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, and so you will already be familiar with many of its core functions, and 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. Appendix V shows you how to generate comparison files from BLAST. Whilst having a local copy BLAST to generate ACT comparison files can be very useful, it means that you are tied to a particular computer. Another way of generating comparison files for ACT is to use the WebACT web resource (see page 16 of this module). This site allows you to cut and paste or upload your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files.

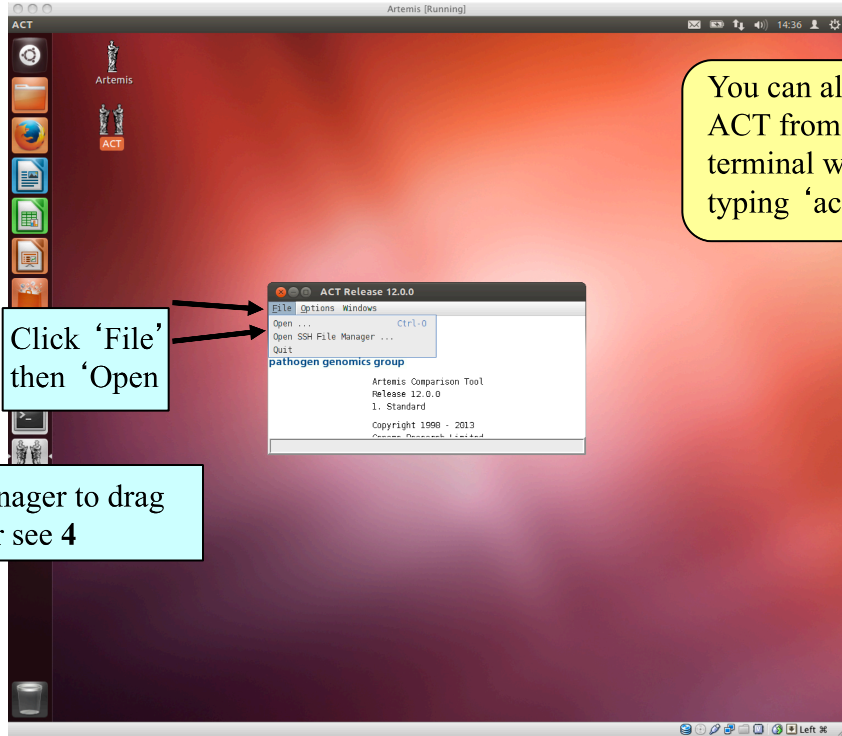
Aims

The aim of this Module is for you to become familiar with the basic functions of ACT by using a series of worked examples. Some of these examples will touch on exercises that were used in previous Modules, this is intentional. Hopefully, as well as introducing you to the basics of ACT, this Module will also show you how ACT can be used for not only looking at genome evolution but also to back up, or question, gene models and so on. In this module you will also use a web resource, WebACT, to generate your own comparison files and view them in ACT.

1. Starting up the ACT software

Double click the ACT icon on the desktop. A small start up window will appear.

The files you will need for this exercise are: *S_typhi.dna.gz*
S_typhi.dna_vs_EcK12.dna.crunch.gz
EcK12.dna.gz

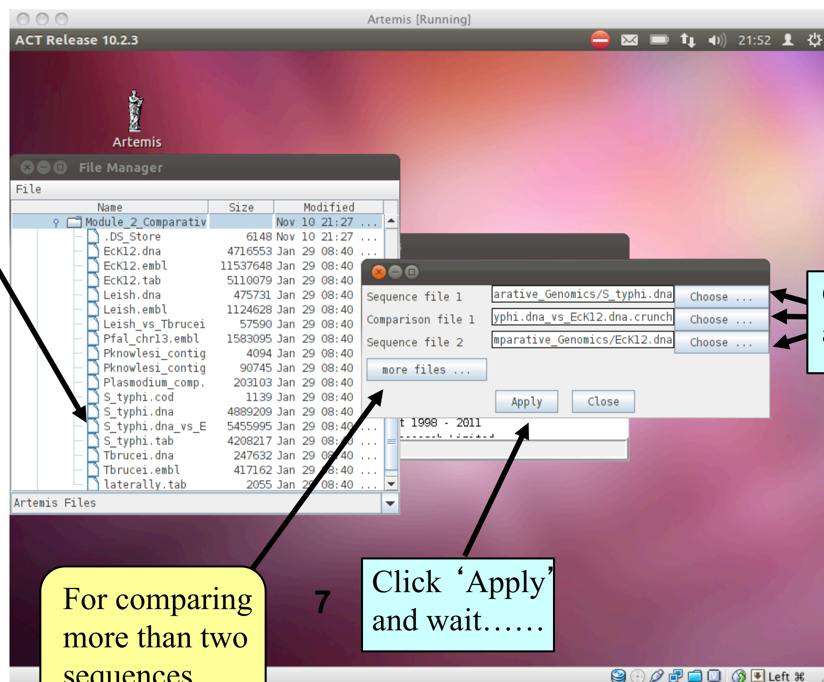


You can also start ACT from the terminal window by typing 'act'

1 **2** Click 'File' then 'Open'

3 Use the File manager to drag and drop files or see **4**

Comparison files end with '.crunch.gz'



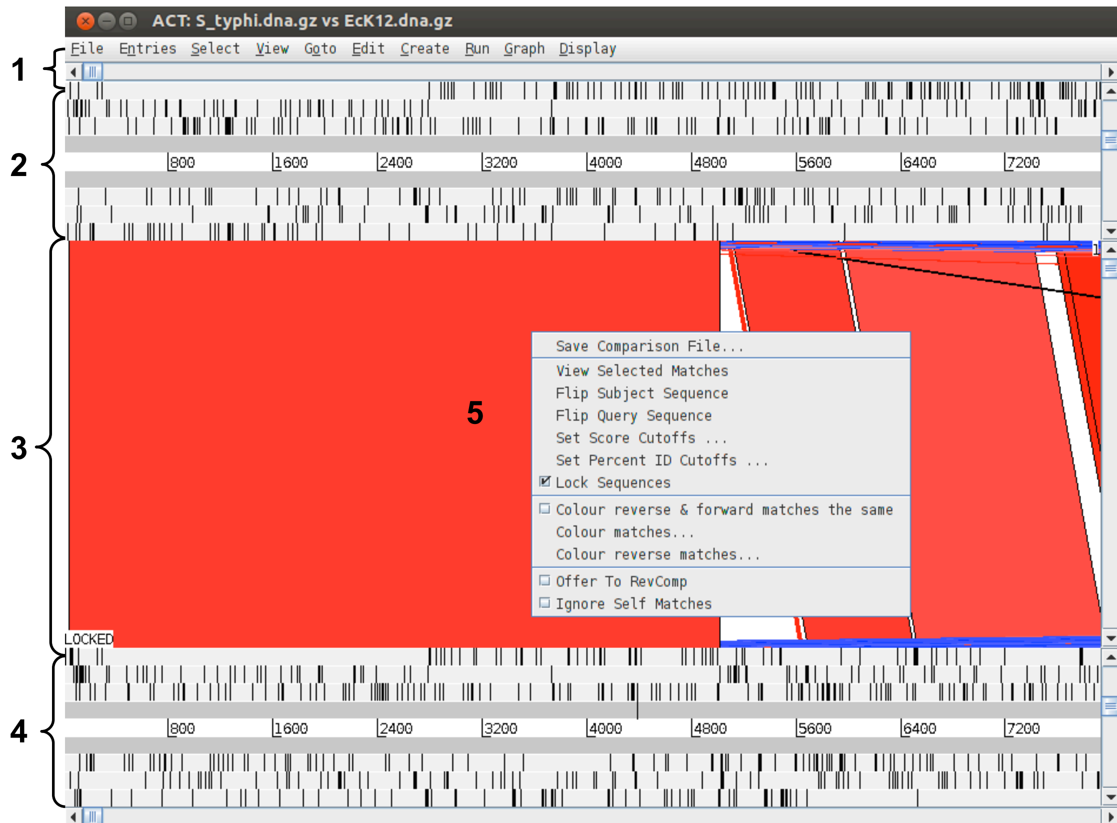
4, 5 & 6 Click and select appropriate files

For comparing more than two sequences

7 Click 'Apply' and wait……

2. The basics of ACT

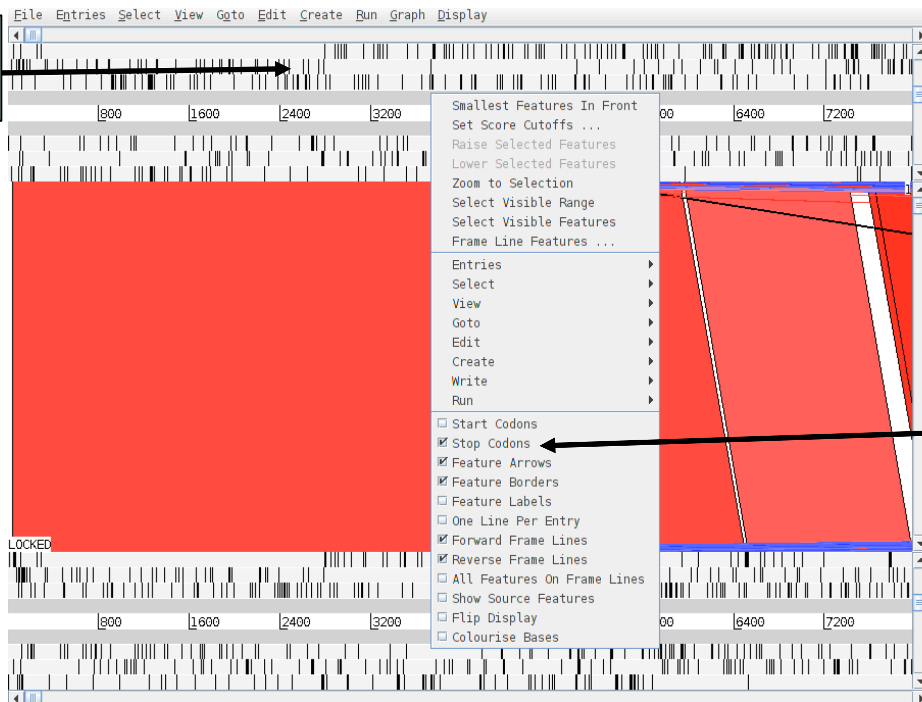
You should now have a window like this so let's see what is there.



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.

1

Right button click here



2

De-select stop codons

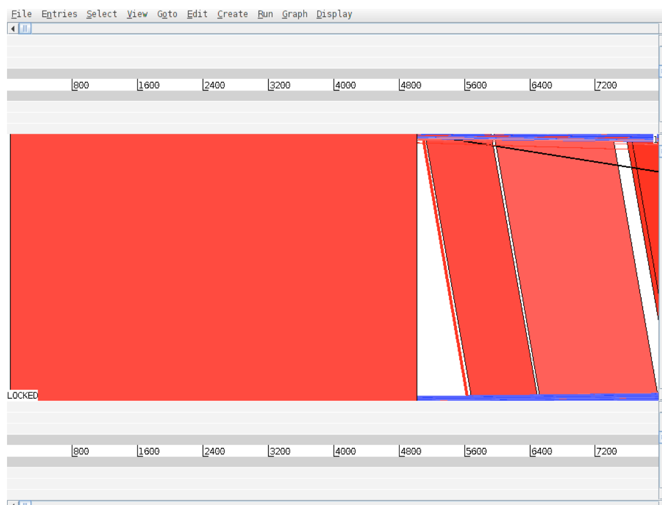
3. Exercise 1

Introduction & Aims

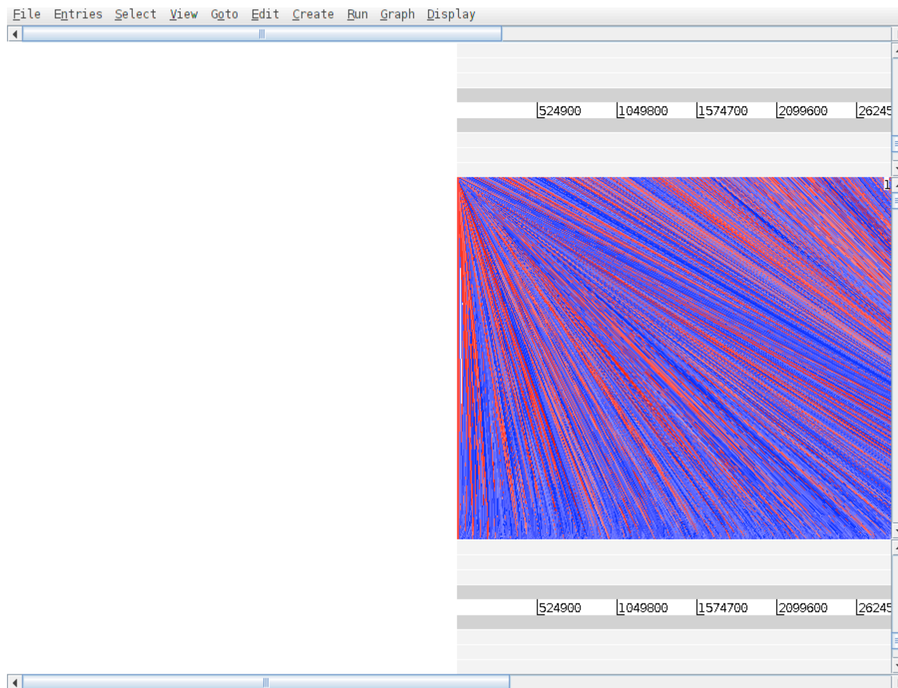
In this first exercise we are going to explore the basic features of ACT. Using the ACT session you have just opened we firstly are going to zoom outwards until we can see the entire *S.* Typhi chromosome compared against the *E. coli* K12 chromosome. As for the Artemis exercises we should turn off the stop codons to clear the view and speed up the process of zooming out.

The only difference between ACT and Artemis when applying changes to the sequence views is that in ACT you must click the right mouse button over the specific sequence that you wish to change, as shown above.

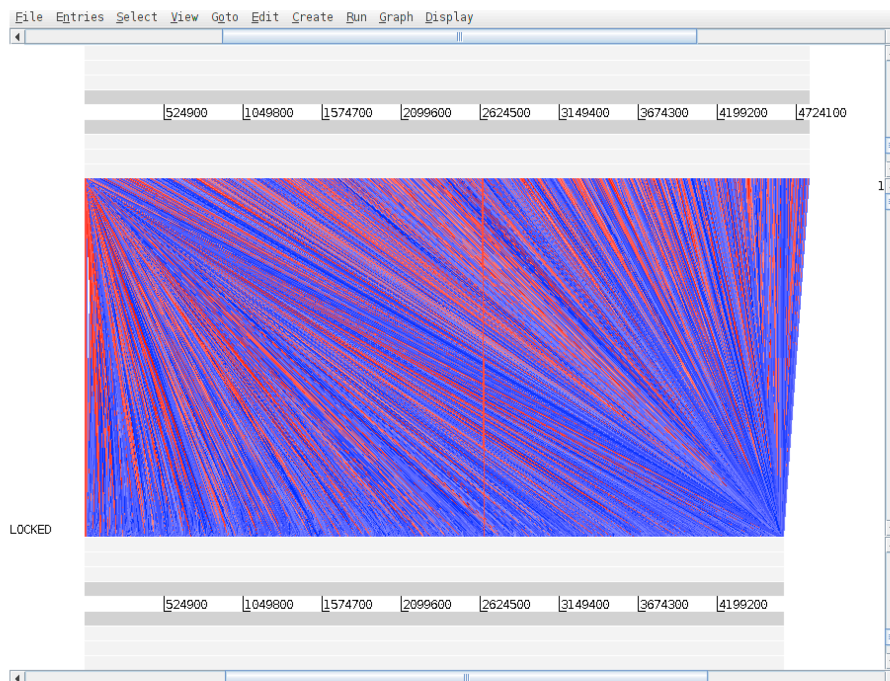
Now turn the stop codons off in the other sequence too. Your ACT window should look something like the one below:



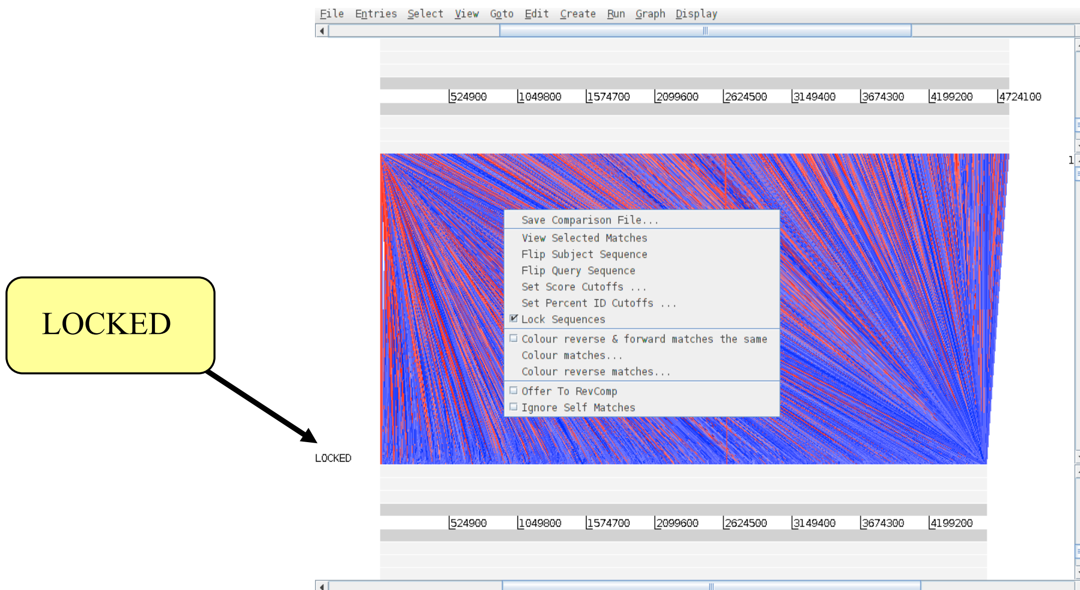
Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in synch.



Once zoomed out, your ACT window should look similar to the one shown above. If the genomes in view fall out of view to the right of the screen, use the horizontal sliders to scroll the image and bring the whole sequence into view, as shown below. You may have to play around with the level of zoom to get the whole genomes shown in the same screen as shown below.



Notice that 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



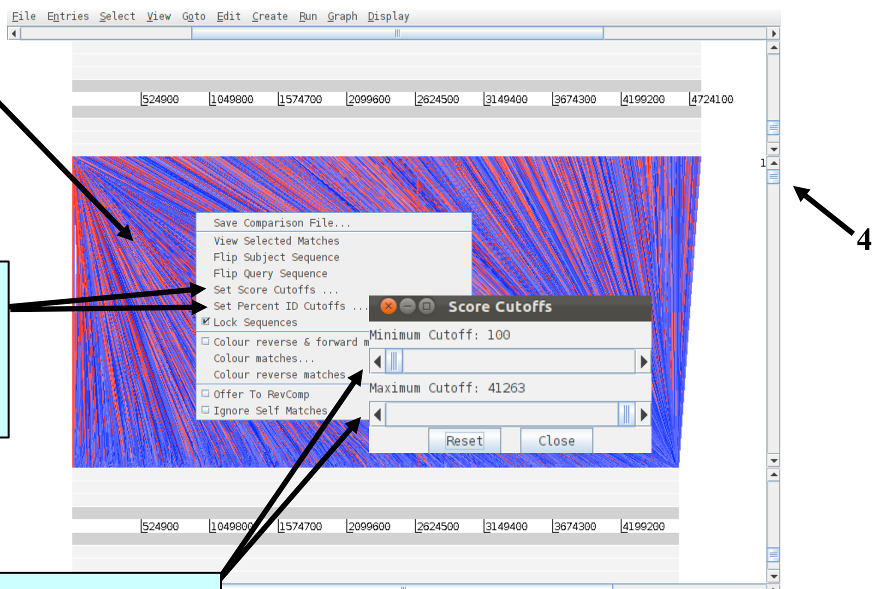
You can optimise your image by either removing ‘low scoring’ (or percentage ID) hits from view, as shown below **1-3** or by using the slider on the the comparison view panel **(4).** The slider allows you to filter the regions of similarity based on the length of sequence over which the similarity occurs,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

1

Right button click in the Comparison View panel

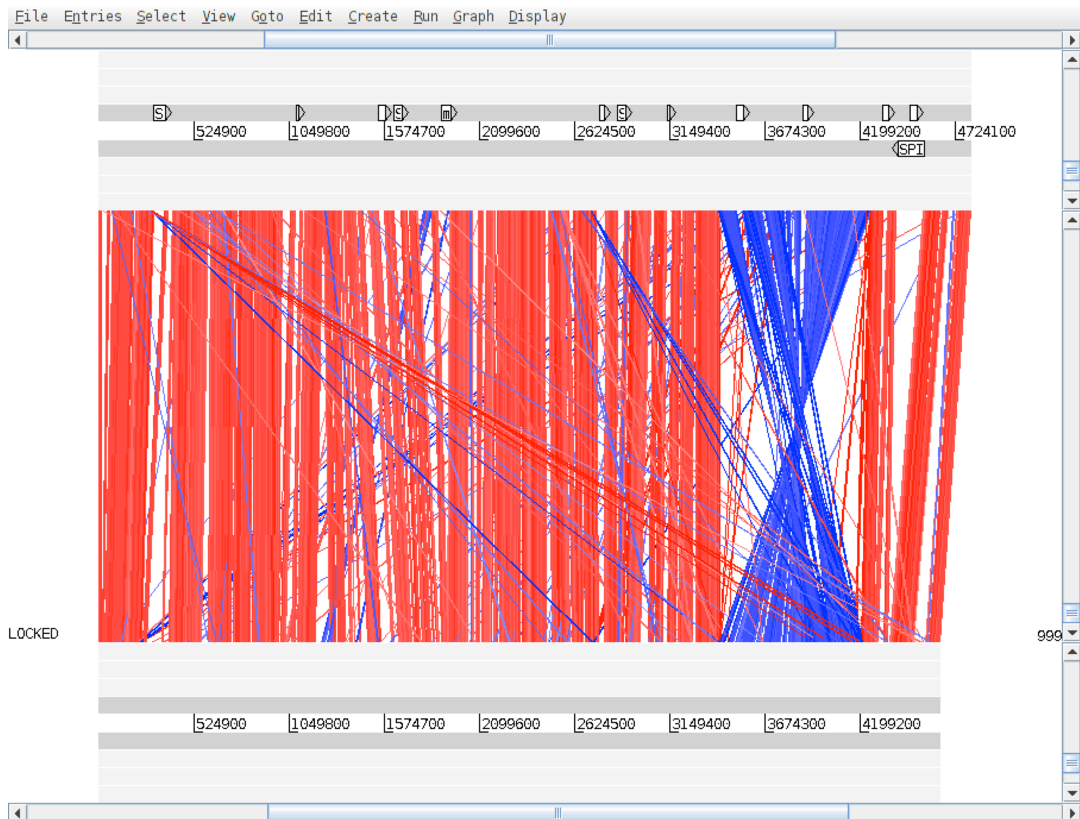
2

Select either **Set Score Cutoffs** or **Set Percent ID Cutoffs**



3

Move the sliders to manipulate the comparison view image



4. Things to try out in ACT

Load into the top sequence (*S. typhi*) a '.tab' file called 'laterally.tab.gz'. You will need to use the 'File' menu and select the correct genome sequence ('S.typhi.dna.gz') before you can read in an entry. If you are zoomed out and looking at the whole of both genomes you should see the above. The small white boxes are the regions of atypical DNA covering regions that we looked at in the first Artemis exercise. It is apparent that there is a backbone sequence shared with *E. coli* K12, plus chunks of *S.* Typhi specific DNA, which appear to be insertions relative to *E. coli* K12.

5. More things to try out in ACT

1. Double click red boxes to centralise them.
2. Zoom right in to view the base pairs and amino acids of each sequence.
3. Load annotation files into the sequence view panels.
4. Use some of the other Artemis features e.g., graphs etc.
5. Find an inversion in one genome relative to the other then flip one of the sequences.

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

Exercise 2

P. falciparum and *P. knowlesi:* Genome Comparison

Introduction

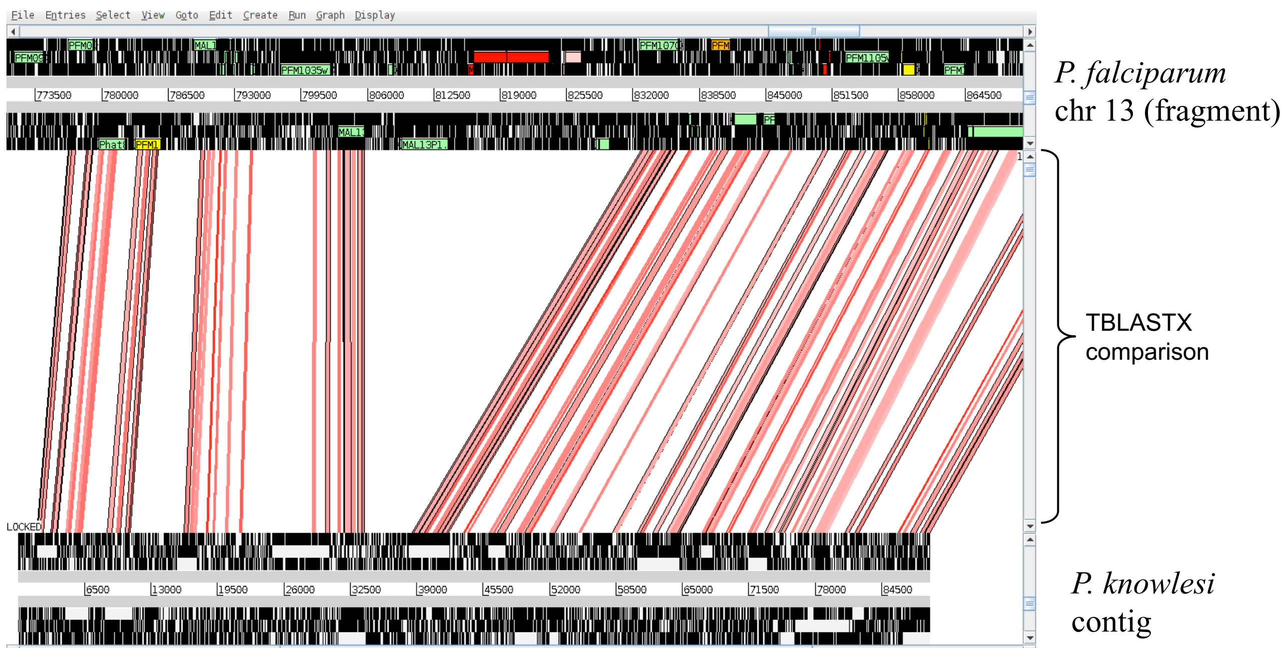
The annotation and analysis of the whole genome of *P. falciparum* 3D7 has been completed and genome sequences of several other malaria parasites are also available. This allows us to perform comparative analysis of the genomes of malaria parasites and understand the basic biology of their parasitism, based on the similarities / dissimilarities between the parasites at DNA / protein level.

Aim

You will be looking at the comparison between a genomic DNA fragment of the primate malaria *P. knowlesi* and the previously annotated chromosome 13 of *P. falciparum*. By comparing the two genomic sequences you will be able to study the degree of conservation of gene order (i.e., synteny) and identify genes in *P. knowlesi* genome. As part of the exercise you will also identify an unique region between the two genomic fragments and finally modify the gene model of a multi-exon gene in *P. knowlesi,* using ACT.

The files that you are going to need are:

Pfal_chr13.embl.gz	- *P. falciparum* annotation file with sequence
Pknowlesi_contig.seq.gz	- *P. knowlesi* DNA file (without annotation)
Pknowlesi_contig.embl.gz	- *P. knowlesi* annotation file
Plasmodium_comp.crunch.gz	- TBLASTX comparison file

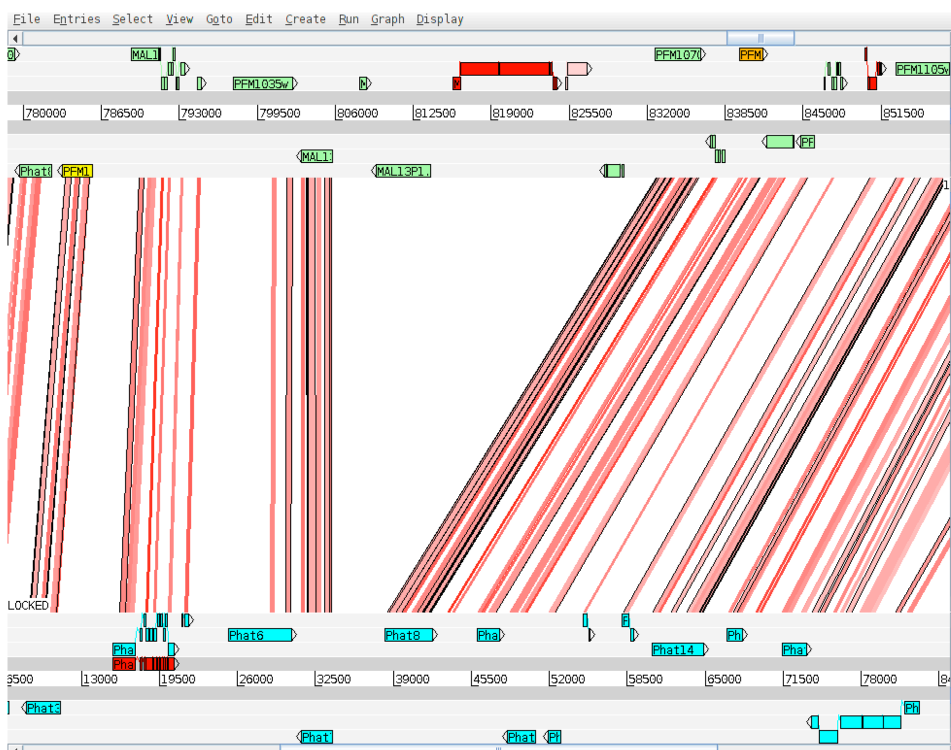


Comparison of *P. knowlesi* contig and the annotated chromosome 13 fragment of *P. falciparum*

Exercise 2 Part I

Conservation of gene order (synteny)

- In the ACT start up window load up the files Pfal_chr13.embl.gz, Pknowlesi_contig.seq.gz and the comparison file Plasmodium_comp.crunch.gz. Add the annotation file 'Pknowlesi_contig.embl.gz' to the Pknowlesi_contig.seq.gz sequence.
- Use the slider on either sequence view panel to obtain a global view of the sequence comparison. Also use the slider on the comparison view panel to remove the 'shorter' similarity hits. What effects does this have?
- Can you see conserved gene order between the two species?
- Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within this unique region in file: Pfal_chr13.embl.gz (top sequence)
- Example location: Pfal_chr13.embl.gz, 815823..829969
- What are the predicted products of the genes assigned to this unique location? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.
- Can you identify genes in conserved regions that have not been annotated in the *P. knowlesi* contig, but are present in the *P. falciparum* chromosome 13? This will allow you to see any potential protein coding regions.
- Any thoughts about the possible biological relevance of the comparison?



P. falciparum
Pfal_chr13.embl.gz

P. knowlesi
Pknowlesi_contig.embl.gz

Exercise 2 Part II

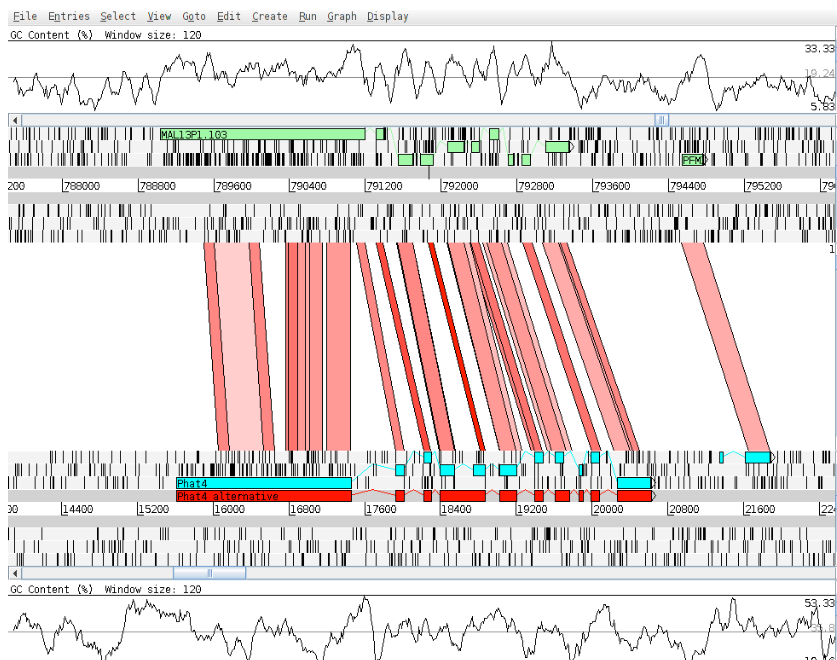
Prediction of gene models:

There are several computer algorithms, covered earlier in the course, that predict gene models, based on training the algorithm with previously known gene sets with previously known experimentally verified exon-intron structures (in eukaryotes). However, no single programme can predict the gene structure with 100% accuracy and one needs to curate / refine the gene models, generated by automated predictions. We have generated automated gene models for the *P. knowlesi* contig, using PHAT (Pretty Handy Annotation Tool, a gene finding algorithm, see in Mol. Biochem. Parasitol. 2001 Dec;118(2):167-74) and the automated annotations are saved in Pknowlesi_contig.embl.gz.

- Zoom into the *P. falciparum* gene labelled PFM1010w shown below. Can you compare the 2 gene models and identify the conserved exon(s) between the 2 species?
- Use the slider on the comparison view panel to include some 'shorter' similarity hits. Can you now identify all the conserved exons of the PFM1010w orthologue in the *P. knowlesi* contig? (For the time being, disregard the misc_feature for 'Phat4', coloured in red in the 'Pknowlesi_contig.embl.gz' file)
- Open the '*GC Content (%)*' window from '*graph*' menu for both the entries. Can you relate the exon-intron boundaries to GC-content for the *P. falciparum* gene labelled PFM1010w? Is it also applicable to the gene model 'Phat4' in the *P. knowlesi* contig?
- Example regions:

  Pfal_chr13.embl.gz, 789034..793351

  Pknowlesi_contig.embl.gz, 15618..20618



P. falciparum
Pfal_chr13.embl.gz

P. knowlesi
Pknowlesi_contig.embl.gz

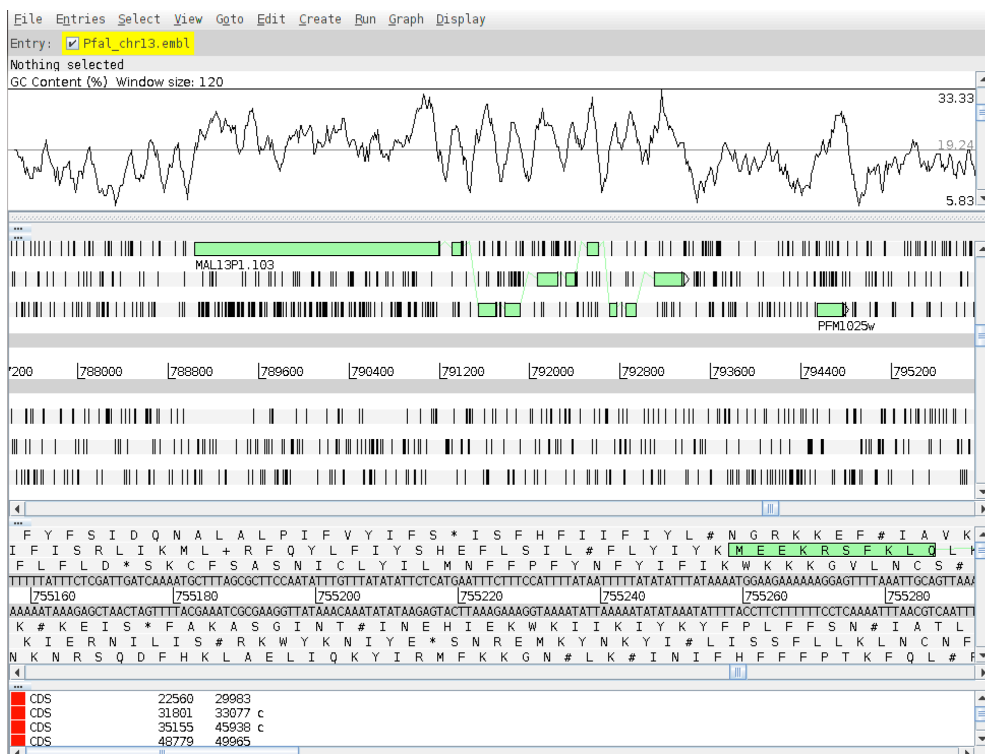
Comparison between orthologous genes in *P. falciparum* and *P. knowlesi*

Exercise 2 Part III (OPTIONAL)

Gene models for multi-exon genes in *P. falciparum*:

- Use '*File*' menu to select entry 'Pfal_chr13.embl.gz' and select '*Edit In Artemis*' to bring up an Artemis window.
- In Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)*' window.
- Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This Qualifier Value*' and type 'PFM1010w', click then close the dialogue box.
- Go through the annotated gene model for 'PFM1010w' and have a look at the the exon-intron boundaries and compare with the splice site sequences from *P. falciparum* given in **Appendix XI**.
- Also have a glance through a few other gene models for multi-exon genes and have a look at the intron sequences as well. Can you find any common pattern in the putative intron sequences? Hint – look at the complexity of the sequence
- You can delete exon(s) of any gene by selecting the exon(s) and then choosing '*Delete Selected Exons*' from '*Edit*' menu. Similarly, you can add an exon to a particular gene by co-selecting the exon and the gene (CDS features) followed by selecting '*Merge Selected Features*' from the '*Edit*' menu.

- Example regions:
  Pfal_chr13.embl.gz, 789034..793351, 657638..660023, 672361..673753



Example location: 789034..793351, in Pfal_chr13.embl.gz

Exercise 2 Part IV (OPTIONAL)

Curation of gene models in *P. knowlesi*:

We are now going to edit the gene model for *P. knowlesi*.

•Use '*File*' menu from the ACT displaying *P. falciparum* and *P. knowlesi to* select entry 'Pknowlesi_contig.embl.gz' and select '*Edit In Artemis*' to bring up an Artemis window.

•Within the Artemis window, use '*Graph*' menu and switch 'on' the '*GC Content (%)*' window.

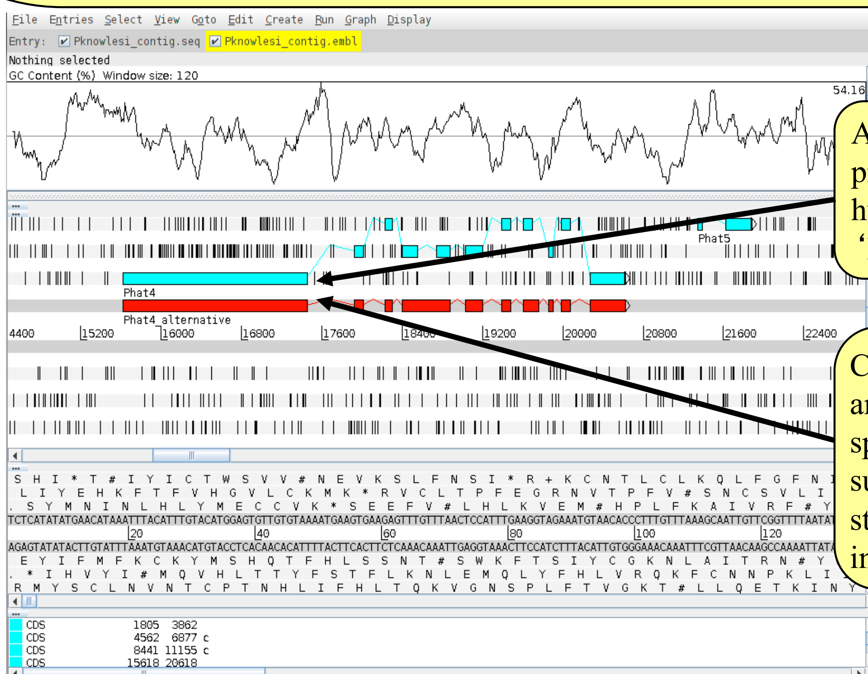
•Use '*Goto*' menu to select '*Navigator*' window and within the Navigator window, select '*Goto Feature With This Qualifier Value*' and type 'Phat4'.

•Go to the first ACT window (first small window that appears when starting up ACT), and use the '*Options*' menu to select '*Enable Direct Editing*'

•Go through the gene model of 'Phat4' and have a glance through the exon-intron boundaries. Can you suggest any alternative gene model, after consulting the Table provided in **Appendix XI**, containing several examples of experimentally verified splice site sequences for *P. falciparum*?

•Example modifications:

Have a look at the '*misc_feature*', coloured in red (location: 15618..20618). Can you spot any difference in the red gene model of 'Phat4' at the exon-intron boundaries? Select the red feature, click on '*Edit*' menu and select '*Edit Selected Features*' and in the new window that pops out, change the 'Key' from misc_feature to 'CDS' and click on 'OK' button to close the window. Now you can compare the automatically created blue gene model and the curated red gene models at protein level and predict any alternative splicing pattern.



Automated gene prediction for hypothetical gene 'Phat4'

Can you suggest any alternative splicing pattern such as the gene structure shown in red?

Example location: 15618..20618, in Pknowlesi_contig.embl.gz

Exercise 3

Introduction

Having familiarised yourselves with the basics of ACT, we are now going to use it to look at a region of synteny between *T. brucei* and *Leishmania*.

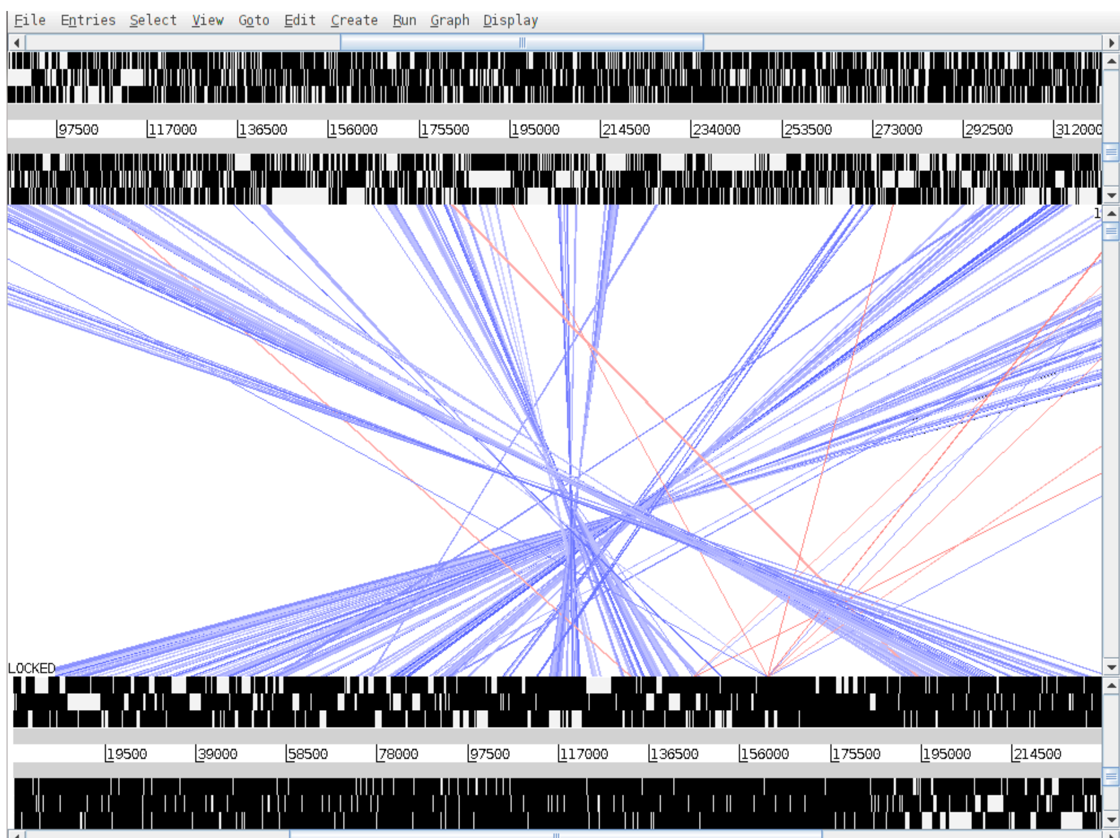
Aim

By looking at a comparison of the annotated sequences of *T. brucei* and *L. major* you will be able to analyse, in detail, those genes that are found in both organisms as well as spot the differences. You will also see how ACT can be used to study the different chromosome architecture of these two parasite species.

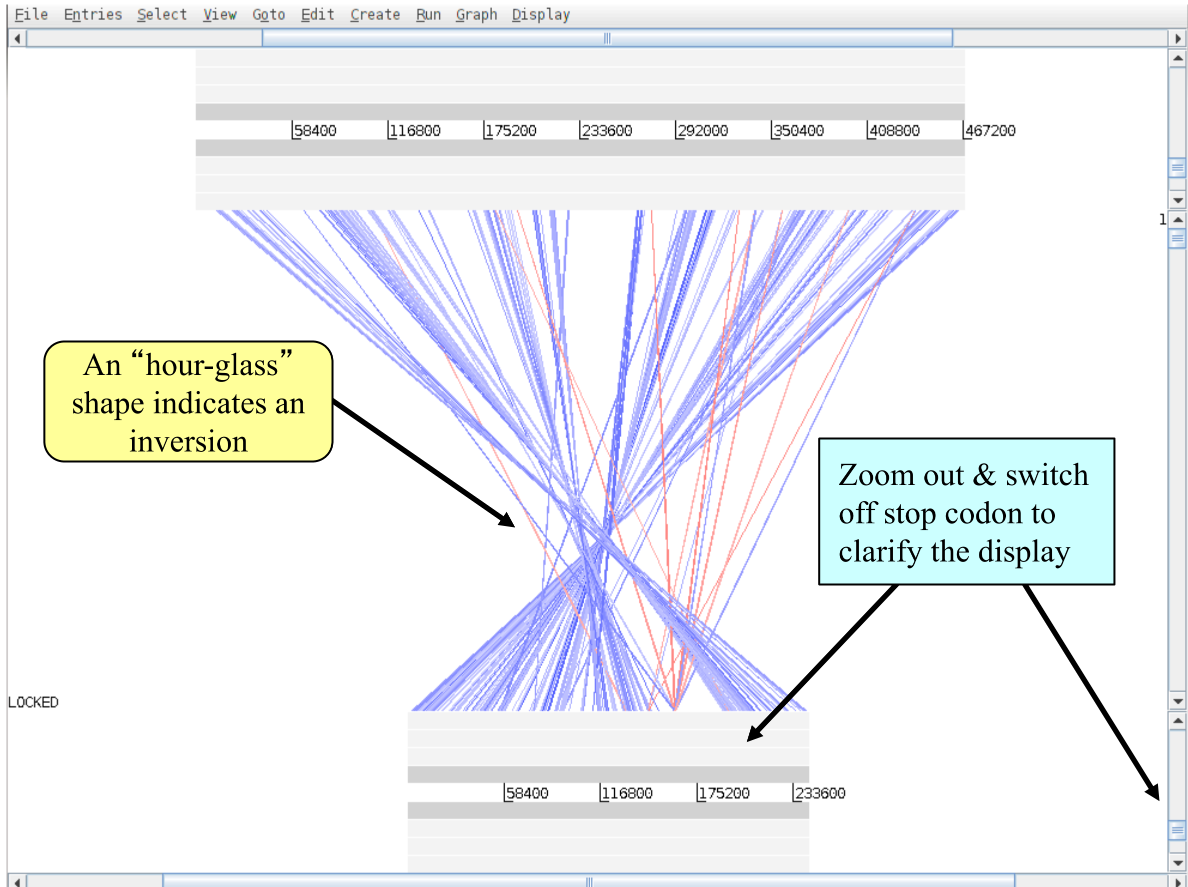
The files that you are going to need are:

File	Description
Tbrucei.dna.gz	- *T. brucei* sequence
Tbrucei.embl.gz	- *T. brucei* annotation
Leish_vs_Tbrucei.tblastx.gz	- comparison file
Leish.dna.gz	- *L. major* sequence
Leish.embl.gz	- *L. major* annotation

First, load up the sequence files for *T. brucei* and *L. major* and the comparison file in ACT.



Next, you need to find the regions of synteny between the sequences.



When you have determined where there is synteny, zoom in to the region for a detailed look. At this point you can add the annotation from the files called **Leish.embl.gz** and **Tbrucei.embl.gz**.

Can you see conserved gene order between the two species?

Can you see any region where similarity is broken up? Zoom in and look at some of the genes encoded within these regions.

What are the predicted products of the genes assigned to these locations? View the details by clicking on the feature, and then select *`Edit selected feature'* from the *'Edit'* menu after selecting the appropriate CDS feature.

Can you identify any genes in one organism that have hits to, or are similar to, regions in the other organism but which don't appear to be predicted? If so, add these to your annotation.

Exercise 4

Introduction

If you do not have access to BLAST software running on a local computer, there is a web resource WebACT (**Appendix VI** for the URL) that can be used for generating ACT comparison files. WebACT allows you to cut and paste, or upload, your own sequences, and generate ACT readable BLASTN or TBLASTX comparison files. WebACT also has a large selection of recomputed comparison files for bacterial genomes, which can be downloaded along with the EMBL sequence entries and viewed in ACT.

For the purposes of this exercise we are going to focus on the Gram-negative bacterial pathogens *Burkholderia pseudomallei* and *Burkholderia mallei.* Both of these organisms are category B bio-threat agents and cause the diseases Melioidosis and Glanders respectively. The two species are closely related (DNA-DNA identity is >99%, multi locus sequence typing (MLST) predicts that *B. mallei* is a clone of *B. pseudomallei*), however they differ markedly in the environmental niches that they occupy.

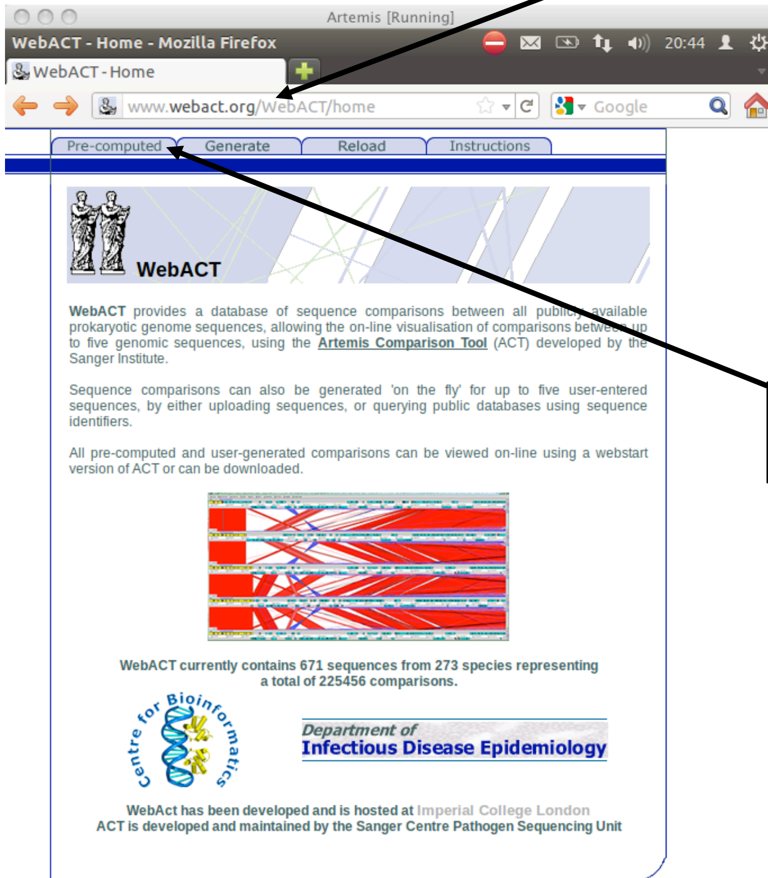
B. pseudomallei is found in S.E. Asia and northern Australia, and is prevalent in the soil in Melioidosis endemic areas. Inhalation, or direct contact with cuts or breaks in the skin, by soil-borne *B. pseudomallei* is the cause of Melioidosis in humans and higher mammals. In contrast, *B. mallei* is a zoonotic pathogen that is host restricted to horses and cannot be isolated from the environment. Comparative genomic analysis has provided insights into evolution of these two pathogens and the genetic basis for ecological and pathological differences of these two pathogens.

The genomes of these two organisms both consist of two circular chromosomes. Comparisons of the genomes reveals that the genome of *B. pseudomallei* is ~1.31 Mb larger than that of *B. mallei*; 16% of chromosome 1, and 32% of chromosome 2, are unique in *B. pseudomallei* with respect to *B. mallei*.

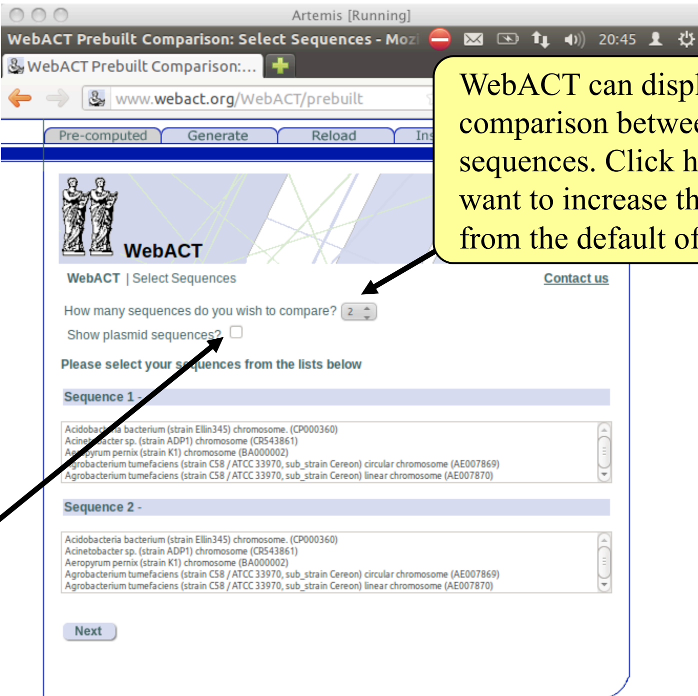
Aim

You are going to use a web resource, WebACT, to generate a comparison file of the smaller chromosomes of *B. pseudomallei* and *B. mallei*. From the WebACT site you will download a pre-computed ACT comparison comparison file, along with the appropriate EMBL sequence and annotation files, which you will then open in ACT. Using this comparison you can then investigate some of the the genotypic differences that differentiate these closely related pathogens, and look for the basis of structural differences in these chromosomes. We have not provided files for this exercise - you are on your own.

Open up a web browser and go to the URL: **www.webact.org**



Click on the Pre-computed tab.



WebACT can display pairwise comparison between up to 5 sequences. Click here if you want to increase the number from the default of 2.

The 'Pre-computed' page contains genomic sequences that have been compared using BLASTN to each other. By selecting the desired sequences from the sequence lists, the appropriate sequence and comparison files can be downloaded

In addition to the chromosome sequences, plasmids can also be displayed by clicking in this box

You are going to compare the smaller chromosomes of *B. pseudomallei* and *B. mallei*.

In the **Sequence 1** list select *Burkholderia pseudomallei* chromosome 2 (accession number BX571966)

In the **Sequence 2** list select *Burkholderia mallei* chromosome 2 (accession number CP000011)

Once you have selected the sequences click the **Next** button

In this window you can specify the regions in the selected sequences to generate the comparison over. It is possible to query the sequences on gene name or coordinates. The default setting is for the whole sequence, and this is what we want for this exercise as you are going to compare the whole chromosomes.

Click the **Next** button

In the **Overview of Selection** you can see a schematic representation of the relative size of the two sequence that have been chosen to be compared.

The Expect (E) value cut-off can be changed in this box. The default value is 0.01, but the range is from 10.0 to 0.0001.

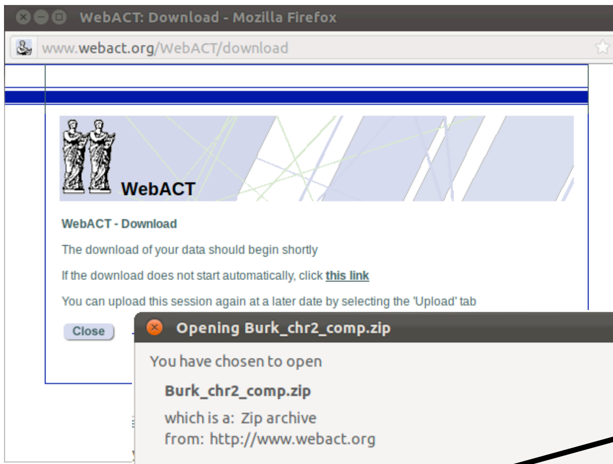
Click the **Download files** button

In addition to downloading the comparison files and sequence file it is also possible to view the comparison in a webstart version of ACT. This will run locally on your machine and does not require ACT to be previously loaded, as a webstart version of ACT will be included in the download. You are not going to use this option in this exercise.

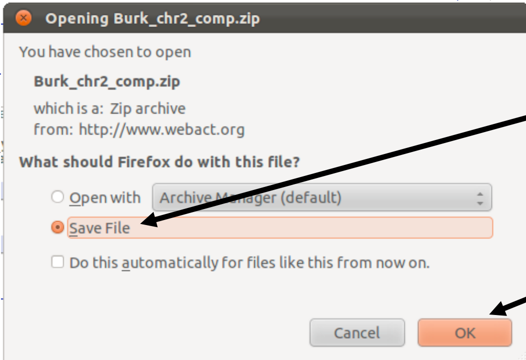
The comparison file and sequences files will contained in a folder. For the ease of downloading the folder is zipped.

In the filename box you can type the file name of the zip file containing the sequence and comparison files. For this exercise call the file: **Burk_chr2_comp.zip**

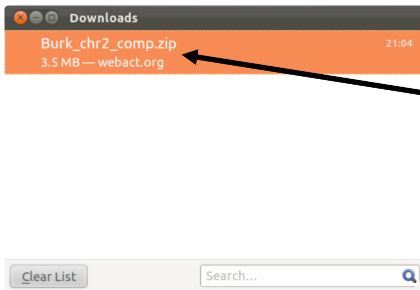
Click the **Download files** button



You may get a window appearing asking you what Firefox should do with the Burk_chr2comp.zip file? Save the file to disk.

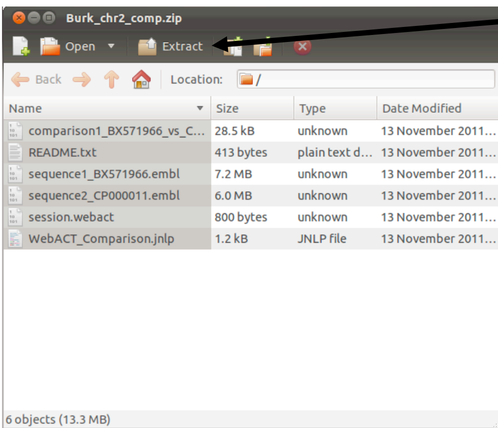


Click the **OK** button



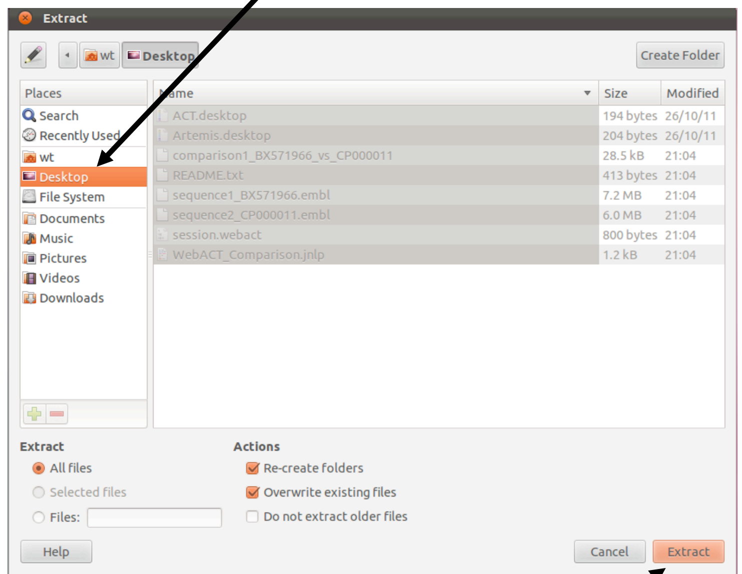
Burk_chr2_comp.zip should now be in the **Downloads** directory

To unzip the file, double click with the left mouse button on the file name



Click the **Extract** button

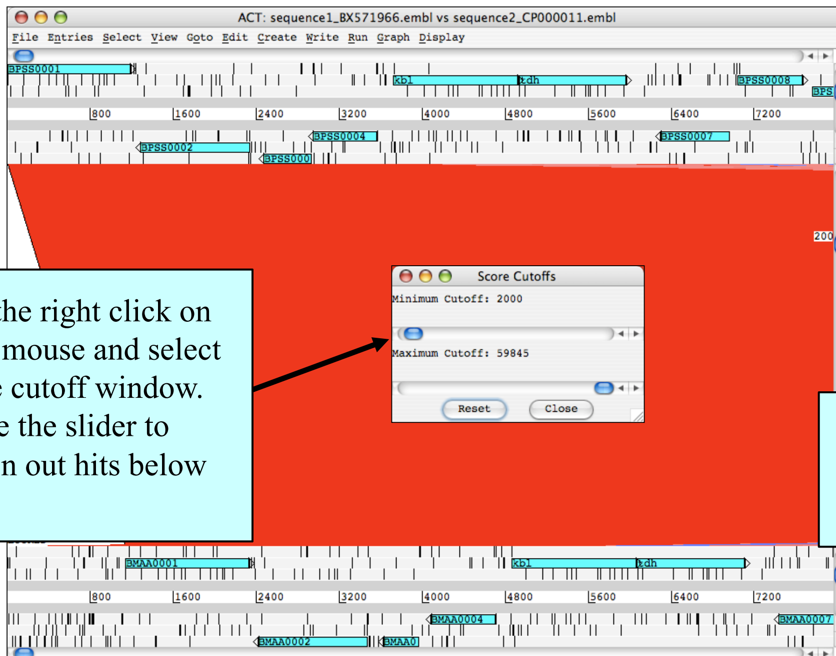
Select a location to extract the files to, such as **Desktop**



The files contained in the unzipped directory should include: comparison1_BX571966_vs_CP000011, sequence1_BX571966.embl and sequence2_CP000011.embl. These are the ACT comparison file and the *B. pseudomalllei* and *B. mallei* chromosome 2 EMBL annotation and sequence files respectively.

Click the **Extract** button

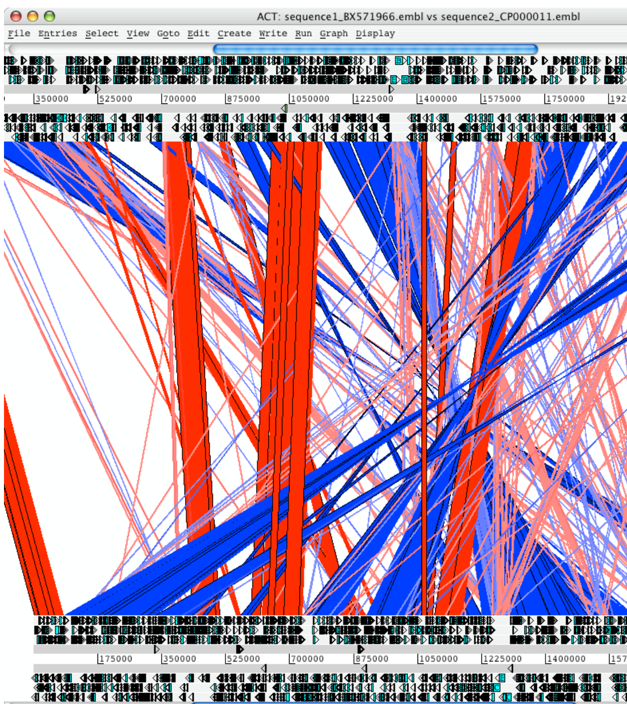
Open up ACT, and load up the comparison (comparison1_BX571966_vs_CP000011) along with the two EMBL sequence and annotation files (sequence1_BX571966.embl and sequence2_CP000011.embl). If you get a warnings window asking if you want to read warning, click **No**.



Use the right click on your mouse and select score cutoff window. Move the slider to screen out hits below 2000

Move the slider to 200 to show only BLASTN matches greater than 200

Now remove the stop codons for both entries, and then zoom out you will see the overall conservation of the structure of the small chromosomes is poor.



If you were to look at the comparison for the large chromosomes you would see a similar picture. The lack of conservation is the result of intra-chromosomal rearrangements. What do you think caused this? Zoom into the regions on the edge of the rearranged matches and look at the annotation in the *B. mallei* chromosome.

What is the function of the CDSs consistently found in these regions. Are there matches in the *B. pseudomallei* chromosome?

Try selecting CDSs in *B. pseudomallei* that match these regions and look how many matches there are in *B. mallei*. Are these regions repeated throughout the chromosome?

If you have time, you may want to generate, and view in ACT, comparisons for your own sequences. If you do not have any loaded on your workshop computer, why not try and download some. Sequence in various formats can be cut and pasted, or up loaded onto the WebACT site. In addition, if you know the accession number of the sequence that you want to compare, you can use that. As the web site will have to run BLAST to generate your comparison file, you may want to limit the size of the sequence that you submit for this exercise to <100 kb. The the web site can handle larger sequences, but it will just take longer.

Artemis [Running]

WebACT: Enter Query - Mozilla Firefox

WebACT: Enter Query

www.webact.org/WebACT/generate

Pre-computed | Generate | Reload | Instructions

WebACT

WebACT | Enter Query

Contact us

How many sequences to you wish to compare? 2

Send e-mail notification on job completion?

e-mail address:

For each sequence below, please either paste a sequence, upload a sequence file or enter an EMBL or Refseq Accession number i.e. NTCAD19MR

Sequence 1 -

Paste sequence (raw, EMBL or FASTA format)

Upload File (raw, EMBL or FASTA format)

Browse...

Enter an EMBL or Refseq Accession number

Sequence 2 -

Paste sequence (raw, EMBL or FASTA format)

Upload File (raw, EMBL or FASTA format)

Enter an EMBL or Refseq Accession number

Blast Search Options [show]

Submit | Clear

Clicking on the 'Generate' tab will take you to this page

Number of sequences to compare

Cut and paste sequence

Upload file

Type accession number

Click here for BLAST options, such as changing from the default BlastN to TBlastX, and altering the BLAST cutoffs

Once you added the relevant sequence information, submit your query. The comparison file or files are down loaded as shown in the example, and can them be loaded in to ACT.