SEQUENCHER™

Tutorial for Windows and Macintosh

Working with Features

Gene Codes Corporation

TCA GENE
AGT CODES

Gene Codes Corporation
775 Technology Drive, Ann Arbor, MI 48108 USA
1.800.497.4939 (USA) +1.734.769.7249 (elsewhere)
+1.734.769.7074 (fax)
www.genecodes.com info@genecodes.com

Working with Features

Working with Features

GenBank database sequences are often used as a reference or exemplar in sequencing applications. Some GenBank sequences have been the subject of intense scrutiny and have been determined to be scientifically valid. GenBank sequences are often accompanied by detailed annotations that have been derived from laboratory experimentation or by comparison to similar sequences. Biologically important regions are described according to a set of standardized Feature Keys. These annotations are set out in a Feature Table. When additional information is needed, it is described using Feature Qualifiers.

Sequencher can import GenBank sequences with their associated Features Tables. When you use a GenBank sequence in your project, you will see a subset of features displayed according to either default or user preferences.

In the Overview, the Sequence and Contig Editors, the Variance Tables and in the Summary views, Sequencher's Reference Sequence function sets the base numbering and orientation of any contig into which it is assembled. Using a GenBank sequence as a Reference Sequence lets you combine the properties of both features and a Reference in your analyses.

GETTING STARTED

In this tutorial, you will take an imported GenBank sequence, change some of its display properties, edit features and use the sequence as a Reference Sequence. You will then assemble all the sequences in the project and compare them. You will first need to open a project.

- Launch Sequencher.
- Go to the File menu and select **Import > Sequencher Project…**
- Navigate to the Sample Data folder inside the Sequencher application folder.
- Choose the **Working with Features** project and select **Open**.

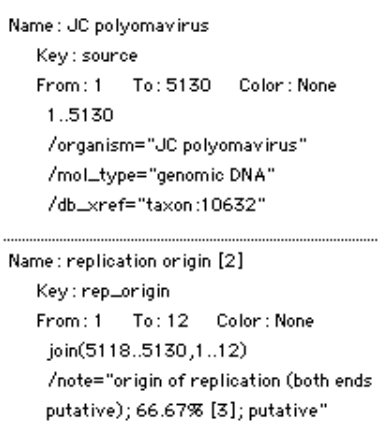
The project contains 4 small viral genomes. All of the sequences have been downloaded from GenBank. Hence there was no trace data for these sequences. Each one was accompanied by a Feature Table. Sequencher read the Feature Table for each sequence when the sequences were imported into the project and assigned default Display Styles.

REVIEWING A FEATURE LISTING

Each of the sequences in this project has a series of features. The first step will be to review the list of features for one particular sequence.

- Click on the icon for sequence NC_001699.
- Go to the **Sequence** menu and select **Feature Listing**.
- Scroll down the **Feature Listing** window and review some of the features.

Notice how each feature is separated from the next one by a dotted line. Each feature has a Name, a Feature Key, a location, a Display Style and optional qualifiers. The display of a feature outside of the **Feature Listing** is determined by settings in either the **Feature Editor** or the **Feature Default Style** in the **Feature, Motif** panel of **User Preferences...**



- Scroll down the **Feature Listing** window until you reach Name: CDS at location 227.

This feature has more detail. The display color is Blue and the Display Style is Protein Translation, so this feature will be displayed in blue text and will be accompanied by a protein translation in the Sequence Editor.

There are 8 Feature Qualifiers. The first tells you that the locus tag is "Jvgp1" and the db_xref tags alert you of a cross-reference to an external database. The remaining qualifiers provide information about the encoded protein.

```
Key : CDS
From : 277    To : 492    Color : Blue    Style : Protein Translation
 277..492
 /locus_tag="Jvgp1"
 /note="agnoprotein"
 /codon_start="1"
 /product="hypothetical protein"
 /protein_id="NP_043508.1"
 /db_xref="GI:9628643"
 /db_xref="GeneID:1489519"
 /translation="MVLRQLSRKASVKVSKTWSGTKKRAQRILIFLLEFLLD
 FCTGEDSVDGKKRQRHSGLTEQTYSALPEPKAT"
```

- Close the **Feature Listing** window.

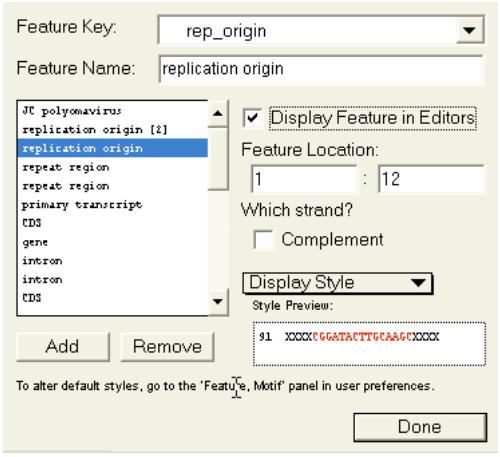
EDITING DISPLAY STYLES

- Ensure that NC_001699 is still selected in the Project Window.
- Go to the **Sequence** menu and select **Edit Features…**

You will see a list of features in the left hand pane. You can edit as many of the existing features as you wish. You can also use this editor to add sequence feature information. For instance features with a common Feature Key assignment will share the same default display characteristics.

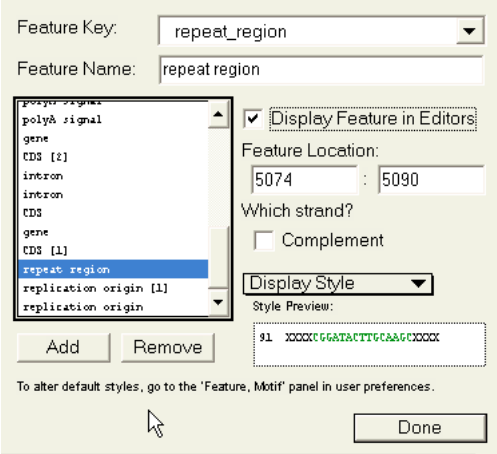
You are now going to change the **Display Style** for one feature.

- Select the third feature in the list: replication origin at location 1 to 12.
- Click in the **Display Feature in Editors** checkbox.
- Choose **Red** from the **Display Style** drop down menu



Now change the **Display Style** for another feature.

- Scroll down the list of features until you reach repeat region at location 5074 to 5090 (toward the end of the list).
- Click in the **Display Feature in Editors** checkbox.
- Choose **Green** from the **Display Style** drop down menu

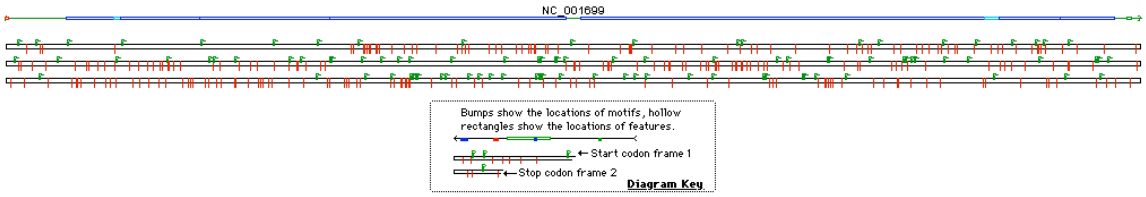


- Click the **Done** button to close the **Feature Editor**.
- Double click on the icon of NC_001699.
- Scroll down the **Sequence Editor**.

You will notice that the first 12 bases in the Sequence Editor are now displayed in red. This is the Replication Origin of the virus. You may have noticed that there are several origins described in the **Feature Listing**. The blue display marks the first CDS. The protein translation is displayed automatically. Scroll to the end of the sequence. You will see a short stretch of sequence in green. This is the second feature you edited.

Before proceeding, look at how Sequencher displays Features in other views.

- From the button bar in the Sequence Editor click on the **Overview** button.



You can see Features marked in blue. These are coding sequences; their Feature Key is CDS. You can also see the Replication Origin in red at the 5' end and the Repeat Region in green at the 3' end of the genome. Below this you can see the Start-Stop map. In that map, the red bars represent Stop codons and green flags represent Start codons. Notice how the third reading frame is heavily blocked with Stop codons.

- Close the **Overview** window.

Making your GenBank sequence into a Reference Sequence

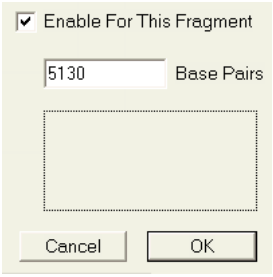
You are now ready to make the GenBank sequence serve as a Reference Sequence. The Reference Sequence function sets the base numbering and the orientation of the contig you are about to assemble. This allows you to compare a feature or a sequence difference to a fixed position. The Reference Sequence does *not* contribute to the consensus sequence.

- Click the sequence NC_001699 to select it.

- Choose **Sequence > Reference Sequence**.

You will notice that the icon of the Reference Sequence now contains an R. The sequence is now protected from editing. If you attempt to edit it you will receive a warning. This warning can be overridden.

The sequence you have marked as a Reference Sequence is from a DNA tumor virus. The viral genome is circular. You can enable circular numbering for the sequence.



- Click on the Reference Sequence NC_01699 icon to select it.
- Choose **Sequence > Set Circular Genome Size…**
- Click in the **Enable For This Fragment** checkbox.
- Type 5130 in the **Base Pairs** text box and click the **OK** button.

Circular numbering is now set. This feature is useful when you are working with circular genomes and need to work with bases that span the termini of the genome.

You can also change the start numbering of your sequence to a base position different than the one originally assigned.

- Double click on the Reference Sequence icon.
- Select base number 22
- From the **Sequence** menu choose **Set Base Number > As Base 1**.

Notice how the base numbering changes.

```
5110    GCCTCGGCCT CCTGTATATA TTAAAAAAAG GGAAGGGATG

20      GCTGCCAGCC AAGCATGAGC TCATACCTAG GGAGCCAACC

60      AGCTAACAGC CAGTAAACAA AGCACAAGGC TGTATATATA
```

This exercise also demonstrates the Circular Numbering feature.

- Restore the original numbering by selecting base 5110.
- From the **Sequence** menu choose **Set Base Number > As Base 1**.

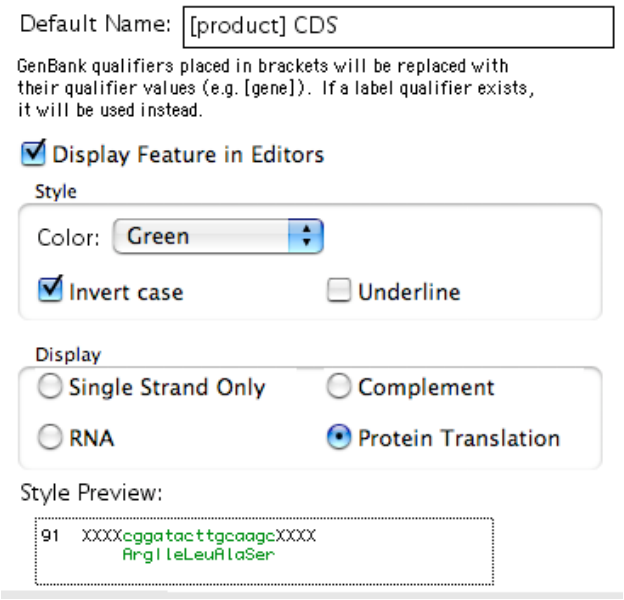
CHANGING DEFAULT FEATURE STYLES ON THE REFERENCE SEQUENCE

As you have seen, you can change your Display Styles one at a time in the **Feature Editor**. The default **Display Styles** are set as **User Preferences**. You can simultaneously change the display and the naming conventions of features that have the same Feature Key. While the display of some Feature Keys is turned off by default, you can choose to turn these on in **User Preferences**. You can also change the case of the text in the Sequence and Contig Editors, and you can underline text.

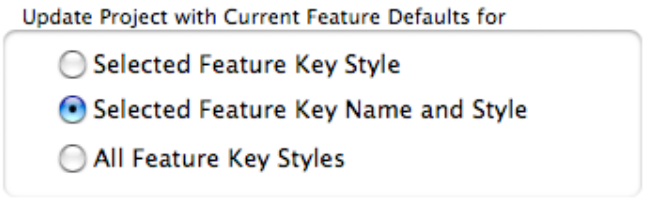
In this section of the tutorial you will change the style and name for all the features with a CDS Feature Key.

- Choose **Window > User Preferences…**
- Choose the **Display** tab.
- Click on **Feature, Motif**.
- Click on the **Define Feature Key Default Styles…** button.

- Scroll down in the **Feature Key:** pane until you find **CDS**. Select **CDS**.
- Change the color to **Green**. Click on **Invert Case**.
- Change the **Default Name** to [product] CDS. Ensure your all settings match the image below.



- Click on the **Update Project…** button.
- Select the **Selected Feature Key Name and Style** radio button. Click **OK**.



- Dismiss the window confirming that the new styles have been applied by clicking **OK**.
- Scroll up in the **Feature Key:** pane until you find **gene**. Select **gene**.

Notice that gene features have no color or text style and **Display Feature in Editors** is off. Updating the project with this default Feature Key Style will hide all of the gene features in every sequence.

- Click on the **Update Project…** button.
- Ensure that the radio button **Selected Feature Key Style** is selected. Click **OK**.
- Click **Done**. Close the **User Preferences** window.

Scrolling through the Sequence Editor window, you will notice that the translated segments of the sequence are green and in lower case.

- From the button bar in the **Sequence Editor** click on the **Overview** button.

The features which were previously blue are now green. You have successfully changed the **Display Style** of all the features with either a gene or CDS Feature Key.

- Choose **Sequence > Feature Listing** to see the updated names in the **Feature Listing**.
- Scroll to the first feature that has a CDS Feature Key.

```
Name: hypothetical protein CDS
    Key: CDS
    From: 277    To: 492    Color: Green    Style: Invert Case, Protein Translation
     277..492
     /locus_tag="Jvgp1"
     /note="agnoprotein"
     /codon_start="1"
     /product="hypothetical protein"
     /protein_id="NP_043508.1"
     /db_xref="GI:9628643"
     /db_xref="GeneID:1489519"
     /translation="MVLRQLSRKASVKVSKTWSGTKKRAQRILIFLLEFLLDFCTGEDSVDGKKRQRHSGLTEQTYSALPEPKAT"
```

Note that the new feature Name, **Hypothetical protein CDS**, is derived from the combination of the **/product** field, which you placed in brackets in the **User Preferences** dialog, followed by the text **CDS**. All features that share the Feature Key CDS will follow this naming convention.

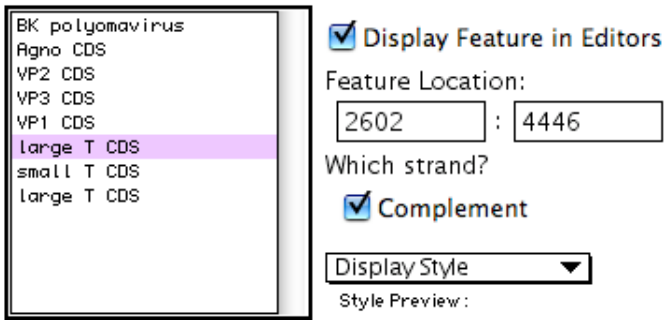
- Close the **Feature Listing** Window.
- Close the **Overview** window.

Follow these steps whenever you want to change the style or naming of features that have a common Feature Key.

JOINED FEATURES

Certain features are composed of several elements joined together in a specified order. Examples include exons or CDSs. In this next section of the tutorial, you will see how these elements are identified in a **Feature Listing** or the **Feature Editor** and how you can create these for yourself.

- Click on the icon for sequence AB211385.
- Go to the **Sequence** menu and select **Edit Features…**
- Select the feature called large T CDS.
- Change the **Feature Name** of the feature to large T CDS [2].



This feature is on the Complement strand and is the second of a two-part feature. You can see the first element at the bottom of the list in the **Feature Editor** window. It is also called **large T CDS**.

- Select the second feature called large T CDS.
- Change its **Feature Name** to large T CDS [1]

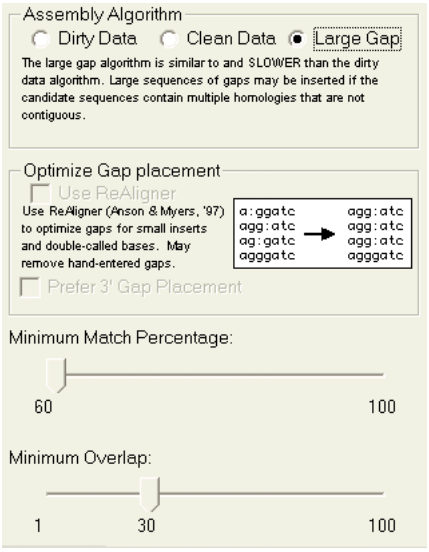
You can create your own joined features using this convention. When you import joined features from GenBank, the imported sequences will already have the bracketed numbers incorporated into the name of each element of the join. When you create joined features in Sequencher, each feature must have the same Feature Key and share the same name with sequential numbers enclosed in square brackets, as in CDS [01], CDS [02], CDS [03] etc. Any time the feature is harbored on the Complement strand, the sequential numbering should be reversed.

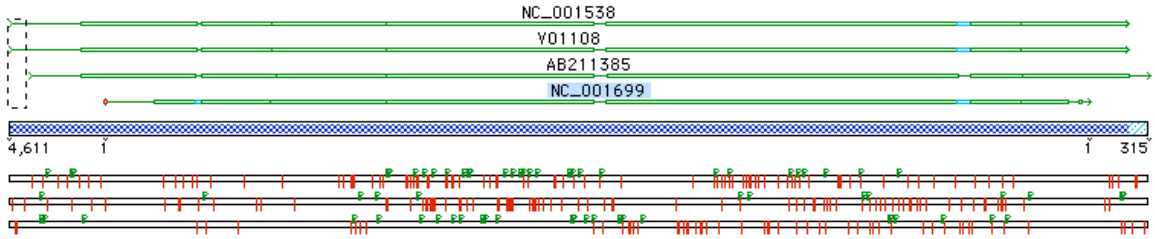
ASSEMBLING YOUR SEQUENCES

The next step is to start using your Reference Sequence. In this tutorial, you will compare a Reference Sequence to three strains of the virus and identify similarities between the strains.

Sequencher provides several algorithms for sequence assembly. Each algorithm has been devised for a specific purpose and contains parameters you can control. Since it is likely that the virus strains will not match perfectly with your exemplar, you are going to use the **Large Gap** algorithm.

- Click on the **Assembly Parameters** button in the **Project Window** button bar.
- Change the settings to match the image, choosing the **Large Gap** algorithm, a **Minimum Match Percentage** of 60 and a **Minimum Overlap** of 30.
- Click the **OK** button.
- Select all the sequences.
- From the Project Window button bar click on the **Assemble Automatically** button.
- In the **Assembly Completed** window click on the **Close** button.
- Double click on the contig icon to open the **Overview**.



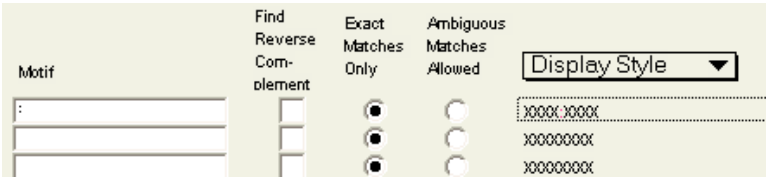


Your sequences have been assembled and are now ready for further examination and analysis. From this early view of the assembly it appears as if the CDSs, in green, are mostly aligned.

ANALYSING YOUR CONTIG

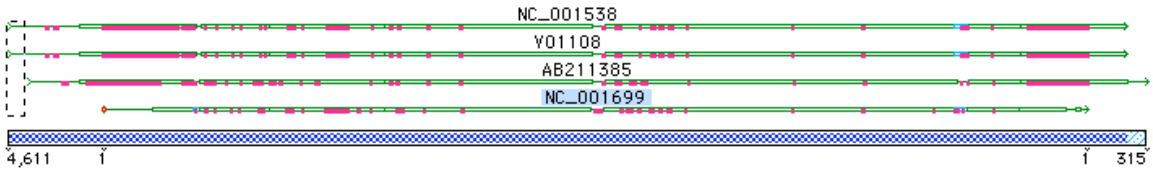
There is a quick way to display gaps in the **Overview** using **Motif Definitions**. A Motif is a short sequence of user-defined bases. You can add a **Display Style** to your Motif. Motifs are displayed in a similar fashion to Features. You will use the : (colon) character to represent a gap.

- From the **Window** menu click on the **Motif Definitions…** menu item.
- Create a motif by typing : into the first **Motif** text box.
- Choose **Magenta** from the **Display Style** drop down menu.
- Close the **Motifs** editor.



- From the **View** menu click on the **Display Motifs** menu item.

Notice how the display is updated with the magenta gap motif.



You can see from the display that there are regions where large gaps have been introduced but overall there is a high degree of similarity.

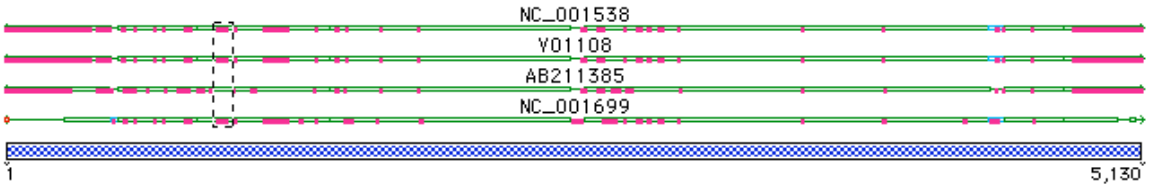
The Reference Sequence sets the numbering of the contig. Occasionally, sample sequences contain bases that are absent in the Reference Sequence. In the assembled contig this appears as a gap in the Reference Sequence. A special *decimal* numbering is used where there are insertions relative to the Reference Sequence.

- From the **Overview** window button bar click the **Bases** button.
- Select the any base in the consensus sequence.
- Choose **Select > Bases by Number**.
- Enter 502.1 into both boxes. Then click the **OK** button.

The highlight is now located on a position where there are two gaps in the Reference. These gaps have been inserted into the Reference Sequence because the three sample sequences have two additional "C" bases. The first "C" has the consensus position of 502.1 and the second "C" has the consensus position of 502.2. Wherever you see this *decimal* numbering you know that a base is inserted relative to the Reference Sequence.

The Reference Sequence is the shortest of the four viral sequences. You can use one simple command to trim off the overhanging sequences.

- From the **Contig** menu click on the **Trim to Reference Sequence** menu item.
- Click on the **Overview** button.



Notice how the overhanging sequences have been trimmed off all sequences relative to the NC_001699 Reference.

You can compare the features on the Reference Sequence with those of your samples. This can often indicate a biological function prior to any experimentation.

- Open the **Contig Editor** by clicking on the **Bases** button.
- Select a base in the consensus sequence.
- From the **Select** menu click on the **Bases by Number…** menu item.
- Type in 4427 in both boxes.

Sequencher has highlighted the first base of an intron. The default color for an intron is Cyan and the display style is lower case for the characters. You will notice that two of the other sequences also have an intron. One sequence, AB211385, does not have any feature marked. By analogy with the Reference Sequence and the two sample sequences you can infer that the intron for this sequence commences at the analogous base. You can also infer that the intron continues until the final "C" of the sequence "TTACC".

CONCLUSION

In this tutorial you have learned how to use imported GenBank sequences. You edited individual **Display Styles** and default **Display Styles**. You enabled circular numbering and reset the numbering of your genome. You used the **Large Gap** algorithm to assemble the sequences. You used a **Motif** to show gaps in the contig. Once the sequences were assembled, you learned how to remove unwanted sequence by using the **Trim to Reference** command. You learned how a Reference Sequence controls the numbering of the contig that contains it. You learned about the special decimal numbering that is used to indicate inserted bases relative to the Reference Sequence. Finally you learned how to use the Reference Sequence to make inferences about your sample sequences.

You can learn more about the use of the Reference Sequence in other tutorials in this series and in the manual.