Gene Modeling Tutorial using Artemis

We are working on gene structures for *Ichthyophthirius multifilliis G5*, or “Ich”—a parasite that afflicts fish. Ich is a ciliate. Ciliates use a different codon set for stop codons; TGA is the only codon encoding a ‘stop’ in ciliates, unlike the universal codons of TGA, TAG, TAA, which most organisms use.

Introduction: Modeling genes in Artemis

We know from our experience working with the various kinds of evidence that some are more reliable than others. In our experience with Ich, we trust them in this order: transcripts & Training set >> TTA1_gw > paramecium > oxytricha & glimmerHMM > genezilla > augustus.

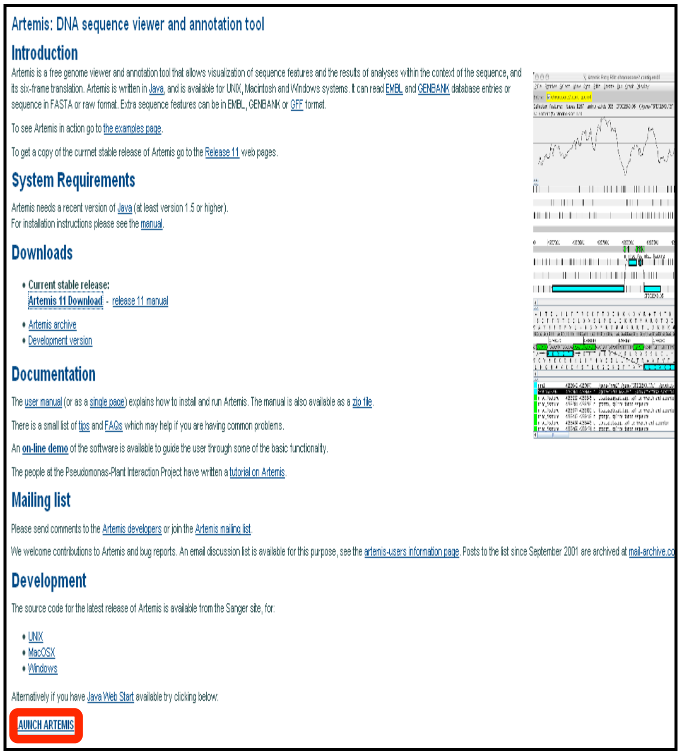
We ran Evidence modeler (EVM) after we ran genefinders, protein alignments and PASA. EVM combined the evidence using this ranking scheme, via its weights file:

TRANSCRIPT	alignAssembly-img1_122008_pasa	20
PREDICTION	Train	20
PROTEIN	genewise-Paramecium_peptide_v1.fa	7
PROTEIN	genewise-oxytricha_all.fasta	6
PROTEIN	genewise-TTA1.pep	8
PREDICTION	augustus	4
PREDICTION	genezilla	5
PREDICTION	glimmerHMM	6

The output of EVM is a series of gff3 files containing the evidence for each scaffold. EVM is the automated gene prediction for each locus. We will now examine these gene predictions in Artemis, and model some genes.

SECTION I. Open and configure Artemis.

Open Artemis: Using The Artemis link in course_links, bring up the Artemis page at the Sanger Institute. Locate the link to the Java Web Start version of Artemis. Click on the link outlined in red in the graphic, and a Welcome screen will come up.



Artemis: DNA sequence viewer and annotation tool

Introduction

Artemis is a free genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of the sequence, and its six-frame translation. Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA or raw format. Extra sequence features can be in EMBL, GENBANK or GFF format.

To see Artemis in action go to the examples page.

To get a copy of the currnet stable release of Artemis go to the Release 11 web pages.

System Requirements

Artemis needs a recent version of Java (at least version 1.5 or higher).
For installation instructions please see the manual.

Downloads

- Current stable release:
  Artemis 11 Download - release 11 manual
- Artemis archive
- Development version

Documentation

The user manual (or as a single page) explains how to install and run Artemis. The manual is also available as a zip file.

There is a small list of tips and FAQs which may help if you are having common problems.

An on-line demo of the software is available to guide the user through some of the basic functionality.

The people at the Pseudomonas-Plant Interaction Project have written a tutorial on Artemis.

Mailing list

Please send comments to the Artemis developers or join the Artemis mailing list.

We welcome contributions to Artemis and bug reports. An email discussion list is available for this purpose, see the artemis-users information page. Posts to the list since September 2001 are archived at mail-archive.co

Development

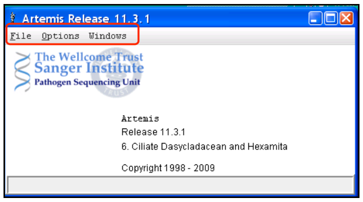
The source code for the latest release of Artemis is available from the Sanger site, for:

- UNIX
- MacOSX
- Windows

Alternatively if you have Java Web Start available try clicking below:

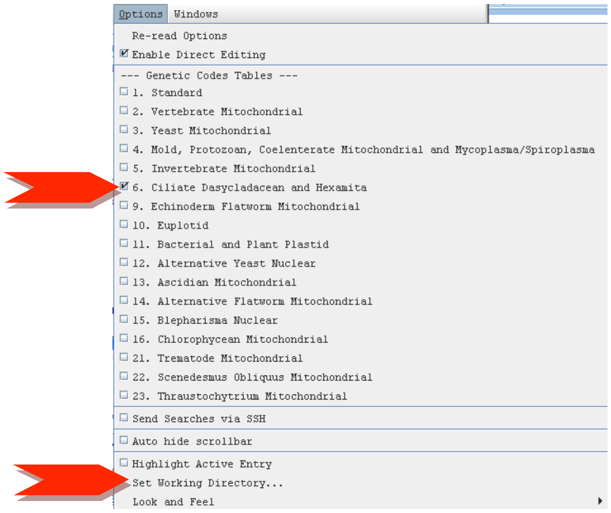
Setting up the options

The Artemis Welcome screen has 3 menu items: File, Options, and Windows.



Artemis
Release 11.3.1
6. Ciliate Dasycladacean and Hexamita

Copyright 1998 - 2009


Options Windows
Re-read Options
☑ Enable Direct Editing
--- Genetic Codes Tables ---
☐ 1. Standard
☐ 2. Vertebrate Mitochondrial
☐ 3. Yeast Mitochondrial
☐ 4. Mold, Protozoan, Coelenterate Mitochondrial and Mycoplasma/Spiroplasma
☐ 5. Invertebrate Mitochondrial
☑ 6. Ciliate Dasycladacean and Hexamita
☐ 9. Echinoderm Flatworm Mitochondrial
☐ 10. Euplotid
☐ 11. Bacterial and Plant Plastid
☐ 12. Alternative Yeast Nuclear
☐ 13. Ascidian Mitochondrial
☐ 14. Alternative Flatworm Mitochondrial
☐ 15. Blepharisma Nuclear
☐ 16. Chlorophycean Mitochondrial
☐ 21. Trematode Mitochondrial
☐ 22. Scenedesmus Obliquus Mitochondrial
☐ 23. Thraustochytrium Mitochondrial
☐ Send Searches via SSH
☐ Auto hide scrollbar
☐ Highlight Active Entry
Set Working Directory...
Look and Feel

- Open Options menu, check the box next to ”6. Ciliate Dasycladacean and Hexamita”
- Also click on the “Set working directory” option in the Options menu, near the end of the options list. Set the working directory to your flash drive’s data directory.

Open up the Ich scaffold sequence in Artemis

Open the File menu, and click on “Open...” A window labeled “Select a file…” will open.

- The first line reads “Look In: and has a menu with a directory. Make sure it is pointing to your data directory.
- The last line reads “Files of Type:”—Select “All Files”.
- Find the file “genome.fasta” in your data directory. Select it. Click the “Open” button.

This will bring up the Artemis viewer. Across the top are the following menus:

- File, Entries, Select, View, Goto, Edit, Create, Run, Graph, Display.

SECTION II: Entering the evidence

We will now open up the GFF3 files containing results of our searches and alignments:

Open, in order, each of the following:

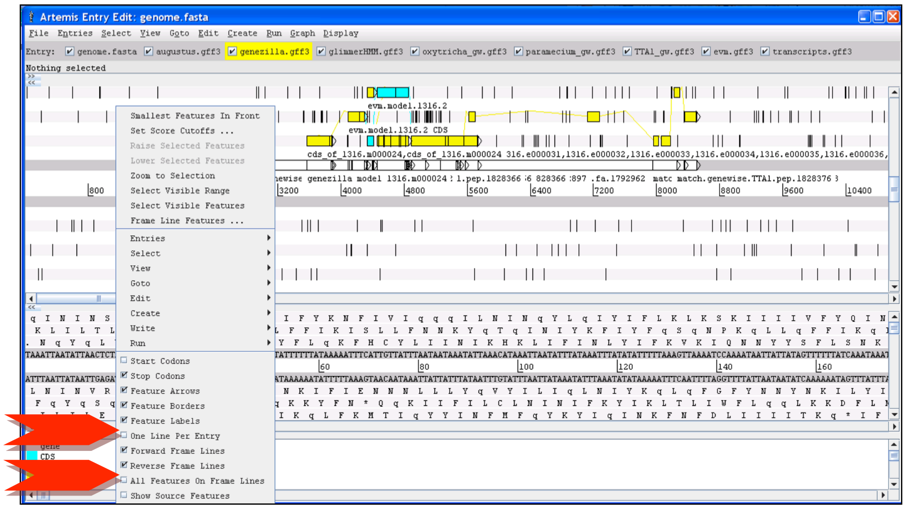
1) augustus.gff3
2) genezilla.gff3
3) glimmer.gff3
4) oxy_gw.gff3
5) para_gw.gff3
6) TTA1_gw.gff3
7) evm.gff3
8) PASA.gff3
9) Train.gff3

Open the File menu in the Artemis viewer, and choose the second line, “Read an entry…”

Change “Files of Type:” to **All Files**.

Using Artemis

You now have all of the evidence generated by genefinders, spliced protein alignments, and PASA loaded into Artemis.



<u>Put each kind of evidence on its own line</u>.

Right-click and check the following:

- One line per entry
- All features on frame lines

As a matter of preference, you might click on:

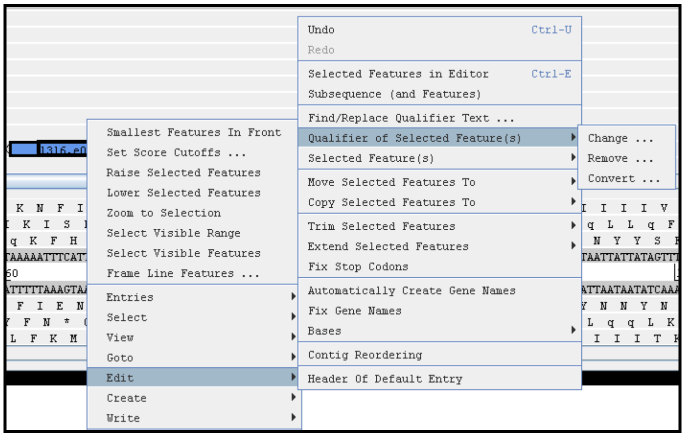
- Feature labels

Coloring tracks (optional)

You might wish to color the tracks so you can easily distinguish them.

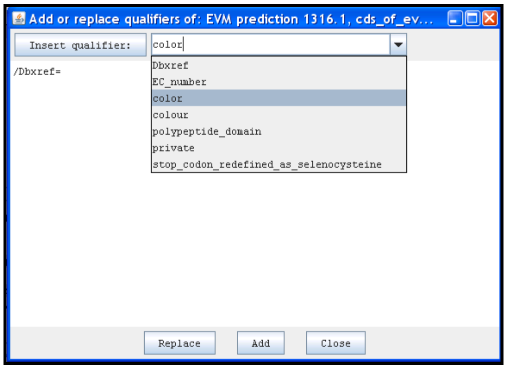
To choose only one track at a time, so each has a different color, you must uncheck all of the others on the Entry menu or on the Entry bar, which is immediately below the main menu. For example, uncheck everything except "augustus.gff3".

Then, on the Select menu, click on "All".
Right-clicking anywhere inside the viewer will bring up a menu.



- Choose "Edit", then
- Choose "Qualifier of Selected Feature(s), then
- Choose "Change …"

A box will pop up. The box is shown below.


Add or replace qualifiers of: EVM prediction 1316.1, cds_of_ev...
Insert qualifier:
color
/Dbxref=
Dbxref
EC_number
color
colour
polypeptide_domain
private
stop_codon_redefined_as_selenocysteine
Replace
Add
Close

Choose 'color ' from the pull-down menu and click the "Insert qualifier" button. This inserts the line:

/color=

Now, type: 100 149 237 [the code for Cornflower Blue.}

The line is now: /color= 100 149 237

Click "Replace" then "Close".

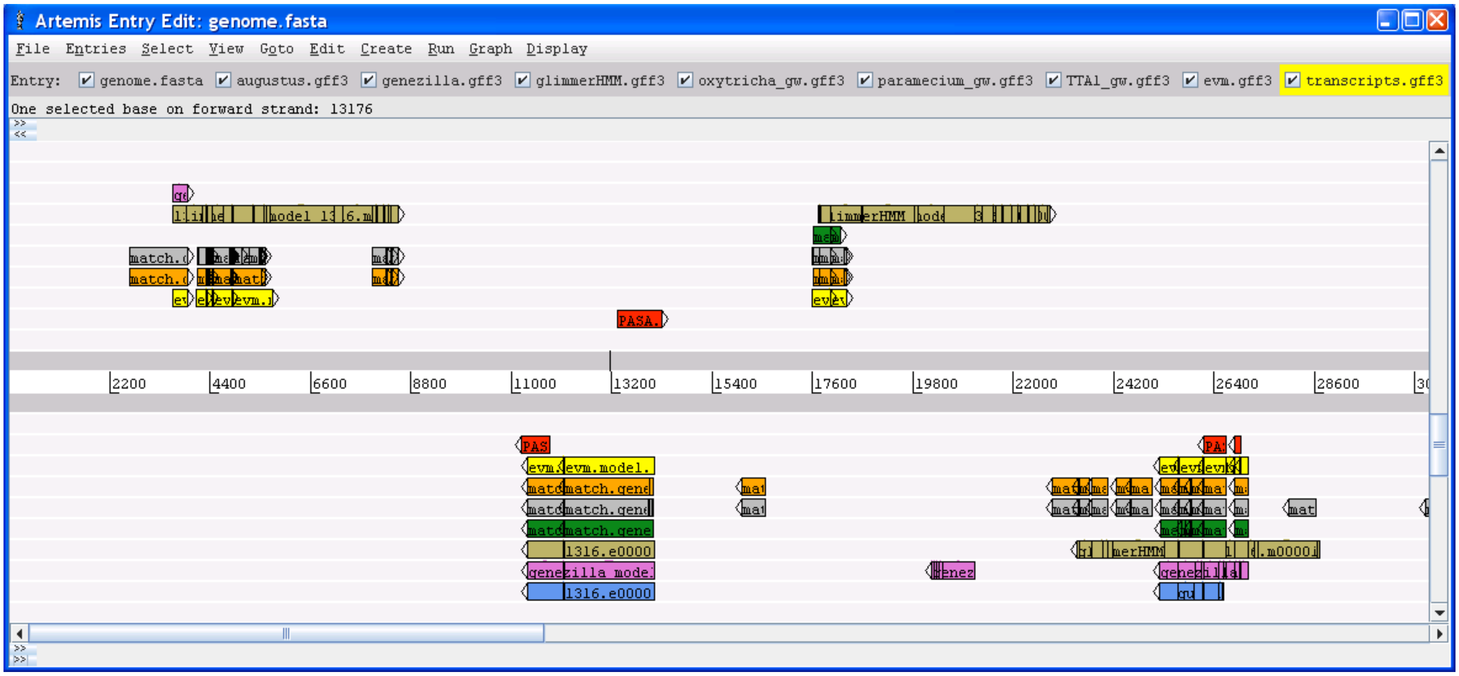
All of the augustus predictions should have turned blue.

Repeat this procedure for each color change. Here are some RGB color codes:

Forest Green	34	139	34
Dark Khaki	189	183	107
Cyan	0	255	255
Yellow	255	255	0
Goldenrod	218	165	32

Indian Red	205	92	92
Orange	255	165	0
Red	255	0	0
Orchid	218	112	214
Purple	160	32	240
Gray	190	190	190
Brown	139	119	101

Your result might look something like this:



Artemis Entry Edit: genome.fasta

Entry: ☑ genome.fasta ☑ augustus.gff3 ☑ genezilla.gff3 ☑ glimmerHMM.gff3 ☑ oxytricha_gw.gff3 ☑ paramecium_gw.gff3 ☑ TTA1_gw.gff3 ☑ evm.gff3 ☑ transcripts.gff3

One selected base on forward strand: 13176

The separation and coloring of lines of evidence, while not necessary, make Artemis tracks much easier to interpret.

Section III: Modeling genes

We have loaded lots of evidence into Artemis. The evidence included genefinder results, Genewise alignments of protein sets, transcript assembly alignments, and a training set of genes. All of these were used by Evidence modeler with the following weights:

TRANSCRIPT	alignAssembly-img1_122008_pasa	20
PREDICTION	Train	20
PROTEIN	genewise-Paramecium_peptide_v1.fa	7
PROTEIN	genewise-oxytricha_all.fasta	6
PROTEIN	genewise-TTA1.pep	8
PREDICTION	augustus	4
PREDICTION	genezilla	5
PREDICTION	glimmerHMM	6

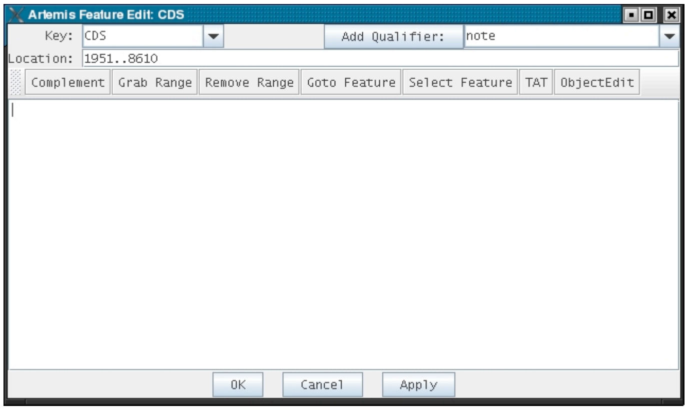
That is, the PASA transcripts and the training set had the most impact on the result, followed by Tetrahymena proteins, etc. Keep in mind that in our experience with Ich, we judge the quality of the evidence in this order: transcripts & Training set >> TTA1_gw > paramecium & glimmerHMM > oxytricha & glimmerHMM > genezilla > augustus. These are reflected in EVM. EVM is the automated gene prediction for each locus. Examine these gene predictions, and model some genes.

Model a gene

Set a default entry to save gene models to. Menu/Entries/Set Default Entry→genes.gff3. Then, save it as a GFF file: File/Save an Entry as… /GFF Format/genes.gff3. The program will ask if you want to overwrite the existing file. Say yes.

Zoom into the region 2000-9000. An easy way to do that is to click on the longest prediction in that region, and hit the "z" key on your keyboard.

Next, drag your mouse from right to left in the positive direction above the coordinate line (above the numbers). Menu/Create/Feature from Base Range (or Control-C). A window pops up:



Artemis Feature Edit: CDS

Key: CDS

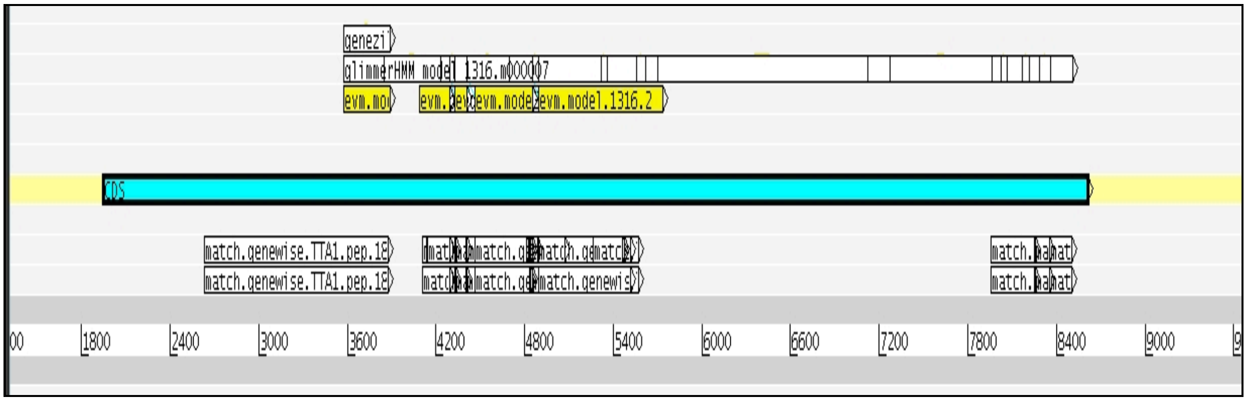
Add Qualifier: note

Location: 1951..8610

Click “Apply.”

Note that it has a line that says “Location:”. Your coordinates will differ from those shown.

Decide on the structure of the gene you are about to model. Is there an EVM model?



You might notice that EVM predicts two genes in this space. Glimmer predicted one, and genewise matches in three regions. What other evidence is there for the structure of this gene? Which is the best evidence, since there is no transcript evidence?

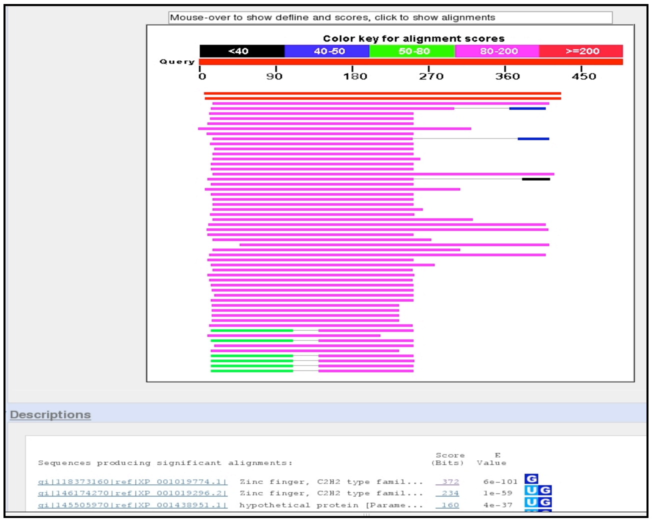
Let’s model the longest gene predicted by EVM in this region first:

1) click on the first EVM predicted exon, with coordinates 4088 to 4286;
2) click on “Grab Range” and the Location line will change to join(1951..8610,4088..4286)
3) Now delete the first two numbers, leaving: join(4088..4286)
4) Click on the next exon, grab range, and the Location line will change to : join(4088..4286,4330..4405)
5) Proceed to grab each exon until the Location line reads: join(4088..4286,4330..4405,4461..4848,4893..5726)
6) Click the “Apply” button. Click “OK”.

Is this a good gene model? How do we know if it should include the exons that are upstream?

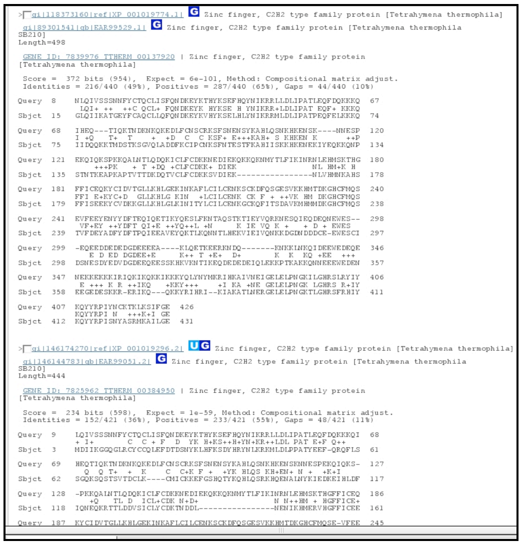
Check the gene model:
One way to gauge how to accurately model this gene is to use BLASTP. If your gene is not a unique gene, you may be able to discern its correct boundaries based upon orthology. Bring up firefox. Go to the BLASTP page at NCBI.

Click on your new gene model, and choose “View/Amino Acids of Selection as FASTA” and copy the sequence that pops up. Paste it into the text box on the BLASTP page.


Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
90
180
270
360
450
Descriptions

```
Sequences producing significant alignments:                    Score     E
                                                               (Bits)  Value

gi|118373160|ref|XP_001019774.1|  Zinc finger, C2H2 type famil...  372    6e-101
gi|146174270|ref|XP_001019296.2|  Zinc finger, C2H2 type famil...  234    1e-59
gi|145505970|ref|XP_001438951.1|  hypothetical protein [Parame...  160    4e-37
```



```
>gi|118373160|ref|XP_001019774.1| Zinc finger, C2H2 type family protein [Tetrahymena thermophila]
 gi|89301541|gb|EAR99529.1| Zinc finger, C2H2 type family protein [Tetrahymena thermophila
SB210]
Length=498

 GENE ID: 7839976 TTHERM_00137920 | Zinc finger, C2H2 type family protein
[Tetrahymena thermophila]

 Score =  372 bits (954),  Expect = 6e-101, Method: Compositional matrix adjust.
 Identities = 216/440 (49%), Positives = 287/440 (65%), Gaps = 44/440 (10%)

Query  8    NLQIVSSSNNFYCTQCLISFQNDKEYKTHYKSEFHQYNIKRRLLDLIPATLEQFDQKKKQ  67
             LQI+ ++   ++C QCL+ FQNDKEYK HYKSE H YNIKRR+LDLIPAT EQF+ KKKQ
Sbjct  15   GLQIIKATGEYFCAQCLLQFQNDKEYKVHYKSELHLYNIKRRMLDLIPATPEQFELKKKQ  74

Query  68   IHEQ---TIQKTNDKNKQKEDLFCNSCRKSFSNENSYKAHLQSNKHKENSK----NNESP  120
            I +Q    T+  T    +  +D  C  C KSF+ E+++KAH+ S KHKEN K       ++P
Sbjct  75   IIDQQKKTMDSTKSGVQLADDFKCIPCNKSFNTESTFKAHIISKKHKENEKIYEQKKQNP  134

Query  121  EKQIQKSPKKQAINTLQDQKICLFCDKKNEDIEKQKKQKNMYTLFIKINRNLEHMSKTHG  180
                +++PK    + T +DQ +CLFCDKK+ DIEK                NL HM+K H
Sbjct  135  STNTKEAPKAPTVTTDKDQTVCLFCDKKSVDIEK----------------NLVHMNKAHS  178

Query  181  FFICEQKYCIDVTGLLKHLGEKINKAFLCILCENKSCKDFQSGESVKKHMTDKGHCFMQS  240
            FFI E+KYC+D  GLLKHLG KIN  +LCILCENK CK F + ++VK HM DKGHCFMQS
Sbjct  179  FFISEEKYCVDKKGLLKHLGLKINITYLCILCENKGCKQFITSDAVKMHMMDKGHCFMQS  238

Query  241  EVFEEYENYYDFTEQIQETIKYQESLFKNTAQSTKTIEYVQRKNESQIEQDEQNEWES--  298
             VF+EY ++YDFT QI+E ++YQ++L +N      K IE VQ K +    + D + EWES
Sbjct  239  TVFDEYADFYDFTPQIEEAVEYQKTLKQNNTLHEKVIEIVQNKKDGDNDDDCE-EWESCI  297

Query  299  -EQEEDDEDEDGDEEEEA----KLQETKEERKNDQ-------KNKKINKQIDEEWEDEQE  346
              E D ED DGDEE+E     K++ T +E+  D+        K  K  KQ +EE  +++
Sbjct  298  DSNESDYEDVDGDEEQEESSKHKVKNTIKEQDEDEDEIQLEKKPTKAKKQNNEEEWEDEN  357

Query  347  NEKKEKKKIRIQKIKQKKIKKKYQLYNYMKRIHKAIVNEIGELELPNGKILGHRSLRYIY  406
             E +++ K R ++IKQ   +KKY+++    +I KA +NE GELELPNGK LGHRS R+IY
Sbjct  358  EEGEDESKKR-ERIKQ---QKKYRIHRI--KIAKATINERGELELPNGKTLGHRSFRHIY  411

Query  407  KQYYRPIYNCKTKLKSIFGE  426
            KQYYRPI N  +++K+I GE
Sbjct  412  KQYYRPISNYASRMKAILGE  431


>gi|146174270|ref|XP_001019296.2| Zinc finger, C2H2 type family protein [Tetrahymena thermophila]
 gi|146144783|gb|EAR99051.2| Zinc finger, C2H2 type family protein [Tetrahymena thermophila
SB210]
Length=444

 GENE ID: 7825962 TTHERM_00384950 | Zinc finger, C2H2 type family protein
[Tetrahymena thermophila]

 Score =  234 bits (598),  Expect = 1e-59, Method: Compositional matrix adjust.
 Identities = 152/421 (36%), Positives = 233/421 (55%), Gaps = 48/421 (11%)

Query  9    LQIVSSSNNFYCTQCLISFQNDKEYKTHYKSEFHQYNIKRRLLDLIPATLEQFDQKKKQI  68
            + I+        C  C + F   D  YK H+KS++H+YN+KR++LDL PAT E+F Q++
Sbjct  3    MDIIKGGQGLRCYCCQLEFDTDSNYKLHFKSDYHRYNLKRKMLDLPPATYEEF-QRQFLS  61

Query  69   HEQTIQKTNDKNKQKEDLFCNSCRKSFSNENSYKAHLQSNKHKENSKNNESPEKQIQKS-  127
              Q  Q T+  +  K    C  C+K F +  +YK HLQS KH+EN+ N +  +K+I
Sbjct  62   SGQKSQSTSVTDCLK----CMICKKEFGSHQTYKQHLQSRKHQENAINYKIEDKEIHLDF  117

Query  128  -PKKQAINTLQDQKICLFCDKKNEDIEKQKKQKNMYTLFIKINRNLEHMSKTHGFFICEQ  186
              +Q   TL D  ICL+CDK N+D+                N N++HM + HGFFICE+
Sbjct  118  IQNEQKRTTLDDVSICLYCDKTNDDL----------------NENIKHMERVHGFFICEE  161

Query  187  KYCIDVTGLLKHLGEKINKAFLCILCENKSCKDFQSGESVKKHMTDKGHCFMQSE-VFEE  245
```

This is an almost full-length match (84%) to a *Tetrahymena thermophila* protein, which is supported by a transcript. It is probably a real gene.

There is one start site just upstream of the start codon chosen by EVM. Is this one a "better" choice? How would you check?

Save your new gene model. Click on the new CDS, File/Save Default Entry (Ctrl-S).

Zoom in to the region on the forward strand at approximately 65kb. Choose the region approximately between 64kb and 66kb on the forward strand. Drag your cursor over that region. While there is a yellow box from dragging,

Click the menu item " Create", then choose "Feature from Base Range …"

The "Artemis Feature Edit: CDS" window will pop up. Under Location: you will see the coordinates that you swiped across with your mouse.

Now, click on the PASA evidence in one of the tiers. Going back to the Feature edit window, click on "Grab Range". The coordinates of that exon are shown in the "join" statement:

join(64173..66333,64830..65583) [Your first two numbers will be different. They are the range you selected with your mouse.]

Click on the next PASA exon in this evidence tier, "grab range" and repeat for the third exon. You should now have:

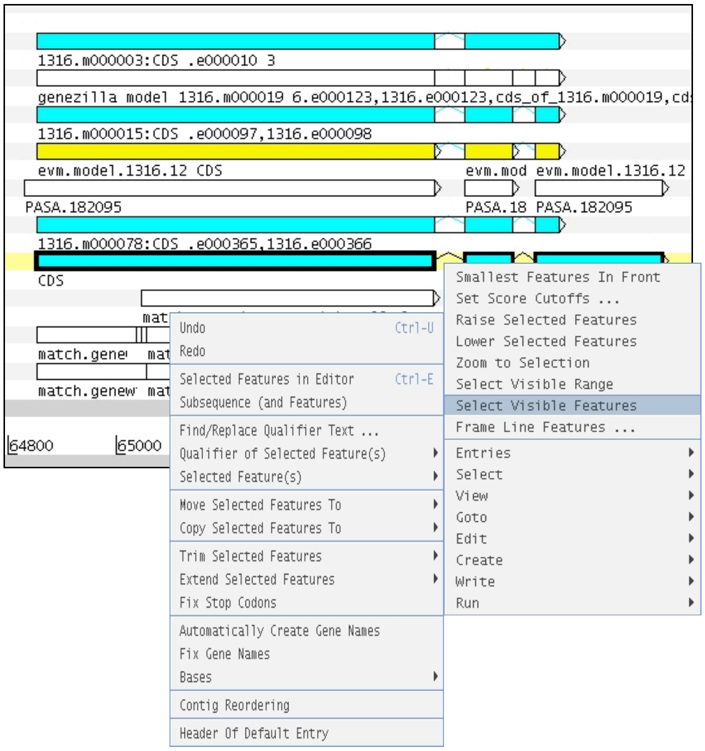
join(64173..66333,64830..65583,65639..65725,65768..66000)

Let's now get rid of the approximate coordinates we chose in the beginning with our mouse.

join(64830..65583,65639..65725,65768..66000)

Click OK.

Because this gene model was based on an EST assembly, it may have untranslated regions (UTRs). Determine the appropriate start and stop codons, if possible (BLASTP is one way to do this). To trim it back to the start and stop codons, right-click the model, select Edit/Trim Selected Features/To Met. To extend to the next stop codon, select Edit/Extend Selected Features/To Previous Stop Codon or To Next Stop Codon. In some cases, it is easier to visually identify the coordinates of the stop codon and enter them in the editor. In this case, the final coordinates would be 64830... 65811.



If you wish to add UTRs, follow the instructions in the Genewise tutorial.

As time permits, continue to model genes.

Save them in the genes.gff3 file as you go.