Module 2
Comparative Genomics

Introduction

The Artemis Comparison Tool (ACT), also written by Kim Rutherford, was designed to extract the additional information that can only be gained by comparing the growing number of sequences from closely related organisms (Carver *et al*. 2005). ACT is based on Artemis, so you will already be familiar with many of its core functions. It is is essentially composed of three layers or windows. The top and bottom layers are mini Artemis windows (with their inherited functionality), showing the linear representations of the DNA sequences with their associated features. The middle window shows red and blue blocks, which span this middle layer and link conserved regions within the two sequences, in the forward and reverse orientation respectively. Consequently, if you were comparing two identical sequences in the same orientation you would see a solid red block extending over the length of the two sequences in this middle layer. If one of the sequences was reversed, and therefore present in the opposite orientation, there would be a blue ‘hour glass’ shape linking the two sequences. Unique regions in either of the sequences, such as insertions or deletions, would show up as breaks (white spaces) between the solid red or blue blocks.

In order to use ACT to investigate your own sequences of interest you will have to generate your own pairwise comparison files. Data used to draw the red or blue blocks that link conserved regions is generated by running pairwise BLASTN or TBLASTX comparisons of the sequences. ACT is written so that it will read the output of several different comparison file formats; these are outlined in Appendix III. Two of the formats can be generated using BLAST software freely downloadable from the NCBI, which can be loaded and run on a PC or Mac. Another way of generating comparison files for ACT is to use the WebACT web resource (http://www.webact.org/). This site allows you to cut and paste or upload your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files.

Aims

The aim of this Module is for you to become familiar with the basic functions of ACT. In the first exercise you will be looking at a comparison between chromosome 11 of *P. falciparum* 3D7 (a human malaria parasite) and chromosome 9 of *P. chabaudi* AS (a rodent malaria parasite). By comparing the two chromosomes you will be able to study the degree of conservation of gene order and identify small and large synteny breaks.

Exercise 1
Starting up the ACT software

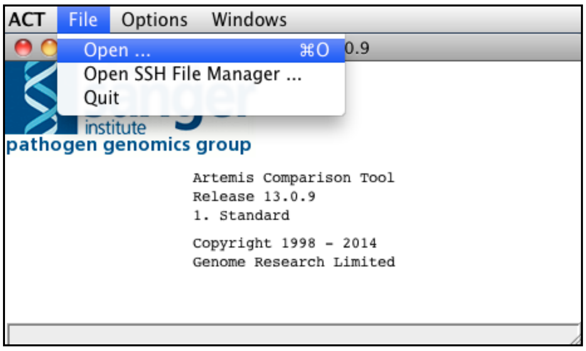
Make sure you're in the **Module_2_Comparative_Genomics, exercise_1** directory.
Then type
act & [return]
A small start up window will appear.

To open ACT you can also double click the ACT icon on your Desktop.

The files that you are going to need are:

Pf3D7_11_v3.embl	- *P. falciparum* chr11
Pf3D7_11_v3.fasta-txchab09.fasta.crunch	- tblastx comparison file
chab09.embl	- *P. chabaudi* chr9

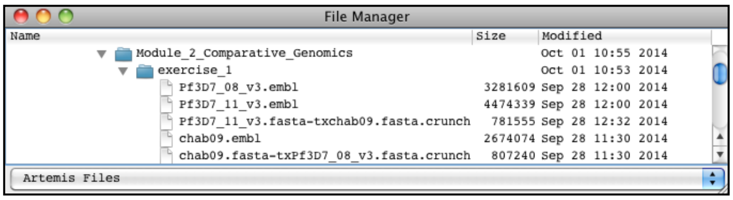
Click 'File' then 'Open'



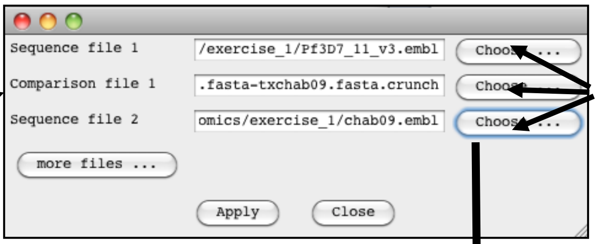
Artemis Comparison Tool
Release 13.0.9
1. Standard

Copyright 1998 - 2014
Genome Research Limited

Use the File manager to drag and drop files.

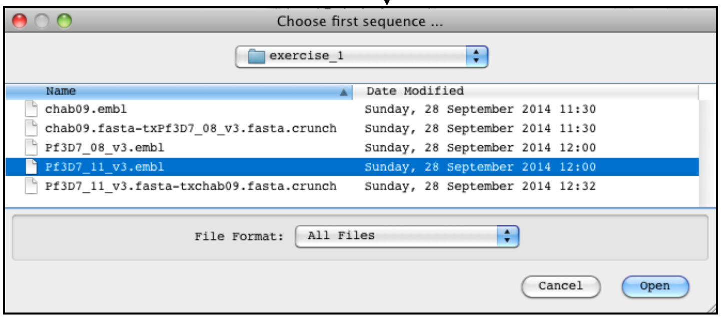


Name	Size	Modified
Module_2_Comparative_Genomics		Oct 01 10:55 2014
exercise_1		Oct 01 10:53 2014
Pf3D7_08_v3.embl	3281609	Sep 28 12:00 2014
Pf3D7_11_v3.embl	4474339	Sep 28 12:00 2014
Pf3D7_11_v3.fasta-txchab09.fasta.crunch	781555	Sep 28 12:32 2014
chab09.embl	2674074	Sep 28 11:30 2014
chab09.fasta-txPf3D7_08_v3.fasta.crunch	807240	Sep 28 11:30 2014



Sequence file 1 /exercise_1/Pf3D7_11_v3.embl
Comparison file 1 .fasta-txchab09.fasta.crunch
Sequence file 2 omics/exercise_1/chab09.embl

Instead of dragging and dropping the files, you can also choose them.



Name	Date Modified
chab09.embl	Sunday, 28 September 2014 11:30
chab09.fasta-txPf3D7_08_v3.fasta.crunch	Sunday, 28 September 2014 11:30
Pf3D7_08_v3.embl	Sunday, 28 September 2014 12:00
Pf3D7_11_v3.embl	Sunday, 28 September 2014 12:00
Pf3D7_11_v3.fasta-txchab09.fasta.crunch	Sunday, 28 September 2014 12:32

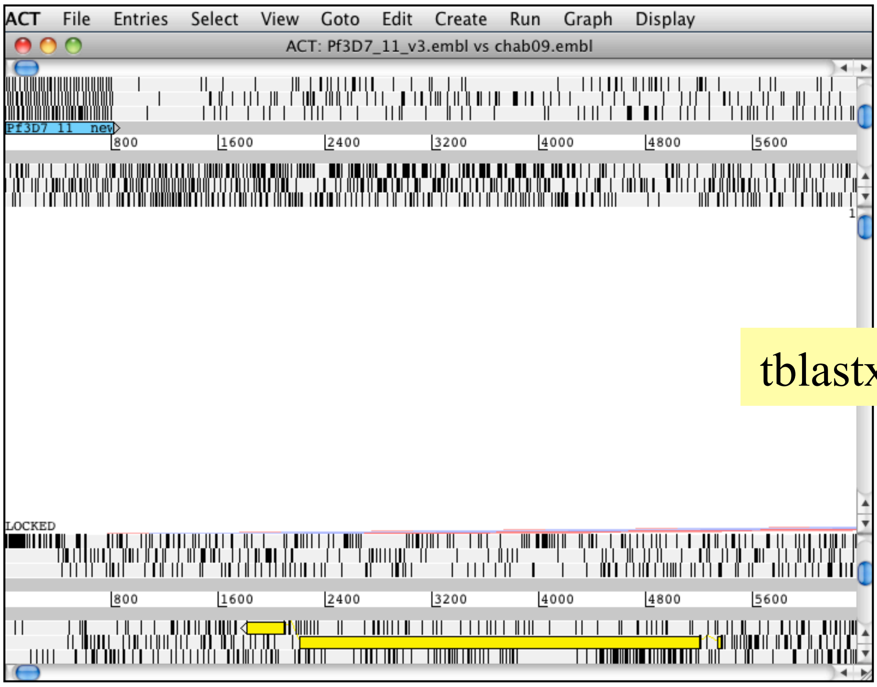
File Format: All Files

For comparing more than two DNA files!

Click 'Apply' and wait

Comparison files end with '.crunch'. For more info on comparison files see Appendix III

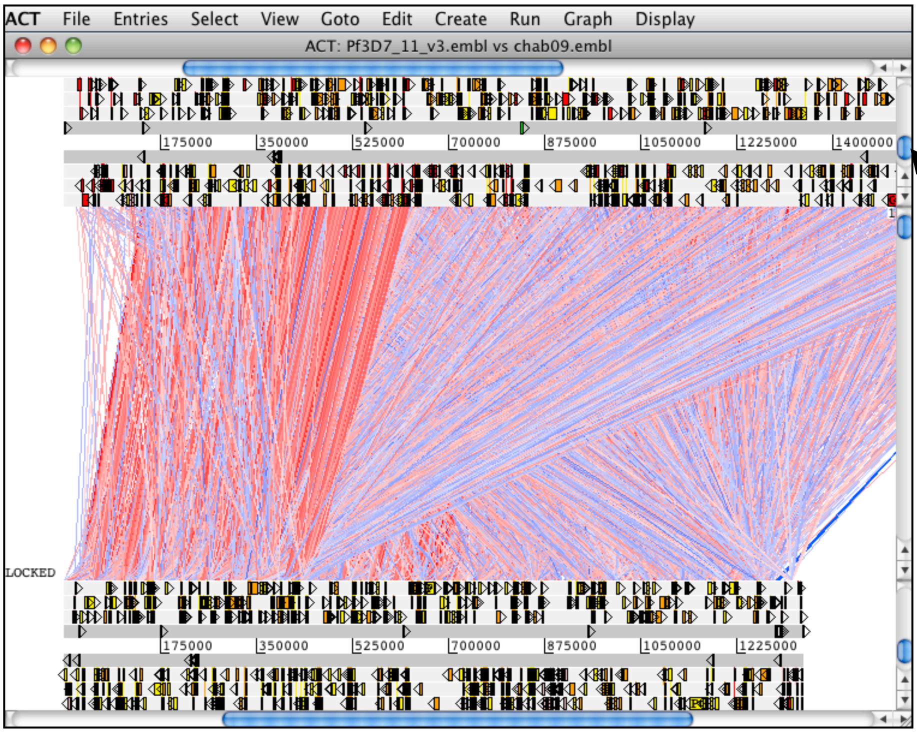
Once you have opened the files you will see a picture like this:


ACT File Entries Select View Goto Edit Create Run Graph Display
ACT: Pf3D7_11_v3.embl vs chab09.embl
Pf3D7_11 new
800
1600
2400
3200
4000
4800
5600
LOCKED

P. falciparum chr11

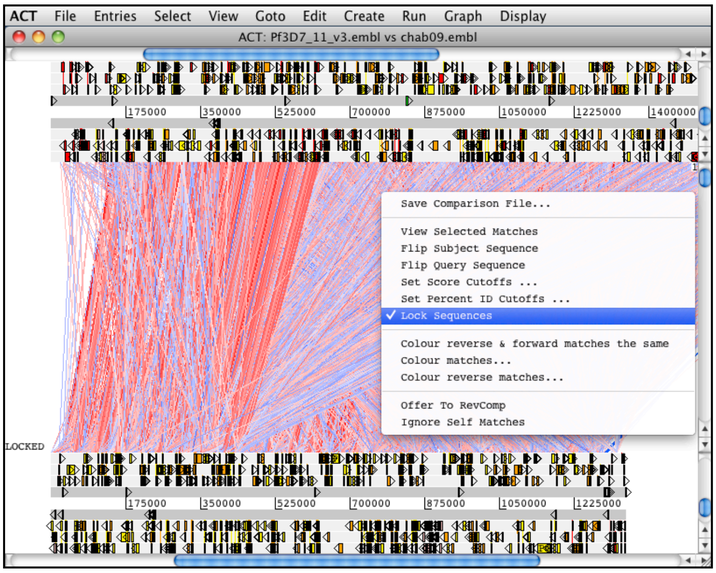
tblastx comparison

P. chabaudi chr9


ACT File Entries Select View Goto Edit Create Run Graph Display
ACT: Pf3D7_11_v3.embl vs chab09.embl
175000
350000
525000
700000
875000
1050000
1225000
1400000
LOCKED

Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.

When you scroll along with either slider both genomes move together. This is because they are ‘locked’ together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that ‘LOCKED’ has disappeared from the comparison view panel and the genomes will now move independently.


ACT File Entries Select View Goto Edit Create Run Graph Display
ACT: Pf3D7_11_v3.embl vs chab09.embl
Save Comparison File...
View Selected Matches
Flip Subject Sequence
Flip Query Sequence
Set Score Cutoffs ...
Set Percent ID Cutoffs ...
✓ Lock Sequences
Colour reverse & forward matches the same
Colour matches...
Colour reverse matches...
Offer To RevComp
Ignore Self Matches
LOCKED

LOCKED

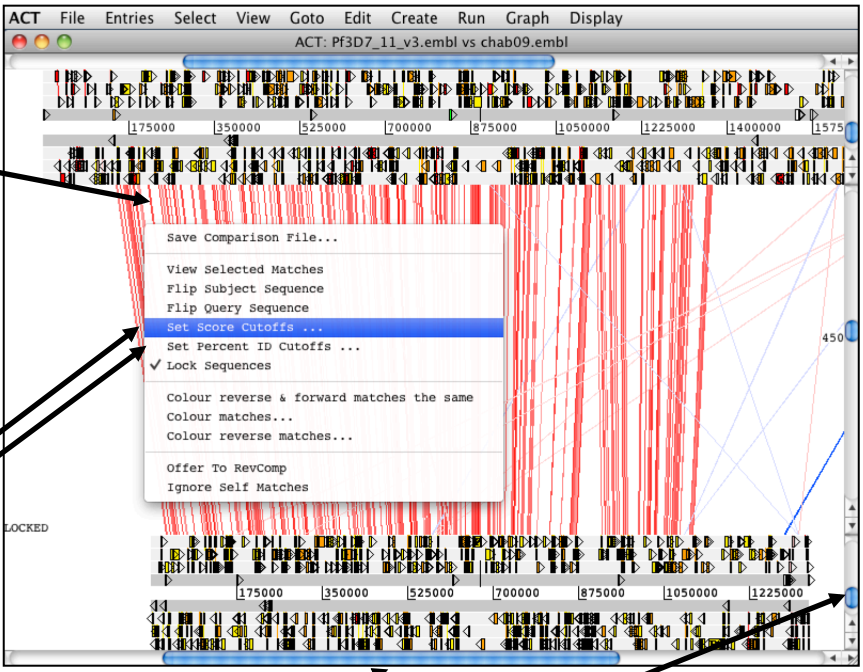
You can optimise your image by either removing ‘low scoring’ (or percentage ID) hits from view, as shown below **1-3** or by using the slider on the comparison view panel (**4**). The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs, sometimes described as the “footprint”.

1

Right button click in the Comparison View panel

2

Select either Set Score Cutoffs or Set Percent ID Cutoffs

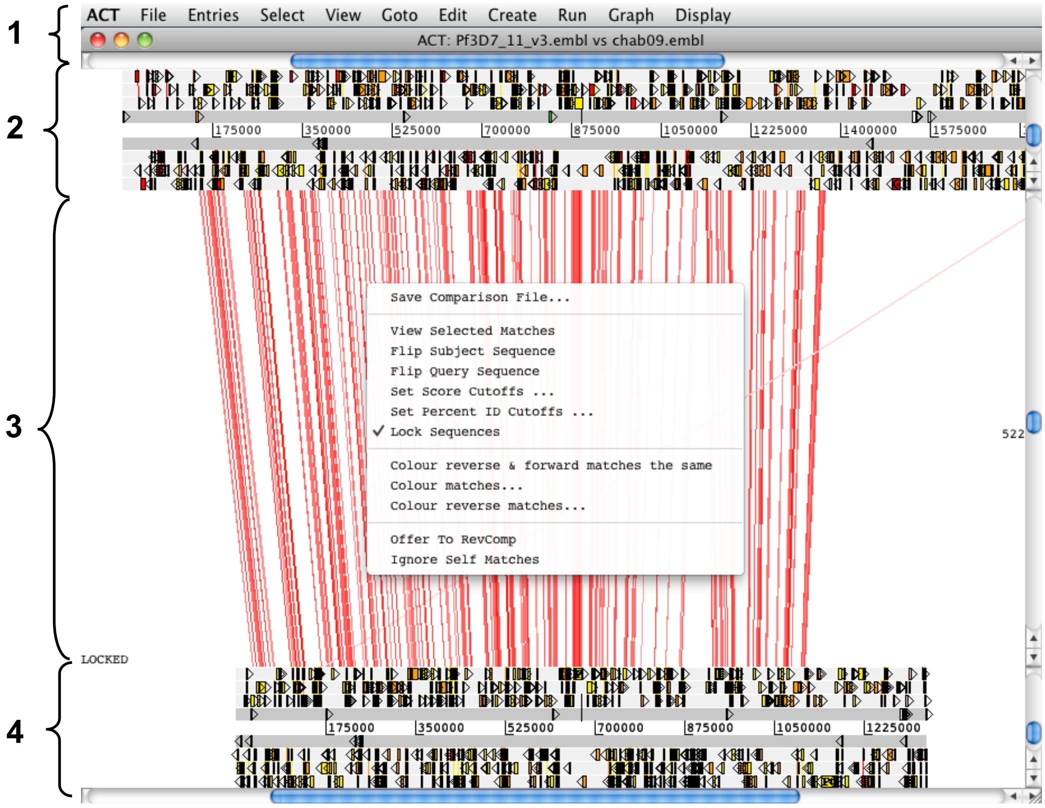

ACT File Entries Select View Goto Edit Create Run Graph Display
ACT: Pf3D7_11_v3.embl vs chab09.embl
Save Comparison File...
View Selected Matches
Flip Subject Sequence
Flip Query Sequence
Set Score Cutoffs ...
Set Percent ID Cutoffs ...
✓ Lock Sequences
Colour reverse & forward matches the same
Colour matches...
Colour reverse matches...
Offer To RevComp
Ignore Self Matches
LOCKED
450

4

3

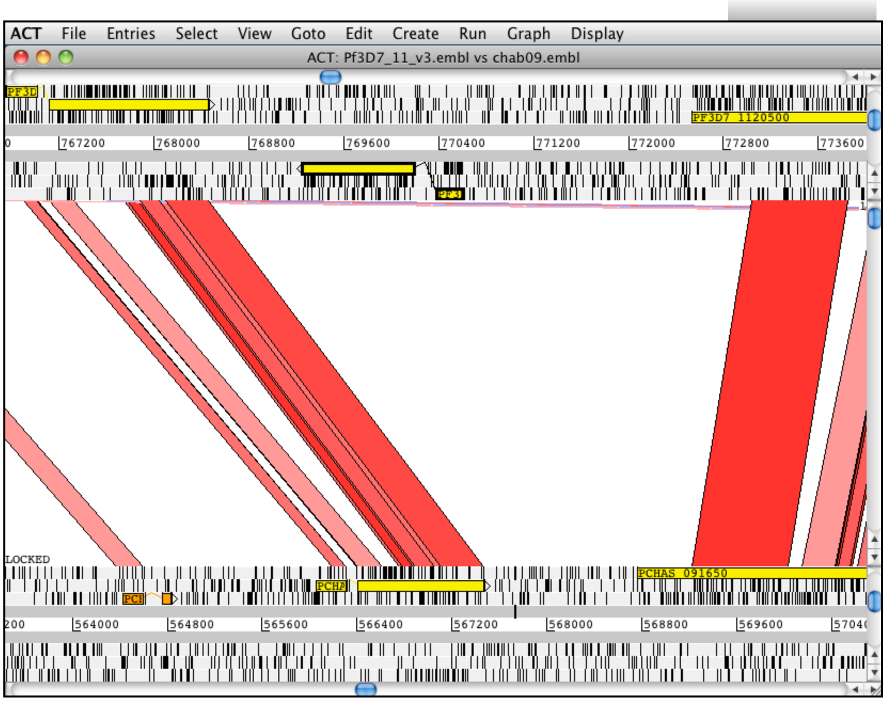
Move the sliders to manipulate the comparison view image

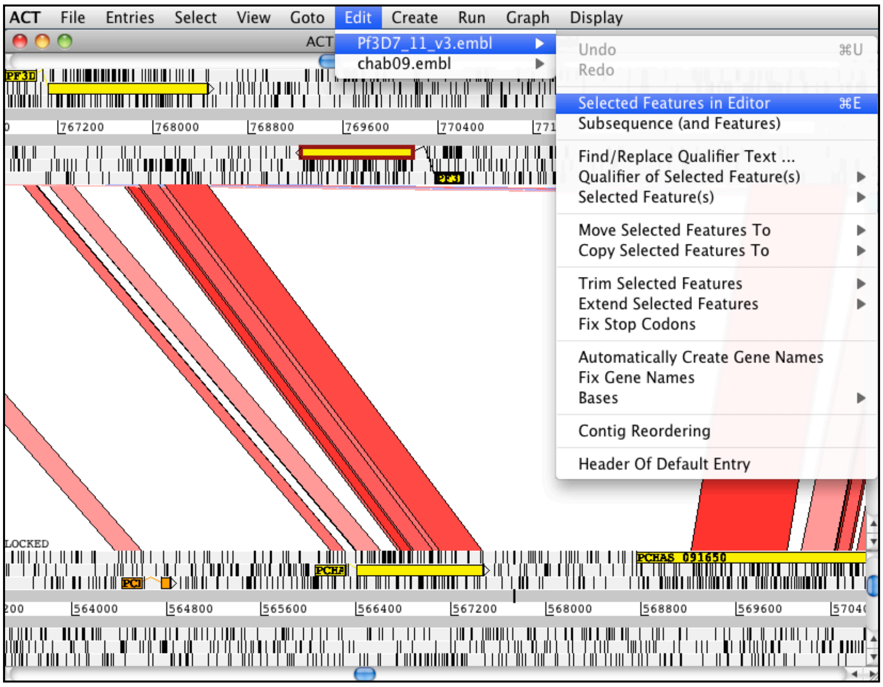
Now that you have an ACT window open let's look what's in there.



1. Drop-down menus. These are mostly the same as in Artemis. The major difference you'll find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you selected earlier. It's a slightly compressed version of the Artemis main view panel. The panel retains the sliders for scrolling along the genome and for zooming in and out.
3. The Comparison View. This panel displays the regions of similarity between two sequences. Red blocks link similar regions of DNA with the intensity of red colour directly proportional to the level of similarity. Double clicking on a red block will centralise it. Blue blocks link regions that are inverted with respect to each other.
4. Artemis-style Sequence View panel for 'Sequence file 2' (Query Sequence).
5. Right button click in the Comparison View panel brings up this important ACT-specific menu which we will use later.

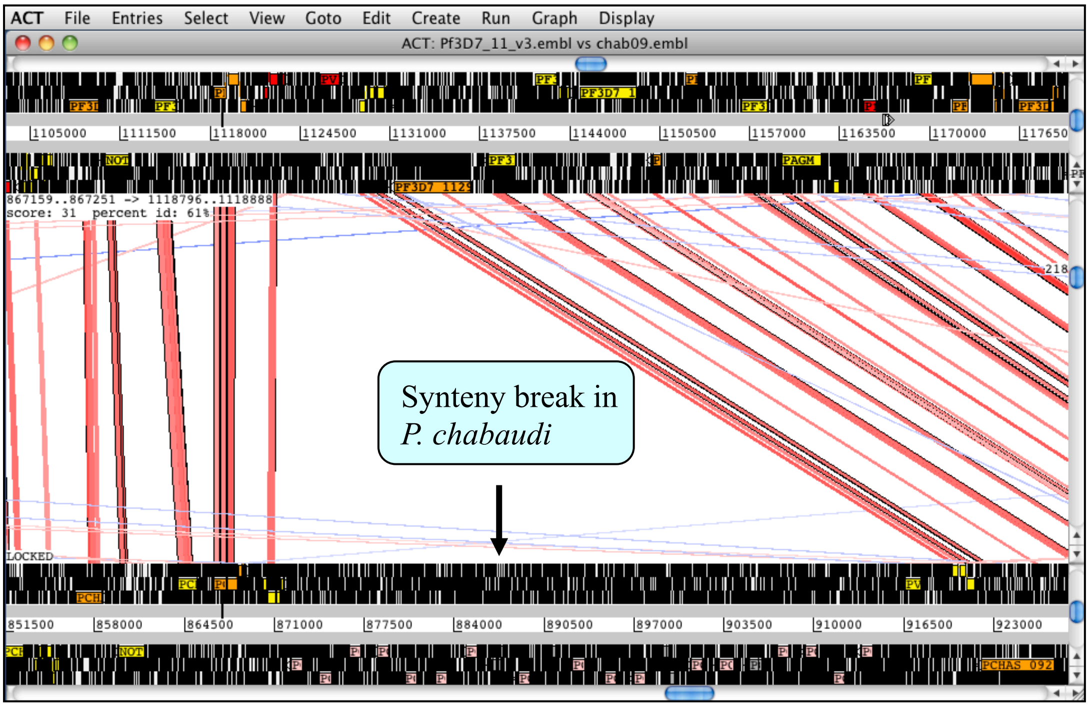
Scroll along the chromosome and look for small synteny breaks between *P. falciparum* and *P. chabaudi*. One example is shown here. Use the 'Goto' option to go to this region. You can either use the option 'Goto base' or 'Goto Feature With Gene Name' (e.g. PF3D7_1120400).




ACT File Entries Select View Goto Edit Create Run Graph Display
Pf3D7_11_v3.embl
chab09.embl
Undo ⌘U
Redo
Selected Features in Editor ⌘E
Subsequence (and Features)
Find/Replace Qualifier Text ...
Qualifier of Selected Feature(s)
Selected Feature(s)
Move Selected Features To
Copy Selected Features To
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry
LOCKED
PCHAS_091650

To get more information about this gene, select the gene, then go to 'Edit' 'Selected Features in Editor'. As a shortcut you can also just press 'E' on your keyboard.

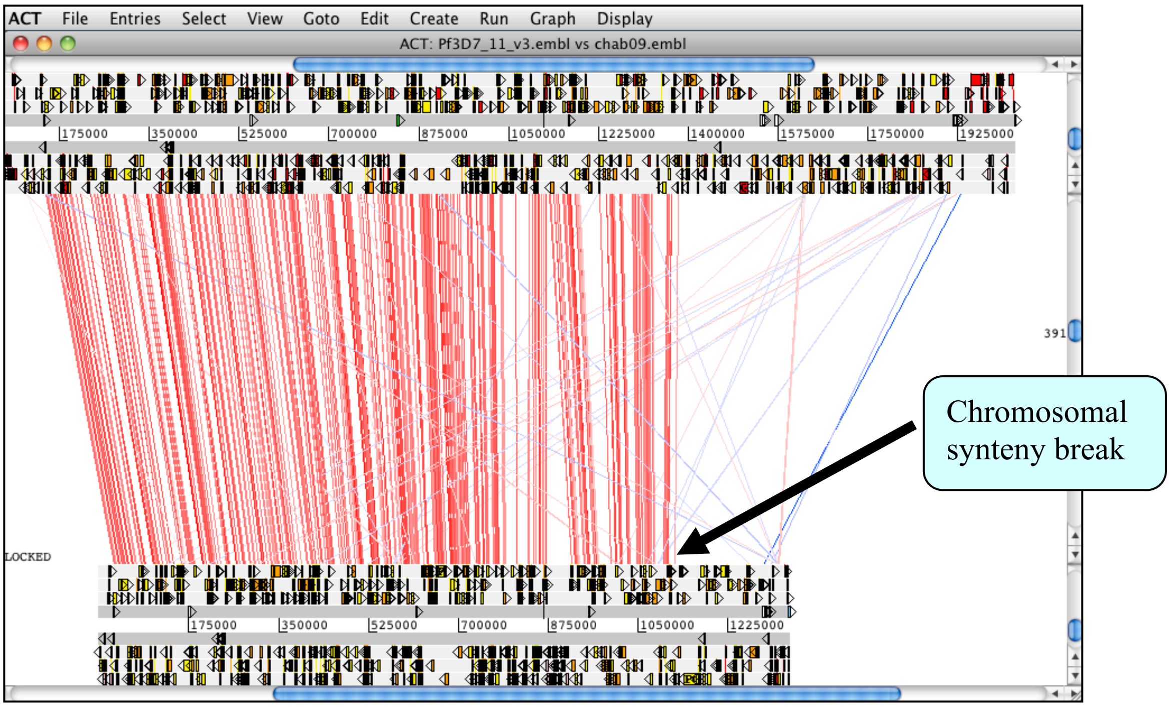
Scroll along the chromosome and try to get an estimate on the number of synteny breaks. Can you find the largest synteny break? Identify the genes that are located in this area.



```
867159..867251 -> 1118796..1118888
score: 31  percent id: 61%
```

Synteny break in *P. chabaudi*

Can you locate the region of a chromosomal synteny break point?

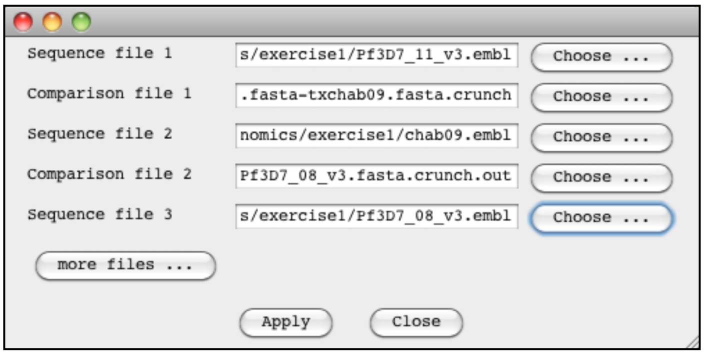


Chromosomal synteny break

In this part of the exercise we will look at a three-way comparison and explore the chromosomal synteny break in *P. chabaudi*.

The files you are going to need are:

Pf3D7_11_v3.embl	- *P. falciparum* chr11
Pf3D7_11_v3.fasta-txchab09.fasta.crunch	- tblastx comparison file
chab09.embl	- *P. chabaudi* chr09
chab09.fasta-txPf3D7_08_v3.fasta.crunch	- tblastx comparison file
Pf3D7_08_v3.embl	- *P. falciparum* chr8



Sequence file 1 s/exercise1/Pf3D7_11_v3.embl

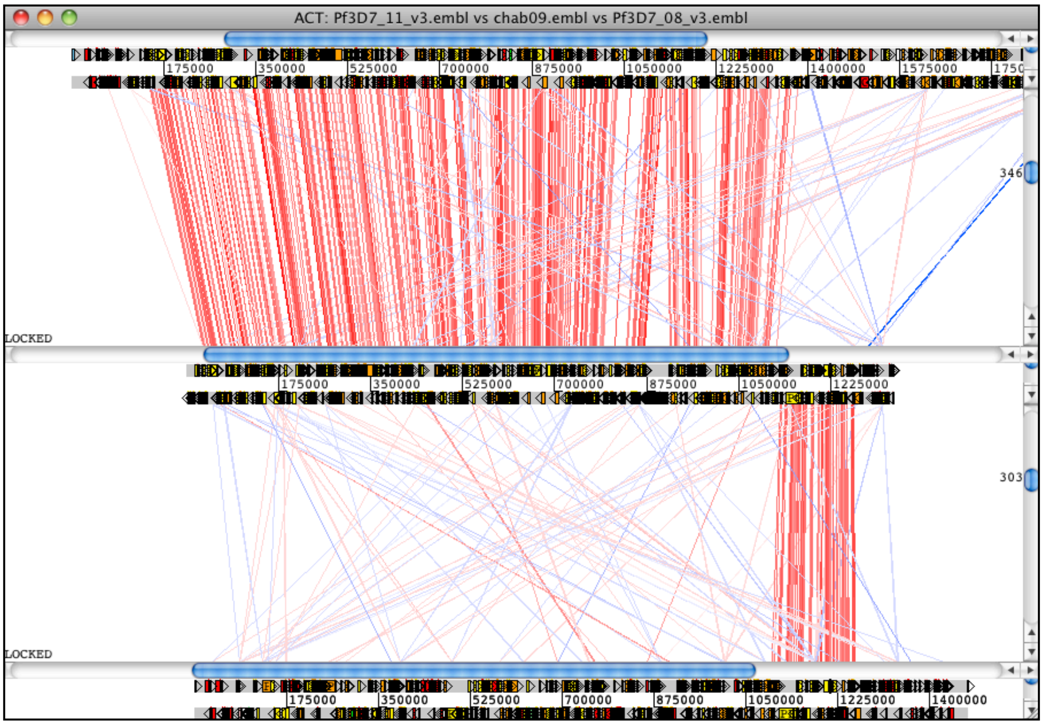
Comparison file 1 .fasta-txchab09.fasta.crunch

Sequence file 2 nomics/exercise1/chab09.embl

Comparison file 2 Pf3D7_08_v3.fasta.crunch.out

Sequence file 3 s/exercise1/Pf3D7_08_v3.embl

Click on ‘more files’ to compare more than 2 files.



P. falciparum chr11

P. chabaudi chr09

P. falciparum chr08

Once you have finished this exercise remember to close this ACT session down completely before starting the next exercise.

Exercise 2

Genomes are often highly similar and one has to look quite closely to find small differences. In this exercise we will be comparing the human malaria genome *P. falciparum* with the recently sequenced chimpanzee malaria genome, *P. reichenowi* and try to identify differences.

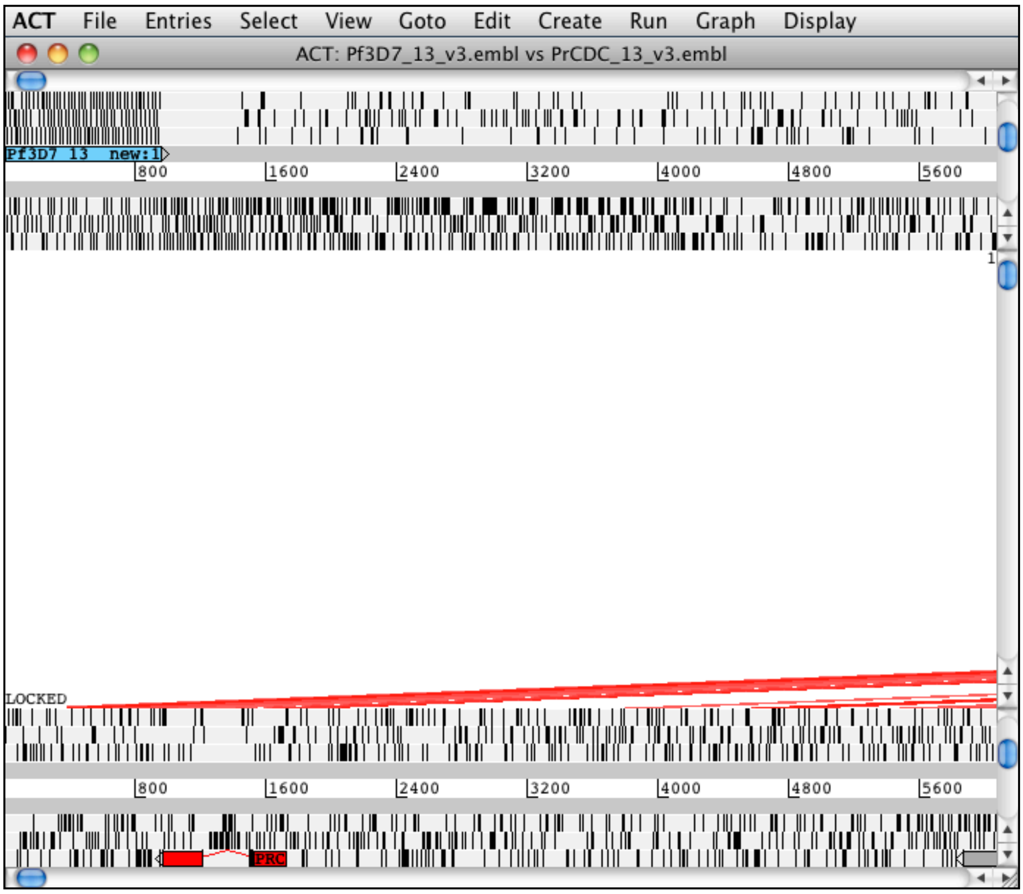
Make sure you're in the **Module_2_Comparative_Genomics, exercise_2** directory.
Then type
act & [return]
A small start up window will appear.

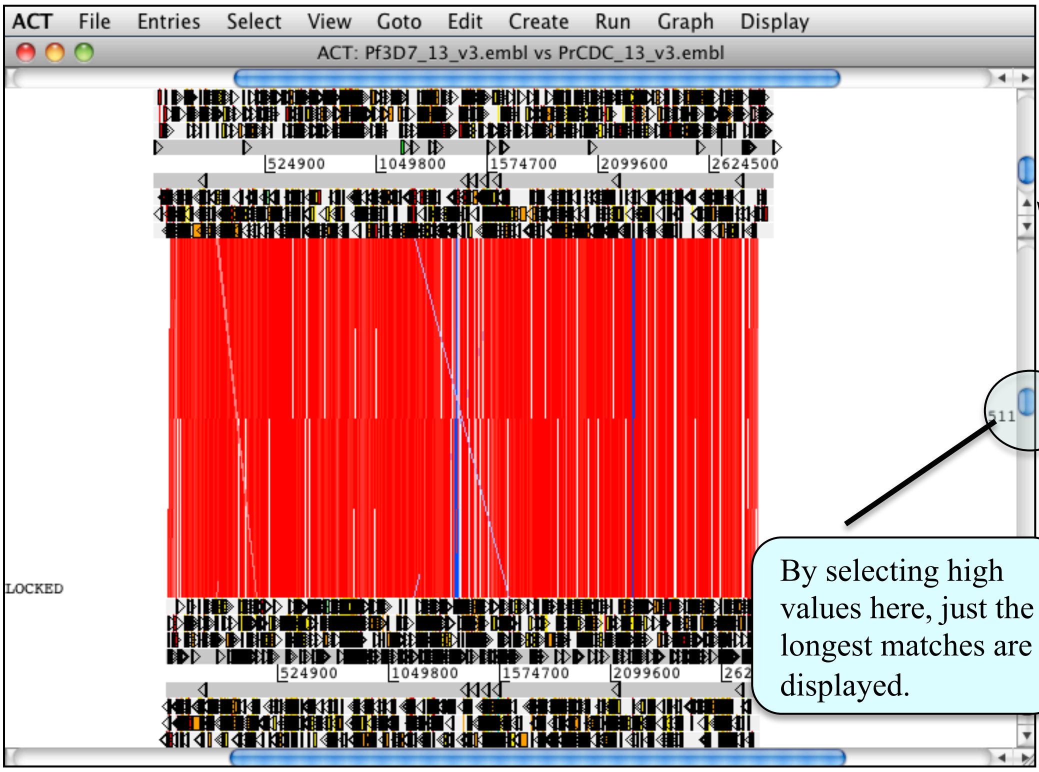
To open ACT you can also double click the ACT icon on your Desktop.

The files that you are going to need are:

Pf3D7_13_v3.embl	- *P. falciparum* chr13
Pf3D7_13_v3.fasta-txPrCDC_13.fasta.crunch	- tblastx comparison file
PrCDC_13_v2.embl	- *P. reichenowi* chr13

Once you have opened the files you will see a picture like this:



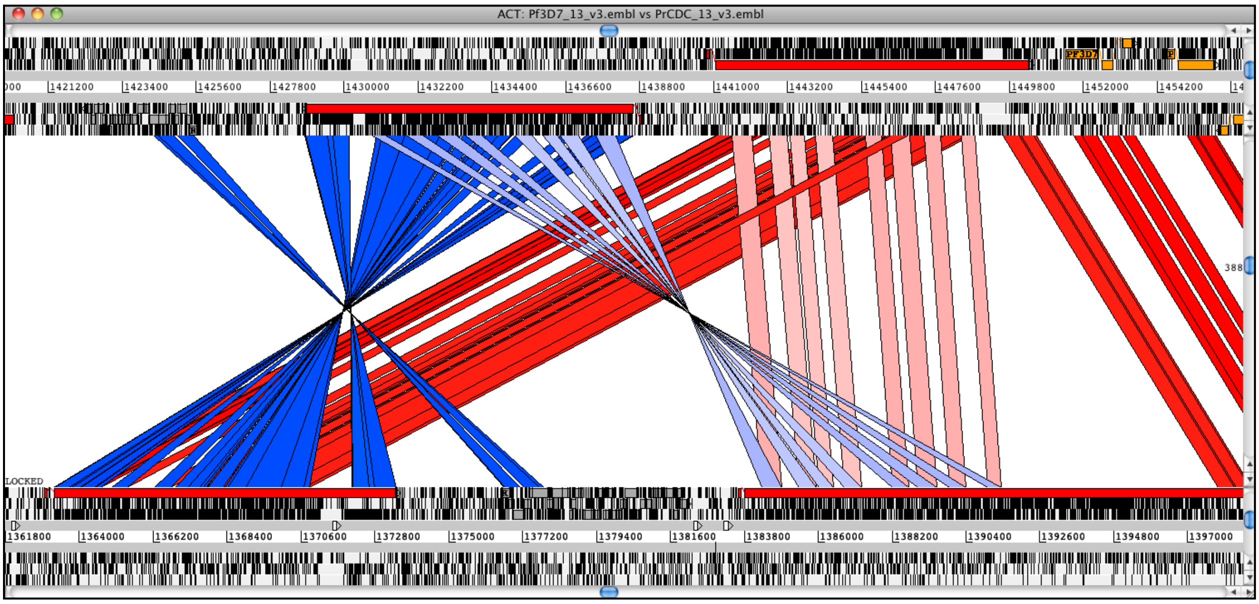

ACT File Entries Select View Goto Edit Create Run Graph Display
ACT: Pf3D7_13_v3.embl vs PrCDC_13_v3.embl
524900
1049800
1574700
2099600
2624500
LOCKED
511
524900
1049800
1574700
2099600
262

To get an overview use the vertical sliders to zoom out.

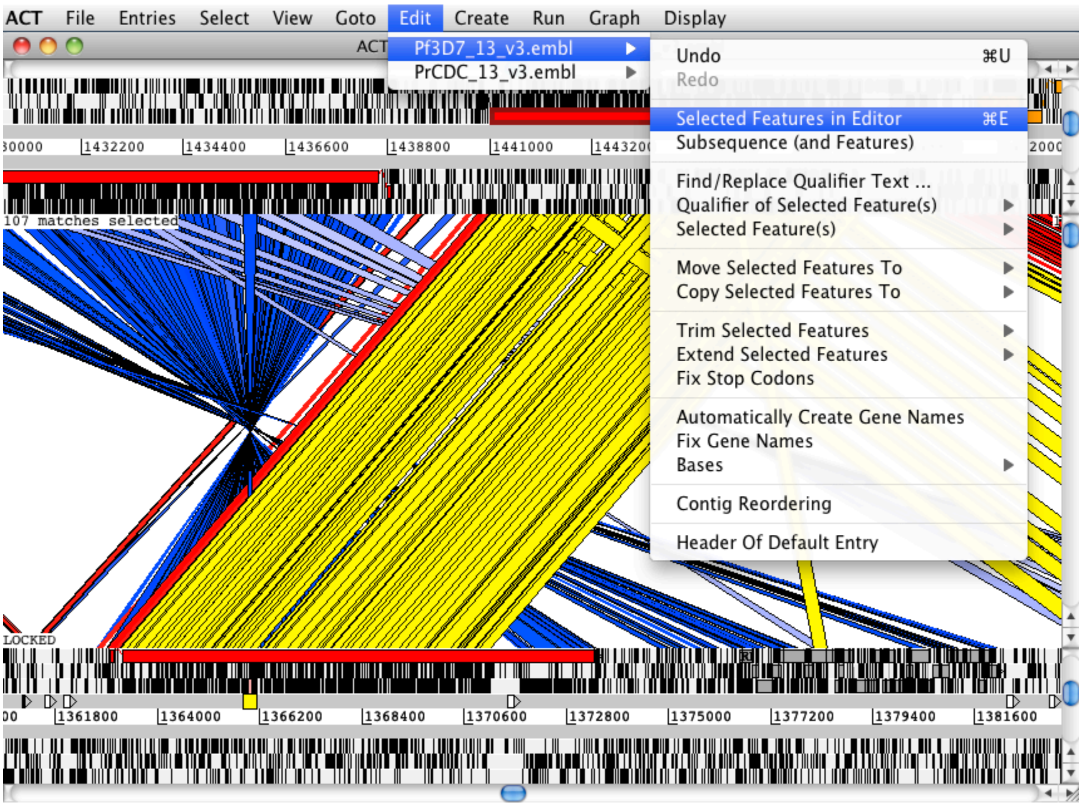
By selecting high values here, just the longest matches are displayed.

How many differences can you identify? Zoom in to see the genomes in more detail.

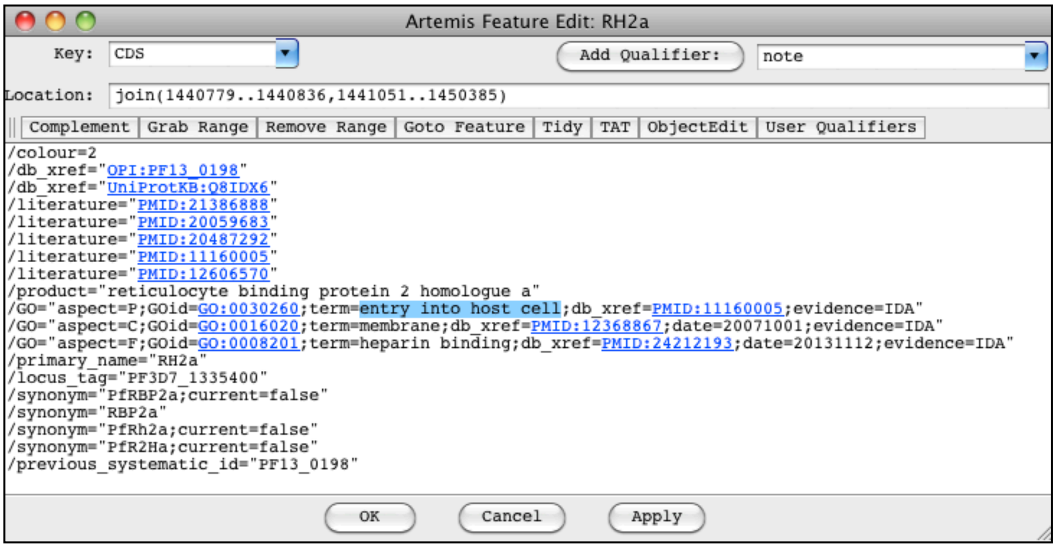
Have you already come across this area? Here are the coordinates: 1422300 - 1450390. Let’s have a more detailed look at the genes.



To look at the annotation, mark one of the genes in this area (e.g. PF3D7_1335400), go to ‘Edit’ and select ‘Selected Features in Editor’. As a shortcut you can just press ’E’.



What are the gene products? Are they important for host specificity? Does the Gene Ontology annotation give any additional information?



Key: CDS

Add Qualifier: note

Location: join(1440779..1440836,1441051..1450385)

```
/colour=2
/db_xref="OPI:PF13_0198"
/db_xref="UniProtKB:Q8IDX6"
/literature="PMID:21386888"
/literature="PMID:20059683"
/literature="PMID:20487292"
/literature="PMID:11160005"
/literature="PMID:12606570"
/product="reticulocyte binding protein 2 homologue a"
/GO="aspect=P;GOid=GO:0030260;term=entry into host cell;db_xref=PMID:11160005;evidence=IDA"
/GO="aspect=C;GOid=GO:0016020;term=membrane;db_xref=PMID:12368867;date=20071001;evidence=IDA"
/GO="aspect=F;GOid=GO:0008201;term=heparin binding;db_xref=PMID:24212193;date=20131112;evidence=IDA"
/primary_name="RH2a"
/locus_tag="PF3D7_1335400"
/synonym="PfRBP2a;current=false"
/synonym="RBP2a"
/synonym="PfRh2a;current=false"
/synonym="PfR2Ha;current=false"
/previous_systematic_id="PF13_0198"
```

More information about Gene Ontology can be found here: http://geneontology.org/

Optional exercise: Comparison of *Leishmania* spp genomes

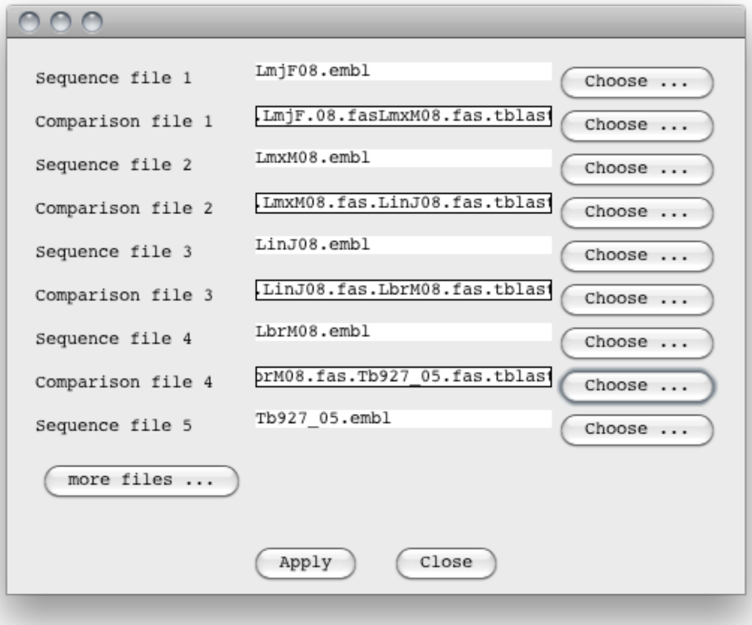
Leishmania are protozoan parasites that, depending on the species, cause a range of disease phenotypes ranging from self-curing lesions to large-scale destruction of facial tissue to potentially deadly visceral disease. In this exercise you will compare the genomes of 4 species:

1. *Leishmania major* (cutaneous, Old World) –the original reference genome species; high quality, manually improved sequence
2. *L. infantum* (visceral, Old World) – An human-improved draft, assembled *de novo* and aligned against *L. major*
3. *L. braziliensis* (mucocutaneous, New World) - An human-improved draft, assembled *de novo* and aligned against *L. major*
4. *L. mexicana* (cutaneous, New World) - An human-improved draft, assembled *de novo and* aligned against *L. major*

Close the previous ACT session.

Restart ACT by typing act & on the command line or double click the icon on the Desktop.

Load sequence and comparison files for 4 *Leishmania* species plus *Trypanosoma brucei* (outgroup)



Sequence file 1 LmjF08.embl

Comparison file 1 LmjF.08.fasLmxM08.fas.tblas

Sequence file 2 LmxM08.embl

Comparison file 2 LmxM08.fas.LinJ08.fas.tblas

Sequence file 3 LinJ08.embl

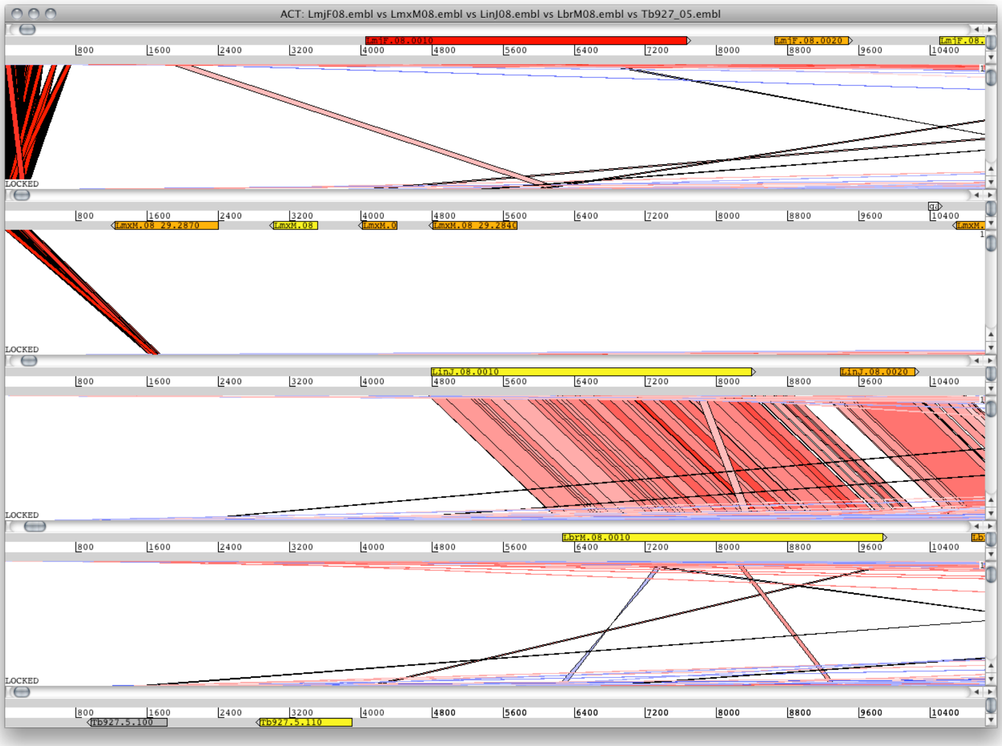
Comparison file 3 LinJ08.fas.LbrM08.fas.tblas

Sequence file 4 LbrM08.embl

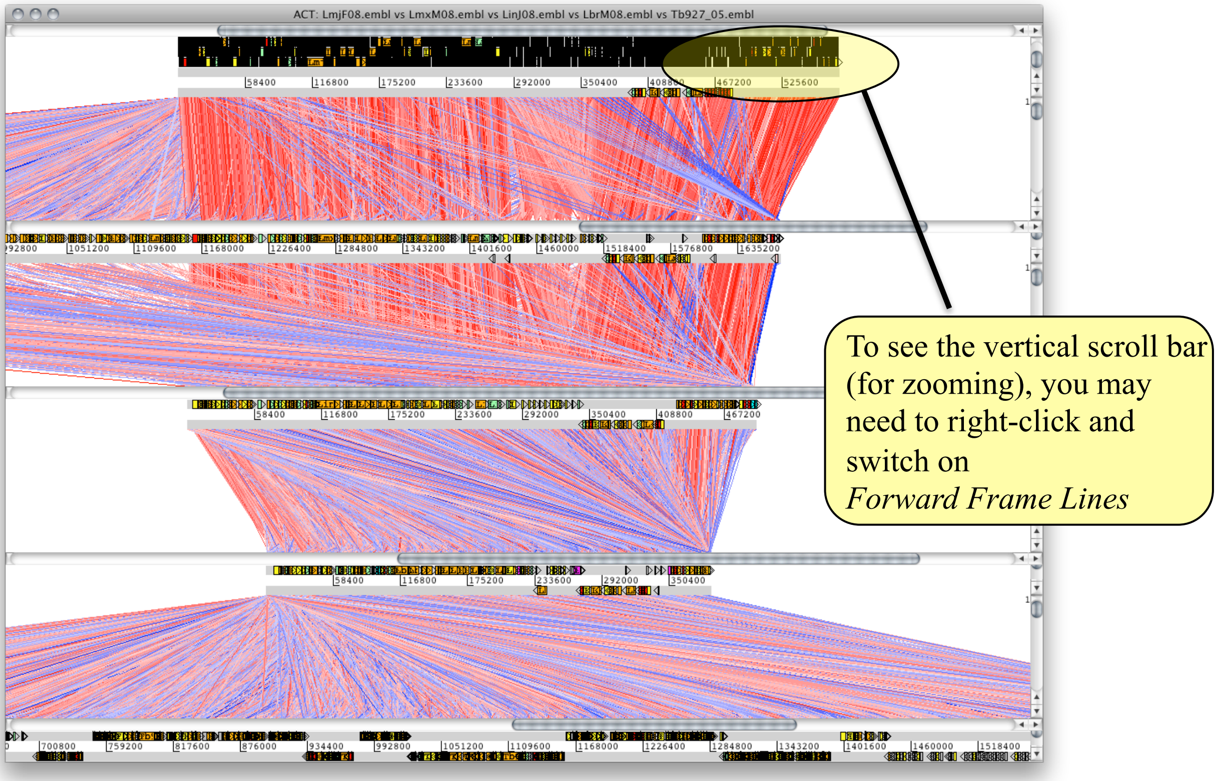
Comparison file 4 brM08.fas.Tb927_05.fas.tblas

Sequence file 5 Tb927_05.embl

Your ACT window should now look something like this:

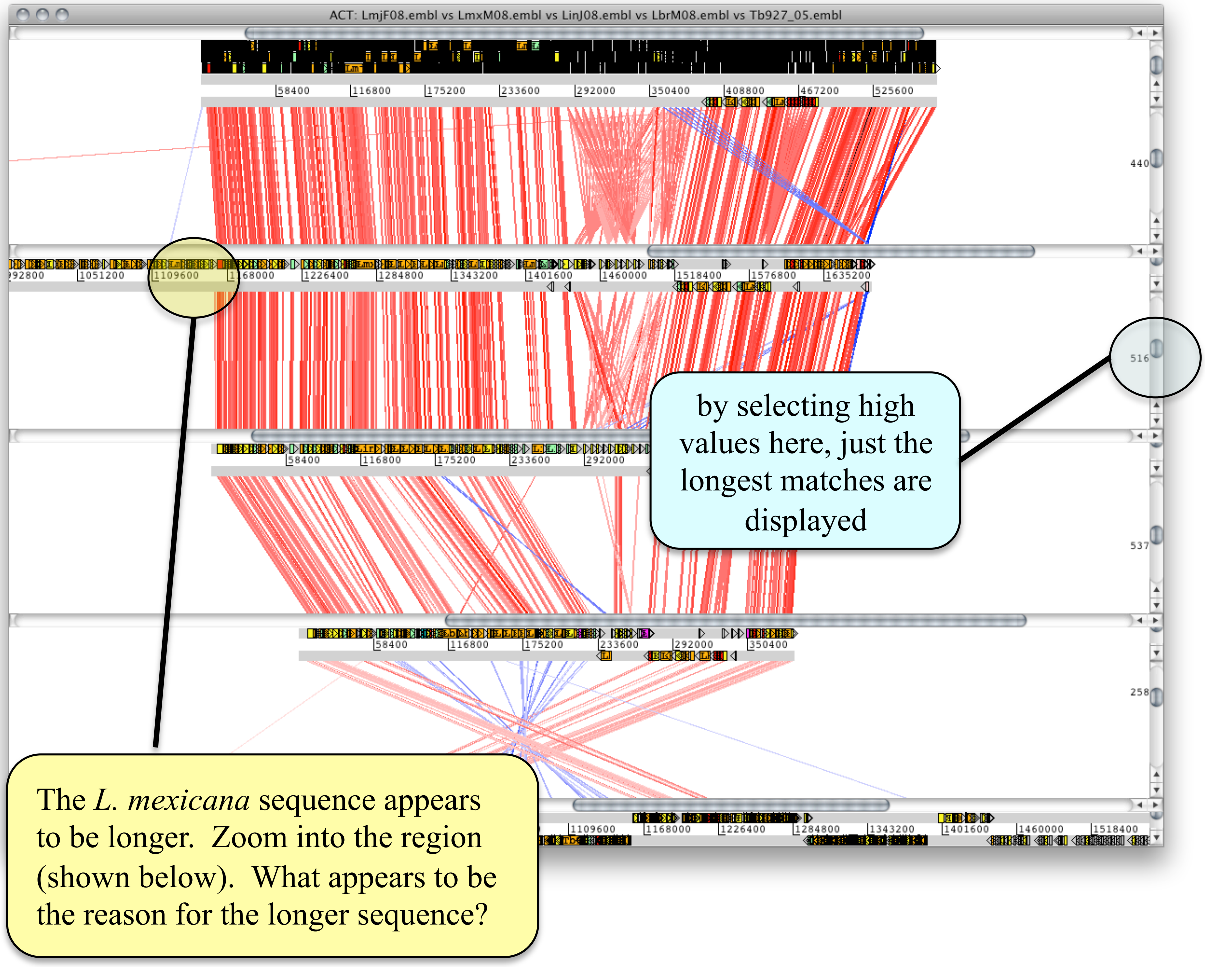

ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl
LOCKED
LinJ.08.0010
LinJ.08.0020
LbrM.08.0010
Tb927.5.100
Tb927.5.110

Zoom out (and scroll) as appropriate to identify the regions of conserved synteny.


ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl

To see the vertical scroll bar (for zooming), you may need to right-click and switch on *Forward Frame Lines*

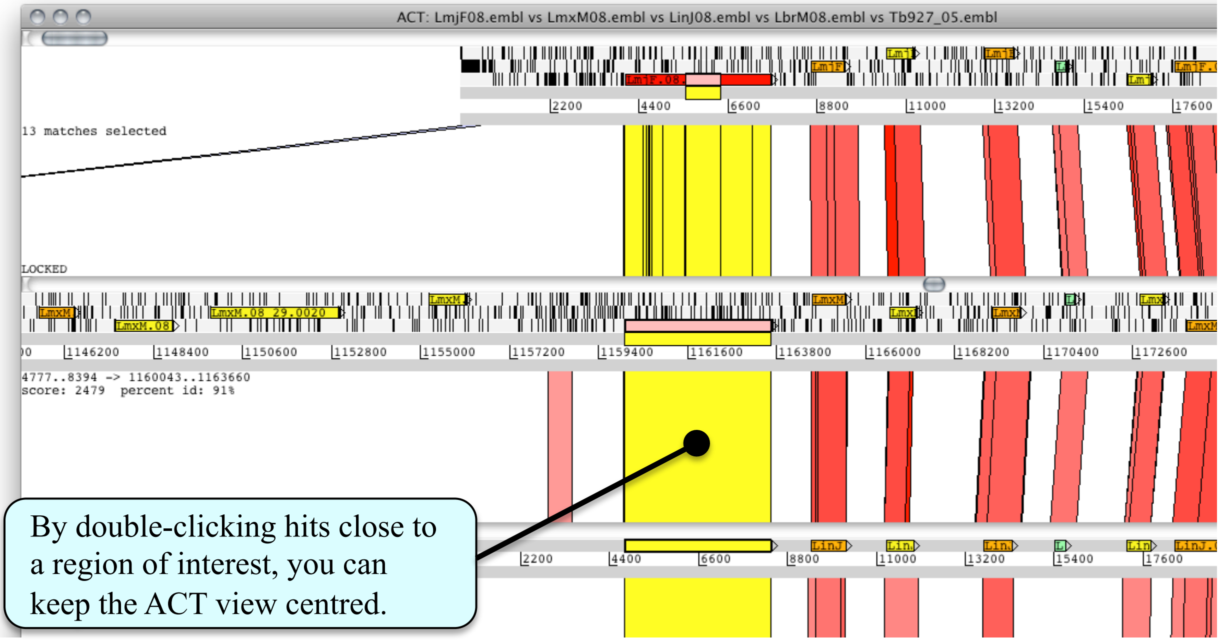
Use the filters (percentage identity, score or length of match) to adjust the signal:noise ratio



ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl

by selecting high values here, just the longest matches are displayed

The *L. mexicana* sequence appears to be longer. Zoom into the region (shown below). What appears to be the reason for the longer sequence?



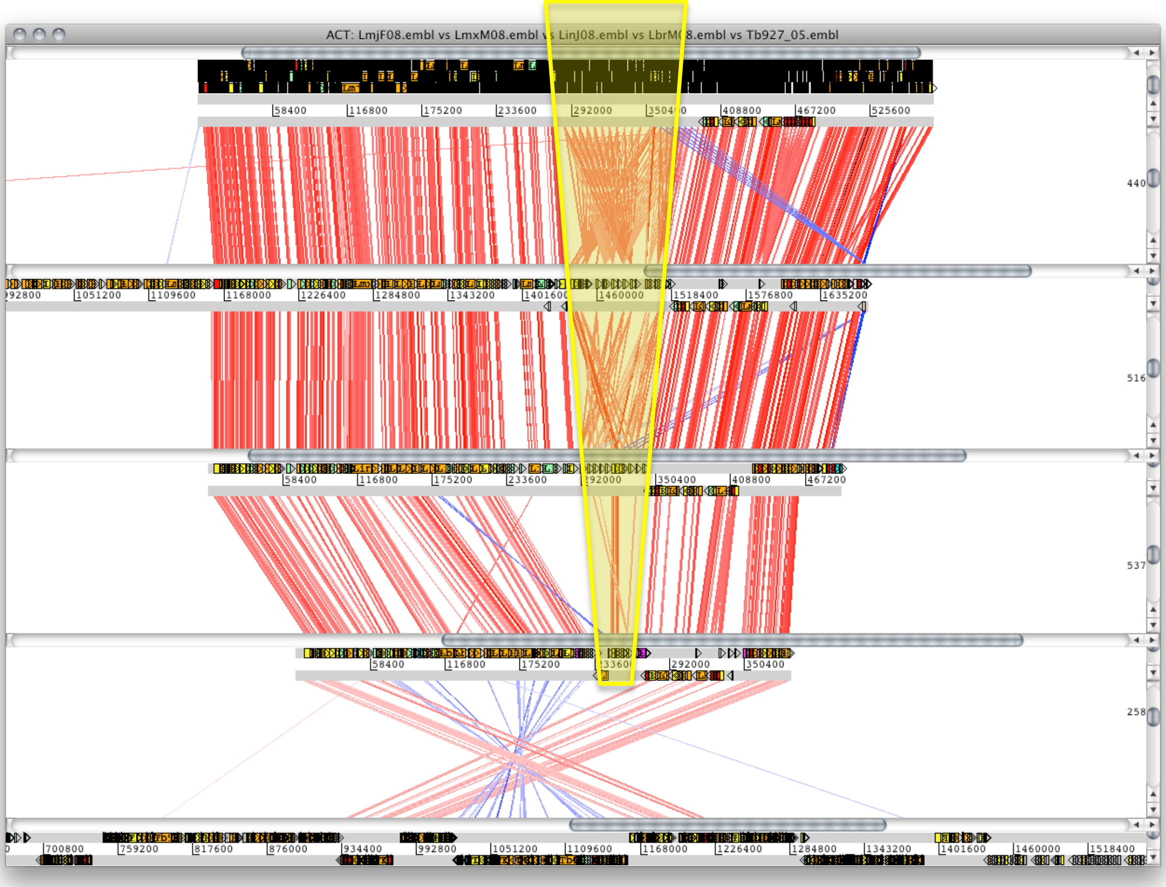
ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl

```
13 matches selected
```

```
LOCKED
```

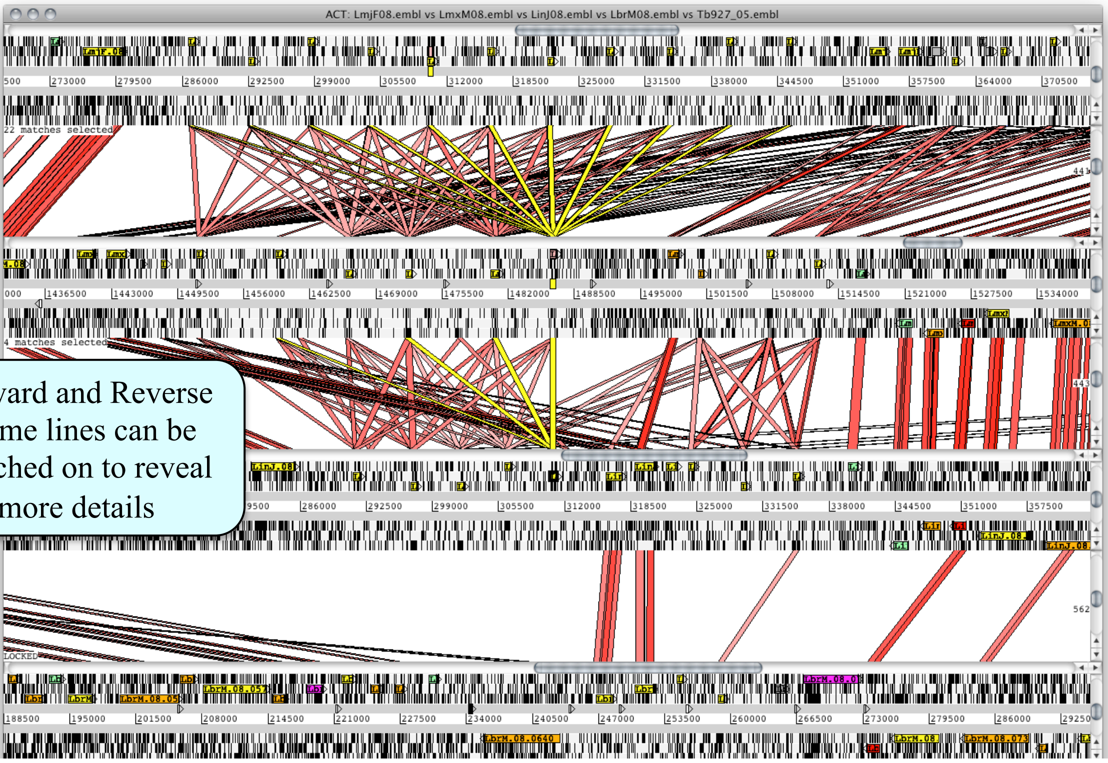
```
4777..8394 -> 1160043..1163660
score: 2479  percent id: 91%
```

By double-clicking hits close to a region of interest, you can keep the ACT view centred.



ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl

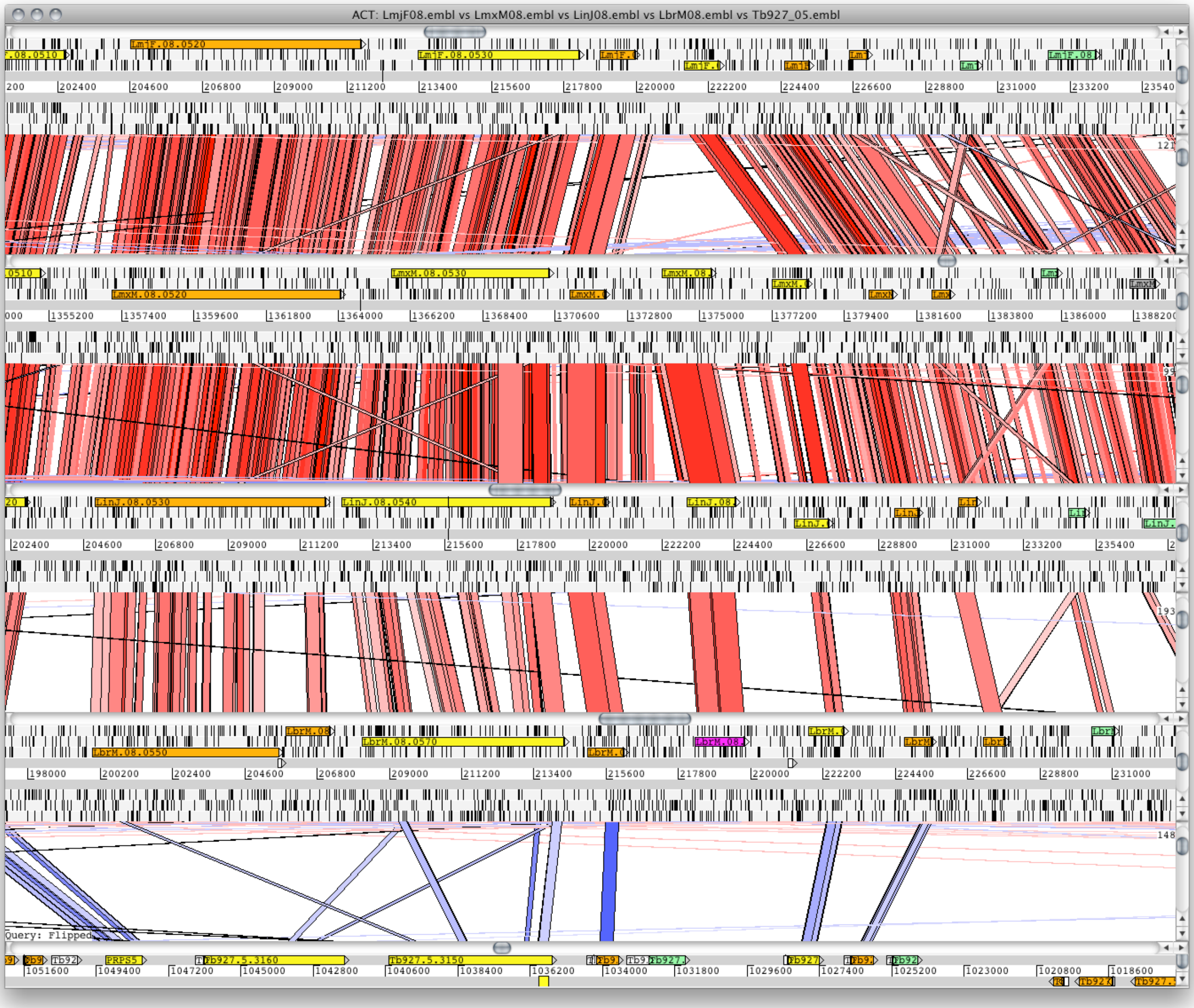
Several regions appear different sizes across different species. One is shown above and below. What is a likely reason?



ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl

Forward and Reverse Frame lines can be switched on to reveal more details

Try to identify this species-specific difference. What is this gene? What additional information can the *T. brucei* outgroup provide about this locus? Can you identify other species-specific differences?


ACT: LmjF08.embl vs LmxM08.embl vs LinJ08.embl vs LbrM08.embl vs Tb927_05.embl
LmjF.08.0520
LmjF.08.0530
LmxM.08.0520
LmxM.08.0530
LinJ.08.0530
LinJ.08.0540
LbrM.08.0550
LbrM.08.0570
PRPS5
Tb927.5.3160
Tb927.5.3150
Query: Flipped