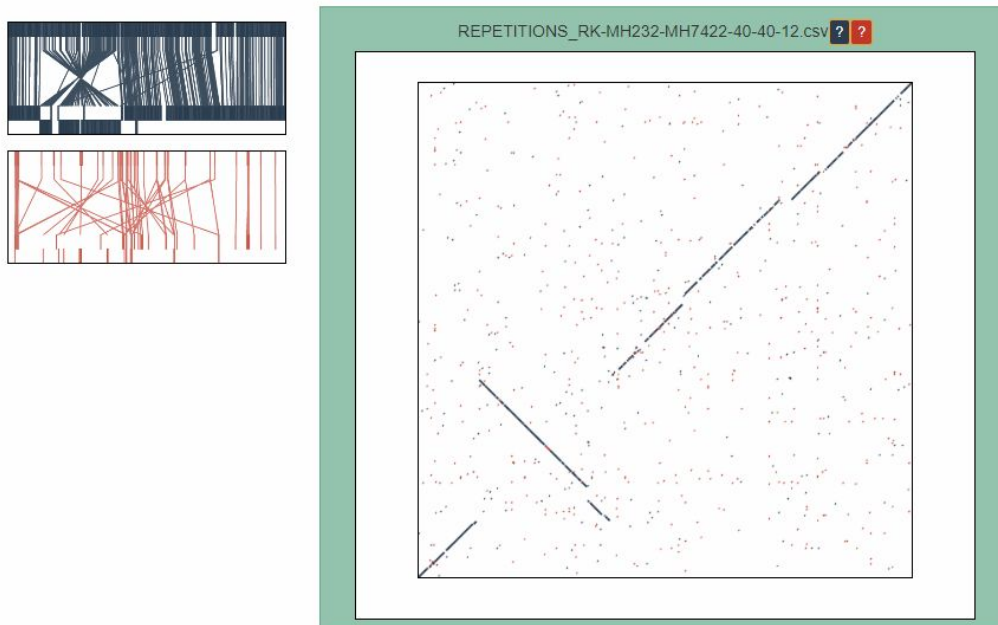
Extract Synteny Blocks:

Synteny Block ID (Optional):

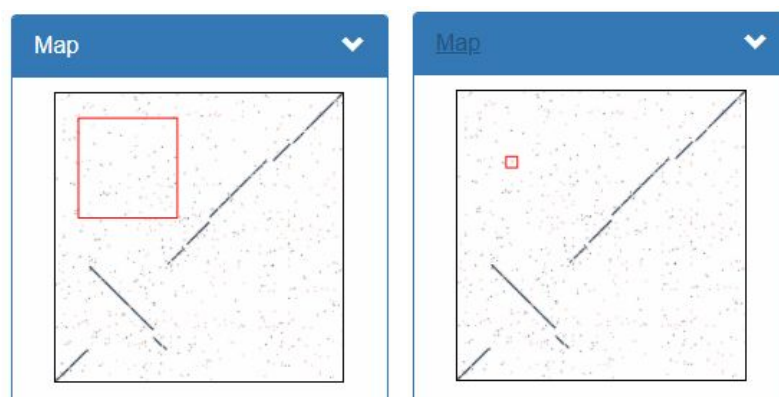
This will produce two output files: CLEAN_RK-MH232-MH7422-40-40-12.csv and REPETITIONS_RK-MH232-MH7422-40-40-12.csv.

5. Now we go back to the main page and open the comparison results generated by the Gecko Workflow from the server using the 'Load frags from server'  icon. Alternatively, if we have downloaded such file from the File Manager, we can 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'. We shall load both output files from the previous step, after extracting the repetitions from the RepKiller CSV (CLEAN_RK-MH232-MH7422-40-40-12.csv and REPETITIONS_RK-MH232-MH7422-40-40-12.csv).


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

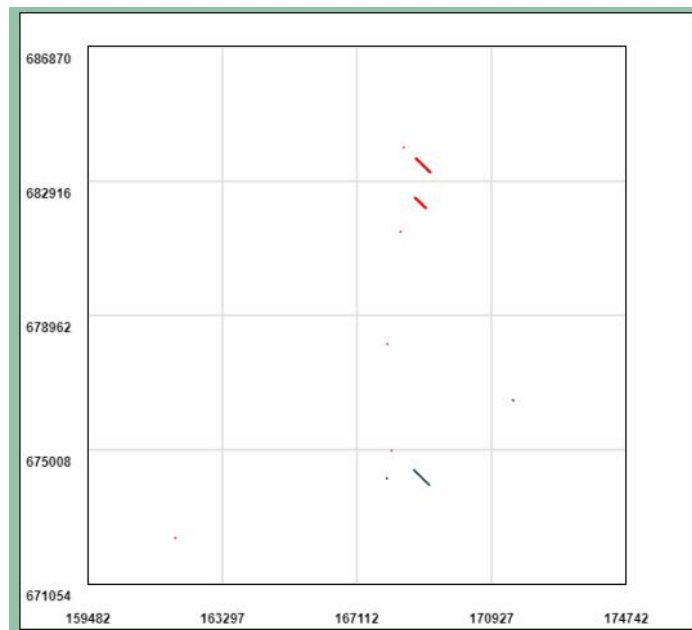



6. Now we will search for a group of repetitions that catch our attention by zooming into the comparison (a zooming example is shown with the images below). Then we will select a group of repetitions by pressing Shift + Click and dragging the mouse to the area of interest in the comparison view.





7. Now, we are going to activate the grid and zoom a bit more, 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.



8. To view the information of 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.

CSB & Frag

Filter:

File 0 File 1

Type	xStart	yStart	xEnd	yEnd	Strand	Block	Leng
Frag	161925	672381	161965	672421	f	281	41
Frag	168327	681444	168368	681403	r	781	42
Frag	168427	683925	168467	683885	r	860	41
Frag	167955	678097	167998	678140	f	1030	44
Frag	168078	674992	168118	674952	r	1034	41
Frag	168737	682440	169073	682104	r	1037	337
Frag	168765	683594	169071	683288	r	1037	307

9. After clicking 'Selected' this information will be only related to the selected fragments, as shown below. Since our objective is to extract them, a noteworthy field is the Block column.

CSB & Frag

Filter:

File 0 File 1

Type	xStart	yStart	xEnd	yEnd	Strand	Block	Length
Frag	168737	682440	169073	682104	r	1037	337
Frag	168765	683594	169071	683288	r	1037	307

10. We could store that selection in the server by clicking 'Upload', but we could be missing some of the repetitions located in other areas that are not visible because of the zoom. Therefore we will inspect the Synteny Block identifier in the selection

information (In this case 1037) and extract the with the Extract Repetitions from CSV service using the following parameters:

CSV file:

Extract Synteny Blocks:

Synteny Block ID (Optional):

This will generate two output files (CLEAN_RK-MH232-MH7422-40-40-12(1).csv and REPETITIONS_RK-MH232-MH7422-40-40-12(1).csv) and we are interested in the one that contains the repetitions.

11. Now we need the primary sequences of the extracted fragments. To obtain them from a CSV we need to execute the Extract Sequences from CSV service, but first we need the reverse complementary sequence of the genome in the Y axis as an input.

12. To generate the reverse complementary sequence we must launch the Reverse Complement service with the subsequent parameters:

Fasta to reverse:

Output file:

13. Now we launch the Extract Sequences from CSV service with following parameters:

CSV frags file:

X Fasta file:

Y Fasta file:


Y-Reversed Fasta file:

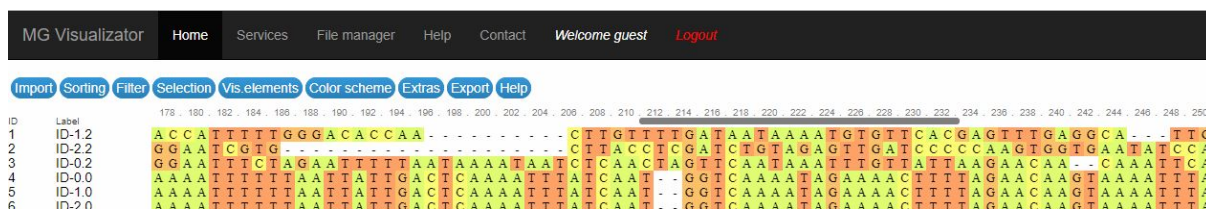
Output FastaFile:

(Output FastaFile: MH232-MH7422-rep_1037.fasta)

Multifasta file:

Multiple Alignment Output:

15. Lastly, we go to the File Manager tab in the navigation bar and click on the View () button for the multiple sequence alignment file generated in the previous step. The results will be shown automatically using the MSA-Viewer from BioJS:



Exercise 2: Multiple pairwise-comparison analysis

2.1 Introduction

This exercise illustrates the concept of layers. We will use the layer concept to visualize how a chromosome from one specie presents similarities with different chromosomes of another specie. Layers might be helpful to compare results from different executions. In this exercise we will load the GECKO results obtained from comparing the first chromosome of the human (*homo sapiens*) against the first five chromosomes of the house mouse (*mus musculus*). This is done in order to visualize the chunks of DNA that are common between different species but not in the same chromosome.

Each execution result will be showed in a new layer, becoming part of the session data. All the layers will be displayed in the main canvas. The size of the X and Y axis will be the one of the first comparison to be loaded.

2.1.a Obtain comparison results

This second exercise starts by comparing the chromosome 1 of the *homo sapiens* (GCF_000001405.13) against the first 5 chromosomes of the *mus musculus* (GCF_000001635.26). The comparison results should be in the guest account in GECKO-MGV. However they can also be downloaded from http://mango.ac.uma.es/compartir/GeckoMGV/comparison-results/homo_sapiens-mus_musculus/. These results were generated from GECKO executions with the parameters:

- Length: 60
- Similarity: 40
- Word Length: 32

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 load all the comparisons at the same time to visualize the similarities of the first *homo sapiens* chromosome with the first five of the *mus musculus*. Each comparison will appear in the canvas as a new layer. After they are all loaded we will visualize the chunks of DNA of a chromosome of one specie are spread into several chromosomes of a different specie

2.2 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:

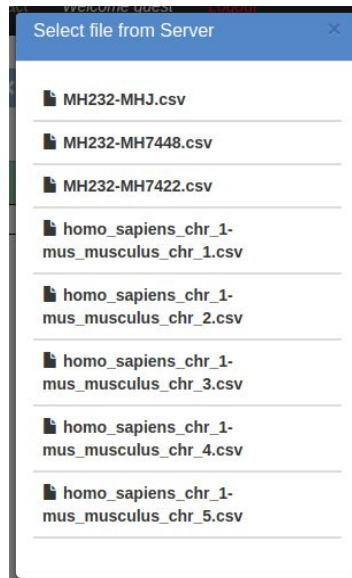
User: guest

Pass: guest

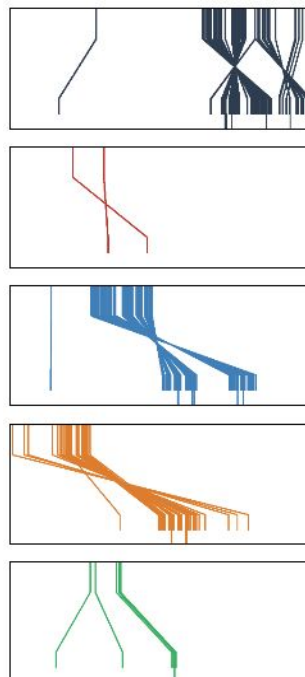


The image shows a login interface with a dark background. On the left is a white 'Sign in' button. Next to it is a white input field containing the text 'guest'. To the right of this field is another white input field containing masked characters '.....'. On the far right is a green 'Sign in' button.

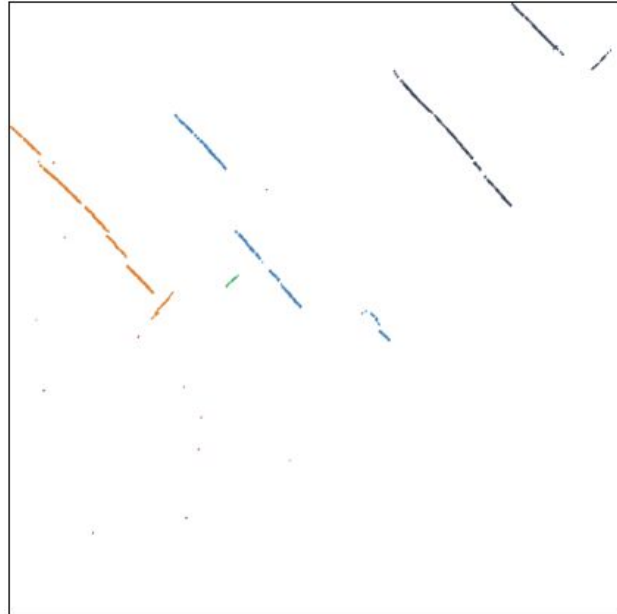
3. In the main page we shall open the five comparison results generated by the GECKO using the 'Load frags from server'  icon. Alternatively, if we have downloaded such file from the server, we can click in the 'Load frags from local'  icon, go through our file system to the folder where the comparison files are in CSV format and click in 'Open'.




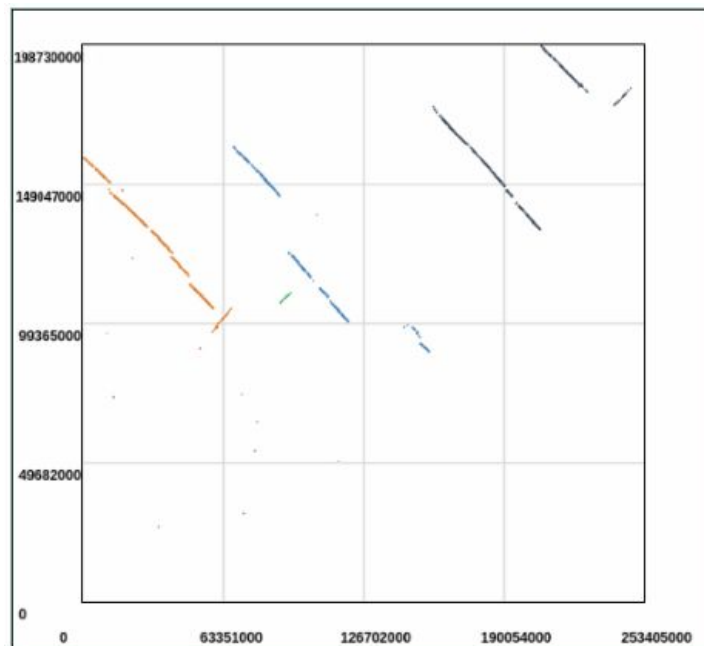
4. After all results are loaded, the canvas should have 5 layers, one for each comparison. The resulting horizontal layers will present in the top the *homo sapiens* chromosome and in the bottom the compared chromosome of the *mus musculus*. It should look as in the following image:



The resulting vertical layers will be displayed in the main canvas, in which the X axis will belong to the *homo sapiens* chromosome, meanwhile the Y axis presents the different chromosomes of the *mus musculus*.

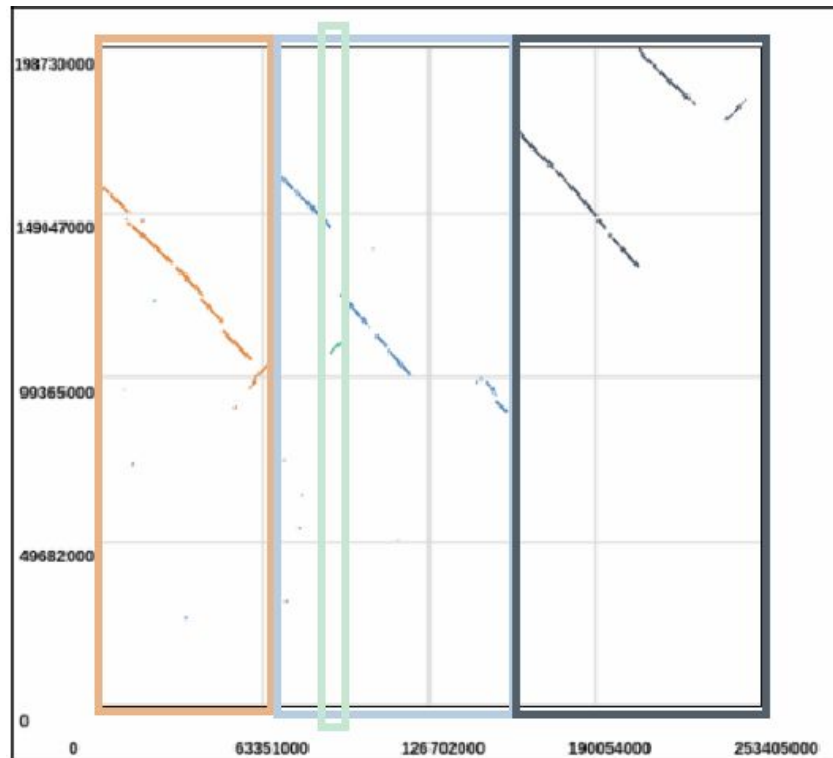


5. Now, we are going to activate the grid to have a better perception of where we are in the genome. For that we click in the 'Grid' () button located at the top menu. As result, the grid is shown.

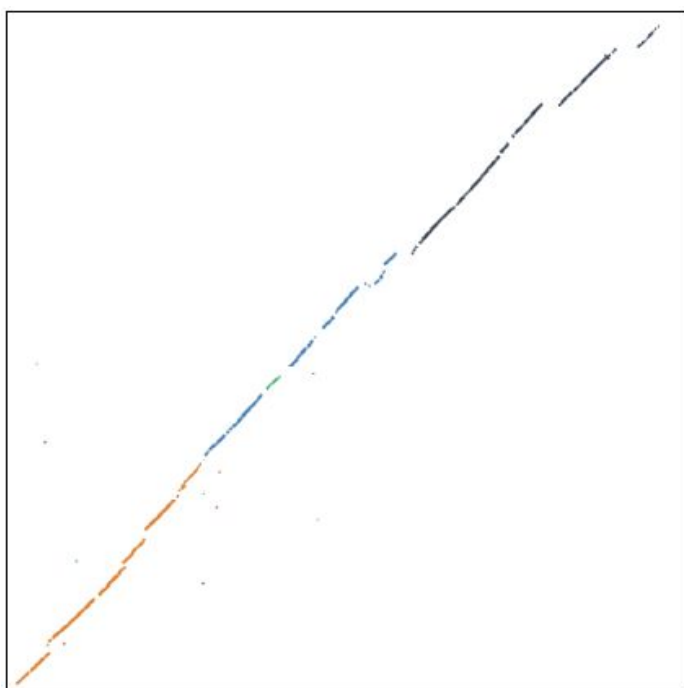


6. After loading all the comparisons we can observe that there are chunks of DNA for each comparison throughout the whole axis of first chromosome of

homo sapiens (X axis). We can observe that the 1st, 3rd and 4th chromosome of *mus musculus* present a noticeable region similar to the 1st chromosome of *homo sapiens*. The 2nd chromosome of *mus musculus* presents a smaller (in comparison to the previous three chromosomes) but visible region similar to 1st chromosome of *homo sapiens*. On the other hand, the 5th chromosome of *mus musculus* is barely noticeable in the canvas.



By now we are visualizing the common regions of DNA that are located in the same chromosome of the *homo sapiens* when compared to *mus musculus*. The resulting canvas is very interesting because it allows us to visualize different chunks of DNA (or groups of genes) that are common between species, but due to evolutionary events they have inherited independently into different chromosomes. The following image represents a mockup of the results if we rearranged the chunks of DNA in order to obtain a full diagonal.



Exercise 3: Pairwise-comparison annotation

3.1 Introduction

In this exercise we will use the multiple genome comparison mode, in which we will identify and extract a collection of repeated fragments that contain the “transposase” annotation. Then we will retrieve the sequences of this fragments and store them in a file. Finally we will perform a multiple sequence alignment with the MUSCLE service and display the results interactively. The aim of this exercise is to use GECKO-MGV as a tool that allows us to annotate regions of a genome that is not annotated.

We will be comparing different strains of the *mycoplasma hyopneumoniae* and searching for the regions annotated with “transposase”. We have selected this annotation because the genes that encode transposases are the most abundant genes known and are widespread in the genomes of most organisms.

Transposases are enzymes that catalyze the movement of DNA sequences that can change its position within a genome, also known as *transposons*. These enzymes can create or reverse mutations, and alter the cell’s genetic identity and genome size. This phenomenon is known as *transposition* and it is important in creating genetic diversity within species and adaptability to changing living conditions.

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


A screenshot of a web application's login interface. It features a dark horizontal bar with a white 'Sign in' button on the left, a white text input field containing the text 'guest', a white password input field with masked characters '.....', and a green 'Sign in' button on the right.

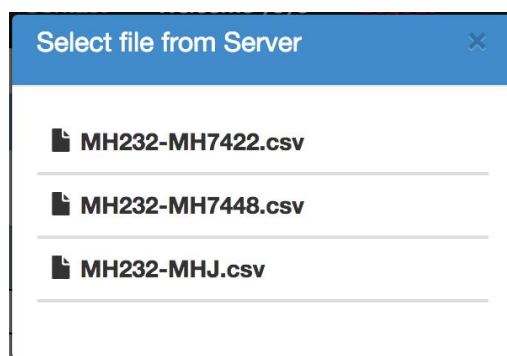
3. We proceed by uploading the files needed for this ecomparisons obtained by the GECKO workflow after. The sequences, annotation files and comparison results for this exercise can be downloaded from <http://mango.ac.uma.es/compartir/GeckoMGV/comparison-results/mycoplasma/>. The parameters of the GECKO workflow to obtain those results are:


- Length: 50
- Similarity: 25
- Word length: 16

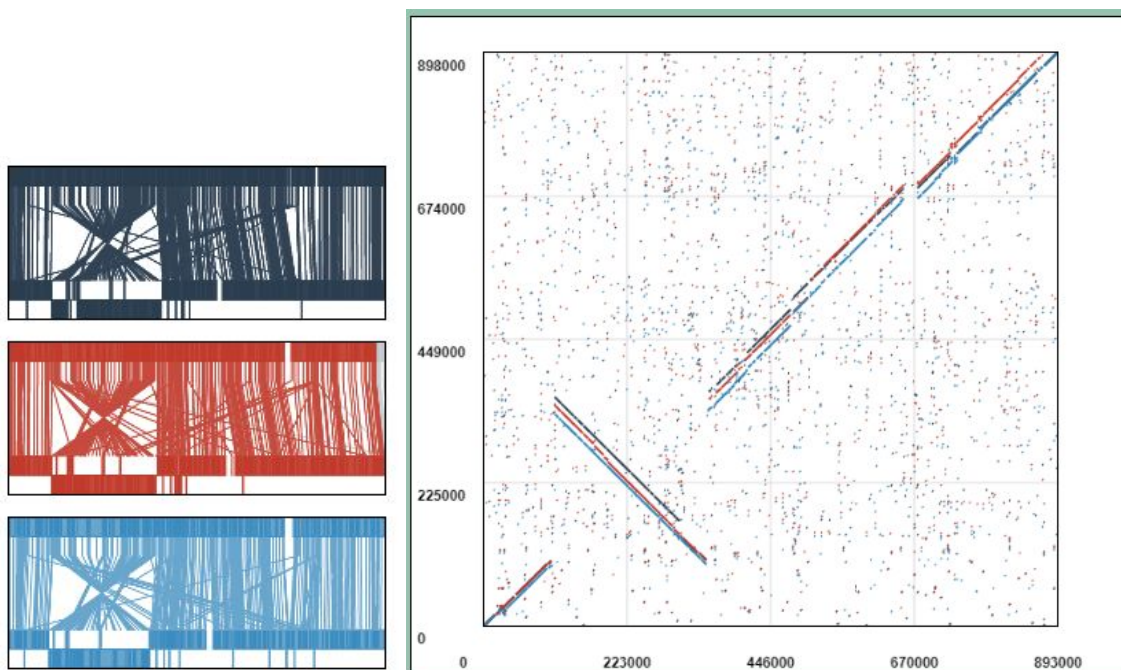
4. If the sequences are already on the user’s file system so we can avoid the next step.

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. We also need to upload the annotation files and the comparison results if you decide to not execute the GECKO workflow service from the 'Services' tab.

5. Now we click on the 'Home' tab and then on the  button on the upper part and then we click on the mycoplasma comparison files displayed on the floating window.



After we have loaded all the comparisons we are going to activate the grid by clicking in the 'Grid' () button in the top menu. The canvas should look something like this for the horizontal view (left) and for the vertical view (right):



6. Now we apply an identity and length filter of 80 and 1000 respectively.

Filters

☐ Statistical Significance
☐ Overlapped (Repetitions)
☐ Duplications
☐ Positive blocks

☒ Identity 80

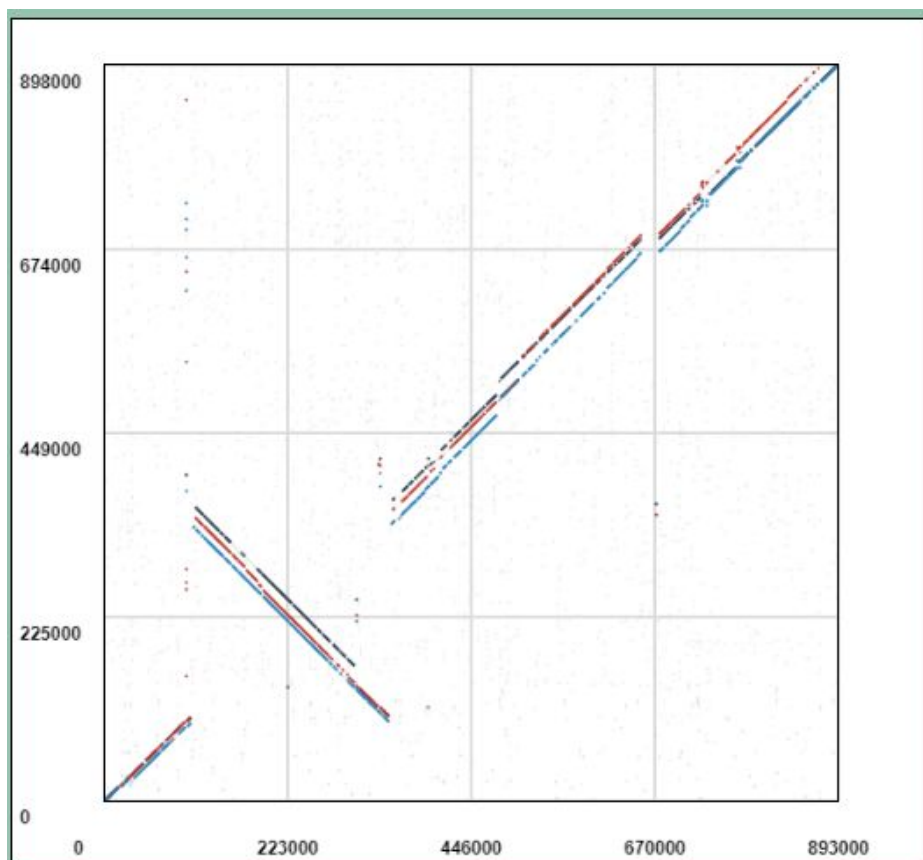
☒ MinLength 200


☐ MinSimilarity 50

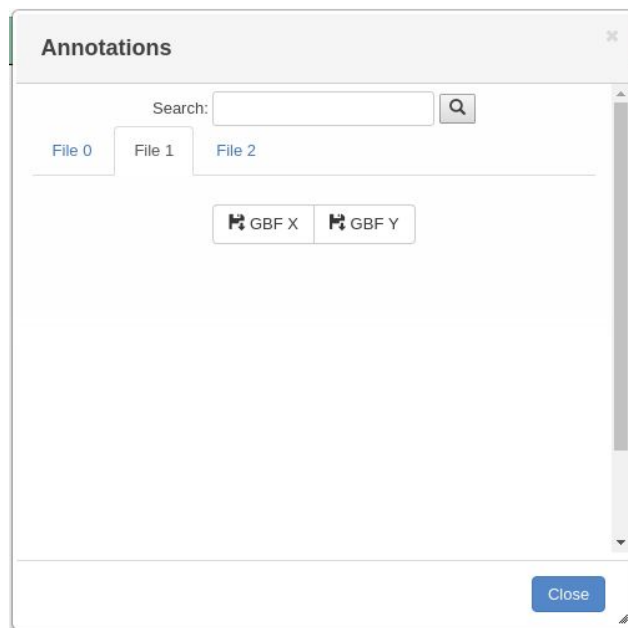
☐ Manual selection

Filter

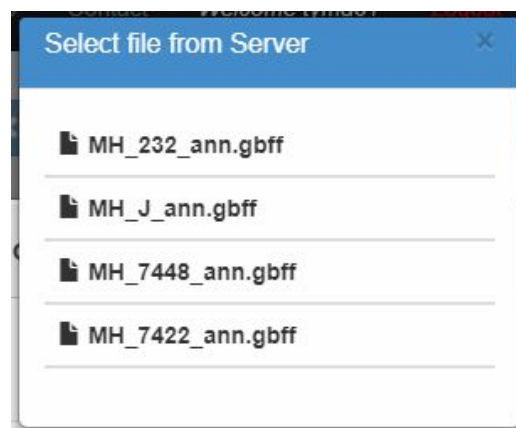
The results should be the following:



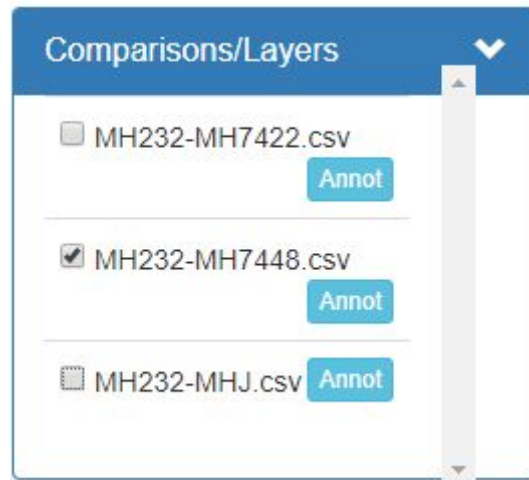
- Now we load the annotations file for the genome in the X axis (in this case for the *Mycoplasma Hyopneumoniae* 232) by pressing 'Annotations Info' () on the top menu. A window will appear in which we select the file tab of the comparison we want to annotate. We will select the 'MH232-MH7448.csv' comparison file (in our case the 'File 1' tab and click on the 'GBF X' button.



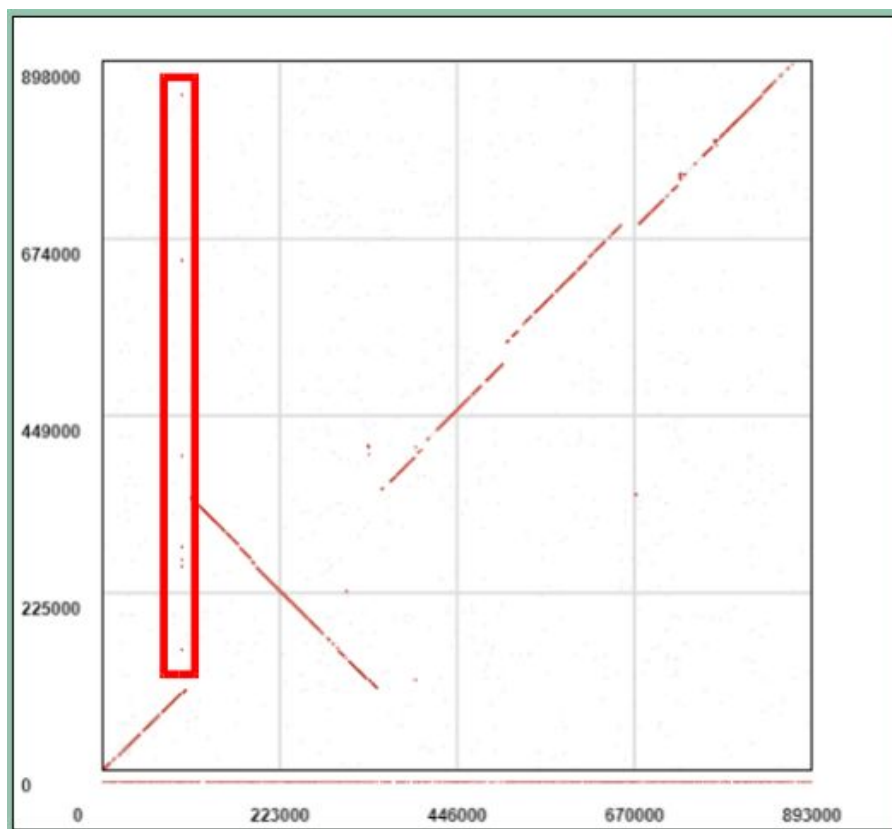
8. This will open up a dialog that contains all the GBFF files uploaded to the server. We select the one belonging to Mycoplasma Hyopneumonia 232.



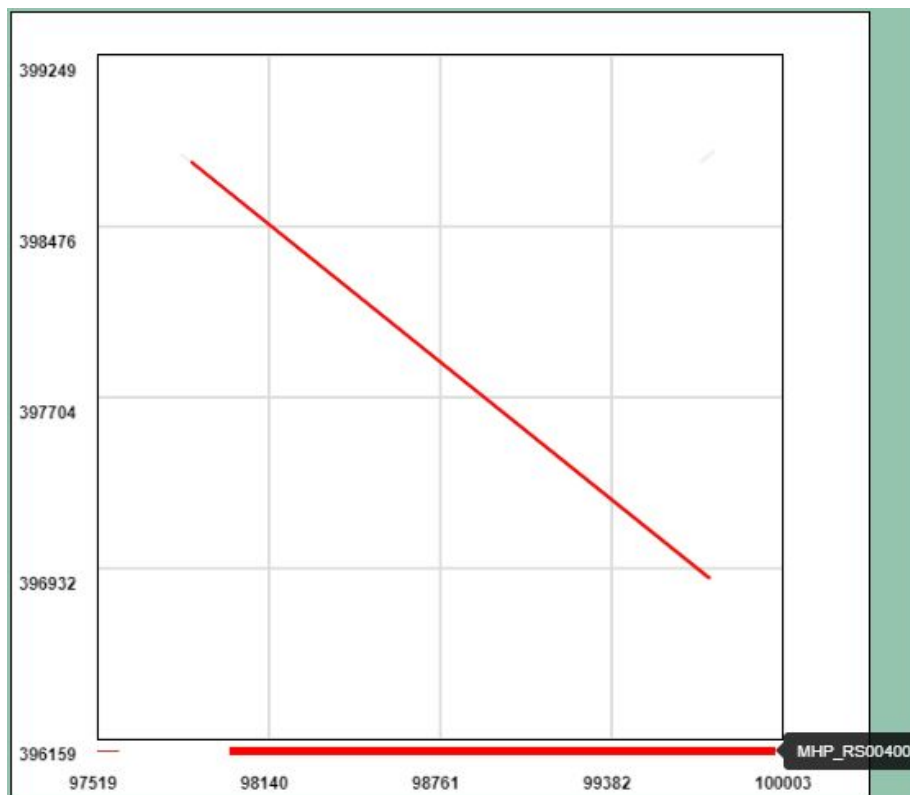
After we have loaded the annotations file, we can activate it by click 'Annot' on the 'Comparisons/Layers' box. We will also deactivate the other layers.



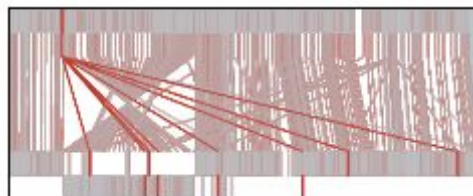
9. In this comparison, there are a series of repetition in the Y axis. We will select them with 'Shift + Click'.

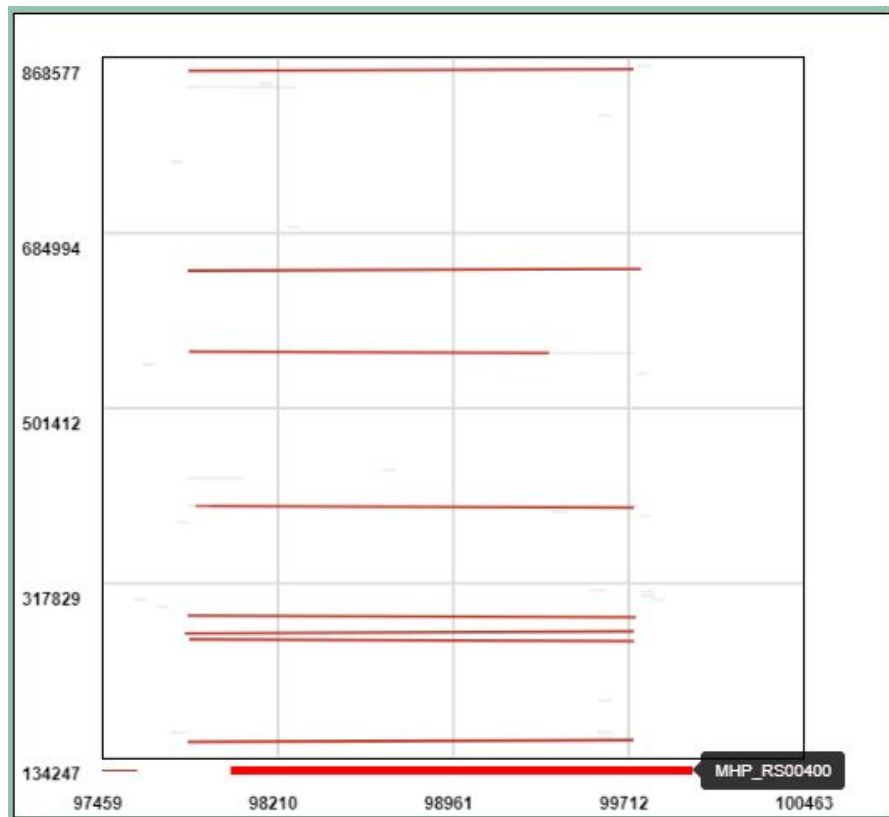


10. We can zoom into one of these fragments to view what is the annotation that corresponds to it.



11. We can also zoom to view multiple of the selected fragments and their corresponding annotation.





In both cases, the annotation that corresponds to such fragments is 'MHP_RS00400'.

12. Now we will search for 'transposase' in the 'Annotations Info' window.

Annotations

Search:

File 0

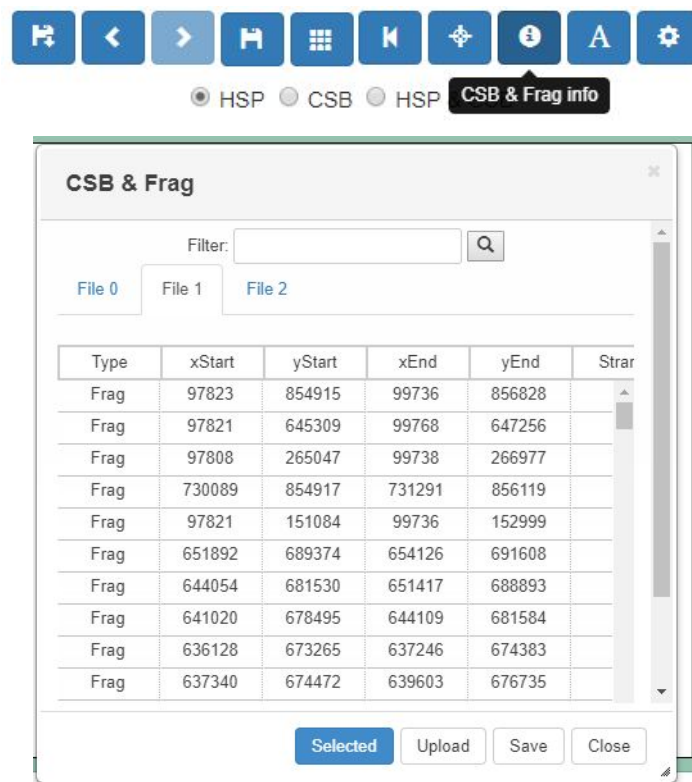
File 1

File 2

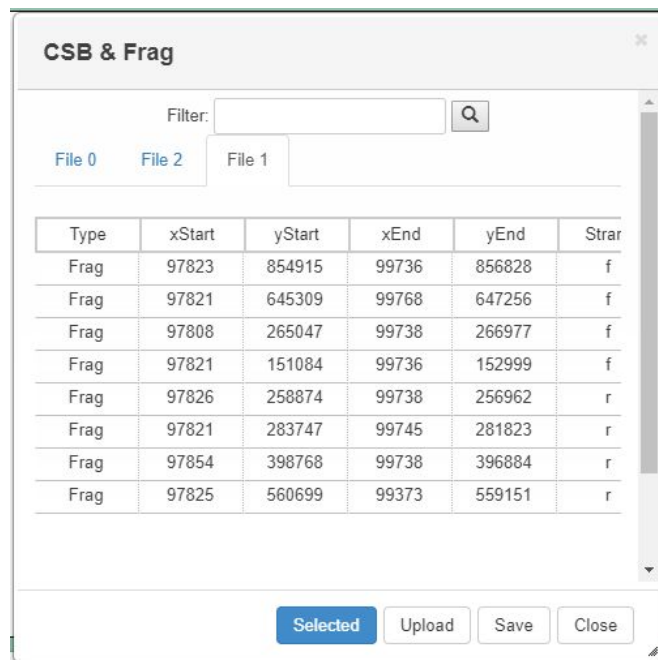
Stop	Strand	Size	Gene	Synonym	Product	File
99579	r	1658	-	MHP_RS00400	IS4 family trans...	X

13. We can see that the synonym belongs to the region visualized. Therefore we will use these 'Transposase' annotated fragments to perform a multiple sequence alignment (MSA) using the MUSCLE service. But to do that we must first obtain a CSV that contains these selection. To do that we click on the 'CSB & Frag info' and select the current file/layer where we have

selected the fragments.



14. Now we shall click on 'Selected' to view the frag information of our selected frags. Then we press the 'Upload' button to upload the current frag information to the server in a new file.



15. When the upload is done, this should appear on the top right corner.

New file available



16. Now we go into the 'Service' tab and execute 'Reverse Complement' to obtain the reverse complementary of the genome in the Y axis (*Mycoplasma Hyopneumoniae* 7488 in this case).

Reverse Complement

Write into the second parameter the reversed complement of the first parameter



Fasta to reverse: MH7448.fasta ▼

Output file: RC_MH7448.fasta

Submit

17. Then we execute the 'Extract sequences from CSV' service using the recently uploaded file that contains the selected frags information.

Extract Sequences from CSV

Extract FASTA sequences from a CSV file



CSV frags file: MH232-MH7448(1).csv ▼

X Fasta file: MH232.fasta ▼

Y Fasta file: MH7448.fasta ▼

Y-Reversed Fasta file: RC_MH7488.fasta ▼

Output FastaFile: multifasta.fasta

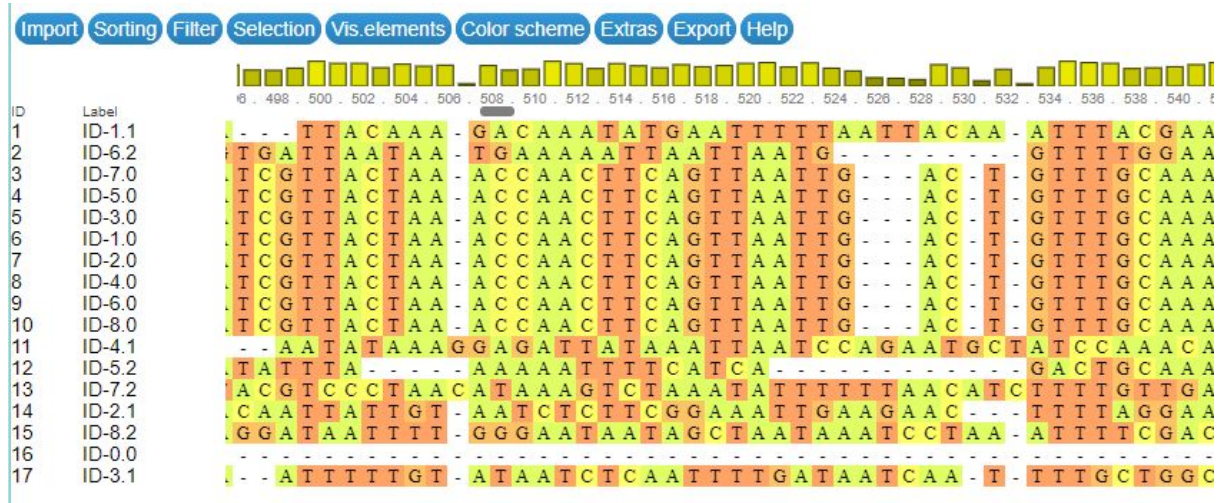
Submit

18. The next step is to execute the 'MUSCLE' service with the multifasta obtained from the previous step.

MUSCLE

MUltiple Sequence Comparison by Log-Expectation. To view results interactively use '.clw' extension for the multiple sequence alignment (MSA) output and '.dnd' for the dendrogram.






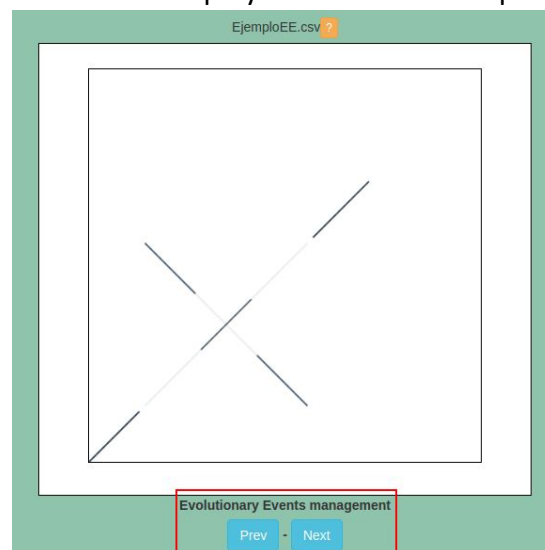
Exercise 4: Evolutionary Events

4.1 Introduction

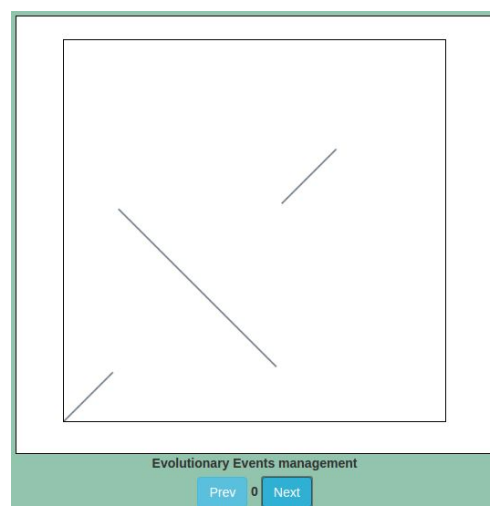
In this exercise we are going to learn how to use the evolutionary events panel, allowing us to navigate through all the evolutionary events occurred on a genome. To perform this exercise you need to download the 'ExampleEE.csv' from <http://mango.ac.uma.es/compartir/GeckoMGV/evolutive-events/>.

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 timeline 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 **Frag**s, 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:

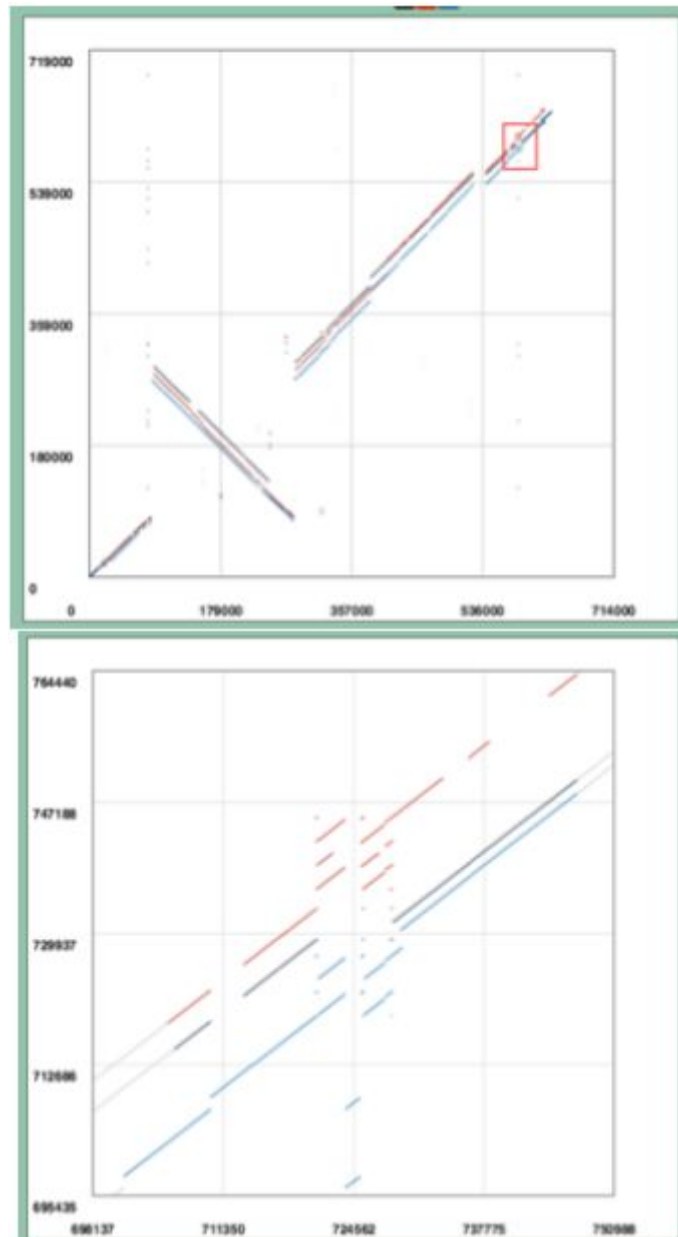
- Input files:
 - o Sequence files: in FASTA format.
 - o Annotations file: in GFF format (optional).
 - o Score matrix (inter residues scores) --optional.
- Intermediate files:
 - o Dictionaries: contains k=32-mers and its repetitions. From these files is possible to obtain repeats, tandem repeats, word frequencies, words appearing over and below the expected values, etc.
 - o Hits: occurrences of the same word in two sequences.
 - o FrequencyKmer: K-mer frequency (k=1) to obtain karlin parameters.
 - o Karpar: Karlin parameters (to calculate p-value).
- Output files:
 - o Fragments file (binary) contains the description and coordinates for each fragment. Usually combines forward and reverse-complementary frags.
 - o *.INF file: metadata information about the procedure to obtain the fragments files (name of the genomes, sequences, parameters, etc.).
 - o 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.
 - o Extended frags: CSB-Master format.

Annex 2:

1 Interacting with the comparison

1.1 Zoom-in and zoom-out

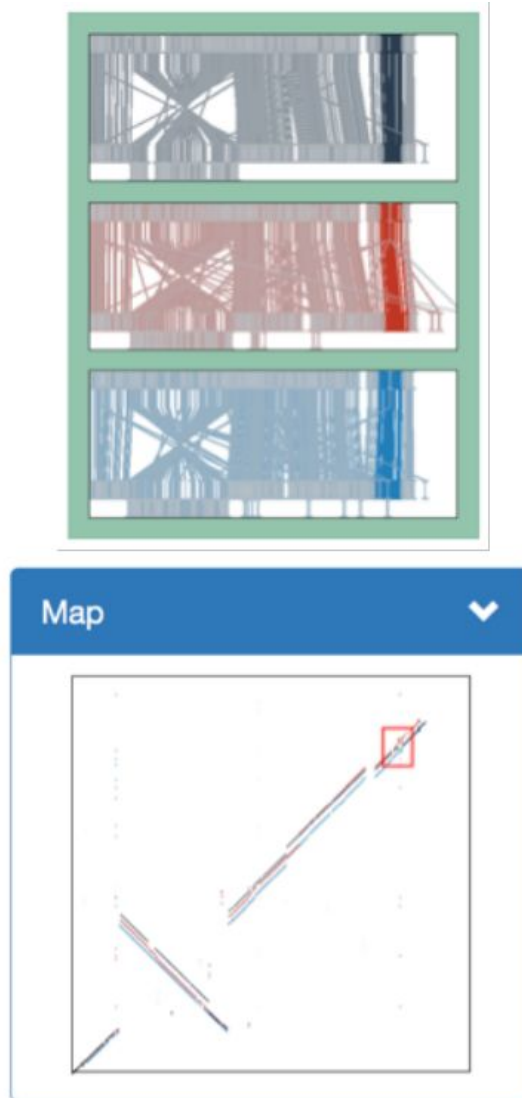
Zooming interaction is done in the main canvas. At first a quadrangle must be drawn while clicking with the left button of the mouse over the zone we want to zoom. After releasing the left button zoom is processed and all the views are updated. Two buttons to go back and go forward have been implemented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