

Artemis

Introduction

Artemis is a free DNA viewer and annotation tool written by Kim Rutherford (Rutherford *et al.*, 2000). It is routinely used by the Parasite Genomics Group at the Wellcome Sanger Institute for annotation and analysis of both prokaryotic and eukaryotic genomes. The program allows the user to view simple sequence files, EMBL/Genbank entries and the results of sequence analyses in a highly interactive and intuitive graphical format. Artemis is designed to present multiple sets/types of information within a single context. This manifests itself as the ability to zoom in to inspect DNA sequence motifs and zoom out to view local gene architecture, several kilobases of a genome or even an entire genome in one screen. It is also possible to perform some analyses within Artemis with the output stored for later access.

Aims

The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. Like all the Modules in this workshop, the key is 'if you don't understand please ask'.

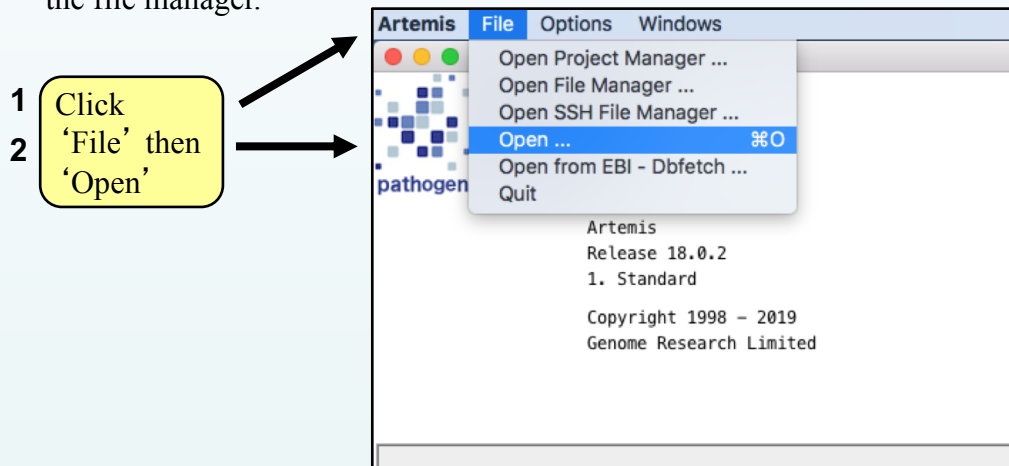
Artemis Exercise 1 Part I

1. Starting up the Artemis software

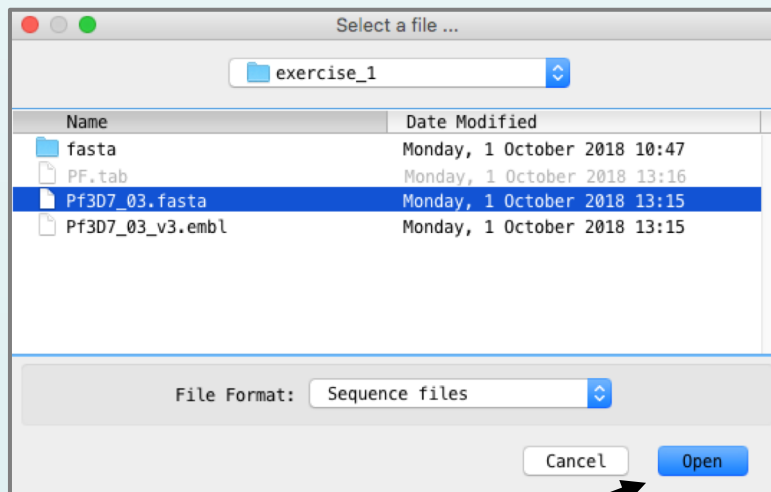
Double click the ARTEMIS Icon on your Desktop

A small start-up window will appear (see below).

Navigate to the directory Module_1_Artemis, exercise_1 containing the file Pf3D7_03.fasta using the file manager.



For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.



3

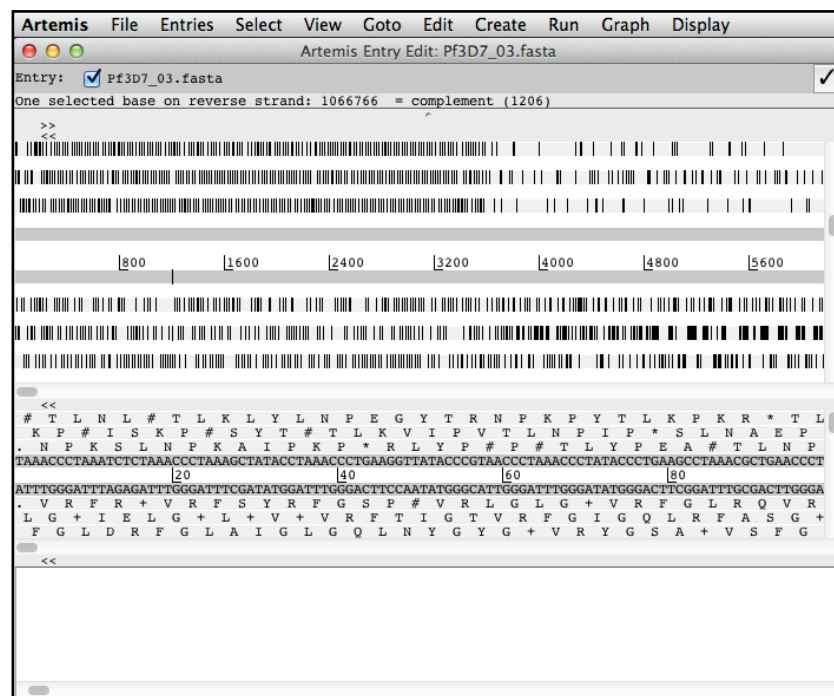
Single click to select the DNA file

4 Single click to open file in Artemis then wait

DNA sequence files will have the suffix '.fasta'. Annotation files end with '.embl', or '.tab'. Use this feature to select the type of file displayed in this panel.

2. Loading annotation files (entries) into Artemis

Hopefully you will now have an Artemis window like this! If not, ask a demonstrator for assistance.



Now follow the numbers to load up the annotation file for *Plasmodium falciparum* 3D7 chromosome 3.

1

Click 'File' then 'Read an Entry'

Entry = file

2

Single click to select Pf3D7_03.embl file

3

Single click to open file in Artemis then wait

What's an "Entry"? It's a file of DNA and/or features which can be overlaid onto the sequence information displayed in the main Artemis view panel.

3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.



1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case an acyl-CoA synthetase (selected line).
3. This is the main sequence view panel. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked as black vertical bars. Genes and other features (eg. Pfam matches) are displayed as coloured boxes. We will refer to genes as coding sequences or CDSs from now on.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene highlighted. The list can be scrolled (8, below).
6. Sliders for zooming view panels.
7. Sliders for scrolling along the DNA.
8. Slider for scrolling feature list.

4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the 'Goto' drop-down menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

4.1 The 'Goto' menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. Most are self-explanatory, so feel free to try any of them.



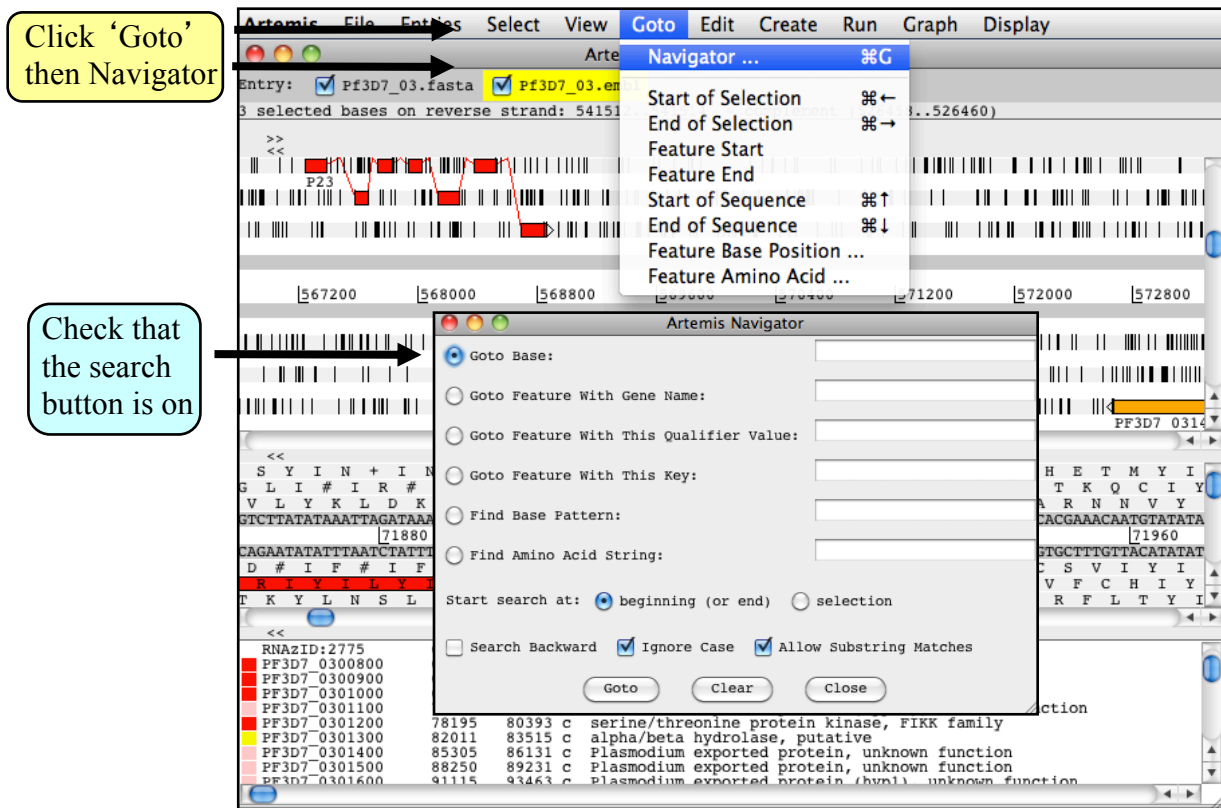
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try! This is a very commonly used feature, so it is worth memorizing the keyboard shortcuts for these, `ctrl<left arrow>` and `ctrl <right arrow>` respectively.

Suggested tasks:

1. Zoom out, highlight a large region of sequence by clicking the left hand button and dragging the cursor, then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.



Suggestions of where to go:

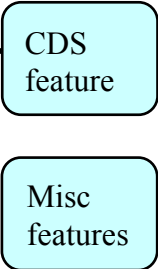
1. Think of a number between 1 and 1067971 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'VAR').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular term. For example using the word 'pseudogene' will take you to the next feature with the word 'pseudogene' in any of its qualifiers. Note how repeated clicking of the 'Goto' button takes you through the pseudogenes as they occur on the chromosome.
4. tRNA genes. Type 'tRNA' in the 'Goto Feature With This Key'.
5. Amino acid consensus sequences (real or made up!). You can use 'X' s. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

What are Keys and Qualifiers? See **Appendix IV**

Clearly there are many more features in Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menus. Hopefully you will find most of them easy to understand.

Artemis Exercise 1 Part II

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region located between bases 134000 to 141000 on the DNA sequence. This region encodes the *CLAG3.1* gene which codes for cytoadherence linked asexual protein. You can use either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.



Misc features

Once you have found this region have a look at some of the information that is available to you:

Information to view:

Annotation

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the 'Edit' menu and select 'Selected Feature in Editor', or simply push 'E'. A window will appear containing all the annotation that is associated with that CDS.

Viewing amino acid or protein sequence

Click on the view menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL or FASTA. This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

Plots/Graphs

Feature plots can be displayed by selecting a CDS feature then clicking 'View' and 'Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

Load additional files

The results from the Pfam protein motif searches are not shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'Selection' or click 'Edit' then 'Selected Features in Editor'. You can also run Pfam by going to the Run menu and selecting 'Pfam search'. For this you need to select one CDS.

Viewing the results of database searches

Click the 'View' menu, then select 'Search Results' and then 'Fasta results'. The results of the database search will appear in a scrollable window.

Further information on specific Pfam entries can be found on the web at <http://pfam.xfam.org/>

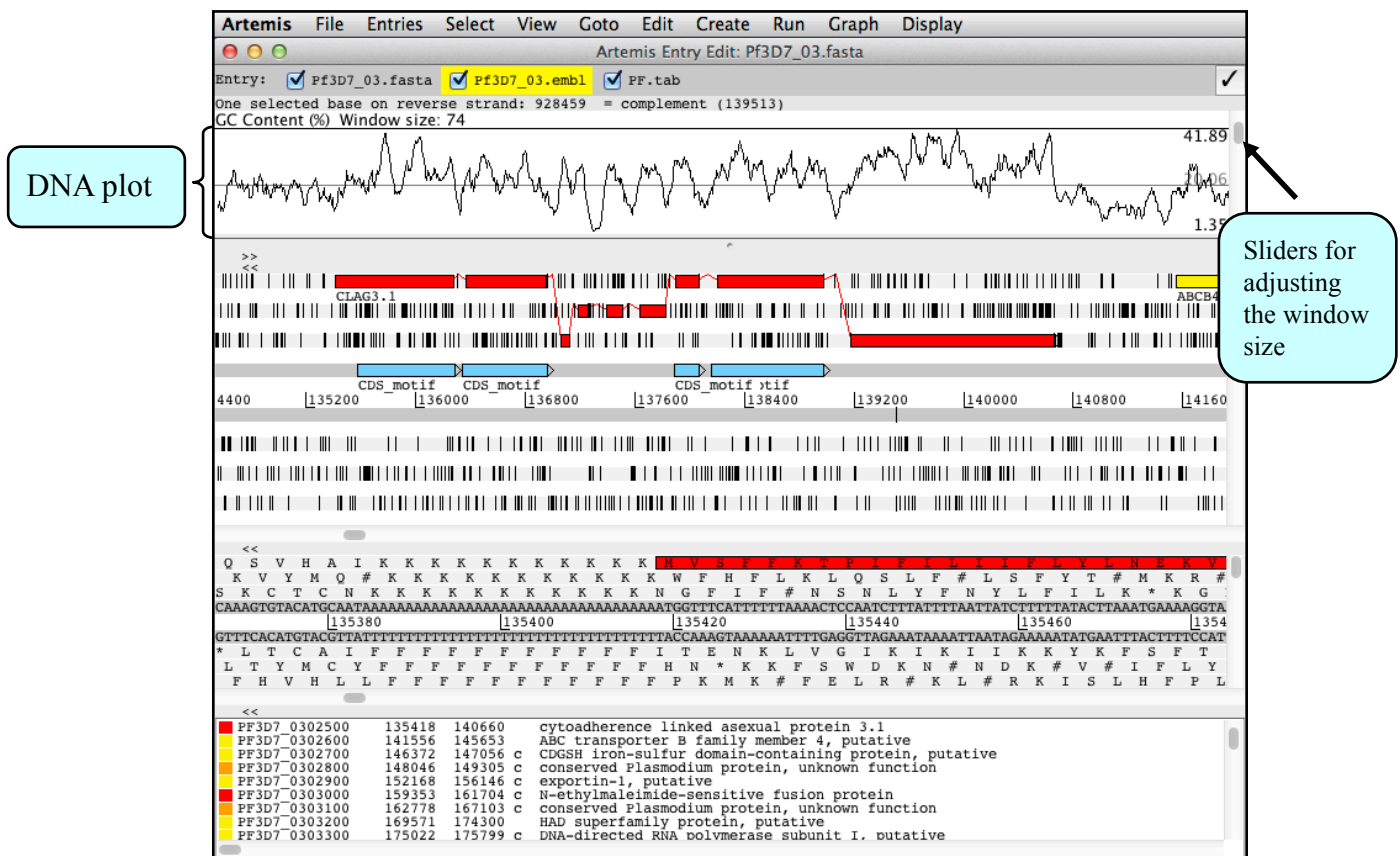
In addition to looking at the fine details of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding to the display various plots showing different characteristics of the DNA.

To view the graphs:

Click on the 'Graph' menu to see all those available. Some of the most useful plots for *P. falciparum* is the 'GC Content (%)' as shown below. G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.

To view the graphs:

Click on the ‘Graph’ menu to see all those available. Some of the most useful plots for *P. falciparum* is the ‘GC Content (%)’ as shown below. G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.



Artemis Exercise 1 Part III

In this part of the Module we will be looking at methods of selecting and extracting features. We are going to extract different genes and regions and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotation and DNA for this region.

In Artemis you can select genes fitting different search criteria. One possibility is to look for a specific product, for example *rifin*, as shown below.

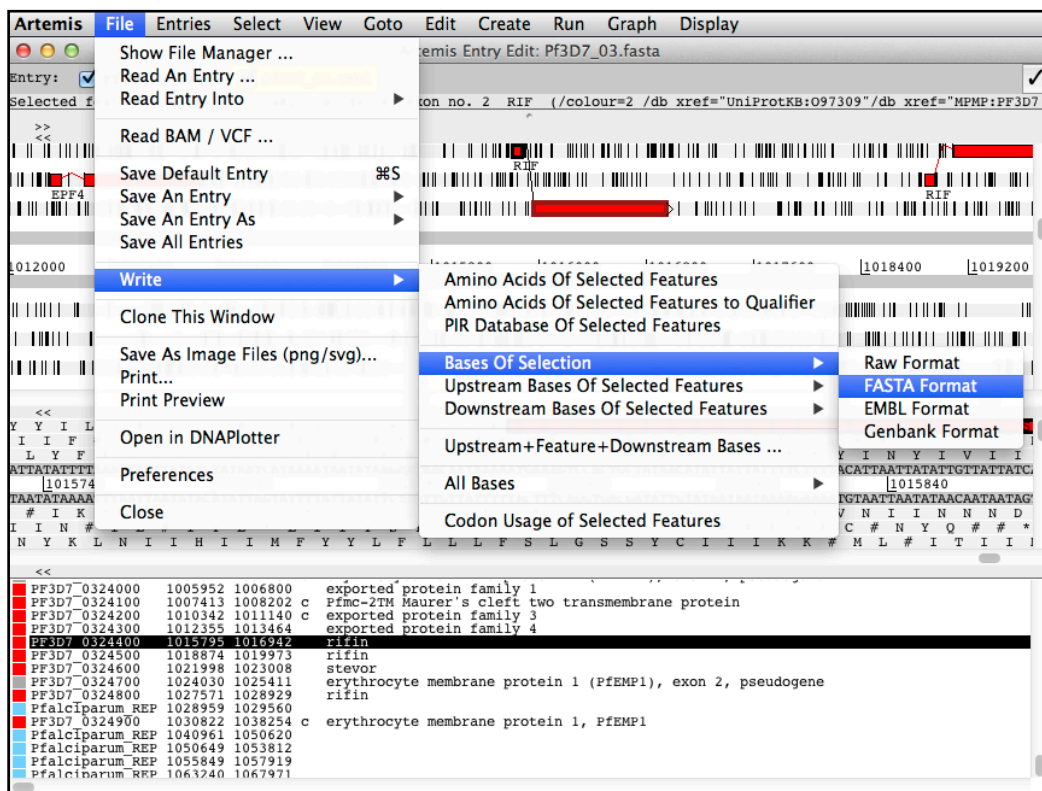
- Click 'Select' then 'Feature Selector'
- Make sure the buttons are down
Set Key to 'CDS' and Qualifier to 'product'
- Type search term
- Click to select features containing search term
- Click to view selected features
- Double click to bring features into main view window.

The screenshot shows the Artemis software interface. The 'Feature Selector' dialog box is open, and the 'CDS' key and 'product' qualifier are selected. The search term 'rifin' is entered in the 'Containing this text' field. The 'Select' button is highlighted. The main window shows a genomic track with various features, including CDS features. A list of features is displayed at the bottom of the main window.

Key	Start	End	Strand	Product	PMID
CDS	46369	47579	c	A-type rifin	(PMID:18197962)
CDS	55390	56584	c	B-type rifin	(PMID:18197962)
CDS	61445	62714	A-type	rifin	(PMID:18197962)
CDS	64572	65783	A-type	rifin	(PMID:18197962)
CDS	1015795	1016942	A-type	rifin	(PMID:18197962)
CDS	1018874	1019973	B-type	rifin	(PMID:18197962)
CDS	1027571	1028929	A-type	rifin	(PMID:18197962)

The genes listed in 6 (on the previous page) are only those fitting your selection criterion. They can be copied or moved in to a new entry so they can be viewed in isolation from the rest of the information within Pf3D7_03.embl. To create a new entry go to 'Create' and choose 'New Entry'.

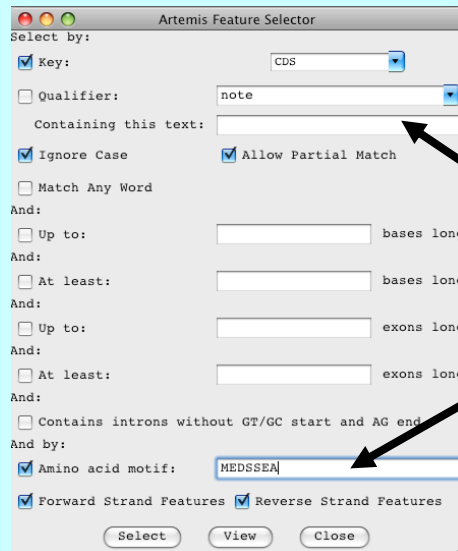
In the next step of the exercise choose one of the selected genes and write out a fasta-file of the sequence.



Click 'File' then 'Write 'Bases of Selection' 'FASTA Format'

Additional methods of selecting/extracting features using the Feature Selector

It is worth noting that the Feature Selector can be used in many other ways to select and extract subsets of features from the genome such as text or amino acid searches.



Artemis Feature Selector

Select by:

☒ Key: CDS

☐ Qualifier: note

Containing this text:

☒ Ignore Case ☒ Allow Partial Match

☐ Match Any Word

And:

☐ Up to: bases long

And:

☐ At least: bases long

And:

☐ Up to: exons long

And:

☐ At least: exons long

And:

☐ Contains introns without GT/GC start and AG end

And by:

☒ Amino acid motif: MEDSSEA

☒ Forward Strand Features ☒ Reverse Strand Features

Select View Close

Space for a search
term or amino acid
motif

In the next part of the exercise we will be looking at the region containing the *rifⁱⁿ* genes in more detail. They are located at the end of the chromosomes, in the subtelomeric region. We are going to extract this region from the whole chromosome sequence. Then we will aim to write and save new EMBL format files which will include just the annotation and DNA for this region.

2 Click 'Edit'

1 Select the region containing rifins by clicking with the left mouse button and dragging.

3 Click 'Subsequence (and Features)'

Artemis Entry Edit: Pf3D7_03.fasta

Entry: ☒ Pf3D7_03.fasta ☒ Pf3D7_03.fasta

21144 selected bases on forward strand 45270..66413

Undo ⌘U
Redo ⌘R

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

repeat_region 1 3600 telomeric repeat
repeat_region 7854 10142 R-CG7
ncRNA 11759 11851 c present in reich
repeat_region 11962 13888 repl1
repeat_region 13934 34268 rep20
RNA 16327 17615 non-coding RNA
RNA 35837 35977 non-coding RNA
RNA 36965 44482 var genes encode
repeat_region 45740 46339 R-FA3
repeat_region 46369 47579 A-type rifin (PM
RNA 47972 48170 non-coding RNA
RNA 49082 49138 non-coding RNA
RNA 49772 51152 c ;query 441-441;c
RNA 52280 53273 c stevor (subtelom
RNA 53671 53751 present in reichenow
RNA 55307 55389 c non-coding RNA
RNA 55390 56584 B-type rifin (PMID:18197962)
RNA 58307 58519 c term=structural;date=20100621;qualifier=added new gene based on similarity to PFA0035c,
term=structural;date=20100621;qualifier=added new gene based on similarity to PFA0035c,

Note the entry names have changed

4

A new Artemis window will appear displaying only the region that you have highlighted.

Note the bases have been renumbered from the first base you selected.

Artemis Entry Edit

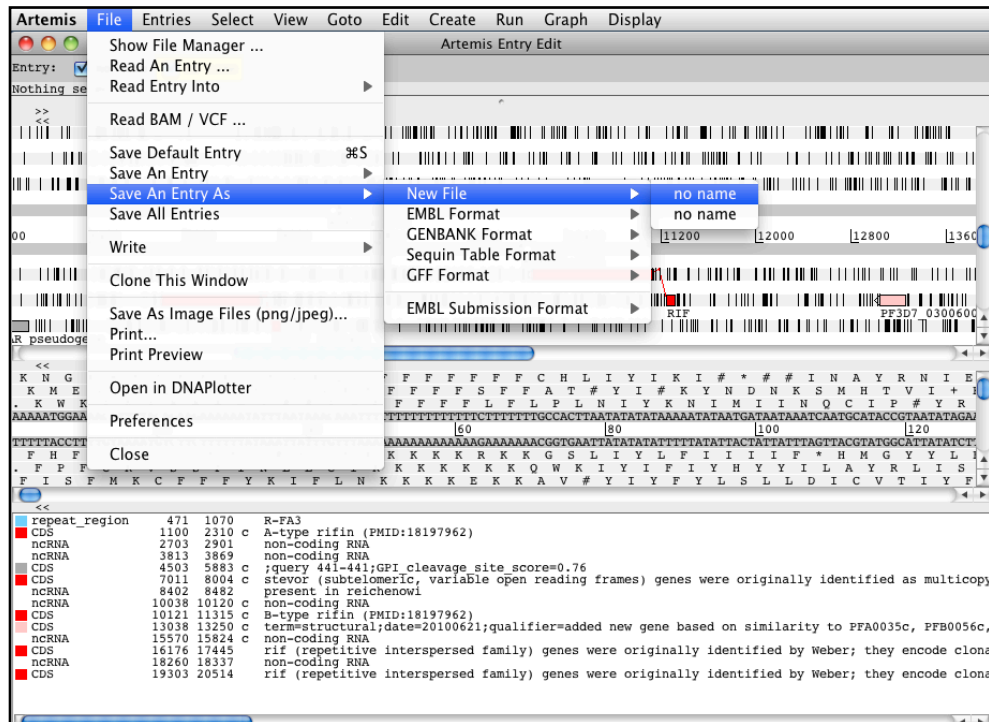
Entry: ☒ no name ☒ no name

Nothing selected

RNAzID:2753
RNAzID:2755

PF3D7_0300400
PF3D7_0300600
PF3D7_0300800
PF3D7_0301000
PF3D7_0301200
PF3D7_0301400
PF3D7_0301600
PF3D7_0301800
PF3D7_0302000
PF3D7_0302200
PF3D7_0302400
PF3D7_0302600
PF3D7_0302800
PF3D7_0303000
PF3D7_0303200
PF3D7_0303400
PF3D7_0303600
PF3D7_0303800
PF3D7_0304000
PF3D7_0304200
PF3D7_0304400
PF3D7_0304600
PF3D7_0304800
PF3D7_0305000
PF3D7_0305200
PF3D7_0305400
PF3D7_0305600
PF3D7_0305800
PF3D7_0306000
PF3D7_0306200
PF3D7_0306400
PF3D7_0306600
PF3D7_0306800
PF3D7_0307000
PF3D7_0307200
PF3D7_0307400
PF3D7_0307600
PF3D7_0307800
PF3D7_0308000
PF3D7_0308200
PF3D7_0308400
PF3D7_0308600
PF3D7_0308800
PF3D7_0309000
PF3D7_0309200
PF3D7_0309400
PF3D7_0309600
PF3D7_0309800
PF3D7_0310000
PF3D7_0310200
PF3D7_0310400
PF3D7_0310600
PF3D7_0310800
PF3D7_0311000
PF3D7_0311200
PF3D7_0311400
PF3D7_0311600
PF3D7_0311800
PF3D7_0312000
PF3D7_0312200
PF3D7_0312400
PF3D7_0312600
PF3D7_0312800
PF3D7_0313000
PF3D7_0313200
PF3D7_0313400
PF3D7_0313600
PF3D7_0313800
PF3D7_0314000
PF3D7_0314200
PF3D7_0314400
PF3D7_0314600
PF3D7_0314800
PF3D7_0315000
PF3D7_0315200
PF3D7_0315400
PF3D7_0315600
PF3D7_0315800
PF3D7_0316000
PF3D7_0316200
PF3D7_0316400
PF3D7_0316600
PF3D7_0316800
PF3D7_0317000
PF3D7_0317200
PF3D7_0317400
PF3D7_0317600
PF3D7_0317800
PF3D7_0318000
PF3D7_0318200
PF3D7_0318400
PF3D7_0318600
PF3D7_0318800
PF3D7_0319000
PF3D7_0319200
PF3D7_0319400
PF3D7_0319600
PF3D7_0319800
PF3D7_0320000
PF3D7_0320200
PF3D7_0320400
PF3D7_0320600
PF3D7_0320800
PF3D7_0321000
PF3D7_0321200
PF3D7_0321400
PF3D7_0321600
PF3D7_0321800
PF3D7_0322000
PF3D7_0322200
PF3D7_0322400
PF3D7_0322600
PF3D7_0322800
PF3D7_0323000
PF3D7_0323200
PF3D7_0323400
PF3D7_0323600
PF3D7_0323800
PF3D7_0324000
PF3D7_0324200
PF3D7_0324400
PF3D7_0324600
PF3D7_0324800
PF3D7_0325000
PF3D7_0325200
PF3D7_0325400
PF3D7_0325600
PF3D7_0325800
PF3D7_0326000
PF3D7_0326200
PF3D7_0326400
PF3D7_0326600
PF3D7_0326800
PF3D7_0327000
PF3D7_0327200
PF3D7_0327400
PF3D7_0327600
PF3D7_0327800
PF3D7_0328000
PF3D7_0328200
PF3D7_0328400
PF3D7_0328600
PF3D7_0328800
PF3D7_0329000
PF3D7_0329200
PF3D7_0329400
PF3D7_0329600
PF3D7_0329800
PF3D7_0330000
PF3D7_0330200
PF3D7_0330400
PF3D7_0330600
PF3D7_0330800
PF3D7_0331000
PF3D7_0331200
PF3D7_0331400
PF3D7_0331600
PF3D7_0331800
PF3D7_0332000
PF3D7_0332200
PF3D7_0332400
PF3D7_0332600
PF3D7_0332800
PF3D7_0333000
PF3D7_0333200
PF3D7_0333400
PF3D7_0333600
PF3D7_0333800
PF3D7_0334000
PF3D7_0334200
PF3D7_0334400
PF3D7_0334600
PF3D7_0334800
PF3D7_0335000
PF3D7_0335200
PF3D7_0335400
PF3D7_0335600
PF3D7_0335800
PF3D7_0336000
PF3D7_0336200
PF3D7_0336400
PF3D7_0336600
PF3D7_0336800
PF3D7_0337000
PF3D7_0337200
PF3D7_0337400
PF3D7_0337600
PF3D7_0337800
PF3D7_0338000
PF3D7_0338200
PF3D7_0338400
PF3D7_0338600
PF3D7_0338800
PF3D7_0339000
PF3D7_0339200
PF3D7_0339400
PF3D7_0339600
PF3D7_0339800
PF3D7_0340000
PF3D7_0340200
PF3D7_0340400
PF3D7_0340600
PF3D7_0340800
PF3D7_0341000
PF3D7_0341200
PF3D7_0341400
PF3D7_0341600
PF3D7_0341800
PF3D7_0342000
PF3D7_0342200
PF3D7_0342400
PF3D7_0342600
PF3D7_0342800
PF3D7_0343000
PF3D7_0343200
PF3D7_0343400
PF3D7_0343600
PF3D7_0343800
PF3D7_0344000
PF3D7_0344200
PF3D7_0344400
PF3D7_0344600
PF3D7_0344800
PF3D7_0345000
PF3D7_0345200
PF3D7_0345400
PF3D7_0345600
PF3D7_0345800
PF3D7_0346000
PF3D7_0346200
PF3D7_0346400
PF3D7_0346600
PF3D7_0346800
PF3D7_0347000
PF3D7_0347200
PF3D7_0347400
PF3D7_0347600
PF3D7_0347800
PF3D7_0348000
PF3D7_0348200
PF3D7_0348400
PF3D7_0348600
PF3D7_0348800
PF3D7_0349000
PF3D7_0349200
PF3D7_0349400
PF3D7_0349600
PF3D7_0349800
PF3D7_0350000
PF3D7_0350200
PF3D7_0350400
PF3D7_0350600
PF3D7_0350800
PF3D7_0351000
PF3D7_0351200
PF3D7_0351400
PF3D7_0351600
PF3D7_0351800
PF3D7_0352000
PF3D7_0352200
PF3D7_0352400
PF3D7_0352600
PF3D7_0352800
PF3D7_0353000
PF3D7_0353200
PF3D7_0353400
PF3D7_0353600
PF3D7_0353800
PF3D7_0354000
PF3D7_0354200
PF3D7_0354400
PF3D7_0354600
PF3D7_0354800
PF3D7_0355000
PF3D7_0355200
PF3D7_0355400
PF3D7_0355600
PF3D7_0355800
PF3D7_0356000
PF3D7_0356200
PF3D7_0356400
PF3D7_0356600
PF3D7_0356800
PF3D7_0357000
PF3D7_0357200
PF3D7_0357400
PF3D7_0357600
PF3D7_0357800
PF3D7_0358000
PF3D7_0358200
PF3D7_0358400
PF3D7_0358600
PF3D7_0358800
PF3D7_0359000
PF3D7_0359200
PF3D7_0359400
PF3D7_0359600
PF3D7_0359800
PF3D7_0360000
PF3D7_0360200
PF3D7_0360400
PF3D7_0360600
PF3D7_0360800
PF3D7_0361000
PF3D7_0361200
PF3D7_0361400
PF3D7_0361600
PF3D7_0361800
PF3D7_0362000
PF3D7_0362200
PF3D7_0362400
PF3D7_0362600
PF3D7_0362800
PF3D7_0363000
PF3D7_0363200
PF3D7_0363400
PF3D7_0363600
PF3D7_0363800
PF3D7_0364000
PF3D7_0364200
PF3D7_0364400
PF3D7_0364600
PF3D7_0364800
PF3D7_0365000
PF3D7_0365200
PF3D7_0365400
PF3D7_0365600
PF3D7_0365800
PF3D7_0366000
PF3D7_0366200
PF3D7_0366400
PF3D7_0366600
PF3D7_0366800
PF3D7_0367000
PF3D7_0367200
PF3D7_0367400
PF3D7_0367600
PF3D7_0367800
PF3D7_0368000
PF3D7_0368200
PF3D7_0368400
PF3D7_0368600
PF3D7_0368800
PF3D7_0369000
PF3D7_0369200
PF3D7_0369400
PF3D7_0369600
PF3D7_0369800
PF3D7_0370000
PF3D7_0370200
PF3D7_0370400
PF3D7_0370600
PF3D7_0370800
PF3D7_0371000
PF3D7_0371200
PF3D7_0371400
PF3D7_0371600
PF3D7_0371800
PF3D7_0372000
PF3D7_0372200
PF3D7_0372400
PF3D7_0372600
PF3D7_0372800
PF3D7_0373000
PF3D7_0373200
PF3D7_0373400
PF3D7_0373600
PF3D7_0373800
PF3D7_0374000
PF3D7_0374200
PF3D7_0374400
PF3D7_0374600
PF3D7_0374800
PF3D7_0375000
PF3D7_0375200
PF3D7_0375400
PF3D7_0375600
PF3D7_0375800
PF3D7_0376000
PF3D7_0376200
PF3D7_0376400
PF3D7_0376600
PF3D7_0376800
PF3D7_0377000
PF3D7_0377200
PF3D7_0377400
PF3D7_0377600
PF3D7_0377800
PF3D7_0378000
PF3D7_0378200
PF3D7_0378400
PF3D7_0378600
PF3D7_0378800
PF3D7_0379000
PF3D7_0379200
PF3D7_0379400
PF3D7_0379600
PF3D7_0379800
PF3D7_0380000
PF3D7_0380200
PF3D7_0380400
PF3D7_0380600
PF3D7_0380800
PF3D7_0381000
PF3D7_0381200
PF3D7_0381400
PF3D7_0381600
PF3D7_0381800
PF3D7_0382000
PF3D7_0382200
PF3D7_0382400
PF3D7_0382600
PF3D7_0382800
PF3D7_0383000
PF3D7_0383200
PF3D7_0383400
PF3D7_0383600
PF3D7_0383800
PF3D7_0384000
PF3D7_0384200
PF3D7_0384400
PF3D7_0384600
PF3D7_0384800
PF3D7_0385000
PF3D7_0385200
PF3D7_0385400
PF3D7_0385600
PF3D7_0385800
PF3D7_0386000
PF3D7_0386200
PF3D7_0386400
PF3D7_0386600
PF3D7_0386800
PF3D7_0387000
PF3D7_0387200
PF3D7_0387400
PF3D7_0387600
PF3D7_0387800
PF3D7_0388000
PF3D7_0388200
PF3D7_0388400
PF3D7_0388600
PF3D7_0388800
PF3D7_0389000
PF3D7_0389200
PF3D7_0389400
PF3D7_0389600
PF3D7_0389800
PF3D7_0390000
PF3D7_0390200
PF3D7_0390400
PF3D7_0390600
PF3D7_0390800
PF3D7_0391000
PF3D7_0391200
PF3D7_0391400
PF3D7_0391600
PF3D7_0391800
PF3D7_0392000
PF3D7_0392200
PF3D7_0392400
PF3D7_0392600
PF3D7_0392800
PF3D7_0393000
PF3D7_0393200
PF3D7_0393400
PF3D7_0393600
PF3D7_0393800
PF3D7_0394000
PF3D7_0394200
PF3D7_0394400
PF3D7_0394600
PF3D7_0394800
PF3D7_0395000
PF3D7_0395200
PF3D7_0395400
PF3D7_0395600
PF3D7_0395800
PF3D7_0396000
PF3D7_0396200
PF3D7_0396400
PF3D7_0396600
PF3D7_0396800
PF3D7_0397000
PF3D7_0397200
PF3D7_0397400
PF3D7_0397600
PF3D7_0397800
PF3D7_0398000
PF3D7_0398200
PF3D7_0398400
PF3D7_0398600
PF3D7_0398800
PF3D7_0399000
PF3D7_0399200
PF3D7_0399400
PF3D7_0399600
PF3D7_0399800
PF3D7_0400000
PF3D7_0400200
PF3D7_0400400
PF3D7_0400600
PF3D7_0400800
PF3D7_0401000
PF3D7_0401200
PF3D7_0401400
PF3D7_0401600
PF3D7_0401800
PF3D7_0402000
PF3D7_0402200
PF3D7_0402400
PF3D7_0402600
PF3D7_0402800
PF3D7_0403000
PF3D7_0403200
PF3D7_0403400
PF3D7_0403600
PF3D7_0403800
PF3D7_0404000
PF3D7_0404200
PF3D7_0404400
PF3D7_0404600
PF3D7_0404800
PF3D7_0405000
PF3D7_0405200
PF3D7_0405400
PF3D7_0405600
PF3D7_0405800
PF3D7_0406000
PF3D7_0406200
PF3D7_0406400
PF3D7_0406600
PF3D7_0406800
PF3D7_0407000
PF3D7_0407200
PF3D7_0407400
PF3D7_0407600
PF3D7_0407800
PF3D7_0408000
PF3D7_0408200
PF3D7_0408400
PF3D7_0408600
PF3D7_0408800
PF3D7_0409000
PF3D7_0409200
PF3D7_0409400
PF3D7_0409600
PF3D7_0409800
PF3D7_0410000
PF3D7_0410200
PF3D7_0410400
PF3D7_0410600
PF3D7_0410800
PF3D7_0411000
PF3D7_0411200
PF3D7_0411400
PF3D7_0411600
PF3D7_0411800
PF3D7_0412000
PF3D7_0412200
PF3D7_0412400
PF3D7_0412600
PF3D7_0412800
PF3D7_0413000
PF3D7_0413200
PF3D7_0413400
PF3D7_0413600
PF3D7_0413800
PF3D7_0414000
PF3D7_0414200
PF3D7_0414400
PF3D7_0414600
PF3D7_0414800
PF3D7_0415000
PF3D7_0415200
PF3D7_0415400
PF3D7_0415600
PF3D7_0415800
PF3D7_0416000
PF3D7_0416200
PF3D7_0416400
PF3D7_0416600
PF3D7_0416800
PF3D7_0417000
PF3D7_0417200
PF3D7_0417400
PF3D7_0417600
PF3D7_0417800
PF3D7_0418000
PF3D7_0418200
PF3D7_0418400
PF3D7_0418600
PF3D7_0418800
PF3D7_0419000
PF3D7_0419200
PF3D7_0419400
PF3D7_0419600
PF3D7_0419800
PF3D7_0420000
PF3D7_0420200
PF3D7_0420400
PF3D7_0420600
PF3D7_0420800
PF3D7_0421000
PF3D7_0421200
PF3D7_0421400
PF3D7_0421600
PF3D7_0421800
PF3D7_0422000
PF3D7_0422200
PF3D7_0422400
PF3D7_0422600
PF3D7_0422800
PF3D7_0423000
PF3D7_0423200
PF3D7_0423400
PF3D7_0423600
PF3D7_0423800
PF3D7_0424000
PF3D7_0424200
PF3D7_0424400
PF3D7_0424600
PF3D7_0424800
PF3D7_0425000
PF3D7_0425200
PF3D7_0425400
PF3D7_0425600
PF3D7_0425800
PF3D7_0426000
PF3D7_0426200
PF3D7_0426400
PF3D7_0426600
PF3D7_0426800
PF3D7_0427000
PF3D7_0427200
PF3D7_0427400
PF3D7_0427600
PF3D7_0427800
PF3D7_0428000
PF3D7_0428200
PF3D7_0428400
PF3D7_0428600
PF3D7_0428800
PF3D7_0429000
PF3D7_0429200
PF3D7_0429400
PF3D7_0429600
PF3D7_0429800
PF3D7_0430000
PF3D7_0430200
PF3D7_0430400
PF3D7_0430600
PF3D7_0430800
PF3D7_0431000
PF3D7_0431200
PF3D7_0431400
PF3D7_0431600
PF3D7_0431800
PF3D7_0432000
PF3D7_0432200
PF3D7_0432400
PF3D7_0432600
PF3D7_0432800
PF3D7_0433000
PF3D7_0433200
PF3D7_0433400
PF3D7_0433600
PF3D7_0433800
PF3D7_0434000
PF3D7_0434200
PF3D7_0434400
PF3D7_0434600
PF3D7_0434800
PF3D7_0435000
PF3D7_0435200
PF3D7_0435400
PF3D7_0435600
PF3D7_0435800
PF3D7_0436000
PF3D7_0436200
PF3D7_0436400
PF3D7_0436600
PF3D7_0436800
PF3D7_0437000
PF3D7_0437200
PF3D7_0437400
PF3D7_0437600
PF3D7_0437800
PF3D7_0438000
PF3D7_0438200
PF3D7_0438400
PF3D7_0438600
PF3D7_0438800
PF3D7_0439000
PF3D7_0439200
PF3D7_0439400
PF3D7_0439600
PF3D7_0439800
PF3D7_0440000
PF3D7_0440200
PF3D7_0440400
PF3D7_0440600
PF3D7_0440800
PF3D7_0441000
PF3D7_0441200
PF3D7_0441400
PF3D7_0441600
PF3D7_0441800
PF3D7_0442000
PF3D7_0442200
PF3D7_0442400
PF3D7_0442600
PF3D7_0442800
PF3D7_0443000
PF3D7_0443200
PF3D7_0443400
PF3D7_0443600
PF3D7_0443800
PF3D7_0444000
PF3D7_0444200
PF3D7_0444400
PF3D7_0444600
PF3D7_0444800
PF3D7_0445000
PF3D7_0445200
PF3D7_0445400
PF3D7_0445600
PF3D7_0445800
PF3D7_0446000
PF3D7_0446200
PF3D7_0446400
PF3D7_0446600
PF3D7_0446800
PF3D7_0447000
PF3D7_0447200
PF3D7_0447400
PF3D7_0447600
PF3D7_0447800
PF3D7_0448000
PF3D7_0448200
PF3D7_0448400
PF3D7_0448600
PF3D7_0448800
PF3D7_0449000
PF3D7_0449200
PF3D7_0449400
PF3D7_0449600
PF3D7_0449800
PF3D7_0450000
PF3D7_0450200
PF3D7_0450400
PF3D7_0450600
PF3D7_0450800
PF3D7_0451000
PF3D7_0451200
PF3D7_0451400
PF3D7_0451600
PF3D7_0451800
PF3D7_0452000
PF3D7_0452200
PF3D7_0452400
PF3D7_0452600
PF3D7_0452800
PF3D7_0453000
PF3D7_0453200
PF3D7_0453400
PF3D7_0453600
PF3D7_0453800
PF3D7_0454000
PF3D7_0454200
PF3D7_0454400
PF3D7_0454600
PF3D7_0454800
PF3D7_0455000
PF3D7_0455200
PF3D7_0455400
PF3D7_0455600
PF3D7_0455800
PF3D7_0456000
PF3D7_0456200
PF3D7_0456400
PF3D7_0456600
PF3D7_0456800
PF3D7_0457000
PF3D7_0457200
PF3D7_0457400
PF3D7_0457600
PF3D7_0457800
PF3D7_0458000
PF3D7_0458200
PF3D7_0458400
PF3D7_0458600
PF3D7_0458800
PF3D7_0459000
PF3D7_0459200
PF3D7_0459400
PF3D7_0459600
PF3D7_0459800
PF3D7_0460000
PF3D7_0460200
PF3D7_0460400
PF3D7_0460600
PF3D7_0460800
PF3D7_0461000
PF3D7_046

Note that the two entries on the grey Entry line are now denoted 'no name', they represent the same information in the same order as the original Artemis window but simply have no assigned name. So click on the File menu then 'Save an entry as' and then 'New file'. Another menu will ask you to choose one of the entries listed. At this point they will both be called 'no name'. Left click on the top entry in the list. A window will appear asking you to give this file a name. The new files can be saved in different formats.



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

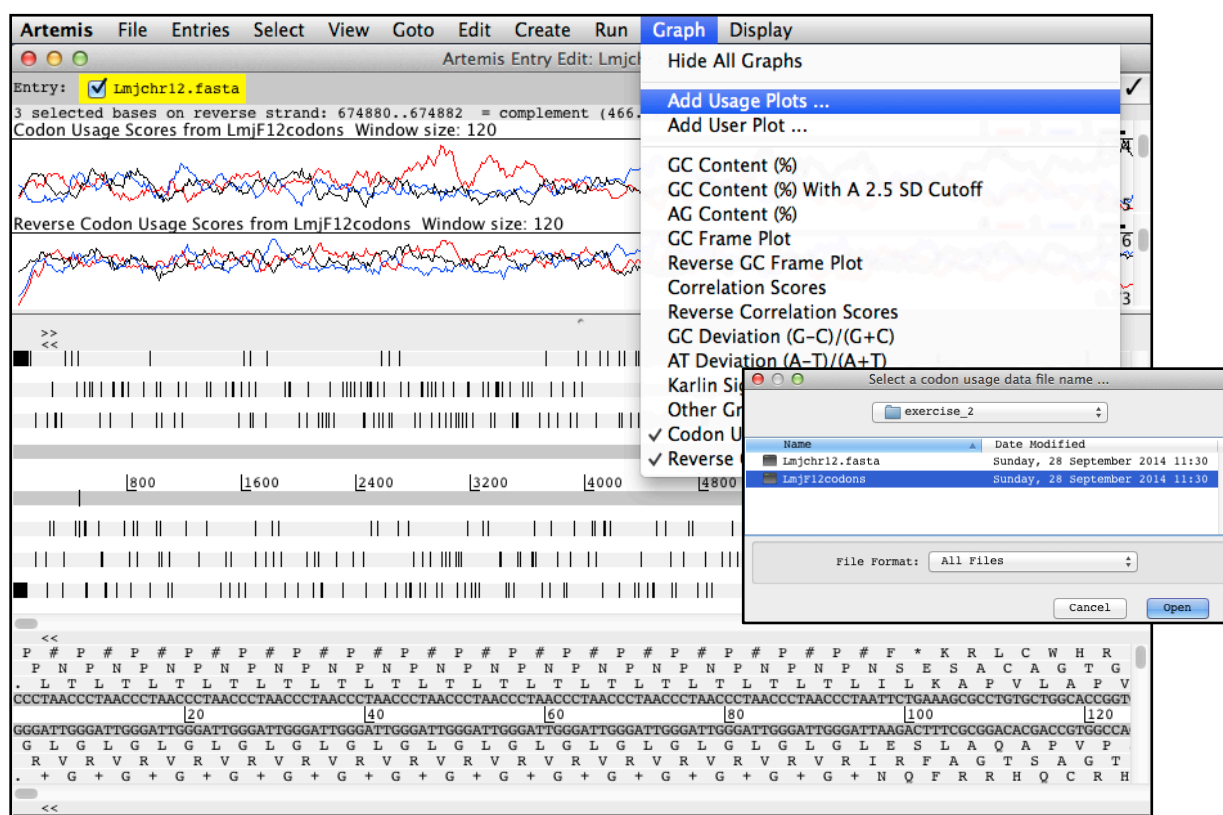
Artemis Exercise 2

We are now switching to a different organism. The following exercise demonstrates how to use Artemis as a tool for structural annotation. Given a length of chromosome with no existing annotation Artemis can mark up ORFs above a given size. This also shows how codon usage plots can be exploited in gene model prediction.

If you haven't already closed the previous session of Artemis, do so now. Double click the ARTEMIS Icon on your Desktop and navigate to the directory Module_1_Artemis, exercise_2 and open the sequence file Lmjchr12.fasta.

Next, open the codon usage table file LmjF12codons by selecting 'Add Usage Plots' from the Graph menu. Codon usage is a very good indicator of coding capacity in *Leishmania* genomes where there is a much more prominent codon bias for some amino acids.

Note, we will cover the use of RNAseq data in gene prediction later on during the course.



Select the first 100 kbs of sequence on the positive strand either by highlighting the sequence in the sequence window (use shift and click to select the final base) or choose the 'Base Range' option in the select menu and enter '1..100000'.

With this region selected, select 'Mark ORFs in Range' from the Create menu. When prompted for minimum ORF size enter 100. Note that this results in the creation of a new entry called 'ORFS_100+'. You can experiment with a range of ORF sizes by de-selecting this entry and repeating the first steps in this process.

Note that the marked up ORFs vary in colour from pale to navy blue. This colouring reflects the codon usage support for this model with darker blue being highly supported by codon usage.

Try selecting some of the newly created features in the gene window. Double clicking on one of these will bring up the predicted peptide sequence in the bottom window. You can rapidly move to the N- or C-terminus of the predicted peptide by holding down ctrl, and then left or right arrow respectively.

Note that we have chosen only to generate ORFs for the positive strand for this example. In a genome not organized into transcription units we would normally do likewise for the reverse strand as well.

The screenshot shows the Artemis genome browser interface. The 'Create' menu is open, and 'Mark ORFs in Range ...' is selected. The interface displays a sequence window with a 'New Entry' label, a gene window with a 'Predicted ORF' label, and a bottom window showing a protein sequence and a table of ORF statistics.

Protein Sequence:

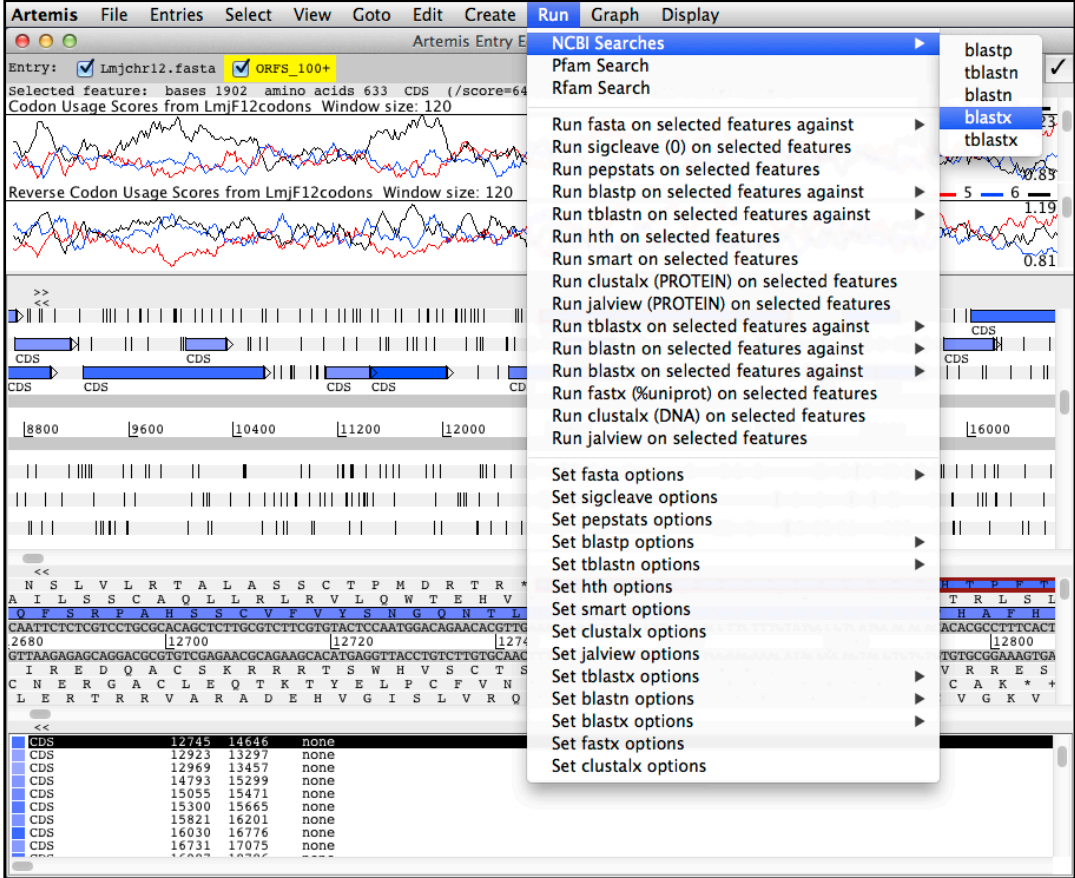
```

<<
R W R L P Q M W F S K P F V C S S Y A P # S L A E S L E V A E S S C H T R R L L C C A
A D G G S R K C G S R S R S C A L V M R H N L S L L L C L S L H H L Y I L E C C F A A
Q M A A P A N V L E A V R V L + L C A I I S R S F S A C R S I I Y T Y S N A A L L
C A G A T G C G G C T C C G C A A A T G G T T C T C G A A G C C G T C G T G T C T A G T A T G C C C A T A A T C T C G C T C C T T C T C G C T T G C G T C C A T C A T C A T A C T C G A A T G C T T T G C T G C
[1720] [1740] [1760] [1780] [1800] [1820] [1840]
G T C T A C C G C C G A G G G C G T T T A C A C C A A G A G C T T C G G C A A G C A C A C A G A T C A A T A C G C G G T A T T A G A G A G C G A G G A A G A G A C G A A C A G C G A G G T A G T A G A T A T G T A T G A G C T T A C G A C G A A C G A C G
S P P E R L H P E R L R E H A R T I R W L R E S R R Q K D S W * R Y M S S H Q K A A
C I A A G A F T T R S A T R T S + N H A M I E R E K E A Q R E M M + V Y E F A A K S C
L H R S G C I H N E F G N T H E L # A G Y D R A G E R S T A G D D I C V R I S S Q Q
  
```

ORF Statistics Table:

ORF	Start	End	Score
CDS	460	960	none
CDS	964	1617	none
CDS	1188	1583	none
CDS	1783	2577	none
CDS	2707	3735	none
CDS	3986	4798	none
CDS	4965	6044	none
CDS	5917	6288	none
CDS	6405	8081	none

Although some of these predictions are likely to be correct, there is considerable overlap between predicted ORFs, and many are small and unsupported by codon usage. To validate/negate our predicted models we need to do further sequence comparison. This can be done with a tool such as ACT (to be discussed later in the Comparative Genomics Module), or with one of the integrated Blast options in Artemis. Select the ORF at position 12745, click on it, then select RUN>NCBI Searches>blastx. This will open a browser window with NCBI results.



conserved hypothetical protein [Leishmania major strain Friedlin]									
Sequence ID: ref XP_001681612.1 Length: 620 Number of Matches: 1									
▶ See 1 more title(s)									
Range	1 to 320	GenPept	Graphics		Next Match	◀	Previous Match		
Score	Expect	Method		Identities	Positives		Gaps		Frame
1238 bits (3203)	0.0	Compositional matrix		620/620(100%)	620/620(100%)		0/620(0%)		+1
Query	40	MHTPTPTSSPSFFPYVASLP	SSPIAHLDAHAGGLLRQV	PLVNRSGESFLLQHLQGV				219	
Sbjct	1	MHTPTPTSSPSFFPYVASLP	SSPIAHLDAHAGGLLRQV	PLVNRSGESFLLQHLQGV					
Query	220	DKCDSADALPTTHSAKAA	RVPWATQTPPSSCKTAE	CTCVLKKRKHGHSVADPTGE				399	
Sbjct	61	DKCDSADALPTTHSAKAA	RVPWATQTPPSSCKTAE	CTCVLKKRKHGHSVADPTGE					
Query	400	VLLRKNKGVOVQIVYNSKI	PSVYGNNQLAKMKEREKRENS	PLFKYPLAVGEGAAQEE				579	
Sbjct	180	VLLRKNKGVOVQIVYNSKI	PSVYGNNQLAKMKEREKRENS	PLFKYPLAVGEGAAQEE					
Query	520	ARKVLQLERHCQQAHRMK	EGLEGRKARLAIEAAVQV	KACETARAEADARKIKID				759	
Sbjct	181	ARKVLQLERHCQQAHRMK	EGLEGRKARLAIEAAVQV	KACETARAEADARKIKID					
Query	760	GEAVSETAAKLTLEERAD	NADVAVSGWNGKEEDRRRLAAET	RTQLAEENFA				939	
Sbjct	241	GEAVSETAAKLTLEERAD	NADVAVSGWNGKEEDRRRLAAET	RTQLAEENFA					
Query	940	AEGRRAAKGQAEQAEAR	AVRVEHMQRLQAE	RVKDLGERRNNAALRGQAS	RRERN			1119	
Sbjct	301	AEGRRAAKGQAEQAEAR	AVRVEHMQRLQAE	RVKDLGERRNNAALRGQAS	RRERN				
Query	1120	RANSADVRLQAGPSSWLS	DAVERNRQDLQARQKMETD	AVNVLRLAQKADQAQRRN				1299	
Sbjct	361	RANSADVRLQAGPSSWLS	DAVERNRQDLQARQKMETD	AVNVLRLAQKADQAQRRN					
Query	1300	DROYAAEYAKENLQFREV	YHARORRQEQEQLQAAEAT	ATVYRQAHADAAARRG	QSV			1479	
Sbjct	421	DROYAAEYAKENLQFREV	YHARORRQEQEQLQAAEAT	ATVYRQAHADAAARRG	QSV				
Query	1480	PLFLFWAAGQGAZKAIK	DANFRFEDLRQKQKDEAR	QEEAERADRALVETDRL				1659	
Sbjct	481	PLFLFWAAGQGAZKAIK	DANFRFEDLRQKQKDEAR	QEEAERADRALVETDRL					
Query	1660	AREAVERRKREKREKEL	RKTLKLEAIJAEKRGV	GDRACAAADVTPVATAE	NLYRC			1839	
Sbjct	541	AREAVERRKREKREKEL	RKTLKLEAIJAEKRGV	GDRACAAADVTPVATAE	NLYRC				
Query	1840	PTVTEGLPASATDFQVRR	K					600	
Sbjct	601	PTVTEGLPASATDFQVRR	620						

Not surprisingly, the top hit is to a gene on chromosome 12 in *L. major*, a hypothetical protein.

Now that we know that this is a real gene we can make a few adjustments. First, open the gene builder window by selecting the ORF and pressing E. This will open a text window where we can add annotations on the gene. Start by deleting the current 'automatic' annotations in this window. Try entering the text in the gene builder shown below to record gene ID, predicted product and a colour code that will distinguish this gene from the automatically generated ORFs.

Press 'E' to open the gene builder for this ORF

This is a coding sequence (CDS). To get an idea of other feature types available, open this pull-down menu.

When done, push the apply button.

Based on the NCBI blast results we can adjust the N-terminus of this model to the correct start codon. To automatically position the sequence window at the N-terminus of the gene model push ctrl-<left arrow>.

Go to Edit>Trim Selected Feature>To Next Met (or ctrl-T), then reposition the sequence window at the new start as described above. Continue until the start resembles the NCBI blast results. If trimmed passed the desired start codon the model can be reset through Edit>Extend Selected Feature>To Previous Stop Codon, or ctrl-Q.

Artemis File Entries Select View Goto Edit Create Run Graph Display

Artemis Entry Edit: Lmjchr12.fasta

Entry: ☒ Lmjchr12.fasta ☒ ORFS_100+

Selected feature: bases 1902 amino acids 633 LmjF12.0070 (/systematic id="LmjF12.0070" /product="hypothetical protein, conserved")

Codon Usage Scores from LmjF12codons Window size: 172

Reverse Codon Usage Scores from LmjF12codons Window size: 150

1. Move to the N-terminus of the gene model with ctrl - <left arrow>.

2. Trim to the next start codon with ctrl-T

Sequence window showing protein sequence: KRRTTRWRPSTLYAVNMHTHTPT

Feature	Start	End	Score
CDS	12969	13457	none
CDS	14793	15299	none
CDS	15055	15471	none
CDS	15300	15665	none
CDS	15821	16201	none
CDS	16030	16776	none
CDS	16731	17075	none

There are more than 20 protein coding genes in the first 100 kbs of chromosome 12. See how many of these you can find by repeating the steps in the past slides.

IMPORTANT!! Any changes made to the predicted ORFs will be written to an entry file called ORFS_100+. When you're done with gene predictions follow the steps below to save these entries to the sequence file instead. Make sure all of the annotated features have a /colour=10 in their gene builder window.

The screenshot shows the Artemis genome browser interface. The 'Select' menu is open, showing options like 'All', 'All Bases', 'None', 'By Key', 'CDS Features without /pseudogene', 'All CDS Features', 'Same Key', 'Features Matching Qualifier', 'Open Reading Frame', 'Features Overlapping Selection', 'Features Within Selection', 'Base Range ...', 'Feature AA Range ...', and 'Toggle Selection'. A yellow callout box points to the 'Features Matching Qualifier' option with the text: '1. Select an annotated gene model'. Another yellow callout box points to the 'Features Matching Qualifier' option with the text: '2. Select Features Matching Qualifier from Select menu'. A third yellow callout box points to the 'Select a qualifier name' dialog box, which has 'colour' selected in the dropdown, with the text: '3. Select colour as a qualifier This will select all features of the same colour.' The dialog box has 'OK' and 'Cancel' buttons. The background shows a genomic track with various features and a list of CDS features at the bottom.

Feature	Start	End	Qualifier
CDS	11445	12026	none
CDS	12504	12809	none
CDS	12745	14646	none
CDS	12923	13297	none
CDS	12969	13457	none
CDS	14793	15299	none
CDS	15055	15471	none
CDS	15300	15665	none
CDS	15821	16201	none

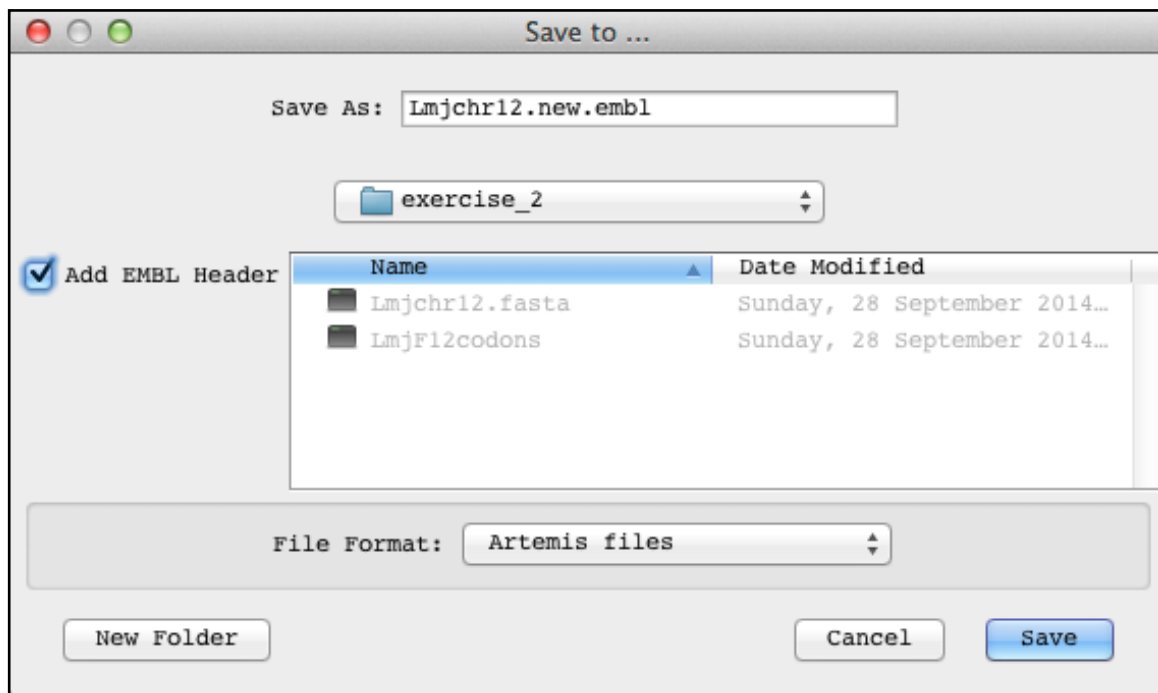
4. From the Edit menu, select 'copy selected features', then select the sequence file Lmjchr12.fasta

5. After the features have been copied to Lmjchr12.fasta, de-select ORFS_100+. Only annotated ORFs should remain.

Feature	Start	End	Score
CDS	11445	12026	
CDS	12504	12809	none
CDS	12745	14646	
CDS	12923	13297	none
CDS	12969	13457	none
CDS	14793	15299	none
CDS	15655	15671	none

6. From the File menu, select save an Entry as > EMBL format > Lmjchr12.fasta.

Feature	Start	End	Score
CDS	11445	12026	
CDS	12745	14646	
CDS	16030	16776	



Save the sequence file as Lmjchr12.new.embl.

Optional exercise

This optional exercise demonstrates how to use Artemis to construct queries for features within features. In the exercise below we will load a file containing SNP data, then retrieve a list of all CDS that overlap with SNP features.

Files required:

Tb927_01_v4.embl - Contains sequence and annotation for *T. brucei* chromosome 1

Tb927_01_v4snps.embl - Contains SNP features for *T. brucei* chromosome 1

Navigate to the directory Module_1_Artemis, optional_exercise.

Use the file manager to open Tb927_01_v4.embl, then as shown in previous exercises select File>Read Entry >Tb927_01_v4snps.embl.

Artemis Entry Edit: Tb927_01_v4.embl

Entry: ☒ Tb927_01_v4.embl ☒ Tb927_01_v4snps.embl

Selected feature: bases 1 variation (/note="Majority of WCS reads call C, One WCS read calls T"/note="type")

SNP features

After loading in the Tb927_01_v4snps.embl file the SNP features will appear in the feature list below the last feature in the Tb927_01_v4.embl file.

repeat unit	1064634	1064639	telomeric repeat hexamer	TTAGGG
repeat unit	1064640	1064645	telomeric repeat hexamer	TTAGGG
repeat unit	1064646	1064651	/label=Trpt	
repeat unit	1064652	1064657	telomeric repeat hexamer	TTAGGG
repeat unit	1064658	1064663	telomeric repeat hexamer	TTAGGG
repeat unit	1064664	1064669	/label=Trpt	
variation	10628	10628	Majority of WCS reads call C, One WCS read calls T	
variation	10685	10685	Majority of WCS reads call C, One WCS read calls A	
variation	10806	10806	Majority of WCS reads call C, One WCS read calls T	
variation	39982	39982	Majority of WCS reads call G, One WCS read calls T	
variation	40880	40880	Majority of WCS reads call C, One WCS read calls T	
variation	40910	40910	Majority of WCS reads call A, One WCS read calls C	
variation	40954	40954	Majority of WCS reads call G, One WCS read calls A	
variation	40965	40965	Majority of WCS reads call T, One WCS read calls C	
variation	40981	40981	Two WCS reads call C, Two WCS reads call T	

Artemis File Entries **Select** View Goto Edit Create Run Graph Display

Entry: ☒ Tb927_01_v4.embl

Selected feature: bases 1

Feature Selector ...

- All %A
- All Bases
- None %N
- By Key
- CDS Features without /pseudogene
- All CDS Features
- Same Key**
- Features Matching Qualifier
- Open Reading Frame
- Features Overlapping Selection
- Features Within Selection
- Base Range ...
- Feature AA Range ...
- Toggle Selection

1. Select any 'variation' feature from the feature selector

2. Select all features with the same key (variation)

variation	10628	10628	Majority of WCS reads call C, One WCS read calls T
variation	10805	10805	Majority of WCS reads call C, One WCS read calls A
variation	39986	39986	Majority of WCS reads call C, One WCS read calls T
variation	40880	40880	Majority of WCS reads call C, One WCS read calls T
variation	40910	40910	Majority of WCS reads call C, One WCS read calls T
variation	40954	40954	Majority of WCS reads call C, One WCS read calls T
variation	40965	40965	Majority of WCS reads call C, One WCS read calls T
variation	40981	40981	Two WCS reads call C, Two WCS reads call T

An alternative way to select all SNP features is to select 'By Key', then select 'variation'.

Artemis File Entries **Select** View Goto Edit Create Run Graph Display

Entry: ☒ Tb927_01_v4.embl

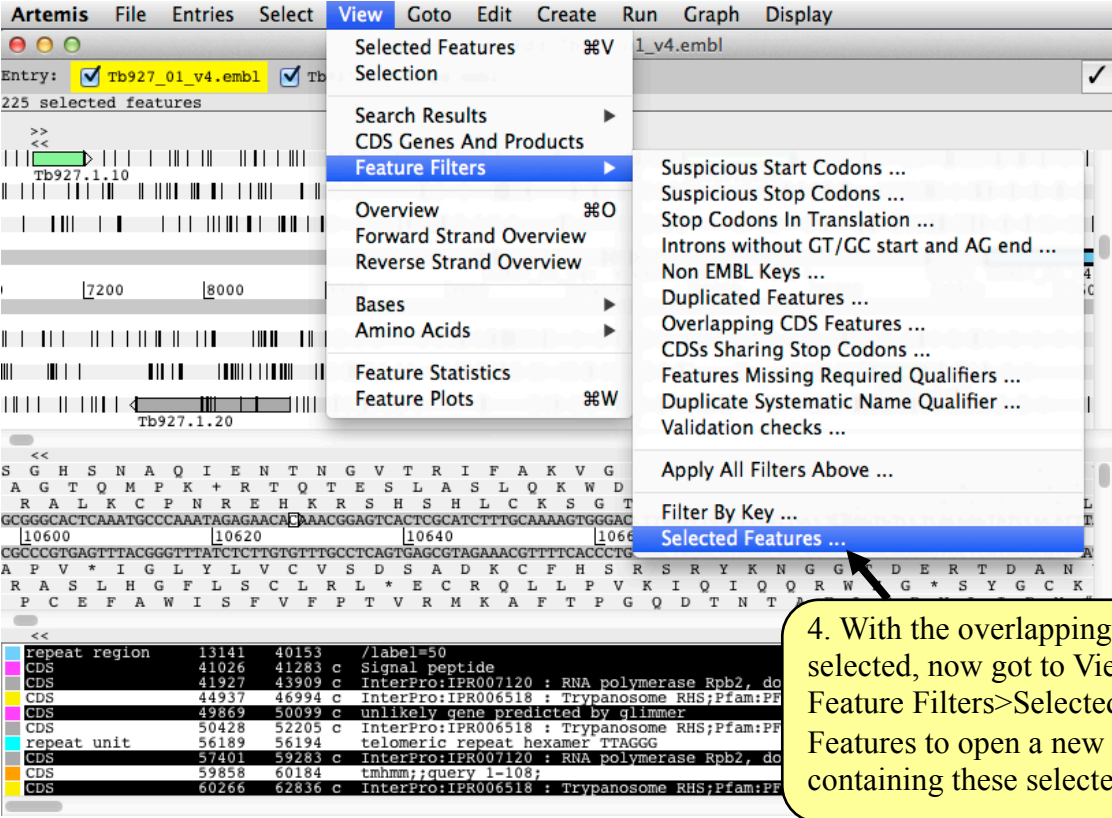
963 selected features

Feature Selector ...

- All %A
- All Bases
- None %N
- By Key
- CDS Features without /pseudogene
- All CDS Features
- Same Key
- Features Matching Qualifier
- Open Reading Frame
- Features Overlapping Selection**
- Features Within Selection
- Base Range ...
- Feature AA Range ...
- Toggle Selection

3. With all SNP features selected, now select 'Features Overlapping Selection' from the Select menu.

repeat_unit	1064634	1064639	telomeric repeat hexamer TTAGGG
repeat_unit	1064640	1064645	telomeric repeat hexamer TTAGGG
repeat_unit	1064646	1064651	/label=Trpt
repeat_unit	1064652	1064657	telomeric repeat hexamer TTAGGG
repeat_unit	1064658	1064663	telomeric repeat hexamer TTAGGG
repeat_unit	1064664	1064669	/label=Trpt
variation	10628	10628	Majority of WCS reads call C, One WCS read calls T
variation	10805	10805	Majority of WCS reads call C, One WCS read calls A
variation	39986	39986	Majority of WCS reads call C, One WCS read calls T
variation	40880	40880	Majority of WCS reads call C, One WCS read calls T
variation	40910	40910	Majority of WCS reads call C, One WCS read calls T
variation	40954	40954	Majority of WCS reads call C, One WCS read calls A
variation	40965	40965	Majority of WCS reads call C, One WCS read calls A
variation	40981	40981	Two WCS reads call C, Two WCS reads call T



Artemis File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Tb927_01_v4.embl ☒ Tb927_01_v4.embl

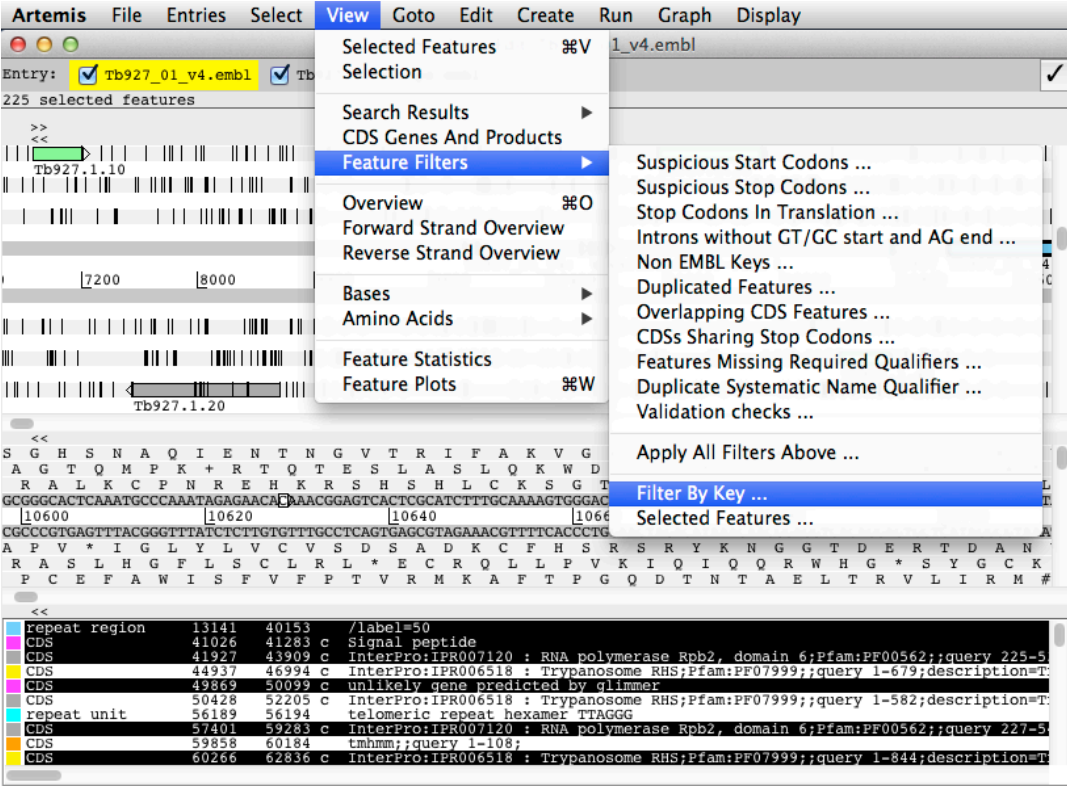
225 selected features

Feature Filters

- Suspicious Start Codons ...
- Suspicious Stop Codons ...
- Stop Codons In Translation ...
- Introns without GT/GC start and AG end ...
- Non EMBL Keys ...
- Duplicated Features ...
- Overlapping CDS Features ...
- CDSs Sharing Stop Codons ...
- Features Missing Required Qualifiers ...
- Duplicate Systematic Name Qualifier ...
- Validation checks ...
- Apply All Filters Above ...
- Filter By Key ...
- Selected Features ...**

4. With the overlapping features selected, now got to View select Feature Filters>Selected Features to open a new window containing these selected results.

All the features overlapping with the SNP features should now appear in a new window. Note that this window contains not only CDS features, but features such as 5' UTRs, repeat regions and other miscellaneous features that overlap with SNPs. To see only CDS features we need to apply a second filter. With the non-overlapping features still selected, select View>Feature Filters>Filter by Key. Select CDS for the Key, and only CDS containing SNPs should appear in the filter window.



Artemis File Entries Select View Goto Edit Create Run Graph Display

Entry: ☒ Tb927_01_v4.embl ☒ Tb927_01_v4.embl

225 selected features

Feature Filters

- Suspicious Start Codons ...
- Suspicious Stop Codons ...
- Stop Codons In Translation ...
- Introns without GT/GC start and AG end ...
- Non EMBL Keys ...
- Duplicated Features ...
- Overlapping CDS Features ...
- CDSs Sharing Stop Codons ...
- Features Missing Required Qualifiers ...
- Duplicate Systematic Name Qualifier ...
- Validation checks ...
- Apply All Filters Above ...
- Filter By Key ...**
- Selected Features ...

Filter by Key

Key: CDS

Filter by Key

Repeat region 13141 40153 /label=50

CDS 41026 41283 c Signal peptide

CDS 41927 43909 c InterPro:IPR007120 : RNA polymerase Rpb2, domain 6;Pfam:PF00562;query 225-5

CDS 44937 46994 c InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;query 1-679;description=T

CDS 49869 50099 c unlikely gene predicted by glimmer

CDS 50428 52205 c InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;query 1-582;description=T

repeat unit 56189 56194 telomeric repeat hexamer TTAGGG

CDS 57401 59283 c InterPro:IPR007120 : RNA polymerase Rpb2, domain 6;Pfam:PF00562;query 227-5

CDS 59858 60184 tmhmm;query 1-108;

CDS 60266 62836 c InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;query 1-844;description=T

Artemis Entry Edit: Tb927_01_v4.embl

Entry: ☒ Tb927_01_v4.embl ☒ Tb927_01_v4snps.embl

225 selected features

misc_feature

- 10_signal
- 35_signal
- 3' UTR
- 5' UTR
- BLASTCDS
- BLASTN_HIT
- BLASTX_HIT
- CAAT_signal
- CDS
- CDS_AFTER
- CDS_BEFORE
- CDS_after
- CDS_before
- CDS_motif
- CRUNCH_D
- CRUNCH_X
- C_region
- D-loop
- D_segment
- GC_signal
- GFF
- J_segment
- LTR
- N_region

Feature Type	Start	End	Description
repeat region	13141	40153	/label=50
CDS	41026	41283	c Signal peptide
CDS	41927	43909	c InterPro:IPR000000
CDS	44937	46994	c InterPro:IPR000000
CDS	49869	50099	c unlikely gene
CDS	50428	52205	c InterPro:IPR000000
repeat unit	56189	56194	telomeric repeat
CDS	57401	59283	c InterPro:IPR000000
CDS	59858	60184	c tmhmm; query 1
CDS	60266	62836	c InterPro:IPR006518 : Trypanosome RHS; Pfam:PF07999; query 1-844; description=T

Other Queries to Try:

1. Try performing the 'reverse' query, select all SNPs that overlap with CDS features.
2. Save a list of features to a file by right clicking on the feature filter window and Selecting 'Save List to File'
3. Use the Select Menu to select all features with the same 'key'
4. Use the Filter menu to look for suspicious gene models - missing start codons, missing stop codons, stop codons in translation and duplicated features.
5. Search for a qualifier value (try 'hypothetical protein'), in the Edit menu, select 'Find/Replace Qualifier Text'. Try doing a boolean search in the same way (try 'hypothetical AND conserved, or 'hypothetical AND unlikely').
6. Using the same option, find features containing duplicate qualifiers (more than one qualifier with the same value)