

# Viewing the Companion output in Artemis

## Learning objectives:

- Download embl files from Companion
- Open embl files in Artemis
- Viewing and interpreting the results of Companion in Artemis

In the next exercise we will examine the Companion output in more detail. We will use Artemis to have a closer look. First, we need to download the files. Go to the tab ‘Result files’ and download the embl file ‘Pseudochromosome level sequence and annotation’. The file is called embl.tar.gz.

Pcoa-Pkno (PCOA)
Completed

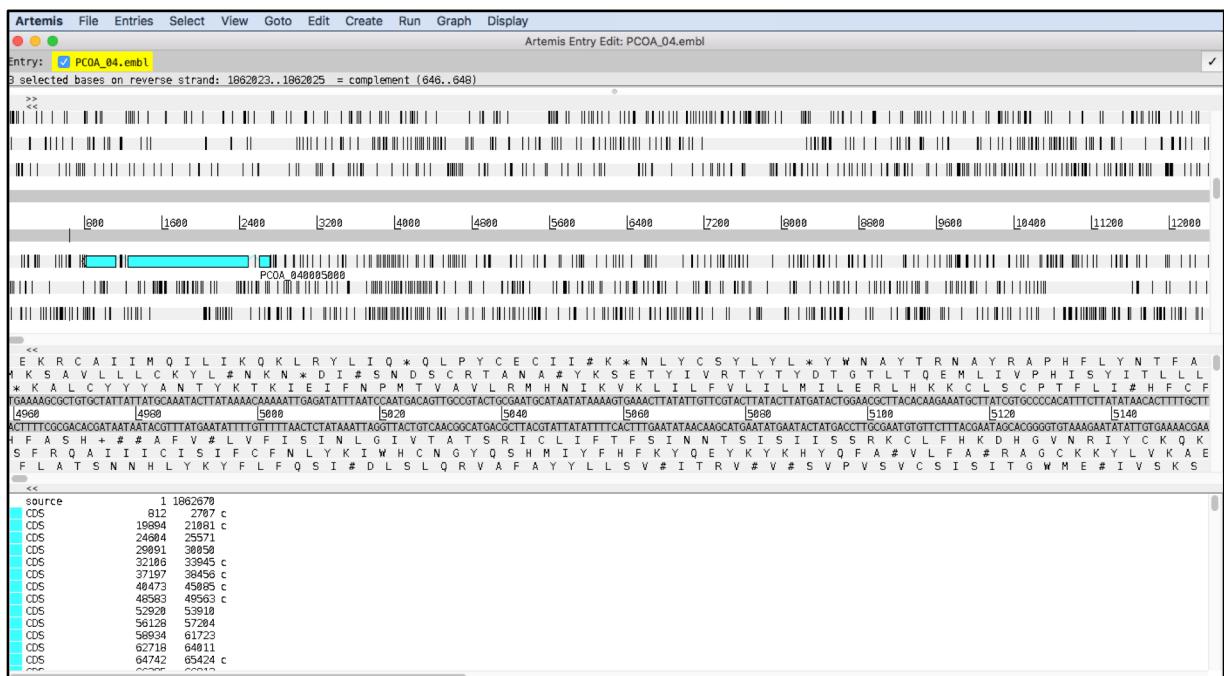
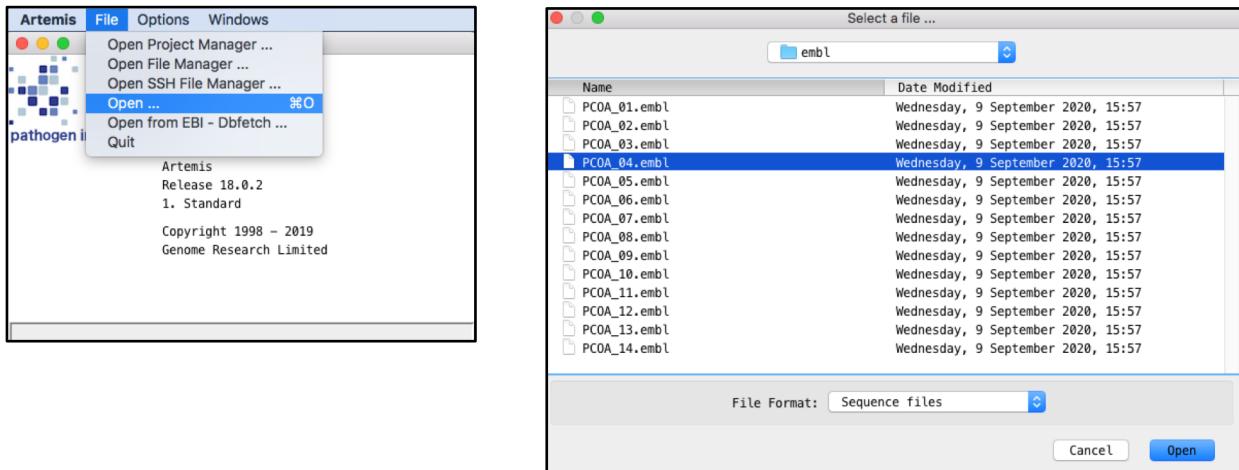
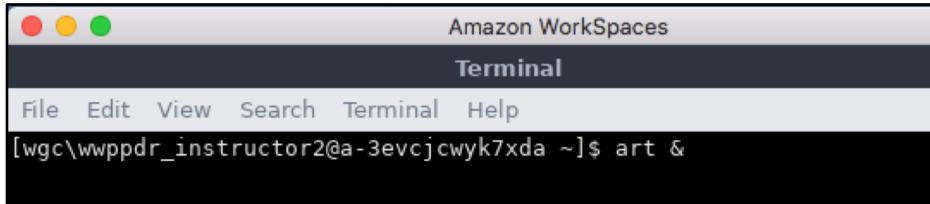
This job was submitted 2 days ago and ran for about 3 hours, finally finishing at 2020-09-09 15:33:07 UTC.

	Format	MD5	Size
<a href="#">Pseudochromosome level genomic sequence</a>	FASTA		7.79 MB
<a href="#">Pseudochromosome level gene annotations</a>	GFF3		5.44 MB
<a href="#">Pseudochromosome layout</a>	AGP		618 Bytes
<a href="#">Scaffold level genomic sequence</a>	FASTA		7.79 MB
<a href="#">Scaffold level gene annotations</a>	GFF3		5.47 MB
<a href="#">Scaffold layout</a>	AGP		825 Bytes
<a href="#">Pseudochromosome level sequence and annotation</a>	EMBL		13.4 MB
<a href="#">Gene Ontology function assignments</a>	GAF1		1.65 MB
<a href="#">Protein sequences</a>	FASTA		3.99 MB

Locate the file called embl.tar.gz in your download folder. Double click on the file. A new folder called ‘embl’ will be created that contains all embl files. Here is the output for *P. coatneyi*. The file contains all chromosomes that were assembled and annotated by Companion. Contigs that could not be placed on one of the chromosomes are in the file \*00.embl

PCOA_01.embl	9 Sep 2020 at 15:57	1.7 MB	sequence
PCOA_02.embl	9 Sep 2020 at 15:57	1.4 MB	sequence
PCOA_03.embl	9 Sep 2020 at 15:57	1.8 MB	sequence
PCOA_04.embl	9 Sep 2020 at 15:57	2.9 MB	sequence
PCOA_05.embl	9 Sep 2020 at 15:57	2.3 MB	sequence
PCOA_06.embl	9 Sep 2020 at 15:57	2 MB	sequence
PCOA_07.embl	9 Sep 2020 at 15:57	2.9 MB	sequence
PCOA_08.embl	9 Sep 2020 at 15:57	3.1 MB	sequence
PCOA_09.embl	9 Sep 2020 at 15:57	4 MB	sequence
PCOA_10.embl	9 Sep 2020 at 15:57	2.7 MB	sequence
PCOA_11.embl	9 Sep 2020 at 15:57	4 MB	sequence
PCOA_12.embl	9 Sep 2020 at 15:57	7.7 MB	sequence
PCOA_13.embl	9 Sep 2020 at 15:57	3 MB	sequence
PCOA_14.embl	9 Sep 2020 at 15:57	3.1 MB	sequence

Artemis is a great tool to visualise your Companion output. Choose one of the chromosomes you've just downloaded and open it in Artemis. To do so, open a Terminal on your Workspace type “**art &**” and hit return. As an example chromosome 4 of *P. coatneyi* (PCOA\_04.embl) is shown.



Once you have your Artemis window open, scroll along the chromosome. Do you find any problems in the annotation? Can you see any missing genes? How many pseudogenes can you find? (Hint: search for the qualifier: **pseudo**). Do you think all of the pseudogenes are real or are some misannotated? You can answer this question quite easily by using a tool called ACT, Artemis Comparison Tool. We will show you how to use it in the next step!

# Comparative Genomics

## Visualising the Companion output in ACT

### Learning objectives:

- Opening ACT and loading files into ACT
- Viewing and interpreting Companion results in ACT
- Download of GFF files from VEuPathDB
- Creating ACT comparison files with NCBI blast
- Optional exercise: Open a 3-way comparison in ACT

## Introduction

In the next part of the exercise we will explore the Companion output in more detail with a tool called Artemis Comparison Tool (ACT). ACT was written by Kim Rutherford and 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 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. You can also use the online BLAST web server from NIH-NCBI to produce an alignment file that can be loaded into ACT. This option can only be used with BLASTN. We will cover this option in this Module.

## Aims

The aim of this Module is for you to become familiar with the basic functions of ACT.

In the first part of this exercise you will learn the basic functions of ACT by looking at the companion output of *P. coatneyi* compared to *P. knowlesi*. By comparing two chromosomes you will be able to study the degree of conservation of gene order and identify small and large synteny breaks. You can also look for incorrectly annotated genes.

Once you are familiar with ACT we will show you how to create your own comparison file and explore your Companion output in ACT.

## Part 1: Starting up the ACT software

In the first part of this exercise we will all use the same files. Make sure you're in the **Module\_2\_Comparative\_Genomics** 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:

PCOA\_04.embl

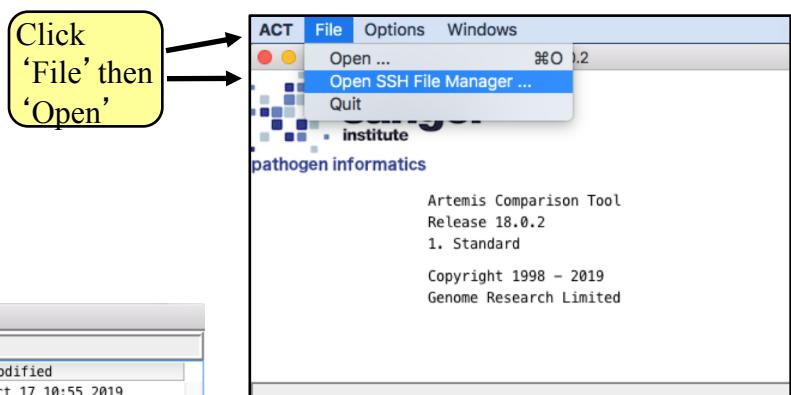
PCOA\_04\_comp\_PKNH\_04

PKNH\_04\_v2.embl

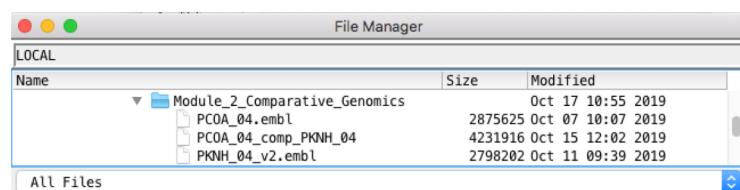
- embl file created by Companion

- tblastx comparison file

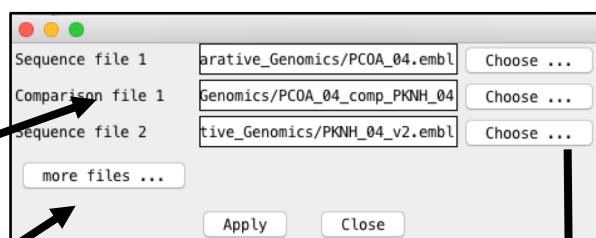
- *P. knowlesi* chr4 (reference used in Companion run)



Use the File manager to drag and drop files.



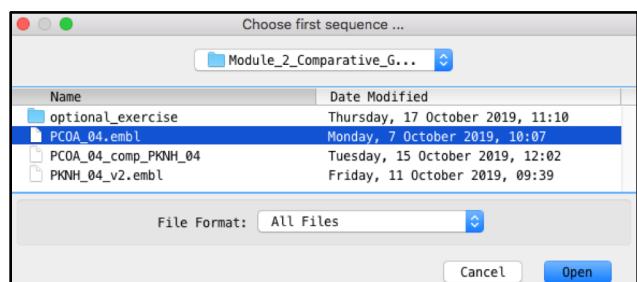
Choose 'All Files'



For comparing more than two DNA files!

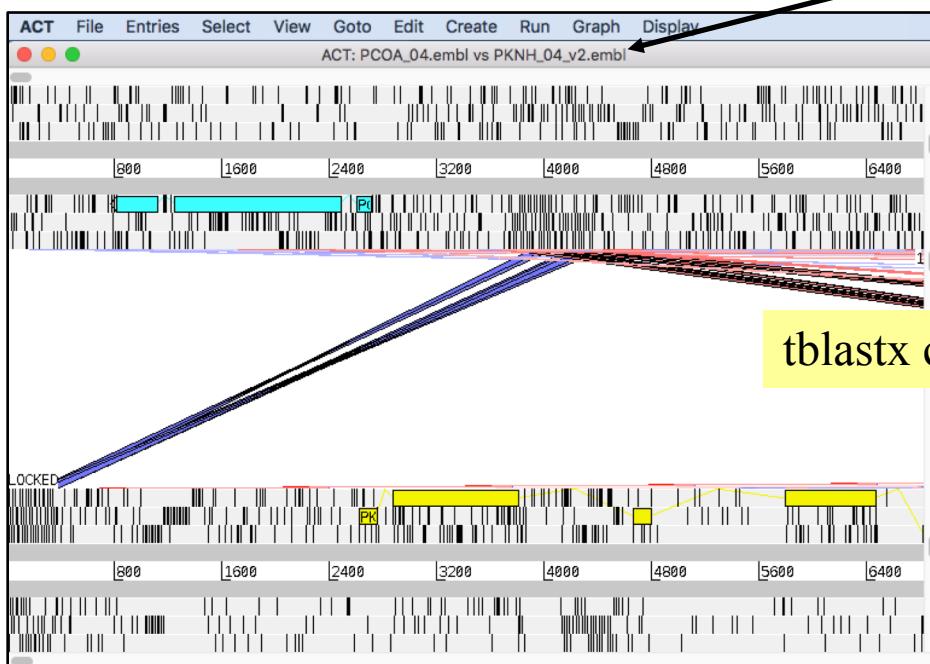
Click 'Apply' and wait

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



For more info on comparison files see Appendix III.

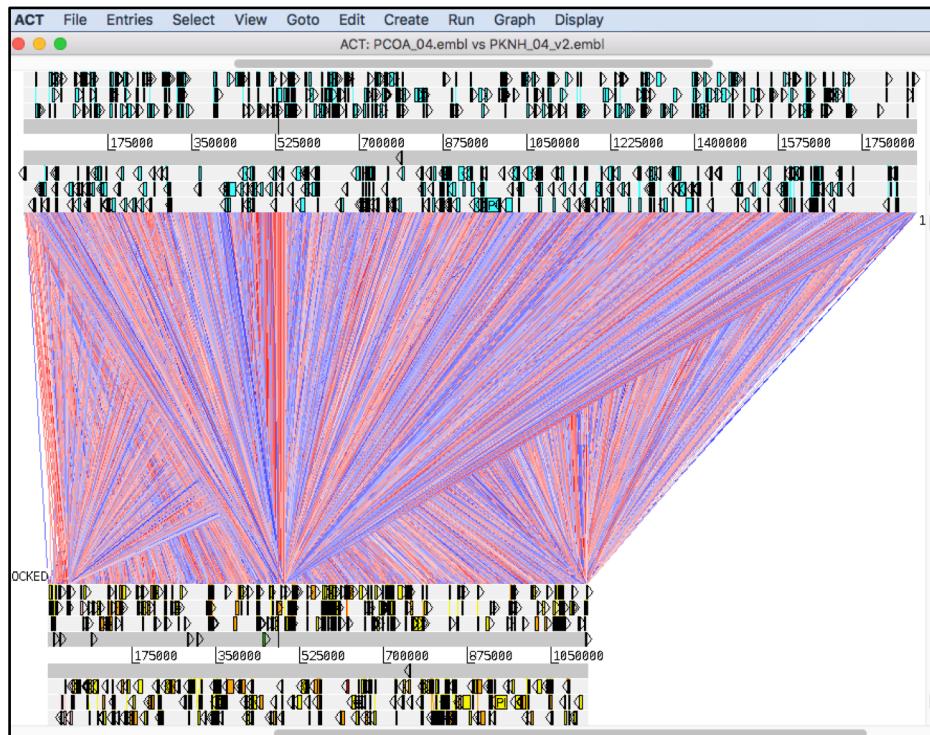
Once you have opened the files you will see a picture like this:



You can see the name of the genomes displayed in ACT on the top of the window.

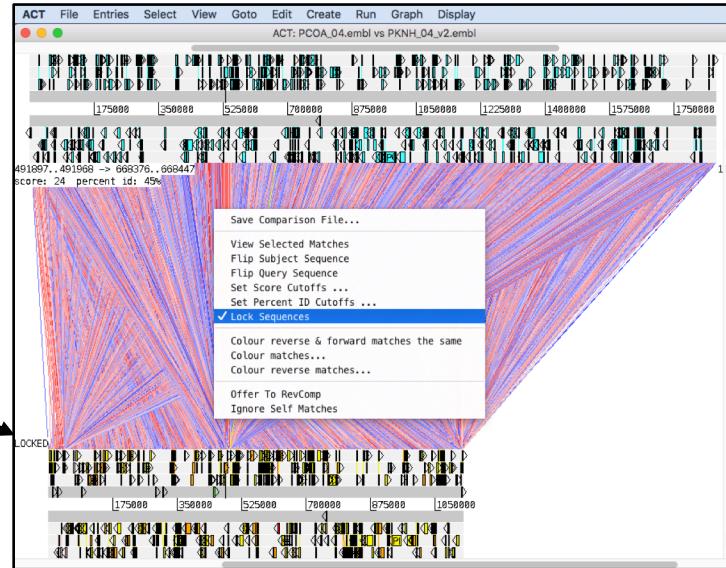
tBLASTx comparison

*P. knowlesi* chr4

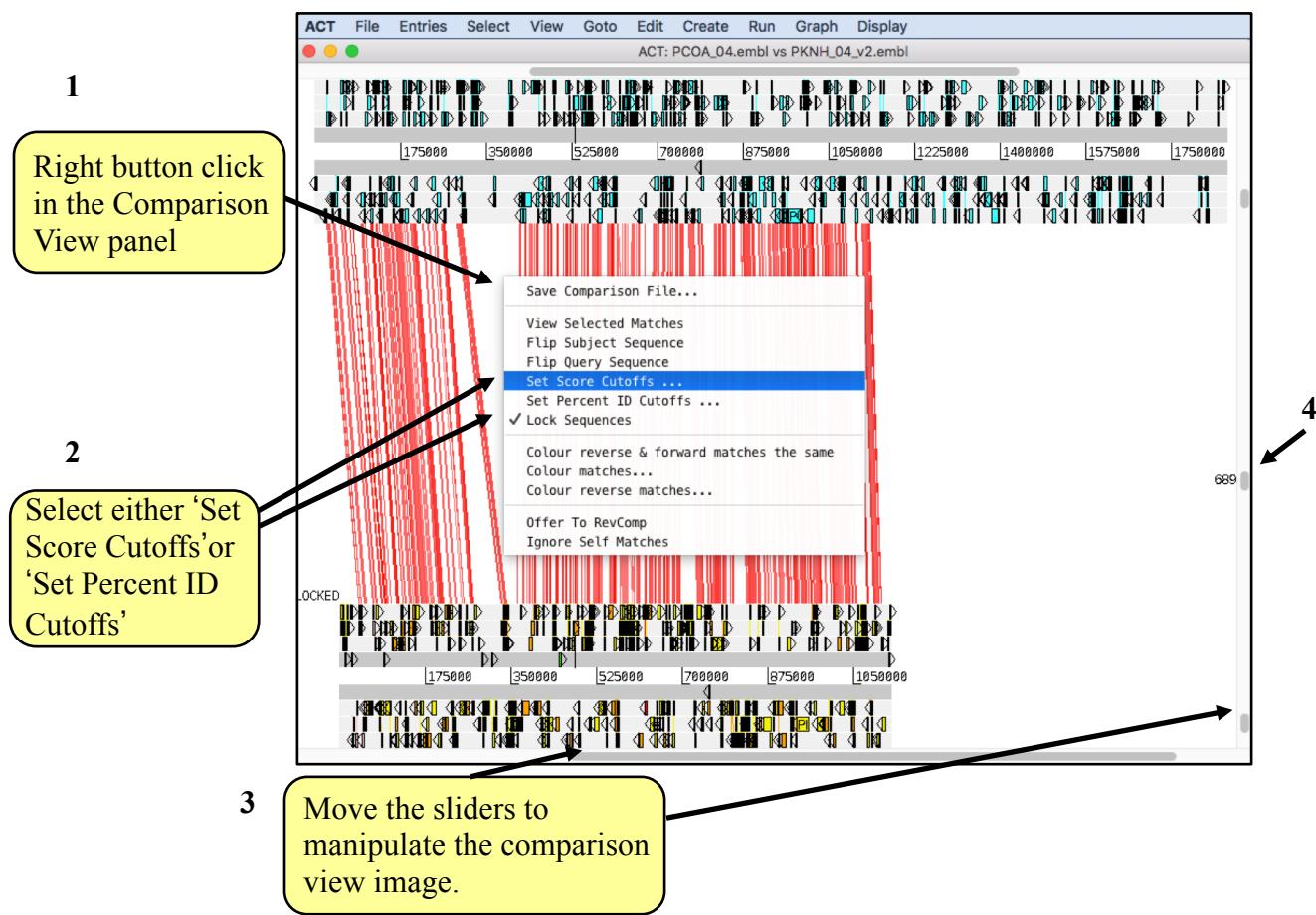


1. Use the vertical sliders to zoom out. Drag or click the slider downwards from one of the genomes. The other genome will stay in sync.

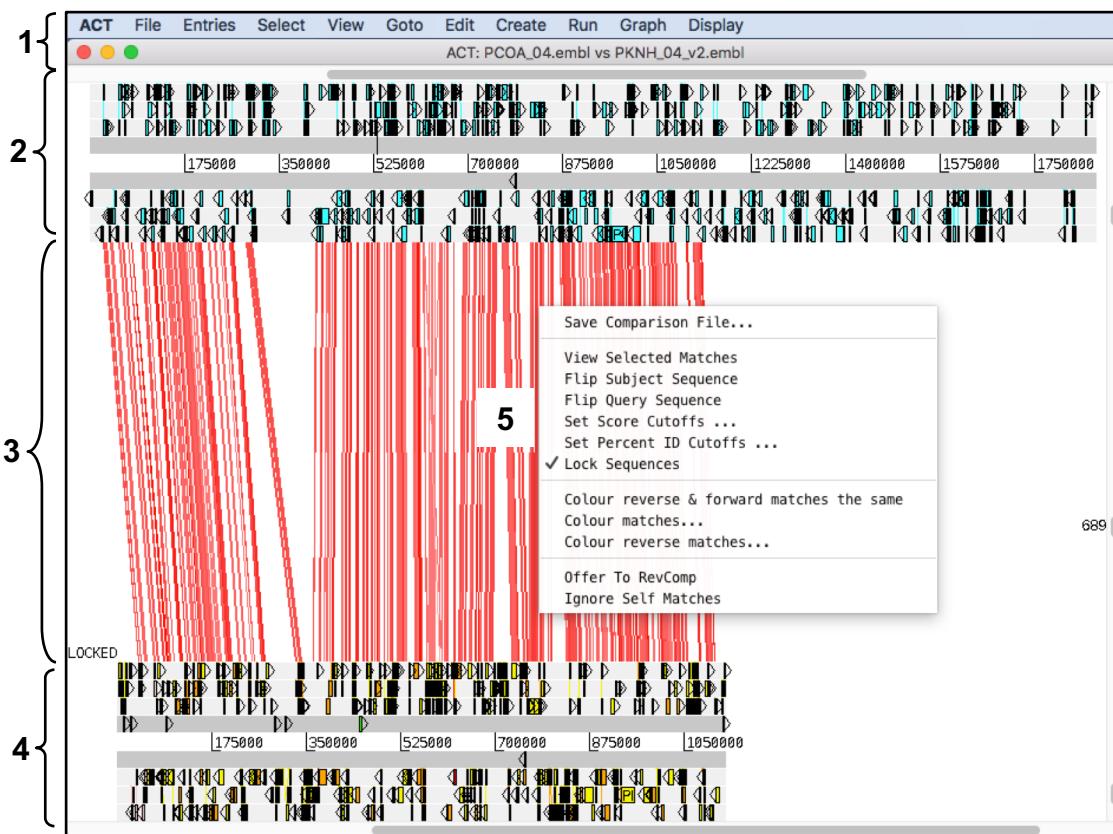
When you scroll along with either slider both genomes move together. This is because they are 'locked' together. Right click over the middle comparison view panel. A small menu will appear, select Unlock sequences and then scroll one of the horizontal sliders. Notice that 'LOCKED' has disappeared from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

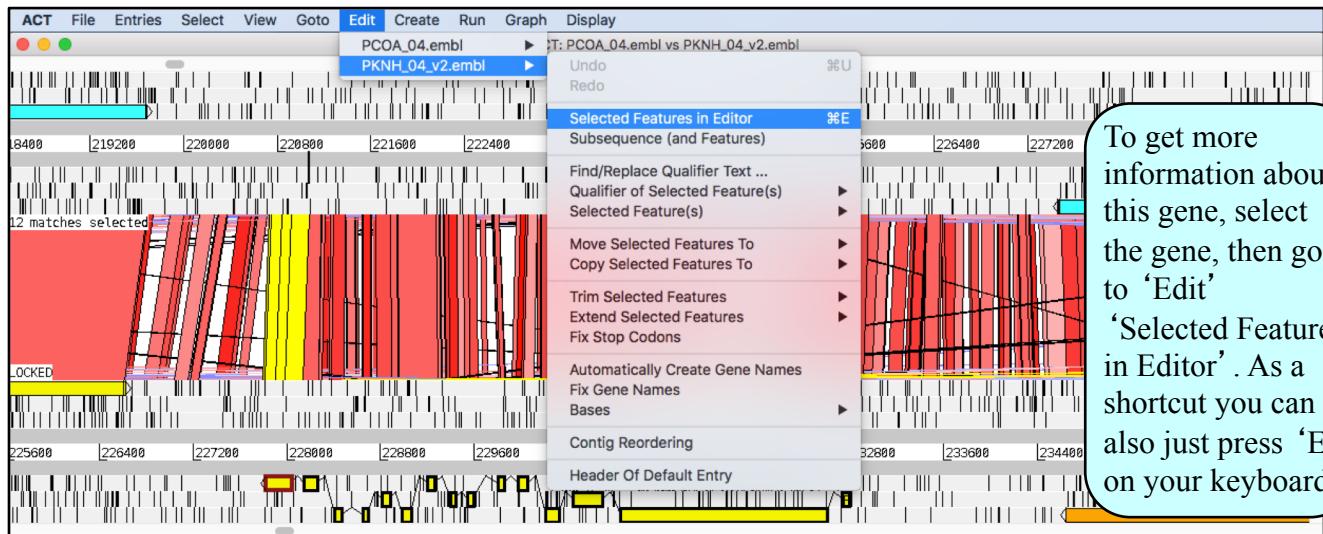
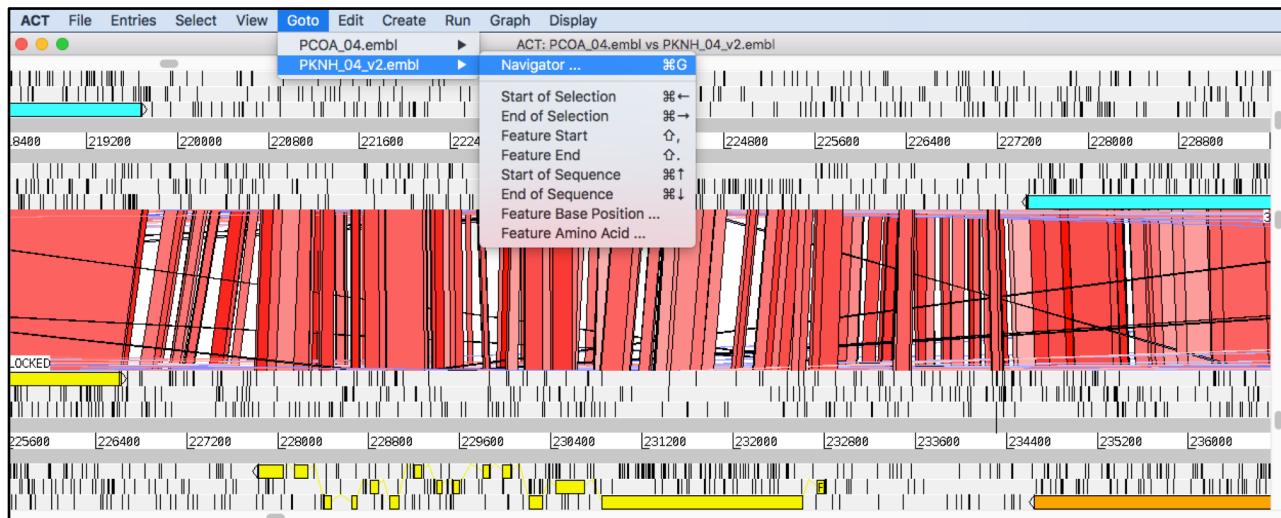


Now that you have an ACT window open let's look what is in there.

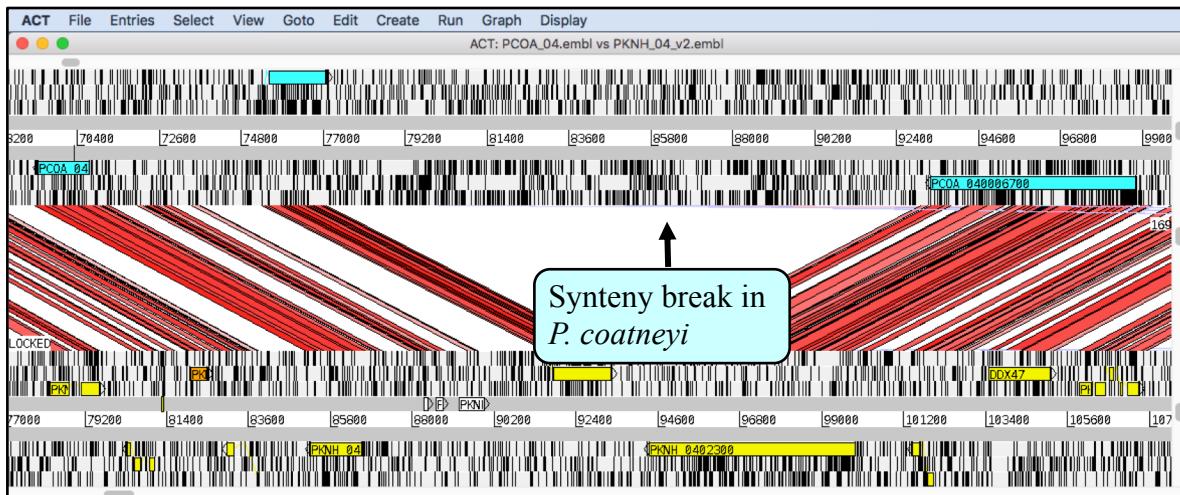


1. Drop-down menus. These are mostly the same as in Artemis. The major difference you will find is that after clicking on a menu header you will then need to select a DNA sequence before going to the full drop-down menu.
2. This is the Sequence view panel for 'Sequence file 1' (Subject Sequence) you 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 ACT-specific menu which we will use later.

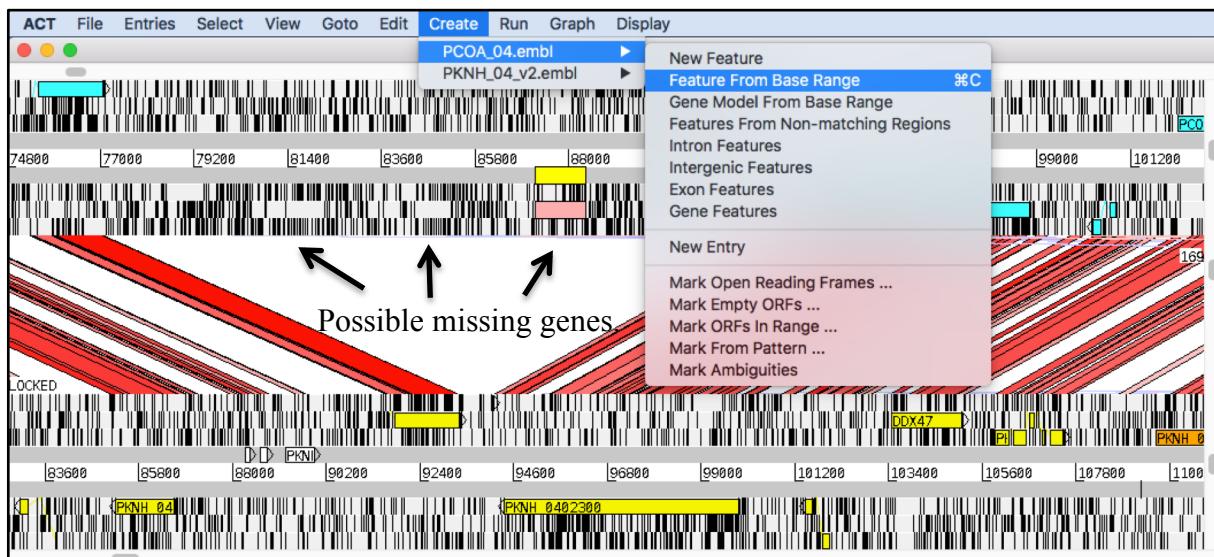
ACT is a great tool to spot any problems in the automatic Companion annotation. Scroll along the genome to find genes that were missed by Companion. Go to the *P. knowlesi* gene PKNH\_0405900 by using the Navigator. Compare it to the *P. coatneyi* annotation on the top. Can you see that this gene has been missed by Companion?



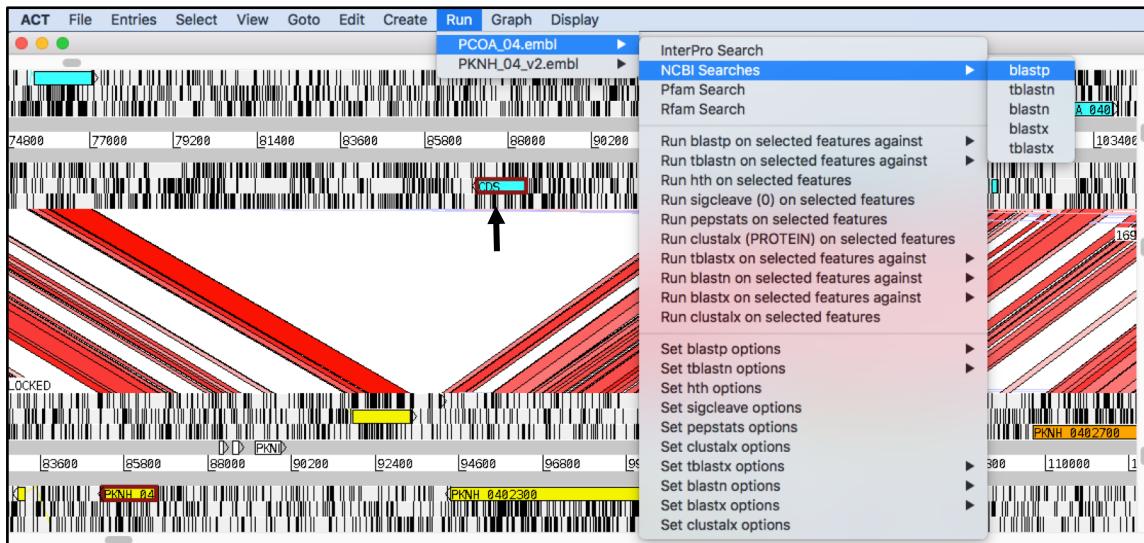
Scroll along the chromosome and try to get an estimate on the number of synteny breaks. Do you think there are any missing genes in the synteny breaks?



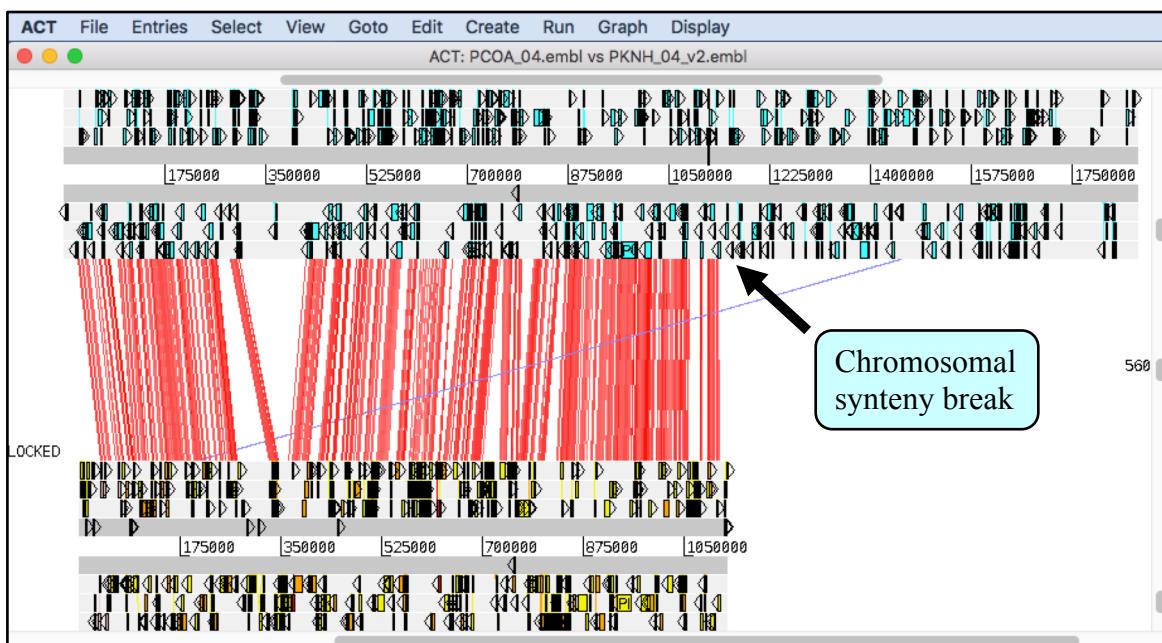
If you think there is a missing gene, you can just mark that area with your mouse. Then go to “Create” and choose “Feature from Base Range”. There is also a shortcut. Just press “C” on your keyboard.



Once you have created a feature, run blast to find out more about the possible missing gene. Can you assign a product?



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



We will show you in the next part how to open a three-way comparison in ACT and explore the synteny break. This is an optional exercise. You can skip it and proceed to part 2, exploring your own Companion output.

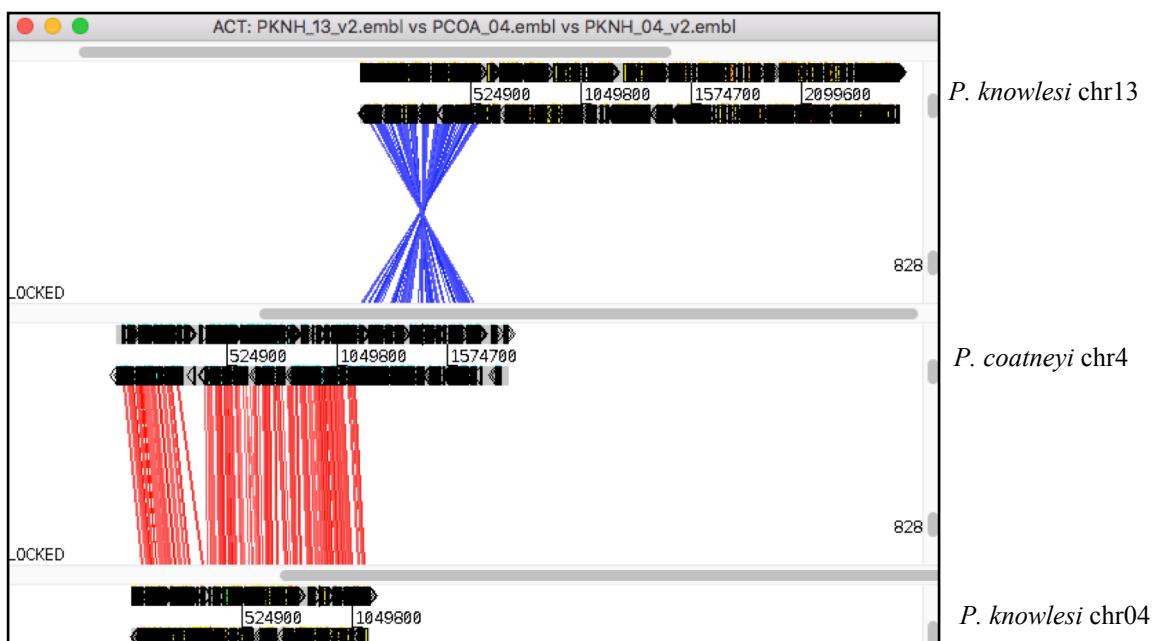
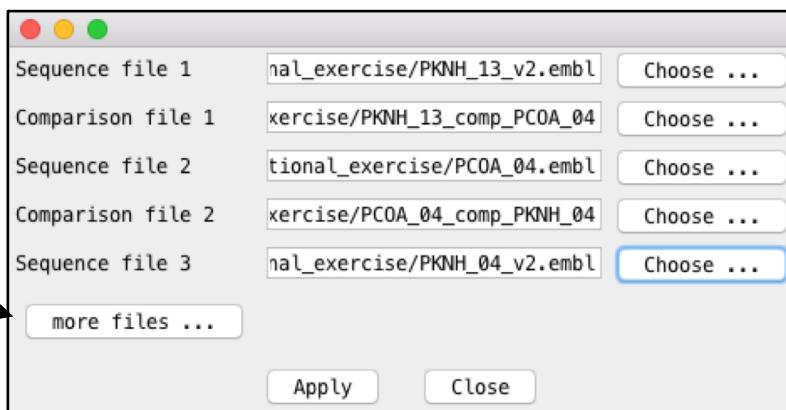
## Optional exercise

In this optional exercise we will show you how to open a three-way comparison in ACT and explore the chromosomal synteny break in *P. coatneyi*.

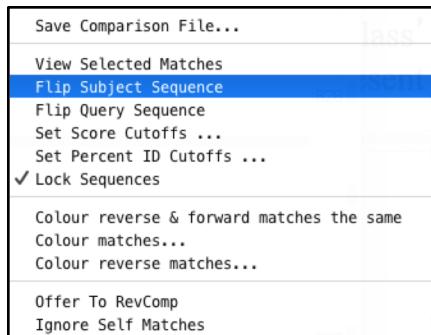
The files you are going to need are:

- |                      |                            |
|----------------------|----------------------------|
| PKNH_13_v2.embl      | - <i>P. knowlesi</i> chr13 |
| PKNH_13_comp_PCOA_04 | - tblastx comparison file  |
| PCOA_04.embl         | - <i>P. coatneyi</i> chr04 |
| PCOA_04_comp_PKNH_04 | - tblastx comparison file  |
| PKNH_04_v2.embl      | - <i>P. knowlesi</i> chr4  |

Click on 'more files' to compare more than 2 files.



The blue 'hour glass' shape indicates that one of the chromosomes is reversed. With a right click in the middle area you can get an additional menu. Select 'Flip Subject Sequence' to flip one of the sequences.



## Part 2: Explore your own Companion output in ACT

Now that you are familiar with the basic functions of ACT, let's explore your own Companion output. To do so, you need the Companion output (sequence and annotation), a comparison file and the reference (sequence and annotation). Have a look at your Companion output and choose one of the chromosomes you want to explore in more detail.

### 1. Download sequence and annotation of your Companion run

You've already downloaded the embl files in the first part of this exercise. To create the comparison file, you also need to download the sequence without annotation. Go to the 'Result files' and download the 'Pseudochromosome level genomic sequence'. This is a sequence file that contains all the chromosomes. Extract by copying and pasting the sequence of the chromosome you are interested in.

File	Format	MD5	Size
Pseudochromosome level genomic sequence	FASTA	██████	2.5 MB
Pseudochromosome level gene annotations	GFF3	██████	2.74 MB
Pseudochromosome layout	AGP	██████	5.12 KB
Scaffold level genomic sequence	FASTA	██████	2.5 MB
Scaffold level gene annotations	GFF3	██████	2.83 MB
Scaffold layout	AGP	██████	2.84 KB
Pseudochromosome level sequence and annotation	EMBL	██████	4.95 MB
Gene Ontology function assignments	GAF1	██████	1.44 MB
Protein sequences	FASTA	██████	2.44 MB

Alternatively, open the embl file you've downloaded from Companion in Artemis and save the sequence as shown below.

Artemis

File Entries Select View Goto Edit Create Run Graph Display

Entry:  One selected

Show File Manager ... Read An Entry ... Read Entry Into >

Read BAM / CRAM / VCF ... Save Default Entry %S

Save An Entry Save An Entry As Save All Entries

**Write** ▶

- Amino Acids Of Selected Features
- Amino Acids Of Selected Features to Qualifier
- PIR Database Of Selected Features
- Bases Of Selection
- Upstream Bases Of Selected Features
- Downstream Bases Of Selected Features
- Upstream+Feature+Downstream Bases ...

All Bases

Codon Usage Of Selected Features

Raw Format

**Fasta Format**

EMBL Format

Genbank Format

DNA sequence viewer showing a sequence from position 6400 to 7200.

## 2. Download sequence and annotation of your reference genome

Downloading the sequence and annotation of your reference, can either be done on VEuPathDB or on one of the main repositories like GenBank or ENA. How to download your sequence in GenBank is shown in the Appendix.

In VEuPathDB you need to download the annotation and sequence for the chromosome you are interested in separately. In the first part of the Companion exercise we showed you how to download a FASTA file. To download the annotation go to Genes on the left hand side, choose ‘Genomic Location’ and select ‘Organism’. Now select the Chromosome and then click on ‘Get Answer’.

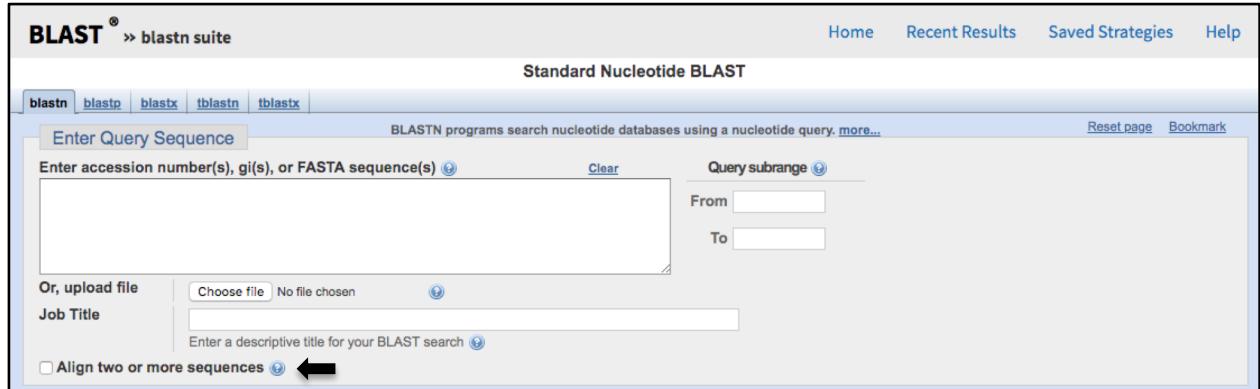
The screenshot shows two panels. The left panel is a sidebar titled "Search for..." with a "Genes" section expanded. Under "Genomic Location", there are options: "Genomic Location" (selected), "Proximity to Telomeres", "Immunology", and "Orthology and synteny". The right panel is titled "Identify Genes based on Genomic Location". It has fields for "Organism" (set to "Cryptosporidium parvum Iowa II"), "Chromosome" (with a dropdown menu "Choose chromosome"), "Start at" (set to "1"), and "End Location (0 = end)" (set to "0"). A red arrow points from the "Get Answer" button at the bottom right to a callout box that says: "Select the chromosome you would like to download and click on “Get Answer”".

This screenshot shows the results of the search. At the top, it says "483 Genes (470 ortholog groups)". The main area displays a table of genes with columns: Gene ID, Transcript ID, Organism, Genomic Location(s), and Product Description. A red circle highlights the "Download" button at the top right of the table header. Other buttons visible include "Add a step", "Revise this search", "Organism Filter", and "Add to Basket".

This screenshot shows the "Download Genes" configuration page. It includes sections for "Choose a Report" (with "GFF3 - gene models and optional sequences" selected), "Generate a report of your query result in GFF3 format", "Download Type" (with "GFF File" selected), and a large red circle highlighting the "Get GFF3 file" button at the bottom right.

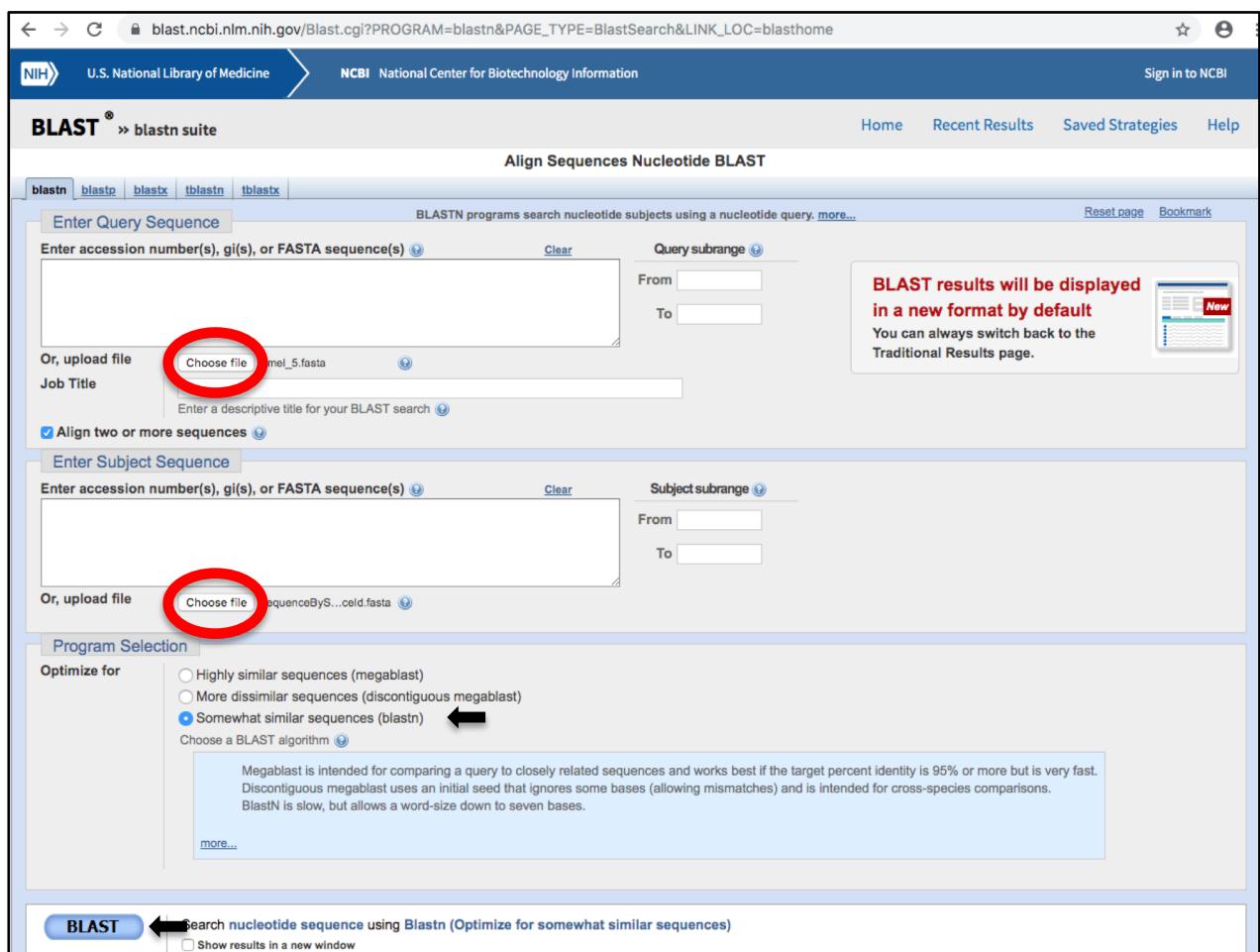
## 2. Create the ACT comparison file

To create the comparison file, go to the following website:  
<https://blast.ncbi.nlm.nih.gov/Blast.cgi> and select ‘Nucleotide Blast’, ‘Align two or more sequences’.



The screenshot shows the BLAST homepage with the title "BLAST® > blastn suite". Below it is the "Standard Nucleotide BLAST" section. In the "Enter Query Sequence" field, there is a placeholder text "Enter accession number(s), gi(s), or FASTA sequence(s)". To the right of this field are "Clear" and "Query subrange" buttons. Below the query field are "From" and "To" input fields. Underneath these are "Or, upload file" and "Job Title" fields. A "Choose file" button is shown with the text "No file chosen". Below these fields is a link "Enter a descriptive title for your BLAST search". At the bottom left of the page, there is a checkbox labeled "Align two or more sequences" with a black arrow pointing to it from the left.

Upload the two sequences you would like to compare and then select a blast option.



The screenshot shows the "Align Sequences Nucleotide BLAST" page. At the top, it says "blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE\_TYPE=BlastSearch&LINK\_LOC=blasthome". The page header includes "NIH U.S. National Library of Medicine" and "NCBI National Center for Biotechnology Information". On the right, there are links for "Sign in to NCBI", "Home", "Recent Results", "Saved Strategies", and "Help". The main form has a "Enter Query Sequence" field with a placeholder "Enter accession number(s), gi(s), or FASTA sequence(s)". To its right are "Clear" and "Query subrange" buttons. Below the query field are "From" and "To" input fields. A red circle highlights the "Choose file" button in the "Or, upload file" field, with the text "Chmel\_5.fasta" displayed next to it. To the right of the query area is a red box containing the text "BLAST results will be displayed in a new format by default" and "You can always switch back to the Traditional Results page". Below the query section is an "Enter Subject Sequence" field with a similar structure. A second red circle highlights the "Choose file" button in the "Or, upload file" field of the subject sequence section, with the text "sequenceByS...celd.fasta" displayed next to it. The "Program Selection" section follows, with a "Optimize for" dropdown containing three options: "Highly similar sequences (megablast)", "More dissimilar sequences (discontiguous megablast)", and "Somewhat similar sequences (blastn)". The third option, "Somewhat similar sequences (blastn)", is selected and highlighted with a black arrow. A detailed description of each program is provided in a box below the dropdown. At the bottom of the page is a large blue "BLAST" button with a black arrow pointing to it from the left, and a link "Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)". There is also a "Show results in a new window" checkbox.

The BLAST run will take a few minutes. Once this is done, select the Download option ‘Hit Table(text)’. This is the comparison file that you can open in ACT.

BLAST® » blastn suite-2sequences » results for RID-U93AK4KU114

Job Title: Cmel\_5  
RID: U93AK4KU114 Search expires on 10-15 22:59 pm  
Program: Blast 2 sequences [Citation](#)  
Query ID: lcl|Query\_4877 (dna)  
Query Descr: Cmel\_5  
Query Length: 1074033  
Subject ID: lcl|Query\_4877 (dna)  
Subject Descr: CM000433 | Cryptosporidium parvum Iowa II | 1 tr  
Subject Length: 1080900  
Other reports: [MSA viewer](#)

Filter Results

Percent Identity: [ ] to [ ]  
E value: [ ] to [ ]

Download All ▾

- Text
- XML
- ASN.1
- JSON Seq-align
- Hit Table(text)** (highlighted with a red circle)
- Hit Table(csv)
- Multiple-file XML2
- Single-file XML2
- Multiple-file JSON
- Single-file JSON
- SAM

Descriptions Graphic Summary Alignments

Sequences producing significant alignments

select all 1 sequences selected

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<a href="#">CM000433   Cryptosporidium parvum Iowa II   1 to 1080900 (reverse-complement)</a>	1.889e+05	1.566e+06	96%	0.0	92.56%	Query_4877

Download Manage Columns Show 100 Graphics

### 3. Open your file in ACT

Let's open the two sequences in ACT.

The files that you are going to need are:

Companion embl file

Comparison file (Hit Table – downloaded from NCBI)

Reference (Fasta file and GFF file downloaded from PlasmoDB)

ACT File Options Windows

Open ... ⌘O 2  
Open SSH File Manager ...  
Quit  
institute  
pathogen informatics

Artemis Comparison Tool  
Release 18.0.2  
1. Standard  
Copyright 1998 - 2019  
Genome Research Limited

Sequence file 1: /Downloads/embl/Cmel\_5.embl  
Comparison file 1: /l/U93AK4KU114-Alignment.txt  
Sequence file 2: /nbl/SequenceBySourceId.fasta

more files ...

Apply Close

Once you've opened ACT load in the GFF file you've downloaded from VEuPathDB.

