Gene Codes Corporation

TCA GENE
AGT CODES

www.genecodes.com

SEQUENCHER® 5 SERIES

Easy... Fast... Powerful DNA Sequence Assembly Software

User Manual for Mac and Windows

Gene Codes Corporation

TCA GENE
AGT CODES

| 775 Technology Drive Suite 100A Ann Arbor, MI 48108 | 734-769-7249 or 800-497-4939 | www.genecodes.com

1. PREFACE

Sequencher is the premier software choice for DNA and RNA sequence assembly and analysis. Its capabilities include:

- Multiple, configurable DNA assembly algorithms
- Comprehensive DNA sequence editing tools
- Complete RNA-Seq workflow for differential expression
- Full support of sequence data confidence values
- Powerful Reference Sequence and Variance Table to find SNPs quickly and easily
- Restriction Mapping
- Extensive data import & export capabilities, including customizable GenBank Feature handling
- Specialized tools for Forensic mtDNA profiling

Over 25 years of daily use by biologists in labs around the world have refined **Sequencher's** tools and interface. You get the power and the speed to get the best, most accurate results from your DNA analysis, and get back into the lab more quickly.

GENE CODES

Gene Codes is a biotechnology-oriented software developer whose headquarters are in Ann Arbor, Michigan. Our goal is to write powerful software tools that are easy to use. If you have any suggestions on how to improve this product, please contact us. You can telephone us at (734) 769-7249, fax us at (734) 769-7074, or write to us at:

Gene Codes Corporation

775 Technology Drive, Suite 100A
Ann Arbor Michigan 48108 http://www.genecodes.com

We welcome your comments or support questions by electronic mail at: support@genecodes.com and general inquiries at: gcinfo@genecodes.com

2. USING THIS MANUAL

This manual is for use on both Windows and Macintosh operating systems. We'll document the differences in conventions or behaviours for each platform where applicable. We'll rotate Windows and Mac images in every other chapter unless the images are different enough to warrant including both the Windows and Mac images.

This chapter will introduce you to the conventions we will be using in this manual and how to get help. We will then introduce you to the button bar and **Sequencher's** menus, which offer you a rich range of options while maintaining a scientist-friendly interface.

CONVENTIONS USED IN THIS MANUAL

This manual follows certain conventions for displaying text and pictures and for alerting you to special information. Where keyboard shortcuts are mentioned in the text, we will show Windows shortcuts like **Ctrl+key** and Mac shortcuts like **Cmd+key**.

INSTALLING SEQUENCHER

Sequencher is easy to install using the installer which you can download from our Free Download website page. Follow the instructions in the installation guide to install your software. You must have Administrator privileges in order to install **Sequencher**.

If your **Sequencher** license is based on a license tied to a **Sequencher** key (dongle), the key, which you will be receiving, will have to be plugged into either the computer serving licenses to other computers, a KeyServer Network key, or to a computer locked to the key, a Standalone key.

SEQUENCHER HELP

Sequencher offers help through a Help viewer. To display Help, go to the **Help** menu and choose **Sequencher Help** or press **F1** (Windows) or **Command+Shift+?** (Mac).

Sequencher Help is organized into two main areas: application screens and the Sequencher User Manual. The **Sequencher** application screens contain context-sensitive links to relevant sections of **Sequencher** Help.

Note: You can leave the **Help** window open and return to **Sequencher** by clicking on an open window.

NAVIGATING SEQUENCHER HELP

The **Sequencher** Help window consists of three main parts. There is a button bar at the top of the window. There is also a navigation pane on the left side of the window and a content pane on the right side of the window.

The button bar has **Back** and **Forward** icons which function like a web browser's Back and Forward buttons. If you click on the **Home** button, you will return to the **Welcome** screen. You can also type keywords into the **Search** input field.

The navigation pane displays a Table of Contents. Click on the relevant icon to display any subtopics. Click on a page link icon in the navigation pane to display the **Sequencher** Help article in the content pane. Some of these pages contain underlined hyperlinks to related topics in the manual and a hyperlink to the following topic in the Table of Contents.

You can also get information on specific topics such as **Assembly Parameters** or the **Sequence Editor** from context-sensitive screens.

To search for a specific topic, type the topic name into the search field at the top right corner of the Help Viewer. When your search is complete, double-click on an item to display it.

CONTEXT-SENSITIVE SCREENS

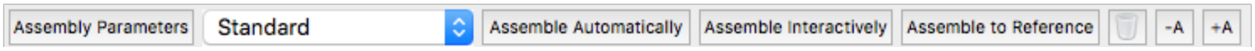
Context-sensitive screens provide easy links to information about the region of **Sequencher** you are currently using. If you have the **Project Window** displayed when accessing Help, then the **Project Window** context page will appear. This will be the case for most of **Sequencher's** major windows.

If you see a flagged icon, this means you can click on a link to get further information. If you place the cursor over this icon, a Tool Tip with the title of the Help article will be displayed.

THE BUTTON BAR

The button bar contains buttons for frequently used commands and a drop-down menu for setting assembly modes. It appears just below the top of many windows. (See Figure 2-1)

Figure 2-1 The Project Window button bar



Clicking on **Assembly Parameters** displays the **Assembly Parameters** dialog where you can change your assembly settings. **Assemble Automatically**, **Assemble Interactively**, and **Assemble to Reference** are one-click buttons which initiate your chosen assembly on the selected items in the **Project Window**. These buttons are context sensitive based on your assembly mode setting.

Clicking on the **Trash Can** icon will remove the selected items in the **Project Window** if you confirm that is what you want to do.

Clicking either of the font sizing icons will increase or decrease the size of the fonts on the items in the **Project Window**.

SEQUENCHER'S MENUS

SEQUENCHER (MAC ONLY)

The **Sequencher** menu gives you information on the version of **Sequencher** you have, contact information for **Gene Codes Corporation, Sequencher** technical support, access to the **User Preferences,** and the **Quit Sequencher** command.

FILE MENU

The commands in the **File** menu let you open, close, and save projects as well as import and export data. You can also open and close windows, specify page setups, and print from the **File** menu.

EDIT MENU

The commands in the **Edit** menu let you cut, copy, and paste selected data. You can also duplicate sequences, mark motifs, reposition sequences, and choose your editing mode from the **Edit** menu.

SELECT MENU

The commands in the **Select** menu cover a variety of items you might want to find and/or highlight to help you in your analysis. These commands use a wide variety of criteria to help you locate items from sequence fragments, subsequences, and individual bases.

ASSEMBLE MENU

The commands in the **Assemble** menu contain the essential items for assembling your Sanger and NGS sequences. These include the assembly parameters, assembly algorithms, and the alignment commands for tools such as Clustal and the next-generation sequencing-alignment algorithms.

CONTIG MENU

The commands in the **Contig** menu facilitate most of your work in contig editing. You can set the consensus mode, dissolve and rename your contigs, remove selected sequences, as well as **Trim to Reference Sequence, Compare Consensus to Reference**, or **Compare Translation to Reference**. The menu also contains commands for creating a new sequence from the consensus and displaying NGS data.

SEQUENCE MENU

The commands in the **Sequence** menu facilitate most of your work in sequence editing. You can set up data entry preferences such as trimming poor quality data or vector contamination, edit features, create and rename sequences, set up a **Reference Sequence**, set up base numbering, and **Revert To Experimental Data.** You can also **Compare Bases** to **Top Sequence, Consensus**, or **Reference Sequence,** these commands will trigger the appearance of the Variance Tables. In the **Analyses** submenu, you will be able to view the results of your Next-Generation sequencing analyses or view the **Frequency Histogram** or **Report**. **NCBI Blast Search** is designed for short, rapid queries of NCBI's nr database. The search is limited to the first 7800bp (Mac OS X) or 1600bp (Windows) of the selected sequence or consensus sequence.

VIEW MENU

The commands in the **View** menu let you control how data is displayed as you work on it. You can add extra information to the display, specify the format and marking of the data, and organize project files from the **View** menu. Some of the functions in the **View** menu have options which you can change by going to **User Preferences** (see Chapter 23 "Customizing **Sequencher** and User Preferences" for more information).

WINDOW MENU

The commands in the **Window** menu offer access to different kinds of user help, to reference information on various aspects of your data, to different windows you have opened, and to **User Preferences**.

HELP MENU

The command in this menu on Mac offers access to user help. The commands in this menu on Windows offer access to user help, gives you information on the version of **Sequencher** you have, and contact information for **Gene Codes Corporation** and **Sequencher** technical support.

CONTEXT-SENSITIVE MENUS AND BUTTONS

Sequencher has a number of menus and buttons that may change depending on the context. Some items will become enabled. The labels of other items will change. This feature is designed so that **Sequencher** can offer you extensive options within the same easy-to-use interface.

CONTEXTUAL MENUS

Most windows in **Sequencher** offer contextual menus. You can invoke these by clicking within the window you are reading while holding down the right hand mouse button. If you select one or more items and then click the right hand mouse button, you will see other menus.

Note: If you have a mouse with only one button, you may be able to mimic the effect of a right hand button with the use of a modifier key.

KEYBOARD SHORTCUTS

Sequencher has a number of keyboard shortcuts for major menu items and other important features and allows you to define some of your own. See Appendix 26 “Keyboard Shortcuts” for a complete listing. See Chapter 23 on “Customizing Sequencher and User Preferences” for more information about how to define your keyboard shortcuts.

3. THE PROJECT WINDOW

In this chapter, we will introduce you to **Sequencher's Project Window**. We will also show you how to get started with **Sequencher** by creating a new project or opening an existing one. This chapter also describes how you view the constituent sequences or contigs in your project and how to annotate information in your project for later editing and record keeping.

SEQUENCHER CONCEPTS

ABOUT THE PROJECT WINDOW

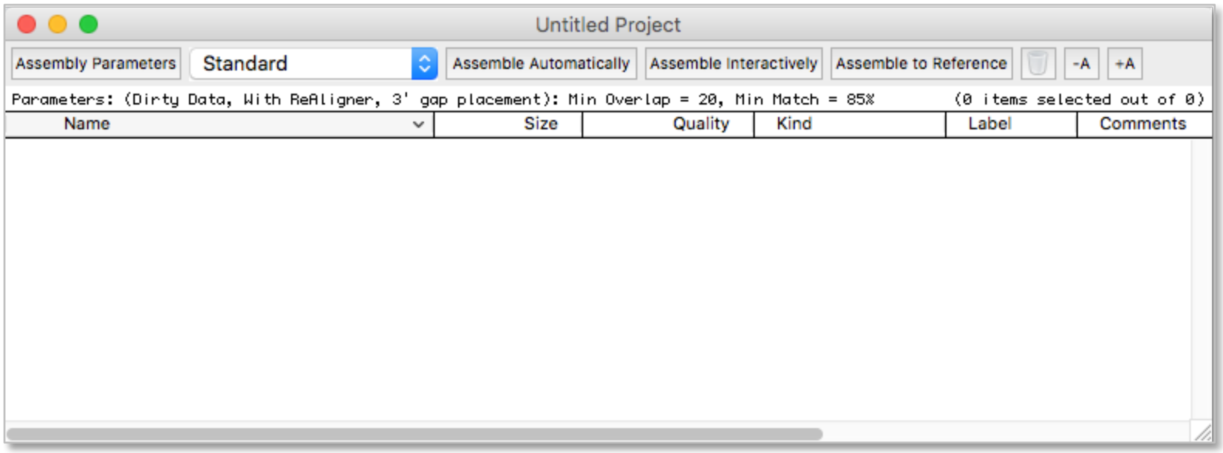
The concept of the project is central to **Sequencher**. Users work within the framework of a project. A **Sequencher** project is comprised of a collection of DNA sequences and contigs (contiguous alignments of overlapping sequences) that are built from those sequences. A project can be as large or as small as you want.

Sequencher stores the sequences you enter, the information on how these sequences fit together to create any contigs you have formed from them, and information on user-specified parameters that control the alignment operations, all combined into a single data file. The **Project Window** displays all of your sequences and contigs.

CREATING A NEW PROJECT

When you launch **Sequencher** by double-clicking the program icon, a new, empty project is created (see Figure 3-1).

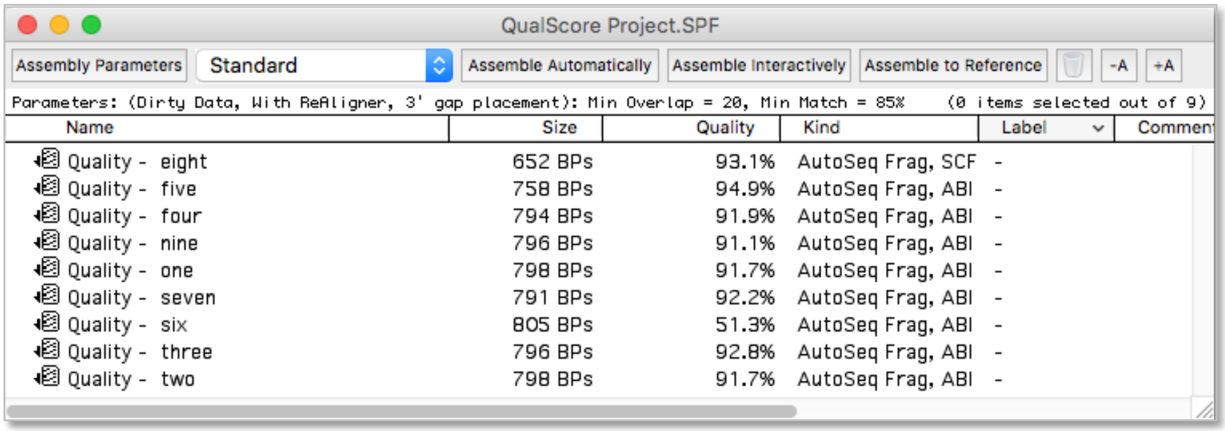
Figure 3-1 An empty Sequencher Project Window



```
Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%        (0 items selected out of 0)
```

Until you create or import sequence fragments, the new project will remain empty. As you add sequences to your project and build contigs from them, the **Project Window** begins to look more like Figure 3-2. All the sequences and contigs added to this particular sequencing project are displayed in this window.

Figure 3-2 A new project



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%   (0 items selected out of 9)

Name	Size	Quality	Kind	Label	Commen
Quality - eight	652 BPs	93.1%	AutoSeq Frag, SCF	-	
Quality - five	758 BPs	94.9%	AutoSeq Frag, ABI	-	
Quality - four	794 BPs	91.9%	AutoSeq Frag, ABI	-	
Quality - nine	796 BPs	91.1%	AutoSeq Frag, ABI	-	
Quality - one	798 BPs	91.7%	AutoSeq Frag, ABI	-	
Quality - seven	791 BPs	92.2%	AutoSeq Frag, ABI	-	
Quality - six	805 BPs	51.3%	AutoSeq Frag, ABI	-	
Quality - three	796 BPs	92.8%	AutoSeq Frag, ABI	-	
Quality - two	798 BPs	91.7%	AutoSeq Frag, ABI	-	

Note: If you have so many open windows on the screen that it's difficult to find the **Project Window**, choose **Window > Project Window**. This brings the current project to the front.

As you work, if you close an existing project and then want to start a new project, go to the **File** menu and choose **New Project**. **Sequencher** will open a new untitled project. **Sequencher** will remember the last projects you opened. You can see these as a list by going to the **File** menu and selecting the **Open Recent** command.

OPENING AN EXISTING PROJECT

To open an existing project, go to the **File** menu and choose **Open Project** and select the project you want to work with. Click on the **Open** button in the lower right area of the dialog, or click the **Enter/Return** key.

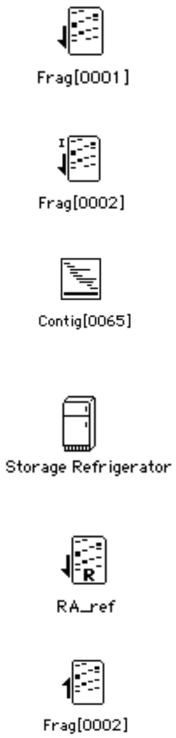
WORKING WITH ITEMS IN THE PROJECT WINDOW

To work with items in the **Project Window**, first select the item(s) you need by clicking on them. Then choose the appropriate menu command.

ICON TYPES

In the **Project Window,** a sequence is represented by an autorad icon. However, there are other data types you may have in your project. To get familiar with some of the icons **Sequencher** uses for these data types, see Figure 3-3.

Figure 3-3 Examples of Sequencher icons



Frag[0001]	A sequence fragment
Frag[0002]	A sequence fragment with edited comments
Contig[0065]	A contig
Storage Refrigerator	A refrigerator
RA_ref	A Reference Sequence
Frag[0002]	A sequence in the inverse-complement orientation

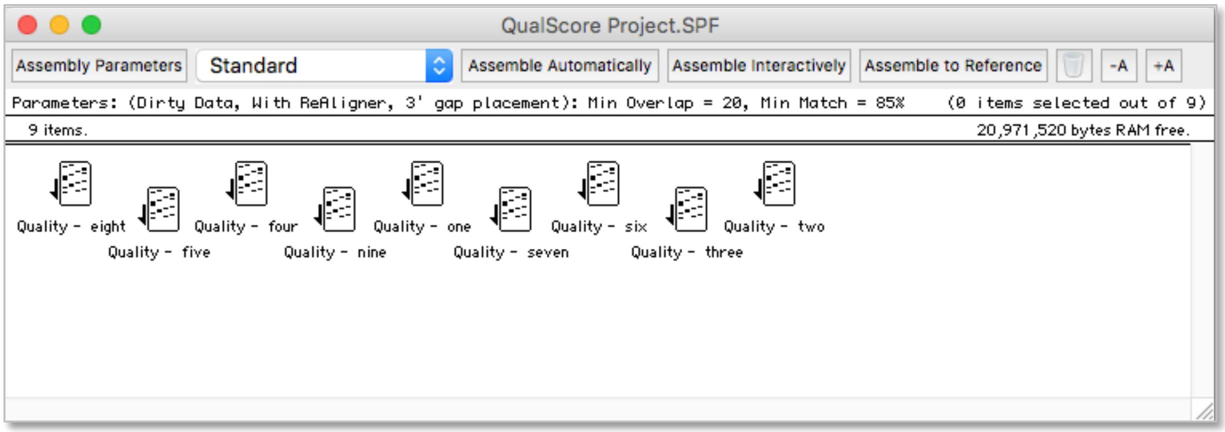
Other information may be displayed in addition to the basic icon. For example, the downward-pointing arrow on a sequencing image icon shows that this sequence is stored in its original orientation (the sequence travels “down” the gel). The arrow points up when the data is inverse-complemented as shown in Figure 3-3 above.

PROJECT VIEWS

VIEW AS LARGE ICONS

To display large icons in your **Project Window**, go to the **View** menu and then choose **Project Icons** As. Then select **Large Icons** from the submenu. The **Project Window** then displays the sequences and contigs as shown in Figure 3-4.

Figure 3-4 Project Window with Large Icons


Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85% (0 items selected out of 9)

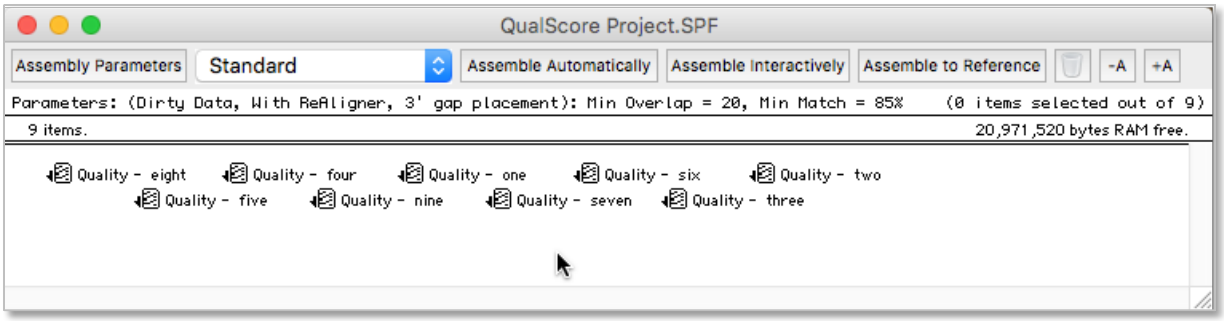
9 items. 20,971,520 bytes RAM free.

Note: A new project ***defaults*** to the icon view unless you previously specified a different view.

VIEW AS SMALL ICONS

To fit more information onto the screen, you can view project items as smaller icons. Go to the **View** menu and choose **Project Icons As** and then choose **Small Icons** from the submenu. The **Project Window** then displays the sequences and contigs as shown in Figure 3-5.

Figure 3-5 Project Window with Small Icons


Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85% (0 items selected out of 9)

9 items. 20,971,520 bytes RAM free.

MOVING ICONS

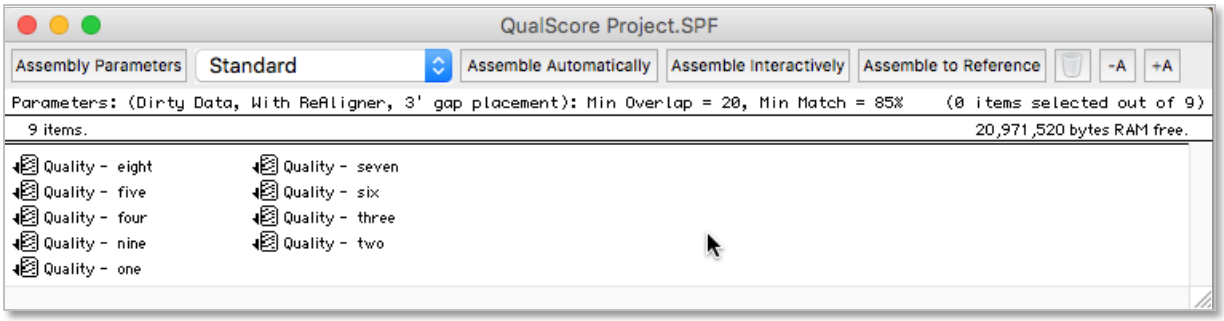
To reposition a sequence or contig icon, click the icon and drag it to the new location.

Note: Icons can be positioned on top of each other, thereby obscuring one or more of them from view.

CLEANING UP

Sequencher has clean-up commands to help you organize the icons in your **Project Window**. Go to the **View** menu and choose the **Sort/Cleanup** command to see a submenu that will let you rearrange the icons according to date edited, name, kind, size, etc. (See Figure 3-6.)

Figure 3-6 Small icons sorted by name



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85% (0 items selected out of 9)

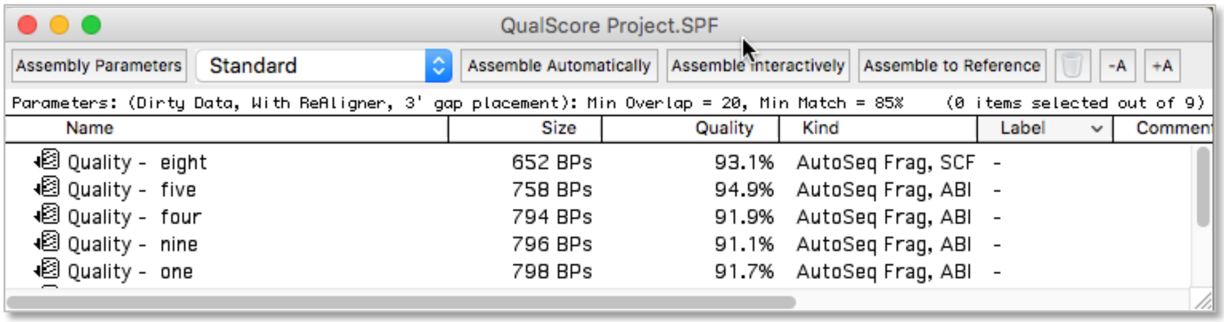
9 items. 20,971,520 bytes RAM free.

Quality - eight
Quality - five
Quality - four
Quality - nine
Quality - one
Quality - seven
Quality - six
Quality - three
Quality - two

VIEW AS A LIST

You can show your project items as a list rather than as icons. Go to the **View** menu and choose **Project Icons As…** and then select **A List** from the submenu. The **Project Window** will look like the one shown in Figure 3-7.

Figure 3-7 Project Window with list view



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85% (0 items selected out of 9)

Name	Size	Quality	Kind	Label	Comment
Quality - eight	652 BPs	93.1%	AutoSeq Frag, SCF	-	
Quality - five	758 BPs	94.9%	AutoSeq Frag, ABI	-	
Quality - four	794 BPs	91.9%	AutoSeq Frag, ABI	-	
Quality - nine	796 BPs	91.1%	AutoSeq Frag, ABI	-	
Quality - one	798 BPs	91.7%	AutoSeq Frag, ABI	-	

You can navigate the list by using the up and down arrow keys. To select a continuous list of sequences, choose the first sequence in the list and then, holding down the **Shift** key, select the last sequence in the list. To select a discontinuous list of sequences, hold down the **Ctrl** (Windows) or **Command** (Mac) key while clicking on the sequences.

PROJECT WINDOW COLUMNS

When you are in the list view, **Sequencher** displays the attributes, such as the item name or kind, of each sequence as a series of columns. You can control which columns are displayed. Go to the **View** menu and choose the **Project Window Columns** command. Then choose your desired option from the submenu.

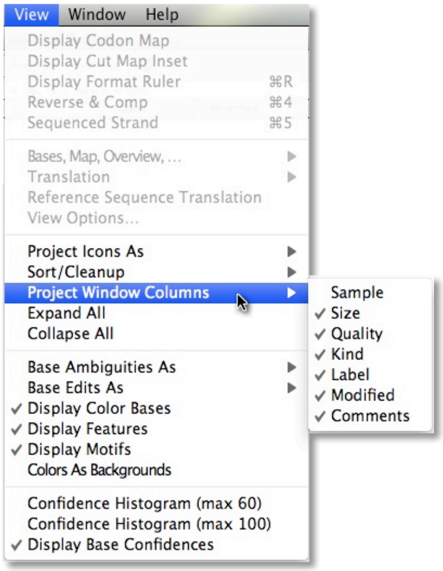
If you are working with ABI sequences, you can view the sample name by selecting **Project Window Columns** and then choosing **Sample**.

The column labeled **Quality** displays a value for each sequence that has confidence scores. The Quality value is the percentage of bases in a sequence that is above the **Low** Confidence Range threshold. The default setting is 20. You can alter the **Low** threshold value in the **Confidence** pane of the **General** settings in User Preferencews. See Chapter 23 “Customizing Sequencher and User Preferences” for more details. If you have used **Consensus by Confidence** for the consensus calculation, then you will see a % quality value in the **Quality** column.

You can sort your list by item name, size, quality, kind, label, or modification date. Click on the title of the column you want to use for your sort. For example, to sort by size, click the word **Size** at the top of that column. **Sequencher** will reorder the list by size in ascending order. If you **shift+click** the top of the column again, the column will re-sort in descending order. You can also sort the list by going to the **View** menu, choosing **Sort/Cleanup,** and then selecting from the options on the submenu.

If you sort by **Kind**, the order of precedence is, first any Contigs which contain Reference Sequence fragments, then any unassembled Reference Sequences, then contigs without Reference Sequences, and finally Sequence Fragments.

Figure 3-8 Project Window Columns submenu



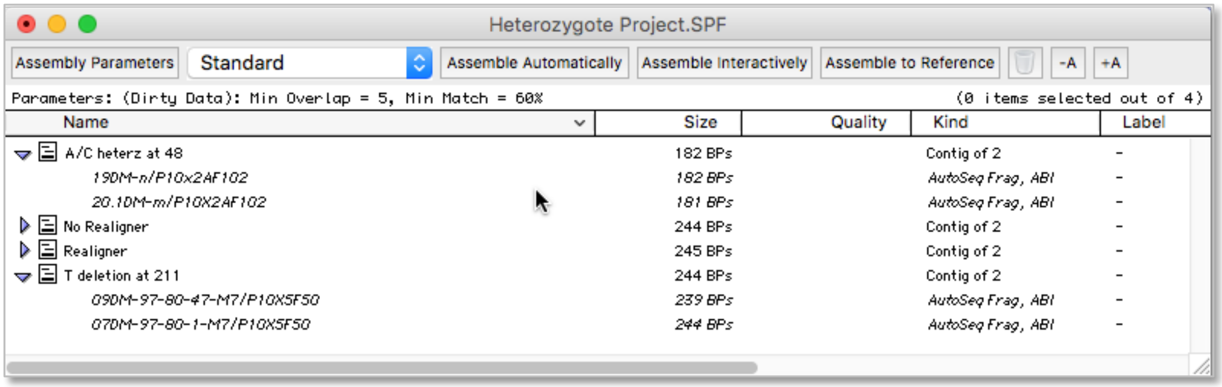
Note: You can toggle between types of views if you hold down the **Alt** (Windows) or **option** (Mac) key and click in the header area, just above the icons and below the parameter information.

VIEWING CONSTITUENT SEQUENCES

When you have contigs in list format, click the triangle to the left of the contig icon (Figure 3-9) to see a sub-list of the sequences incorporated into that contig. That triangle will turn to point downward and the sequence list will expand (see Figure 3-9). To hide a sub-list of sequences, click the triangle again; it turns to point horizontally and the sequence list will collapse.

Go to the **View** menu and click on the **Expand All** and **Collapse All** commands to either display or hide all of the contents of the project's contigs.

Figure 3-9 Viewing a sub-list of a contig's constituents



Parameters: (Dirty Data): Min Overlap = 5, Min Match = 60%   (0 items selected out of 4)

Name	Size	Quality	Kind	Label
A/C heterz at 48	182 BPs		Contig of 2	-
19DM-n/P10x2AF102	*182 BPs*		*AutoSeq Frag, ABI*	-
20.1DM-m/P10X2AF102	*181 BPs*		*AutoSeq Frag, ABI*	-
No Realigner	244 BPs		Contig of 2	-
Realigner	245 BPs		Contig of 2	-
T deletion at 211	244 BPs		Contig of 2	-
09DM-97-80-47-M7/P10X5F50	*239 BPs*		*AutoSeq Frag, ABI*	-
07DM-97-80-1-M7/P10X5F50	*244 BPs*		*AutoSeq Frag, ABI*	-

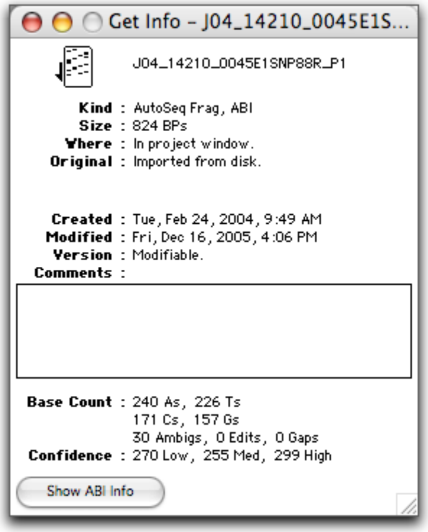
VIEWING THE INFORMATION FOR A SEQUENCE OR CONTIG

The **Get Info** command displays additional information about a selected or open sequence or contig.

To view additional information about an open sequence or a **Contig Editor** window, go to the **File** menu and choose **Get Info**. If you do not have an open sequence or contig editing window, you can select the icons of your sequences or contigs from the **Project Window** and then go to the **File** menu and choose **Get Info.**

Comments from contigs created with Next-Generation sequence alignment algorithms, for example, will contain information about the algorithm used, the data used, and the path to the data together with the data and the time generated.

Figure 3-10 The Get Info window



J04_14210_0045E1SNP88R_P1

Kind : AutoSeq Frag, ABI
Size : 824 BPs
Where : In project window.
Original : Imported from disk.

Created : Tue, Feb 24, 2004, 9:49 AM
Modified : Fri, Dec 16, 2005, 4:06 PM
Version : Modifiable.
Comments :

Base Count : 240 As, 226 Ts
171 Cs, 157 Gs
30 Ambigs, 0 Edits, 0 Gaps
Confidence : 270 Low, 255 Med, 299 High

If you have sequences from an ABI sequencing system, **Sequencher** can display information from the ABI data sheet. In the **Get Info** window shown in Figure 3-10, note the **Show ABI Info** button.

EDITING THE INFORMATION FOR A SEQUENCE OR CONTIG

You can annotate the sequence by clicking anywhere in the **Comments:** box and typing the information. To store more information than the **Comments:** box allows, go to the **Edit** menu and choose **Edit Comments…**. **Sequencher** opens a small text editing window that stores about 250 characters of text for each sequence. After you have added text with **Edits Comments…,** a small "I" (for "Info") appears in the upper left corner of the icon.

If you want to remove the information in the **Comments:** box, select the text and then go to the **Edit** menu and choose **Contents** from the **Clear** menu.

THE PROJECT WINDOW BUTTON BAR

The button bar contains buttons for frequently used commands. It appears just below the top of the **Project Window**. You can use the smaller left-hand button, which has an up arrow on it, to hide the button bar. Click on the button and the button bar will "hide" behind the title bar.

When the button bar is hidden, you can still see the bottom of the left-hand button under the title bar. It will show a down-pointing arrowhead. Click on the arrowhead and the button bar will reappear.

If you click on the **Assembly Parameters** button, the **Assembly Parameters** window will appear. Clicking on the **Assembly Mode** drop-down menu will enable the **Standard**, **Assemble by Name**, or **Multiplex ID** function. **Assemble Automatically**, **Assemble Interactively**, and **Assemble to Reference** are used to assemble sequences you have selected. If **Assemble by**

Name is enabled, then the **Assemble Automatically** and **Assemble to Reference** buttons will have different labels. If you select a sequence and then click on the **Trash Can** button, that sequence will be removed from the project.

TEMPLATES

A template is a special type of project that can be set with your choice of parameters and preferences. Templates can contain ordinary sequences or a Reference Sequence. Preferences can include Feature settings, enzyme sets, and display settings. With templates, you can also set **Assembly Parameters**, **Assemble by Name** handles, and **Trim** parameters. These will be discussed later in this manual.

A template can be used as a new project or imported into an existing project. You can use templates to set up standard operating procedures and methods to reuse or distribute throughout your lab.

SAVE PROJECT AS TEMPLATE

You can save any project as a template by using the **Save Project As Template** command. Set the User Preferences and parameters that you want to use. You can include a Reference Sequence or ordinary sequences.

Once you are satisfied with your choices, go to the **File** menu and click on the **Save Project As Template…** command. A new dialog called **Template Name**: will appear. Type a name for your template in the input field and then click on the **OK** button. Your project will now be saved in a **Templates** folder. The name of the template will appear in the **New Project From Template** submenu.

NEW PROJECT FROM TEMPLATE

The **New Project From Template** command allows you to open a new project containing all the sequences, settings, and preferences associated with your chosen template. Go to the **File** menu and click on **New Project From Template**. Select a template **from** the submenu. A new blank **Project Window** will open.

IMPORT FROM TEMPLATE

You can add preferences, settings, and sequences to an existing project by importing a template. If you have an open project, go to the **File** menu and click on the **Import** command and then choose **From Template**. Select a template from the submenu. The settings and/or sequences from the template will be applied to your open project automatically.

4. IMPORTING DATA

In this chapter, we will discuss how to import data from automated sequencing equipment and other programs into **Sequencher** projects. The different techniques we will discuss include dragging and dropping, copy and paste, and using import commands. We also explain how to deal with projects in specific formats, how to import confidence scores, and how to remove unwanted sequences and contigs from a project.

QUICK AND EASY IMPORTING OF DATA

DRAGGING AND DROPPING FILES

You can drag files from the desktop right into the **Project Window** and **Sequencher** will import and load those files. **Sequencher** will alert you if it cannot read and load the files.

If you drag sequences onto the **Sequencher** icon on your desktop, you will open a new **Project Window** containing your sequences. If **Sequencher** is already open, the sequences will be imported into the open project.

Note: **Sequencher** always shows a newly imported file as selected, so if you do not want to use it right away, de-select it by clicking elsewhere in the project.

COPY AND PASTE

One of the easiest ways to move data between **Sequencher** and other programs is to use the clipboard's **Copy** and **Paste** commands. First, launch both **Sequencher** and the other program. Then, select and copy the bases of interest into the other program.

Go to the **Sequence** menu and choose **Create New Sequence**. Type a name for your sequence into the **New sequence's name**: input field and then click on the **OK** button. Now go to the **Edit** menu and choose **Paste**. When you close the window, you will see a dialog with two buttons. If you have finished entering your data, click on the **Record As Experimental Data** button. Otherwise, click on the **Not Yet Finished** button. Sequencher then removes the window.

HOW TO IMPORT USING MENU COMMANDS

IMPORTING SEQUENCES

Sequencher imports sequences in a variety of text and chromatogram formats. Some of the most popular are ABI, MegaBase, CEQ, SCF, FastA, aligned FastA, FastQ, GenBank, EMBL, DDBJ and plain text files.

To import a sequence, first bring the **Project Window** to the front by clicking on it. Under **File** menu, choose **Import.** Choose **Sequences**… and then browse to and select the file you want to import. Click on the **Open** button in the lower right area of the dialog or press the **Return** key. You may need to select a file format from the drop-down menu called **Enable:** in order to see the files of interest in the dialog. To see all of the files in a particular location, choose **All Documents**.

You can also select from a **Folder of Sequences**…, **ACE Project**…, **CAF Project**…, **CEP Project**…, **FASTA – aligned**…, **GCG Contig**…, **Sequencher Project**…, or **Sequence from VecBase**…. (Working with specific formats is discussed below.)

If **Sequencher** cannot import the file, you will get a warning message to that effect.

IMPORTING GENBANK FEATURES

When you import a sequence in GenBank format, you will automatically import the features, including the location, the feature keys, and feature qualifiers. The feature key used in sequences published in GenBank, EMBL, and DDBJ describes the biological nature of the annotated feature or indicates information about changes to the sequence. (For more information on features and feature keys, see Chapter 19 “Motifs and Features” and Appendix 30 “Feature Keys and Qualifiers”.)

Where feature locations described in the GenBank file are ambiguous, **Sequencher** will warn you that it cannot import these features and will list the features in a dialog you can print out for your records.

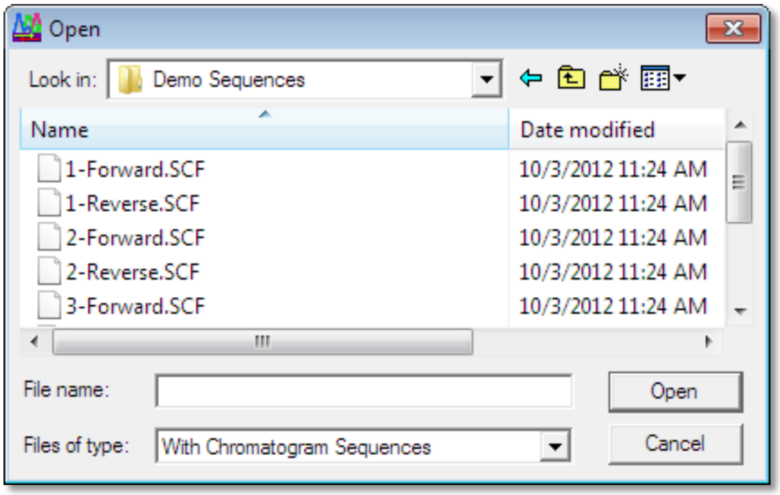
IMPORTING LISTS OF SEQUENCES

Sequencher’s functionality has been modified to take advantage of the file system. If you use the **Import** menu item with the **Sequences…** option, you can select multiple files for a single import command. To select a continuous list of sequences, choose the first sequence in the list and then, holding down the **Shift** key, select the last sequence in the list. The import window selects the two sequences and all of the sequences in between. Click on the **Open** button to import all of the selected sequences into your project.

To select a discontinuous list of sequences, hold down the **Ctrl** (Windows) or **Command** (Mac) key while clicking on the file. Click on the **Open** button to import all of the selected sequences into your project.

The **Files of type** (Windows) or **Enable** (Mac) drop-down menu allows you to filter the files that are displayed in the file browser. On Mac, the filters are **All Importable Documents, All Documents, With Chromatogram Sequences**, and **Without Chromatogram Sequences.** On Windows, the filters are **Supported File Formats, All Files (*.*), With Chromatogram Sequences**, and **Without Chromatogram Sequences.** On Mac, you may see some files becoming grayed out with different file filters because these files do not meet the import criterion of the current filter (see Figure 4-1 where **With Chromatogram Sequences** is set but some of the files are text only files).

Figure 4-1 File import Enable option set to With Chromatogram Sequences



Look in: Demo Sequences

Name	Date modified
1-Forward.SCF	10/3/2012 11:24 AM
1-Reverse.SCF	10/3/2012 11:24 AM
2-Forward.SCF	10/3/2012 11:24 AM
2-Reverse.SCF	10/3/2012 11:24 AM
3-Forward.SCF	10/3/2012 11:24 AM

File name:

Files of type: With Chromatogram Sequences

Note: To import traces, you must import the actual trace file. If you import a text file containing ASCII characters, you will not import traces. Text files will be only a couple of kilobytes in size, whereas trace files will be well over 50 kilobytes.

Sequencher uses a very efficient algorithm to compress the imported traces. This conserves disk space so a project file containing several sequences from an automated sequencer may be smaller than any one of the original trace files.

FOLDER OF SEQUENCES

When you have many sequences to import, you can import all those files from within a single folder. Select the **Project Window** and go to the **File** menu and choose **Import** and **Folder Of Sequences…** from the submenu. The standard **Choose a Folder** dialog appears on Mac and the **Browse for Folder** dialog appears on Windows; select the folder that holds the data you want to import and click on the **Choose** button on Mac and the **OK** button on Windows.

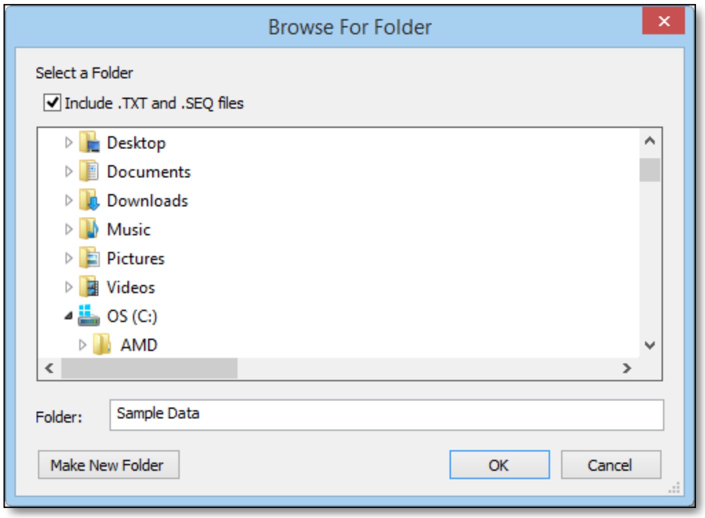
Note: On Mac OS X, the contents of a folder selected for import will be grayed out.

Use the **Include .TXT and .SEQ files** checkbox on Windows or the **Include All Text Files** checkbox on Mac to include or exclude text files from your import.

Sequencher tells you how many files are in the folder and asks whether you want to import them all. If you do, click on **Import All Files in Folder**. **Sequencher** imports all the files, assigning a separate icon to each.

You can specify that you want files imported into the **Project Window** or directly into the trim window (see Chapter 6 “Preparing your Data for Assembly”).

Figure 4-2 Browse for a Folder dialog



Select a Folder

☑ Include .TXT and .SEQ files

Folder: Sample Data

PROJECTS IN SPECIFIC FORMATS

Sequencher supports importing projects as ACE projects, Contig Express projects, GCG contigs, and as CAF projects. The contents of these files include confidence values, feature annotation, and assembly information. Select the **Project Window**, go to the **File** menu and choose **Import,** and then choose the specific project format from the submenu. The standard **Open** dialog appears; select the folder that holds the data you want to import and click on the **Open** button.

Read more about importing confidence values in the section “Importing Confidence Scores” later in this chapter.

SEQUENCHER PROJECTS

At times you may want to merge two projects or combine your own work with that of a colleague. To combine **Sequencher** projects, first open one of the projects. Go to the **File** menu and choose **Import** and then **Sequencher Project…** from the submenu. Select your other project and click on the **Open** button. **Sequencher** displays all the imported projects in the initial **Project Window**, retaining all comments, edits, and even the relative positions of the icons. To keep track of which sequences came from which project, you may want to establish a naming convention for sequences *before* combining them. Since the imported sequences and contigs are highlighted after import, you could also use sequence **Labels** to differentiate them. (See Chapter 23 “Customizing Sequencher and User Preferences”.)

FILES FROM VECBASE

VecBase is a file of vector sequences in your **Sequencher** application folder. You can import a vector from VecBase just as if it were any other DNA sequence. Go to the **File** menu and choose **Import** then the sub-command **Sequence From VecBase**. The **Open** dialog appears. Choose the desired file and click on the **Open** button. If you wish to import more than one file, for continuous selection of files, you can use **Shift+click** (Windows or Mac). For discontinuous selection of files, use **Ctrl+click** (Windows) or **Command+click** (Mac).

CREATING A NEW SEQUENCE

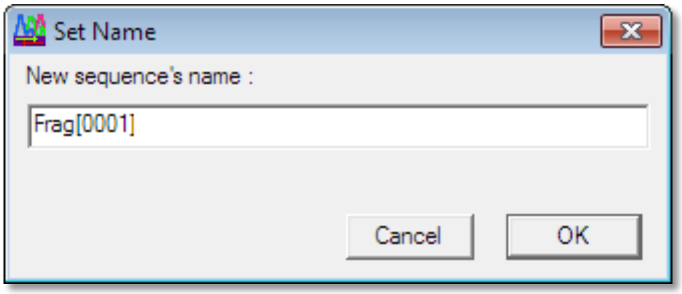
Another way to bring new data into **Sequencher** is to create a new sequence fragment. To do this, go to the **Sequence** menu and choose **Create New Sequence…**

Sequencher displays a dialog asking you to name the new sequence (Figure 4-3). The default name, in this case "Frag[0001]", can be set as a user preference. A sequence name can be a maximum of 255 characters, including any punctuation and spaces, but 31 characters is recommended for export functions.

When you have typed the name you want (or if you accept the default name), click on the **OK** button or press the **Enter** key. A new **Sequence Editor** dialog appears showing the name you selected in the title bar. The new **Sequence Editor** lets you enter and change sequence data. Just type as you would with a standard word processor.

When you are finished entering and/or editing a sequence, click on the close button in the top left corner of the window at the left end of the title bar on Mac or in the top right corner of the window at the right end of the title bar on Windows, or go to the **File** menu and choose **Close Window**. **Sequencher** asks if you have finished editing. At this point, you can record your data as experimental data or you can tell **Sequencher** you have not yet finished editing by selecting the appropriate button. **Sequencher** then removes the **Sequence Editor** window from the screen. The sequence is now represented by a sequencing icon in the Project **Window**.

Figure 4-3 Create New Sequence dialog



Set Name

New sequence's name :

Frag[0001]

DOUBLE-CLICKING A PROJECT ICON

When you double-click on a sequence project icon, this launches **Sequencher** and opens the project. If a **Project Window** is *already* open, when you double-click on a project icon **Sequencher** will ask whether you want to save changes to your current project before closing it. Click on the **Cancel** button if you do not want to proceed. If you want to close the project and save any changes you have made, click on the **Yes** button. If you want close the project without saving your changes, click on the **No** button.

IMPORTING CONFIDENCE SCORES

Many base callers generate confidence values associated with each base call. **Sequencher** supports confidence values from PHRED, ACE, and Trace Tuner files, as well as SCF, FastQ, ABI 3730, and ESD files.

IMPORTING PHRED CALLED DATA

Sequencher can read in the SCF, FASTA, and Qual files generated by PHRED and supports PHRED base calls and confidence values. To view the confidence levels, you must first set the user preferences for displaying confidence values (See Chapter 23 "Customizing Sequencher and User Preferences" for more information). Then go to the **View** menu and choose **Display Base Confidences**.

To import the data into **Sequencher**, it must first be organized in a folder containing three subfolders. These subfolders are **chromat_dir**, which contains the trace files, **edit_dir**, which contains the text files, and **phd_dir**, which contains the quality scores. Select **File** and then go to **Import** and then **Folder of Sequences** to import the **phd_dir** folder. To import the sequences you would like to view, you may also go to the **File** menu and choose the **Import** command with the **Sequences...** submenu.

IMPORTING PHRAP FILES

Sequencher can import the ACE projects created by PHRAP. To import an ACE project select **File** then go to **Import** and then the **ACE Project…** submenu. Browse to a file that ends with ".ace" or ".ace.x" (such as .ace.1,.ace.2, etc.). This file will typically be located in the subfolder "**edit_dir**". Select the highest numbered ACE file to account for all editing and reassembling.

Sequencher automatically imports and incorporates as much auxiliary data as it can find. The auxiliary files must be in the same folder as the ACE file or arranged in the traditional PHRAP assembly hierarchy (chromatogram files in a **chromat_dir** folder and PHD files in a **phd_dir** folder). If the .singlets file, the .qual file, the PHD files, and the chromatogram files all exist and are available, **Sequencher** will import them.

Alternatively, you can import parts of the PHRAP project by going to the **File** menu and choosing **Import** and then **Sequences** or by going to the **File** menu and clicking on **Import** and then **Folder of Sequences**.

Note: When using ftp to transfer your data to your local computer, it is ALWAYS important to transfer these files in *binary* mode. If you are not familiar with ftp, speak with your network administrator.

REMOVING SEQUENCES AND CONTIGS FROM A PROJECT

To remove items from the project, select the items you want to remove. Go to the **Edit** menu and choose **Remove From Project…** or click on the **Trash Can** button in the button bar. Removing an item from a project is permanent so Sequencher asks whether you really want to.

***Note*:** Remember, you cannot back out of this action with a simple **Undo** command, but if you close the project *without saving,* all changes including removed items will be forgotten. The **Revert to Saved Project** command has the same effect.

If you import the chromatogram files from the "**chromat_dir**" folder, these fragments will have PHRED base calls and chromatogram data but no base qualities.

If the "_.phd" files are loaded and "_. fasta.qual" files are available, you will have imported all of the fragments from the PHRAP project. They will have base quality and chromatogram data.

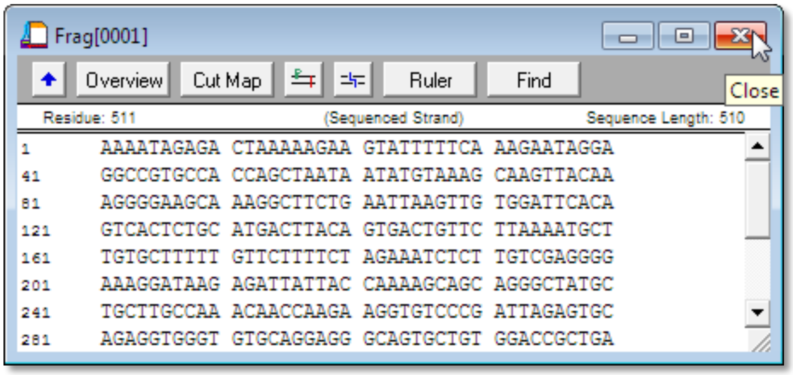
The.qual files alone will not load your sequences.

If you import PHD files from the "phd_dir" and the corresponding chromatogram files are available, these files will import with PHRED base calls, chromatogram data, and base qualities.

CLOSING AN EDITOR WINDOW

When you have finished entering and/or editing a sequence, click on the close button in the top left corner of the window at the left end of the title bar on Mac or in the top right corner of the window at the right end of the title bar on Windows (see Figure 4-4) or choose **Close Window** from the **File** menu. **Sequencher** then removes the **Sequence Editor** window from the screen. Once you have closed the window, a sequencing icon will represent your sequence. (Figure 4-4)

Figure 4-4 Clicking the Close button



```
Residue: 511          (Sequenced Strand)          Sequence Length: 510
1       AAAATAGAGA CTAAAAAGAA GTATTTTTCA AAGAATAGGA
41      GGCCGTGCCA CCAGCTAATA ATATGTAAAG CAAGTTACAA
81      AGGGGAAGCA AAGGCTTCTG AATTAAGTTG TGGATTCACA
121     GTCACTCTGC ATGACTTACA GTGACTGTTC TTAAAATGCT
161     TGTGCTTTTT GTTCTTTTCT AGAAATCTCT TGTCGAGGGG
201     AAAGGATAAG AGATTATTAC CAAAAGCAGC AGGGCTATGC
241     TGCTTGCCAA ACAACCAAGA AGGTGTCCCG ATTAGAGTGC
281     AGAGGTGGGT GTGCAGGAGG GCAGTGCTGT GGACCGCTGA
```

CLOSING A PROJECT

If you have finished working with the project but not with **Sequencher**, choose **Close Project** from the **File** menu. **Sequencher** will ask if you want to save the changes to your project. Click **Yes** to save your changes and close the project. Click **No** to close the project without saving your changes. Click **Cancel** if you wish to continue working with the project. You can then start working with another project by going to the **File** menu and choosing **New Project** or **Open Project**.

If you have finished working with **Sequencher**, go to the **File** menu and choose **Exit** (Windows) or to the **Sequencher** menu and choose **Quit Sequencher** (Mac).

5. ORGANIZING PROJECT VIEWS

In this chapter on organizing project views, you will learn more about the structure of a **Sequencher** project as you begin to work with your data. We will discuss how to select individual and multiple sequences using various techniques, renaming items, and saving your work.

WORKING WITH ICONS AND LISTS

CHANGING BETWEEN PROJECT VIEWS

Sequencher provides you with a variety of ways to view and organize the contigs and unincorporated data in your **Project Window**. You can view your data as large or small icons or in list form (for more information, see Chapter 3 “The Project Window”). You can toggle quickly between types of views by holding down the **Alt** (Window) or **Option** (Mac) key and clicking in the header title area, just above the icons and below the parameter information. You can display a list of sequences in a single contig in the list view by clicking on the triangle to left of the contig icon. You can see a list of sequences in all your contigs by going to the **View** menu and choosing **Expand All**.

COLLECTING SEQUENCES IN REFRIGERATORS

Sequencher lets you create “**refrigerators**” in the **Project Window** to hold certain subsets of your project. Refrigerators hold only raw sequences, not contigs or other refrigerators. To “put away” a few sequences in “cold storage,” select the sequences and then go to the **Edit** menu and choose **Refrigerate. Sequencher** collects the sequences and puts them in the refrigerator.

To remove sequences from the refrigerator, open the refrigerator by double-clicking on it. You will see a list of items. Select the items you want to remove and then click the button called **Move Selected Items To Project Window** to execute the command**.**

SELECTING SEQUENCES

SELECTING SEQUENCES AND CONTIGS

To select sequences you can go to the **Select** menu and use the **Select All**, **Select None**, and **Invert Selection** commands. **Invert Selection** will give you the opposite of your previous selection.

There are also selection functions that are increasingly specific. You can choose data based on **All Items That…** and then further specify the data by choosing narrower classes such as

Contain Subsequence…, Contain Items Named…, or **Have Chromatograms** from the **Select** menu. You can indicate if you want these names to be case sensitive or not.

Note: The subcommand **Contain Items with Names Containing**… only works on contigs.

For more detailed information on these commands, see Chapter 20 “Finding Items”.

SELECTING WITH THE MOUSE

First make sure the **Project Window** is at the front. When you click on an icon, or hold down the **Shift** key while clicking on several icons, **Sequencher** highlights the icon(s). When you have finished selecting a group of icons, release the **Shift** key before you try to do anything with the group.

You can deselect an icon by clicking on it while holding down the **Ctrl** (Windows) or **Cmd** (Mac) key. If you do not hold down the **Ctrl** (Windows) or **Cmd** (Mac) key, **Sequencher** deselects the last highlighted icon(s) as soon as you click on another. If you click in the white space *between* icons and names, **Sequencher** deselects all the highlighted icons.

When you display the list view of a **Project Window** or a **Contig**, you can select a range of sequences by using **Shift+click.** You can make a discontinuous selection by using the **Ctrl** (Windows) or **Cmd** (Mac) key and clicking on the individual sequences.

SELECTING BY TYPING

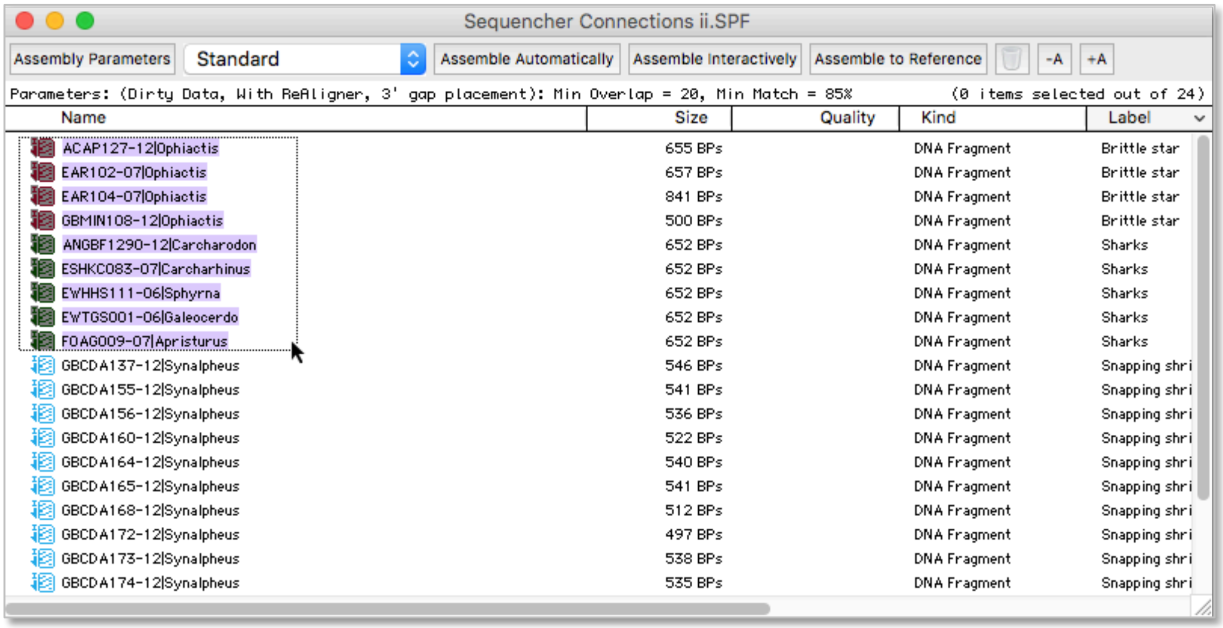
Type the name of a sequence fragment or contig. As you type, **Sequencher** searches the **Project Window** for an icon with that name. If it finds the name, it scrolls to the point at which the icon is visible and highlights the icon.

Note: If you pause for more than a second or two while typing a name, **Sequencher** assumes you have started typing a different name and starts looking for the new one instead.

SELECTING MULTIPLE ITEMS WITH A MARQUEE

You can select multiple sequences by drawing a box around the icons or list items you want. Position the mouse so it is not touching any icons or names. The cursor should be approximately where one corner of the box should be. Hold the mouse button down and drag diagonally across to where you would expect the opposite corner of the box. As you drag, the marquee appears (as shown in Figure 5-1). As you move the mouse, any icons or names touched by the marquee will be selected.

Figure 5-1 Selecting with the marquee



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%   (0 items selected out of 24)

Name	Size	Quality	Kind	Label
ACAP127-12|Ophiactis	655 BPs		DNA Fragment	Brittle star
EAR102-07|Ophiactis	657 BPs		DNA Fragment	Brittle star
EAR104-07|Ophiactis	841 BPs		DNA Fragment	Brittle star
GBMIN108-12|Ophiactis	500 BPs		DNA Fragment	Brittle star
ANGBF1290-12|Carcharodon	652 BPs		DNA Fragment	Sharks
ESHKC083-07|Carcharhinus	652 BPs		DNA Fragment	Sharks
EWHHS111-06|Sphyrna	652 BPs		DNA Fragment	Sharks
EWTGS001-06|Galeocerdo	652 BPs		DNA Fragment	Sharks
FOAG009-07|Apristurus	652 BPs		DNA Fragment	Sharks
GBCDA137-12|Synalpheus	546 BPs		DNA Fragment	Snapping shri
GBCDA155-12|Synalpheus	541 BPs		DNA Fragment	Snapping shri
GBCDA156-12|Synalpheus	536 BPs		DNA Fragment	Snapping shri
GBCDA160-12|Synalpheus	522 BPs		DNA Fragment	Snapping shri
GBCDA164-12|Synalpheus	540 BPs		DNA Fragment	Snapping shri
GBCDA165-12|Synalpheus	541 BPs		DNA Fragment	Snapping shri
GBCDA168-12|Synalpheus	512 BPs		DNA Fragment	Snapping shri
GBCDA172-12|Synalpheus	497 BPs		DNA Fragment	Snapping shri
GBCDA173-12|Synalpheus	538 BPs		DNA Fragment	Snapping shri
GBCDA174-12|Synalpheus	535 BPs		DNA Fragment	Snapping shri

SELECTING USING MENU COMMANDS

SELECTING ALL ICONS

You can change the icons selected in the **Project Window** by going to the **Select** menu and clicking on the **Select All, Select None,** and the **Invert Selection** commands.

SELECTING ALL ITEMS CONTAINING SEQUENCE

To select all items in the **Project Window** that contain a particular subsequence, go to the **Select** menu and choose the **All Items That** command. Then select **Contain Subsequence…** from the submenu. **Sequencher** displays a dialog that lets you type two sequences; the box has buttons to specify how you want the sequences matched to items in the **Project Window**. You can use the drop-down menus to specify restriction enzyme recognition sequences.

FINDING ASSEMBLED SEQUENCES

To locate data assembled in a contig, go to the **Select** menu and click on either the **All Items That>Contain Subsequence**… or **Contain Items With Names Containing**…. **Sequencher** will select the contig(s) that match your selection criteria.

You can also use the **Select** and then **Item Named…** menu option to locate which contig contains a particular fragment. With this command, you must type the full name into the **Find What:** box and then press the **Find** button. The contig containing the fragment will be highlighted.

RENAMING A SEQUENCE OR CONTIG

RENAMING USING THE MOUSE

To rename a sequence sequence or contig with the mouse, click twice on the name of the item. This selects the icon and puts a textbox around the name. Just type in the new name and when you are done, click elsewhere in the **Project Window**.

If you only want to change part of the name, drag the cursor over what you want to replace and then type the replacement text in the highlighted area.

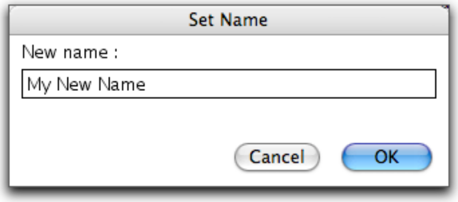
You can also use the **Copy** and **Paste** functions to rename sequences.

RENAMING USING THE MENU

If you wish to rename a sequence, you can go to the **Sequence** menu and choose the **Rename Sequence…** command. To rename a contig, go to the **Contig** menu and click on the **Rename Contig…** command. **Sequencher** displays a dialog (Figure 5-2) with a field for you to enter the new name.

Click the **OK** button when you have finished.

Figure 5-2 Renaming a sequence



Set Name

New name :

My New Name

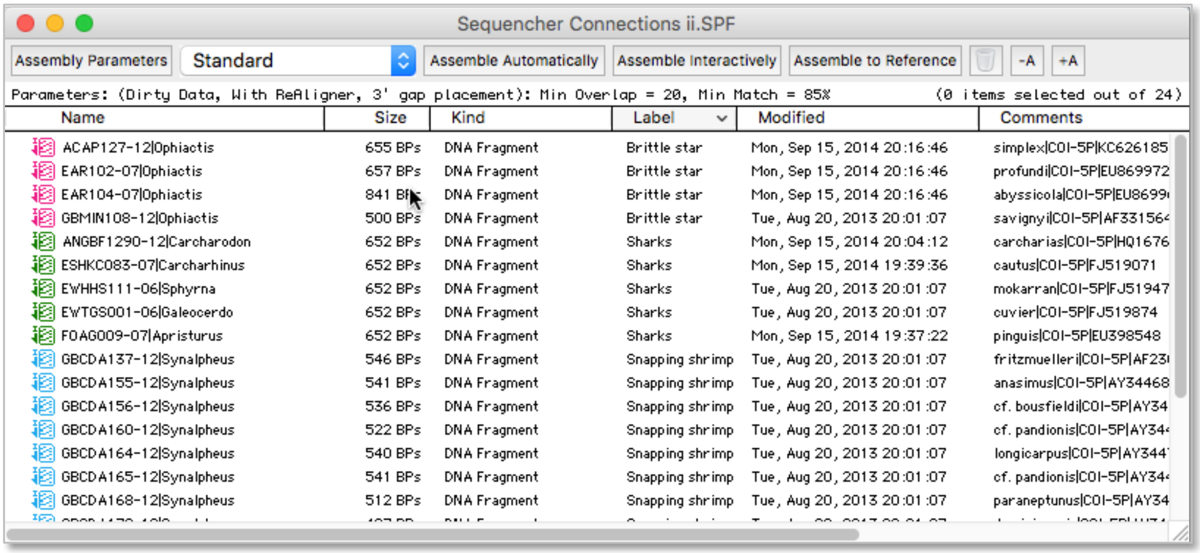
LABELING ITEMS IN THE PROJECT WINDOW

You can organize items visually in the **Project Window** by applying **Labels** to the icons. Before using **Labels** for the first time, you should customize the list of labels under the **General** section of **User Preferences**. You will see a list of **Label & Name** options. You can also select from the default labels available under the **Edit** menu and **Label** submenu. (See Chapter 23 “Customizing Sequencher and User Preferences” for more information.)

When viewing your project as a list, you will see the name of the label in the list. Labelling your icons allows you to sort or find your information easily. For instance, you might define red labels as cDNA data and blue labels as genomic data. Clicking on the label heading in the **Project Window** will sort your project items by label (see Figure 5-3).

You can also find sequences bearing a particular label by going to the Select menu and choosing the All Items That command with the Have Labels Containing… suboption.

Figure 5-3 Viewing labels as a list



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%   (0 items selected out of 24)

Name	Size	Kind	Label	Modified	Comments
ACAP127-12|Ophiactis	655 BPs	DNA Fragment	Brittle star	Mon, Sep 15, 2014 20:16:46	simplex|COI-5P|KC626185
EAR102-07|Ophiactis	657 BPs	DNA Fragment	Brittle star	Mon, Sep 15, 2014 20:16:46	profundi|COI-5P|EU869972
EAR104-07|Ophiactis	841 BPs	DNA Fragment	Brittle star	Mon, Sep 15, 2014 20:16:46	abyssicola|COI-5P|EU8699
GBMIN108-12|Ophiactis	500 BPs	DNA Fragment	Brittle star	Tue, Aug 20, 2013 20:01:07	savignyi|COI-5P|AF33156
ANGBF1290-12|Carcharodon	652 BPs	DNA Fragment	Sharks	Mon, Sep 15, 2014 20:04:12	carcharias|COI-5P|HQ1676
ESHKC083-07|Carcharhinus	652 BPs	DNA Fragment	Sharks	Mon, Sep 15, 2014 19:39:36	cautus|COI-5P|FJ519071
EWHHS111-06|Sphyrna	652 BPs	DNA Fragment	Sharks	Tue, Aug 20, 2013 20:01:07	mokarran|COI-5P|FJ51947
EWTGS001-06|Galeocerdo	652 BPs	DNA Fragment	Sharks	Tue, Aug 20, 2013 20:01:07	cuvier|COI-5P|FJ519874
FOAG009-07|Apristurus	652 BPs	DNA Fragment	Sharks	Mon, Sep 15, 2014 19:37:22	pinguis|COI-5P|EU398548
GBCDA137-12|Synalpheus	546 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	fritzmuelleri|COI-5P|AF23
GBCDA155-12|Synalpheus	541 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	anasimus|COI-5P|AY34468
GBCDA156-12|Synalpheus	536 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	cf. bousfieldi|COI-5P|AY34
GBCDA160-12|Synalpheus	522 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	cf. pandionis|COI-5P|AY34
GBCDA164-12|Synalpheus	540 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	longicarpus|COI-5P|AY344
GBCDA165-12|Synalpheus	541 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	cf. pandionis|COI-5P|AY34
GBCDA168-12|Synalpheus	512 BPs	DNA Fragment	Snapping shrimp	Tue, Aug 20, 2013 20:01:07	paraneptunus|COI-5P|AY34

SAVING YOUR WORK

SAVING YOUR PROJECT

To save a project, go to the **File** menu and choose **Save Project**, which saves all your changes to the current project file. When you use the **Save Project** command for a new project, **Sequencher** will prompt you to name your project with the **Save As** (Windows) or **Save: Sequencher** (Mac) dialog. Enter a name for the new project and then click **Save.**

AUTO-SAVE

If you are working on complex projects, you may find it helpful to have **Sequencher** automatically save your work at pre-set time intervals. You will need to enable the **Auto- Save** user preference from **User Preferences…** from the **Window** menu. (See the Chapter 23 “Customizing Sequencher and User Preferences” for more details.)

SAVING A DUPLICATE OF A CURRENT PROJECT

If you wish to create a second copy of a current project, you can use the **Save Project As…** command to create the new copy of your project with a different name. When you save the new project, a new project icon appears on your disk.

REVERTING TO THE SAVED VERSION OF THE PROJECT

If you are working in a project and want to cancel all the changes you have made since you last saved it, go to the **File** menu and choose **Revert To Saved Project… Sequencher** displays a dialog that tells you when your project was last saved. If you click on the **Yes, Revert to Saved Version** button, you will lose the most recent changes. Click on the **Cancel** button to cancel the command.

Note: Remember, if you click the **Yes, Revert to Saved Version** button, you cannot undo this action.

CLOSING THE PROJECT WINDOW

When you have finished working on an open project and want to work on another, you must first close the open project. You can either close the project by going to the **File** menu and clicking on **Close Project** or you can close the **Project Window** by clicking on the close box. In either case, **Sequencher** will prompt you to save your work if you have made any changes since your last save.

To continue working in **Sequencher**, either go to the **File** menu to create a new project by choosing **New Project** or click on **Open Project** to work on a previously saved file.

Note: If you want to save the data you have entered, you must save the project. Saving a project is different from recording a sequence as experimental data. Only saving the project actually stores data on the disk!

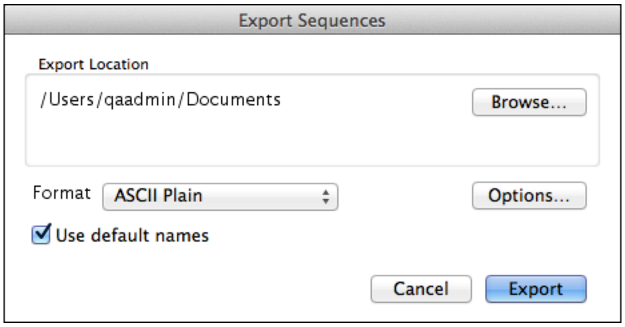
EXPORTING DATA

You can save your work in a file format other than **Sequencher**. You can export the entire file or a subset of the project. Click on the **File** menu and choose an **Export** submenu option. A dialog will allow you to specify the file format in which you want to export your selection (Figure 5-4). For example, if you choose **Selection as Subproject,** you can specify CAF format (for which you can set a number of options), Fasta concatenated format, or older versions of

Sequencher. This is especially helpful if you want to share data with colleagues working with older versions of **Sequencher**.

Note: It is important to remember that any version-specific features you may have used will be lost if you save to an earlier version of **Sequencher**. (See Chapter 21 "Exporting Data " for more information.)

Figure 5-4 Export sequences format menu



Export Sequences

Export Location

/Users/qaadmin/Documents

Format ASCII Plain

☑ Use default names

EXITING THE PROGRAM

When you have finished working with **Sequencher**, make sure you have saved your project. Then go to the **File** menu and choose **Exit** (Windows) or to the **Sequencher** menu and choose **Quit Sequencher** (Mac).

6. PREPARING YOUR DATA FOR ASSEMBLY

In this chapter, we introduce you to **Sequencher**'s powerful tools for trimming poor quality data or vector contamination from your sequences. You will learn how to use these functions when you add sequence data to a project. We discuss the various criteria you can use such as trimming off ambiguities, trimming off low confidence data, how to screen for specific vector contamination, and how to change all the trim tools' criteria to your own specifications.

REMOVING UNDESIRABLE DATA

Sequencher's trimming tools allow you to trim one sequence at a time or thousands of sequences at a time. The two trimming commands under the **Sequence** menu are **Trim Ends…** and **Trim Vector…**. Both of the **Trim** dialogs share a similar interface. All the sequences to be examined are collected in a single table. The criteria that you set determine how much data you trim off.

The trimmed data are fully recoverable because **Sequencher** always stores two copies of every imported file, the original sequence and the edited version. In this way, you can always use the **Revert to Experimental Data** command from the **Sequence** menu to recover the original untrimmed sequence.

TRIMMING POOR QUALITY DATA

SELECTING AND TRIMMING SEQUENCE DATA

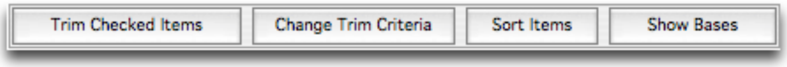
Select one or more sequences(s) in the **Project Window** by clicking on them. Then go to the **Sequence** menu and choose the **Trim Ends…** command. Note that if you select contigs, the **Trim Ends…** function only affects unassembled sequences. All of the sequences are collected into a single table. At the top of the table is a button bar (Figure 6-1) that you use to manage Trim criteria, review which bases will be trimmed, and perform the trim command.

PERFORMING A TRIM WITH DEFAULT SETTINGS

You can specify how much dirty data should be trimmed off as a function of the number of ambiguous base calls per run of bases by altering the **Ends Trimming** criteria. Each pane in the **Ends Trimming** window shows how much sequence should be trimmed off. There will be a checkbox at an end if it requires trimming. Bases to be trimmed will be marked in red.

You can proceed with the trim by clicking on the **Trim Checked Items** button. Once you have trimmed your data, dismiss the window with the command from the **File** menu called **Close Window**.

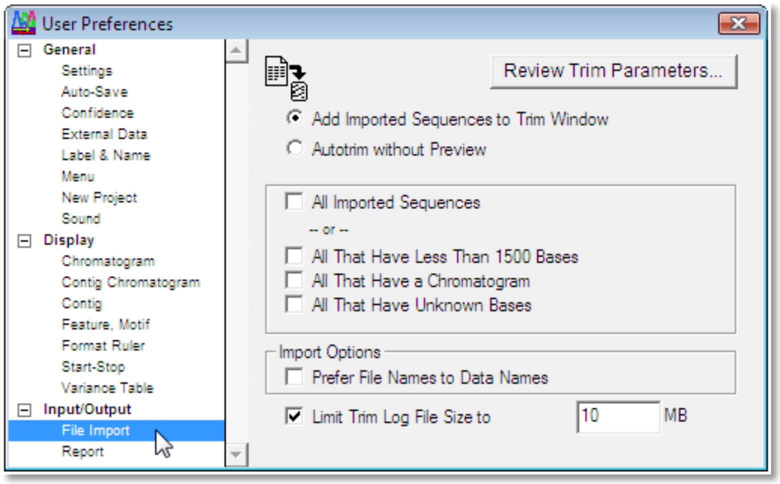
Figure 6-1 The Ends Trimming button bar



AUTOMATICALLY SELECTING SEQUENCES FOR ENDS TRIMMING

You can set the **User Preferences** so that sequence files are automatically added into the trimming window as they are imported. During Import, **Sequencher** loads the sequences that require trimming directly into the trim window. To determine which sequences are directly imported into the trim window, go to the **Window** menu and, under **User Preferences**, click on **Input/Output** and then select the submenu **File Import**. Figure 6-2 shows the criteria you use to specify which sequences are candidates for direct import to the **Ends Trimming** window. (For more information, see Chapter 23 "Customizing Sequencher and User Preferences".)

Figure 6-2 User Preferences for setting trim criteria when files are imported



● Add Imported Sequences to Trim Window
○ Autotrim without Preview

□ All Imported Sequences
-- or --
□ All That Have Less Than 1500 Bases
□ All That Have a Chromatogram
□ All That Have Unknown Bases

Import Options
□ Prefer File Names to Data Names

☑ Limit Trim Log File Size to 10 MB

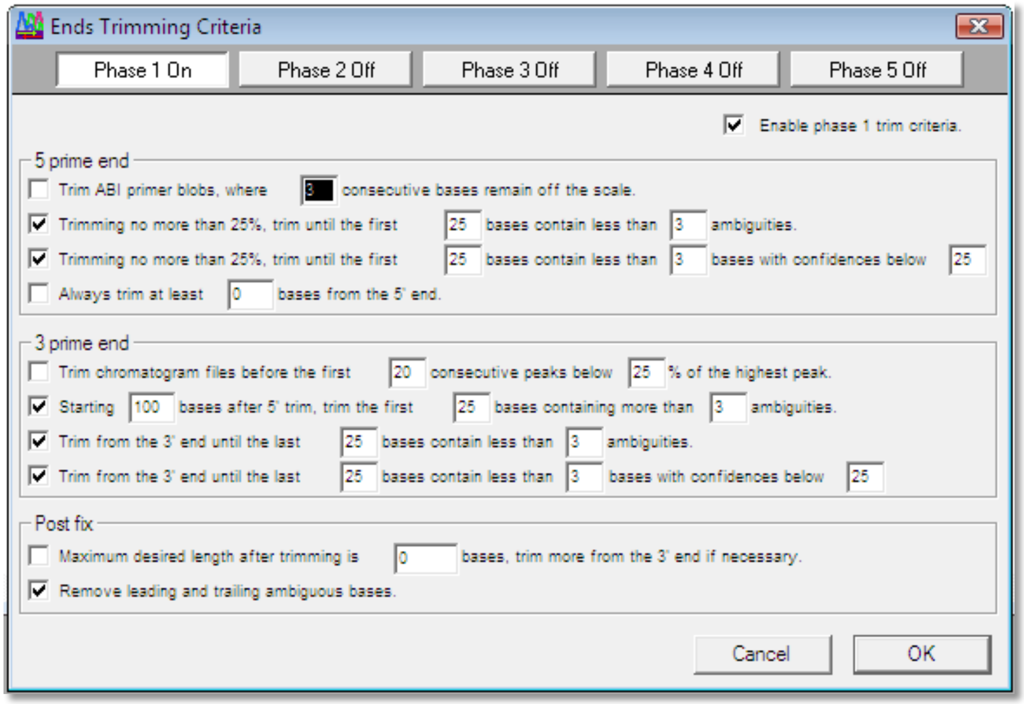
HOW TO SET THE ENDS TRIMMING CRITERIA

The amount to be trimmed is based on the criteria you set for both the 5' and 3' ends of the sequence. These criteria can be set independently of each other. To review or change those criteria, go to the **Ends Trimming** window and click on the **Change Trim Criteria** button.

You will see a series of buttons across the top of the **Ends Trimming Criteria** window. These buttons allow you to a) create a library of different trim criteria thereby saving time when you are managing a series of projects or b) allow controlled trimming using specific criteria in a certain order.

As you can see in Figure 6-3, you can trim off dye-primer peaks, set windows of acceptable reliability, percentage of the highest peak height (3' end), and even set an absolute number of bases. For example, you could specify that sequences must have fewer than 3 N’s in the first 20 and last 30 bases. To activate specific trim criteria, you must first select the checkbox that governs the criteria fields.

Figure 6-3 The Ends Trimming Criteria window



Ends Trimming Criteria

Phase 1 On | Phase 2 Off | Phase 3 Off | Phase 4 Off | Phase 5 Off

☑ Enable phase 1 trim criteria.

5 prime end

☐ Trim ABI primer blobs, where 3 consecutive bases remain off the scale.

☑ Trimming no more than 25%, trim until the first 25 bases contain less than 3 ambiguities.

☑ Trimming no more than 25%, trim until the first 25 bases contain less than 3 bases with confidences below 25

☐ Always trim at least 0 bases from the 5' end.

3 prime end

☐ Trim chromatogram files before the first 20 consecutive peaks below 25 % of the highest peak.

☑ Starting 100 bases after 5' trim, trim the first 25 bases containing more than 3 ambiguities.

☑ Trim from the 3' end until the last 25 bases contain less than 3 ambiguities.

☑ Trim from the 3' end until the last 25 bases contain less than 3 bases with confidences below 25

Post fix

☐ Maximum desired length after trimming is 0 bases, trim more from the 3' end if necessary.

☑ Remove leading and trailing ambiguous bases.

USING CONFIDENCE TO TRIM ENDS

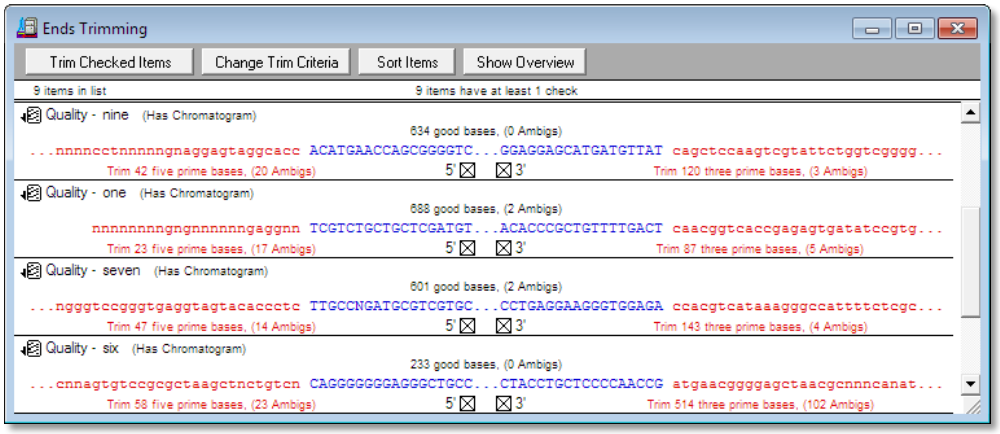
In addition to trimming data based on the number of ambiguities at the 5' and 3' ends, you can trim sequence data based on the confidence scores assigned by your base caller. Confidence scores (also called quality values) are numbers associated with each base call and which define the likelihood that a base call is incorrect. The most common scale is from 1-60, where “60” represents a $1/10^6$ chance of a wrong call and 20 represents a $1/10^2$ chance. Depending on the program used, the confidence score may be based on peak height, the presence of more than one peak, and/or the spacing between the peaks. (See Figure 6-3, the **Ends Trimming Criteria** window for trimming by confidence criteria.)

REVIEWING THE TRIM

Each pane in the trimming windows shows schematically how much sequence should be trimmed off. Blue lines represent the acceptable sequence data. The blue line is then flanked by red scissors, which mark both the 5’ and 3’ suggested trims based on the current **Trim** criteria. The remaining red regions of the sequence define the portion of the sequence to be trimmed.

Click on the **Show Bases** button to see more information on the bases. After you have clicked **Show Bases**, the button toggles to **Show Overview**. **Sequencher** also provides additional control within the **Trim Ends** window. If you want **Sequencher** to ignore the trim command for one or the other ends of the fragment(s) without modifying the overall specifications, just click off the checkbox associated with that trim position. Figure 6-4 shows a portion of the base detail for these sequences. You can double-click on the sequence icon and invoke a **Sequence Editor** to examine the data in more detail.

Figure 6-4 Bases view in Ends Trim window

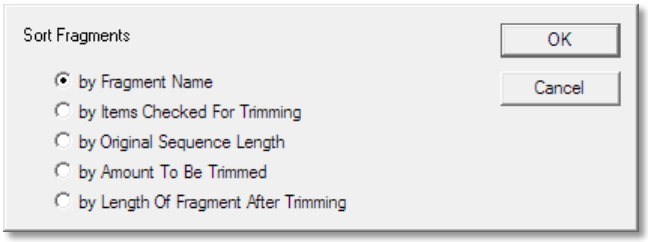


```
9 items in list                                9 items have at least 1 check
Quality - nine  (Has Chromatogram)
                                634 good bases, (0 Ambigs)
...nnnncctnnnnngnaggagtaggcacc ACATGAACCAGCGGGGTC...GGAGGAGCATGATGTTAT cagctccaagtcgtattctggtcgggg...
        Trim 42 five prime bases, (20 Ambigs)      5'☒  ☒3'      Trim 120 three prime bases, (3 Ambigs)
Quality - one  (Has Chromatogram)
                                688 good bases, (2 Ambigs)
   nnnnnnnngngnnnnnnngaggnn TCGTCTGCTGCTCGATGT...ACACCCGCTGTTTTGACT caacggtcaccgagagtgatatccgtg...
        Trim 23 five prime bases, (17 Ambigs)      5'☒  ☒3'      Trim 87 three prime bases, (5 Ambigs)
Quality - seven  (Has Chromatogram)
                                601 good bases, (2 Ambigs)
...ngggtccgggtgaggtagtacaccctc TTGCCNGATGCGTCGTGC...CCTGAGGAAGGGTGGAGA ccacgtcataaagggccattttctcgc...
        Trim 47 five prime bases, (14 Ambigs)      5'☒  ☒3'      Trim 143 three prime bases, (4 Ambigs)
Quality - six  (Has Chromatogram)
                                233 good bases, (0 Ambigs)
...cnnagtgtccgcgctaagctnctgtcn CAGGGGGGGAGGGCTGCC...CTACCTGCTCCCCAACCG atgaacggggagctaacgcnnncanat...
        Trim 58 five prime bases, (23 Ambigs)      5'☒  ☒3'      Trim 514 three prime bases, (102 Ambigs)
```

SORTING ITEMS

In order to assist you with the review of your data prior to trimming, you can sort the sequences. Click on the **Sort Items** button at the top of the **Ends Trimming** window to order the sequences being trimmed. Figure 6-5 shows the sort criteria.

Figure 6-5 Sort Fragments dialog



Sort Fragments

- ● by Fragment Name
- ○ by Items Checked For Trimming
- ○ by Original Sequence Length
- ○ by Amount To Be Trimmed
- ○ by Length Of Fragment After Trimming

EXECUTING THE TRIM

Once you have reviewed your data and are happy with the criteria you have set, you proceed with the trim by clicking on the **Trim Checked Items** button. **Sequencher** warns you that this command cannot be reversed with the **Undo** command. To proceed, click on the **Trim** button.

Any red lines or bases will disappear from the **Ends Trimming** window. Now you can dismiss the window with the command from the **File** menu called **Close Window.**

Remember to save your work at this point by going to the **File** menu and using the **Save Project** command, since none of your trim work will have been recorded onto disk. Alternatively, use the **Auto-Save** function to save your work at user-defined intervals (see Chapter 23 “Customizing Sequencher and User Preferences”).

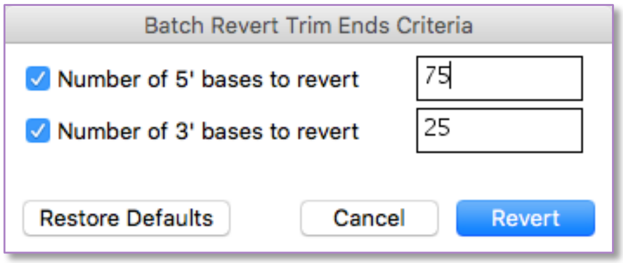
TRIM ENDS WITHOUT PREVIEW

Once you have established a set of trim criteria for your data, you do not need to perform the same process each time. You can use the **Trim Ends Without Preview** command. This command executes the trim without leaving the **Project Window**. If you are working with **Confidence** scores, then you will notice that the **Quality** improves after the trim. This will be the only visible indication that the trim has been successful.

BATCH REVERT TRIM ENDS

If you find that the **Trim Criteria** you used are too stringent, you can use the **Batch Revert Trim Ends…** command to selectively restore bases on either or both of the 5’ or 3’ ends. Select the sequences from the **Project Window** whose bases you want to revert. Choose the **Sequence>Batch Revert Trim Ends…** command and enter the number of bases you wish to revert into the textboxes. Finally click on the **Revert** button.

Figure 6-6 Batch Revert Trim Ends dialog



Batch Revert Trim Ends Criteria

☑ Number of 5' bases to revert 75

☑ Number of 3' bases to revert 25

You should see that the values in the **Project Window Quality** and **Length** columns change**. Batch Revert Trim Ends** does not act on Reference Sequences, Refrigerators or Contigs. If a sequence cannot be reverted using your **Batch Revert** criteria, **Sequencher** will warn you and ask if you want to use **Revert to Experimental** on those sequences.

TRIMMING VECTOR CONTAMINATION

SELECTING SEQUENCES FOR VECTOR TRIMMING

Select one or more sequence(s) in the **Project Window** and go to the **Sequence** menu and choose the **Trim Vector**… command. Note that even if you select contigs, the **Trim Vector** function only affects unassembled sequences.

SPECIFYING A VECTOR USING VECBASE

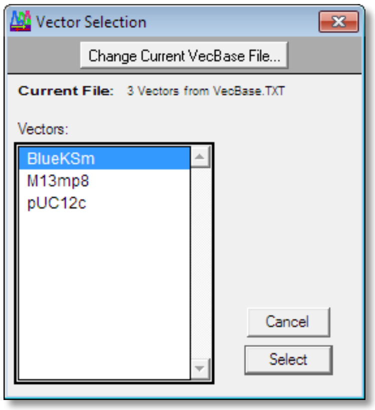
After you have selected the **Trim Vector…** command and if you have not previously specified vector insertion sites, a window will appear. Click on the **Choose Insertion Site Now** button if you are ready to work with the vector trim function. Otherwise, click on the **OK** button to dismiss the window.

You can also go to **Window**>**Specify Vector Insertion Sites**… to see this dialog.

Once you have pressed the **Choose Insertion Site Now** button, the **Vector Insertion Sites** dialog appears. In the bottom part of the window you will see that you can screen for up to five vectors at once.

Load vector information from the VecBase database by clicking the **Use VecBase File** button in the **Vector Insertion Sites** window**. Sequencher** displays a list of the vectors. Select the vector you want and click on the **Select** button.

Figure 6-7 Selecting a vector



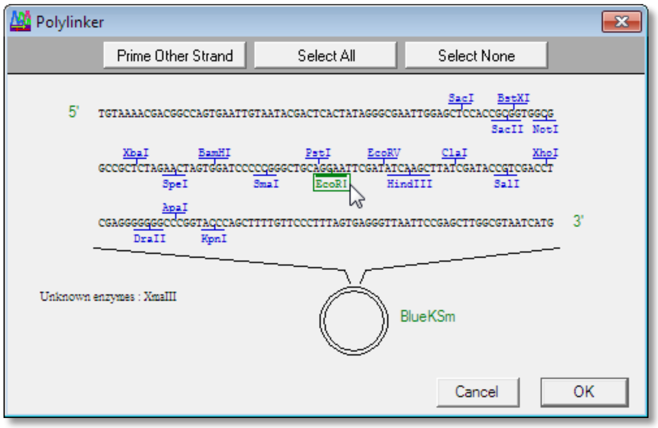
Current File: 3 Vectors from VecBase.TXT

Vectors:

BlueKSm
M13mp8
pUC12c

The polylinker region of the vector you select is displayed in the window shown in Figure 6-7.

Figure 6-8 The polylinker window



```
5' TGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGAGCTCCACCGCGGTGGCG
GCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGGAATTCGATATCAAGCTTATCGATACCGTCGACCT
CGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTCCGAGCTTGGCGTAATCATG 3'
```

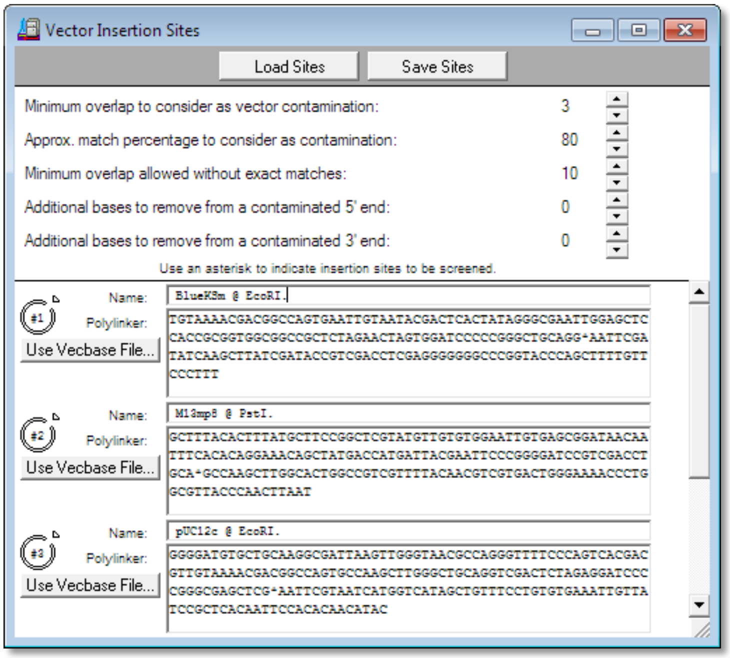
Unknown enzymes : XmaIII

Note: Not all vectors in VecBase have polylinker information. **Sequencher** cannot work with VecBase vectors that do not include polylinker data. Add these manually (see below).

Click the enzyme restriction site (s) where your sequence was inserted into the vector. The name of a selected site(s) changes to green and is boxed. If you change your mind and decide on a different site, you must click the selected site again to deselect it. It will then return to its original color. The bases from the selected site(s) are automatically entered into the **Vector Insertion Sites** dialog, shown in Figure 6-8.

If the cloning technique you are using places your insert between two sites, replace the vector sequence between the two sites with a single bullet (•), just as your insert replaces this sequence.

Figure 6-9 Vector Dialog displaying vector and insertion site



Minimum overlap to consider as vector contamination: 3

Approx. match percentage to consider as contamination: 80

Minimum overlap allowed without exact matches: 10

Additional bases to remove from a contaminated 5' end: 0

Additional bases to remove from a contaminated 3' end: 0

Use an asterisk to indicate insertion sites to be screened.

Name: BlueKSm @ EcoRI.

Polylinker:

```
TGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGAGCTC
CACCGCGGTGGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGG*AATTCGA
TATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTT
CCCTTT
```

Name: M13mp8 @ PstI.

Polylinker:

```
GCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAA
TTTCACACAGGAAACAGCTATGACCATGATTACGAATTCCCGGGGATCCGTCGACCT
GCA*GCCAAGCTTGGCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTG
GCGTTACCCAACTTAAT
```

Name: pUC12c @ EcoRI.

Polylinker:

```
GGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGAC
GTTGTAAAACGACGGCCAGTGCCAAGCTTGGGCTGCAGGTCGACTCTAGAGGATCCC
CGGGCGAGCTCG*AATTCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTA
TCCGCTCACAATTCCACACAACATAC
```

SPECIFYING THE INSERTION SITE MANUALLY

If the vector you select in VecBase does not have polylinker information, or if you are screening for contamination using a sequence not already in VecBase, you must enter the information manually. In the **Specify Vector Insertion Sites** window, paste the sequence (maximum 250 bases) into the **Polylinker** window. Insert a “bullet” (“•”) by holding down the **Shift** (Windows) or **Option** (Mac) key and the 8 key (not the one from the number keypad) to represent the sequence you wish to retain.

SAVING AND LOADING VECTOR SITES

You can save manually entered vector information to a file by clicking on the **Save Sites** button. You can reload the file in the future by using the **Load Sites** button. This allows you to save your preferences and parameters for a later session.

SETTING VECTOR TRIMMING CRITERIA

Sequencher has a set of default criteria to facilitate the recognition of any contaminating sequence. These criteria are listed in the top region of the **Vector Insertion Sites** window. If these parameters do not work for your specific circumstances, you can alter the vector screening parameters. As an example, if there is ambiguity in the region of the vector junction, you can increase the likelihood that **Sequencher** will recognize the vector sequence by decreasing the approximate match percentage. If it is essential to remove even a single base of vector contamination, you can decrease the minimum overlap to be considered as vector contamination. You can change any of the five independent parameters in order to improve your trim. Elevator buttons are used to increase or decrease the parameter value. The following table describes the recognition criteria for contaminating sequence.

Table 6-1 Description of Trim parameters

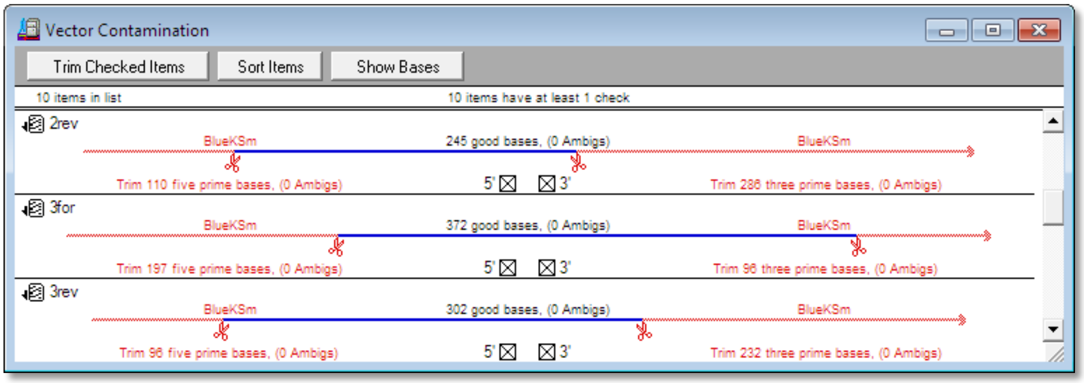
Parameter	**Description**
Minimum overlap to consider as vector contamination	The smallest amount of overlap to consider as vector contamination. The default value is three bases.
Approx. match percentage to consider as contamination	The percentage match **Sequencher** needs to find before considering the matching sequence to be vector contamination. Ambiguities do not count as full matches or mismatches.
Minimum overlap allowed without exact matches	The smallest number of matching bases within which an inexact match is permitted. If fewer bases than this match, they must match exactly.
Additional bases to remove from a contaminated 5' end	Extra material, if any, to be trimmed off the 5' end of the remaining DNA sequence along with the vector contamination. The default is zero bases.
Additional bases to remove from a contaminated 3'	Extra material, if any, to be dropped off the 3' end of the remaining DNA sequence along with the vector contamination. The default is zero bases.

Once you have finished choosing vectors and setting recognition criteria, close the **Vector Insertion Sites** window.

HOW SEQUENCHER SCREENS SEQUENCES

When you execute the **Trim Vector** command, **Sequencher** screens the selected sequences for vector contaminants using the vector sequences you specified. **Sequencher** starts at the designated insertion site, the bullet, and then scans outwards. Any sequences that are found to contain vector bases are collected in the **Vector Contamination** window. (See Figure 6-10.) Please note that any contigs you have selected will be ignored, as will any sequences containing fewer than 21 bases.

Figure 6-10 The Vector Contamination window



10 items in list

10 items have at least 1 check

2rev

BlueKSm — 245 good bases, (0 Ambigs) — BlueKSm

Trim 110 five prime bases, (0 Ambigs) — 5' ☒ ☒ 3' — Trim 286 three prime bases, (0 Ambigs)

3for

BlueKSm — 372 good bases, (0 Ambigs) — BlueKSm

Trim 197 five prime bases, (0 Ambigs) — 5' ☒ ☒ 3' — Trim 96 three prime bases, (0 Ambigs)

3rev

BlueKSm — 302 good bases, (0 Ambigs) — BlueKSm

Trim 96 five prime bases, (0 Ambigs) — 5' ☒ ☒ 3' — Trim 232 three prime bases, (0 Ambigs)

You can print out the **Vector Contamination** window as a record for your lab notebook by going to the **File** menu and choosing **Print**. This is particularly useful if you use the **Show Bases** button to show which sequences are to be trimmed. You can also sort the window using various criteria by clicking on **Sort Items**.

EXECUTING THE TRIM COMMAND

To perform the vector screen and trim, go to the **Sequence** menu and choose **Trim Vector…** Once you have reviewed your data and are happy with the criteria you have set, proceed with the trim by clicking on the **Trim Checked Items** button. **Sequencher** warns you that this command cannot be reversed with the **Undo** command. To proceed, click on the **Trim** button. Any red lines or bases will disappear from the **Ends Trimming** window. Now you can dismiss the window with the command from the **File** menu called **Close Window.**

If you wish to revert some or all of your sequences to their original state, you can go to the **Sequence** menu and click on the **Revert to Experimental Data** command.

Remember to save your work at this point by going to the **File** menu and using the **Save Project** command since none of your trim work will have been recorded onto disk. Alternatively, use the **Auto-Save** function to save your work at user-defined intervals (see Chapter 23 "Customizing Sequencher and User Preferences").

TRIM LOG

The results of any trim will be recorded in a trim log file called **Sequencher**Tr**imEndsHistory** which can be found in the **Home** directory. You can locate the log file from within **Sequencher** using the **Open Trim Log Folder** command from the **Window** menu.

7. THE REFERENCE SEQUENCE

In this chapter, we explain Reference Sequences – sequences or fragments used as baselines or benchmarks for subsequent comparison. We will describe the special features of Reference Sequences. We will then discuss how to designate a sequence as a Reference for your project, how to assemble sequences to the Reference, and how to edit contigs containing a Reference Sequence.

WHAT IS A REFERENCE SEQUENCE?

A Reference Sequence is any sequence that you have selected to act as a model for your subsequent comparisons. You may have obtained this sequence from one of the public DNA databases or you may have characterized it yourself. Once you have marked your sequence as a Reference, **Sequencher** attaches certain properties to it.

WHY USE A REFERENCE SEQUENCE?

There are many situations where you might want to designate one sequence in a project as a Reference Sequence. In forensics, human identification depends on comparing differences with a standard Reference Sequence. In medical genetics, a lab might want to compare the same gene from family members to the specific variant that is known to be part of a disease process. In a population study, a lab might compare the same gene from thousands or tens of thousands of individuals to a wild type sequence. In all of these circumstances, using **Sequencher's Assemble to Reference** command with a Reference Sequence allows you to pinpoint mutations and variants rapidly.

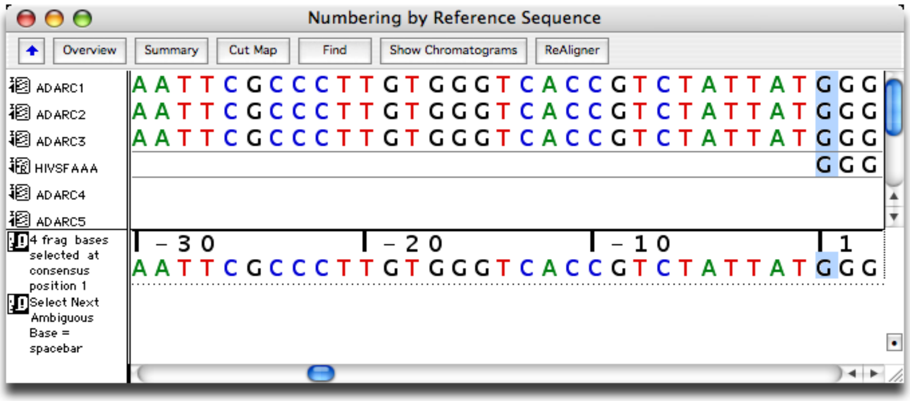
REFERENCE SEQUENCE PROPERTIES

Before you start work with the Reference Sequence, you should be aware of a number of important properties.

BASIC REFERENCE SEQUENCE PROPERTIES

When you assemble sequences, which include a Reference Sequence, the base numbering of a contig will be the same as the base numbering of the Reference Sequence.

Figure 7-1 Contig showing base numbering set by Reference Sequence



```
ADARC1    AATTCGCCCTTGTGGGTCACCGTCTATTATGGG
ADARC2    AATTCGCCCTTGTGGGTCACCGTCTATTATGGG
ADARC3    AATTCGCCCTTGTGGGTCACCGTCTATTATGGG
HIVSFAAA                                GGG
ADARC4
ADARC5
          -30       -20       -10       1
          AATTCGCCCTTGTGGGTCACCGTCTATTATGGG
```

4 frag bases selected at consensus position 1

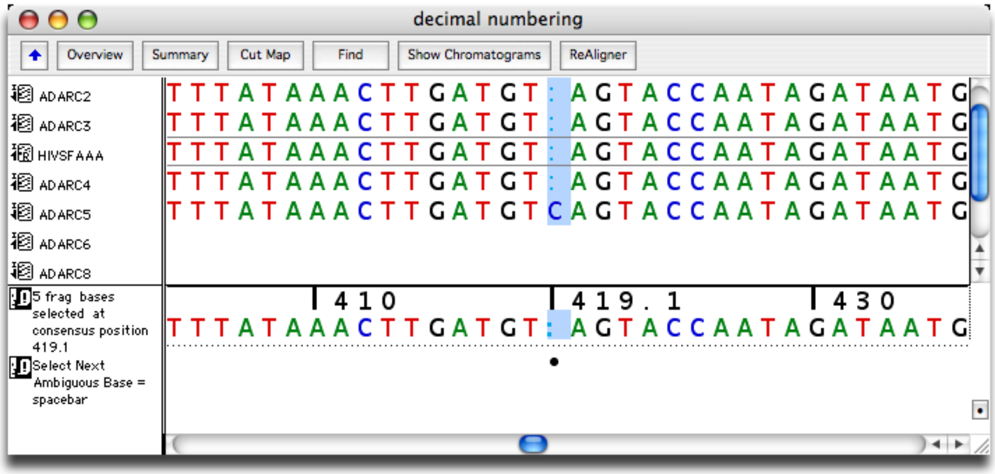
Select Next Ambiguous Base = spacebar

When a contig is built that includes a Reference Sequence, the contig will always remain in the same orientation relative to that Reference. This way you can keep the contig from *flipping* as more sequences are added. Because of this, there can only be one Reference Sequence in a contig. You can easily spot the Reference Sequence in a **Contig Editor** because it is highlighted along its entire length with a gray border.

When you assemble sequences that include a Reference Sequence, you will notice that gaps in the Reference Sequence are given decimal numbers, thereby preserving the absolute numbering of base positions.

The Reference Sequence does not contribute to the calculation of the consensus line of a contig. For example, if a 'T' in the Reference Sequence overlaps a 'G' in another sequence, the consensus line will show a 'G'.

Figure 7-2 Base numbering showing decimal numbers



```
ADARC2    TTTATAAACTTGATGT:AGTACCAATAGATAATG
ADARC3    TTTATAAACTTGATGT:AGTACCAATAGATAATG
HIVSFAAA  TTTATAAACTTGATGT:AGTACCAATAGATAATG
ADARC4    TTTATAAACTTGATGT:AGTACCAATAGATAATG
ADARC5    TTTATAAACTTGATGTCAGTACCAATAGATAATG
ADARC6
ADARC8
                410     419.1      430
          TTTATAAACTTGATGT:AGTACCAATAGATAATG
```

5 frag bases selected at consensus position 419.1

Select Next Ambiguous Base = spacebar

HOW TO MARK A SEQUENCE AS A REFERENCE

To designate a sequence as a Reference Sequence, select the sequence icon in the **Project Window** and, under the **Sequence** menu, select the **Reference Sequence** command**.** From now on, that sequence icon will include a small letter "R" to remind you that it has been marked as a Reference.

ASSEMBLE TO REFERENCE

Assemble to Reference is a powerful command which allows you to assemble all the samples you select to a single Reference Sequence, regardless of any inconsistencies between the individual sequences. Because it is a many-to-one comparison instead of the normal many-to-many comparison, **Assemble to Reference** is much faster than the standard **Assemble Automatically** mode.

To Assemble to Reference, select the sequences you want to assemble. Click on the **Assemble to Reference** button or go to the **Assemble** menu and choose **Assemble to Reference**. You will get a warning dialog if you have not included a properly designated Reference Sequence among your selected sequences.

Note: You can use the **Assembly Parameters** to change the conditions under which the assembly will take place. (See Chapter 9 “Sequence Assembly” for more information.)

TO REFERENCE BY NAME

To **Reference by Name** is a special case where the behavior of the Reference Sequence is modified to work with the **Assemble by Name** command. The **To Reference by Name** button will only be displayed if **Assemble by Name** has been enabled. (For more information on Assemble by Name, see Chapter 9 “Sequence Assembly”.)

When the **To Reference by Name** command is used, each contig created will contain a Reference Sequence and will also retain the numbering of the Reference Sequence.

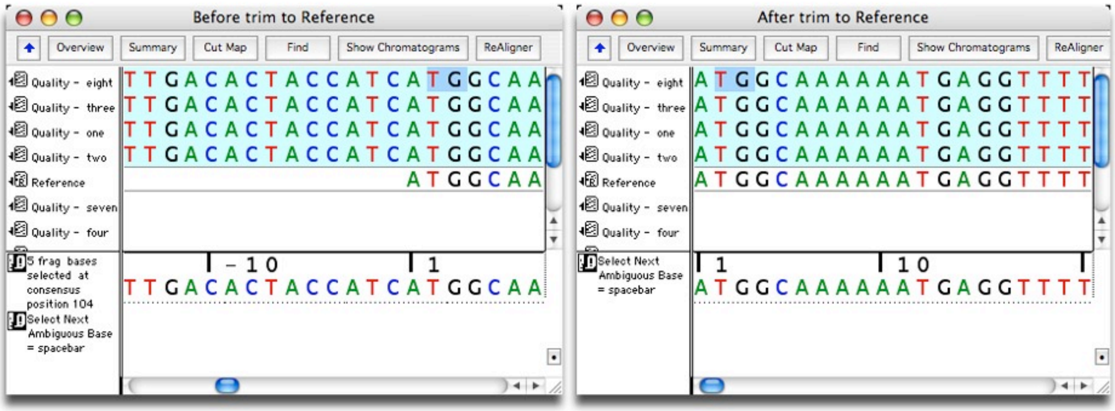
To use **To Reference by Name**, select the sequences and Reference Sequence you want to assemble. Click on the **To Reference by Name** button or go to the **Assemble** menu and choose the **Assemble Contigs** command and then click on **To Reference by Name** from the submenu.

TRIM TO REFERENCE

The **Trim to Reference Sequence** command is used to remove bases in a contig that flank the Reference Sequence on both the 5’ and 3’ ends. To use this command, go to the **Contig** menu and choose **Trim to Reference Sequence**. This command is only available when you are in a contig that contains a Reference Sequence.

Note: You cannot undo this action. If there are any sequences that do not overlap the Reference Sequence, they will be removed from the contig. **Sequencher** will warn you which sequences have been removed in a window, which you can print for your records.

Figure 7-3 Contig with Reference Sequence flanked by overhanging sequence (before and after)



```
Quality - eight   TTGACACTACCATCATGGCAA
Quality - three   TTGACACTACCATCATGGCAA
Quality - one     TTGACACTACCATCATGGCAA
Quality - two     TTGACACTACCATCATGGCAA
Reference                     ATGGCAA
Quality - seven
Quality - four
5 frag bases selected at consensus position 104
Select Next Ambiguous Base = spacebar
                  -10        1
                  TTGACACTACCATCATGGCAA
```

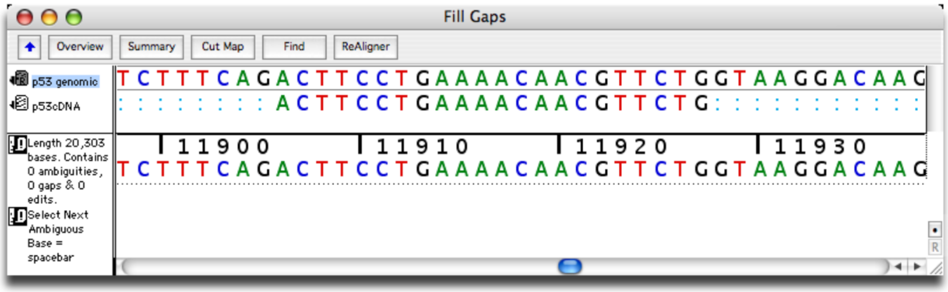
```
Quality - eight   ATGGCAAAAAAATGAGGTTTT
Quality - three   ATGGCAAAAAAATGAGGTTTT
Quality - one     ATGGCAAAAAAATGAGGTTTT
Quality - two     ATGGCAAAAAAATGAGGTTTT
Reference         ATGGCAAAAAAATGAGGTTTT
Quality - seven
Quality - four
Select Next Ambiguous Base = spacebar
                  1        10
                  ATGGCAAAAAAATGAGGTTTT
```

CONTIG EDITING WITH A REFERENCE SEQUENCE

If you try to edit a Reference Sequence which has been assembled into a contig using either the standard **Assemble Automatically** or **Assemble to Reference**, a warning dialog will be displayed asking if you are sure this is what you want to do. You can proceed with the edit by clicking on the **Yes** button. You can cancel the warning and continue to edit by clicking on the **No** button.

The Reference Sequence is protected from any changes you make while you are editing in the consensus line. Therefore, if you delete gaps from the consensus, this will not affect the Reference Sequence unless it also has gaps at the position of the deleted bases. For example, you can delete a range of bases from all sequences by selecting a range within the **Consensus** sequence. When the deleted bases are in the middle of a sequence, they will be replaced with gaps to maintain the alignment with the Reference Sequence. Since the action cannot be reverted using the **Undo** command, you must use the **Revert to Experimental Data** command from the **Sequence** menu to recover.

Figure 7-4 Filling gaps with the Reference Sequence



```
p53 genomic   TCTTTCAGACTTCCTGAAAACAACGTTCTGGTAAGGACAAG
p53cDNA       ::::::::ACTTCCTGAAAACAACGTTCTG:::::::::::
Length 20,303 bases. Contains 0 ambiguities, 0 gaps & 0 edits.
Select Next Ambiguous Base = spacebar
                11900     11910     11920     11930
              TCTTTCAGACTTCCTGAAAACAACGTTCTGGTAAGGACAAG
```

REFERENCE SEQUENCE TRANSLATION

This command is only available when the Reference Sequence is assembled into a contig.

To see a translation of the Reference Sequence in the notation line below the consensus sequence, go to the **View** menu and choose **Reference Sequence Translation**. The translation has a pale gray border above and below the amino acids to distinguish it from the consensus sequence translation.

Note: As you cycle through the Consensus translation button modes, you will notice one mode which translates the consensus sequence in the same frame as the Reference Sequence. In this mode, the button icon is marked with an “r”.

8. THE SEQUENCE EDITOR

In this chapter, we explain basic sequence editing, setting base numbers, and how to reverse and complement your data. We then go on to discuss working with experimental data and how **Sequencher** enables you to find ambiguities, work with codon and restriction maps, and annotate features. You will also learn how to customize your sequence view and use voice verification to help you check base calls.

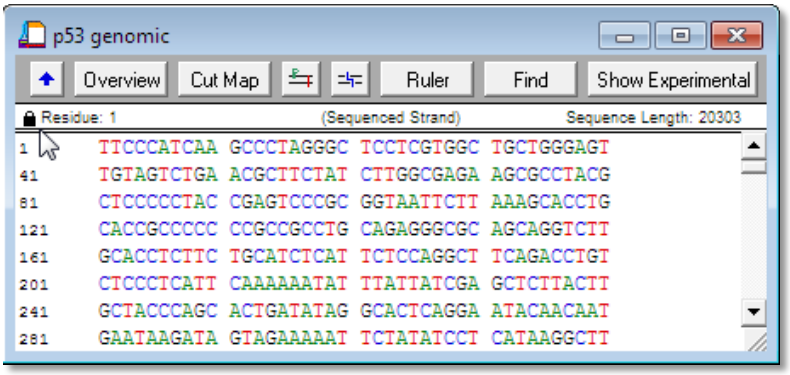
OPENING AN EXISTING SEQUENCE FOR EDITING

To open an existing sequence, select a sequence icon and then go to the **File** menu and choose **Open Window**. You can also just double-click the icon to open its editor window.

LOCKED EDITORS

If an editor displays experimental data or a sequence has been incorporated into a contig, it is locked. A padlock icon (shown to the left of the word "Residue" in the information bar above the sequence in Figure 8-1) shows that the file is locked.

Figure 8-1 A locked editor



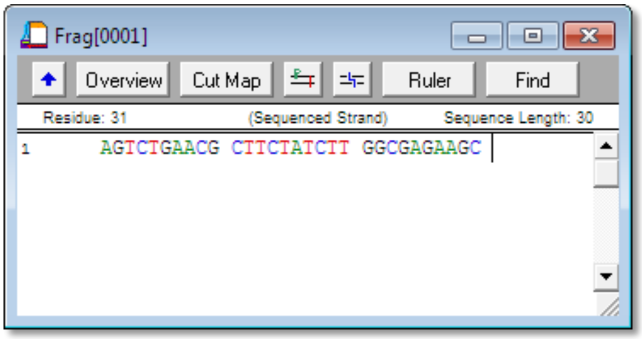
```
Residue: 1                    (Sequenced Strand)            Sequence Length: 20303
1       TTCCCATCAA GCCCTAGGGC TCCTCGTGGC TGCTGGGAGT
41      TGTAGTCTGA ACGCTTCTAT CTTGGCGAGA AGCGCCTACG
81      CTCCCCCTAC CGAGTCCCGC GGTAATTCTT AAAGCACCTG
121     CACCGCCCCC CCGCCGCCTG CAGAGGGCGC AGCAGGTCTT
161     GCACCTCTTC TGCATCTCAT TCTCCAGGCT TCAGACCTGT
201     CTCCCTCATT CAAAAAATAT TTATTATCGA GCTCTTACTT
241     GCTACCCAGC ACTGATATAG GCACTCAGGA ATACAACAAT
281     GAATAAGATA GTAGAAAAAT TCTATATCCT CATAAGGCTT
```

A locked editor is read-only. You cannot directly edit data, you can only view it. However, if the sequence has been incorporated into a contig, you can use the **Contig Editor** to make changes. (See Chapter 12 "Editing Contigs".)

BASIC SEQUENCE EDITING

The **Sequence Editor**, shown in Figure 8-2, is a text processor that is specially optimized to work with DNA sequences.

Figure 8-2 The Sequence Editor



Residue: 31 (Sequenced Strand) Sequence Length: 30

1 AGTCTGAACG CTTCTATCTT GGCGAGAAGC

The default system for coding in **Sequencher** is the IUPAC-IUB coding system. You can define your own coding system for bases and ambiguities. (See Chapter 23, “Customizing Sequencher and User Preferences” for details.) The IUPAC-IUB codes are included in Appendix 28 of this manual.

The **Sequence Editor** works like any other text editor. The insertion point can be moved using either the mouse or the arrow keys.

Note: If you hold down the **Alt** key and use an arrow key, the insertion point moves three bases. If you hold down both the **Ctrl** and **Alt** keys (Windows) or **control** and **Alt** keys (Mac), the insertion point moves nine bases. If you hold down the **Ctrl** (Windows) or **Cmd** (Mac) key and the left arrow, the insertion point moves to the beginning of the sequence. If you hold down **Ctrl** (Windows) or **Cmd** (Mac) key and the right arrow, the insertion point moves to the end of the sequence.

The editor also performs typical operations including **Cut, Copy, Paste,** and **Undo**. Use the mouse to highlight a selection; hold down the **Shift** key and click elsewhere in the sequence with the mouse to extend the selection.

There are also two commands for extending a current selection to the beginning or end of a sequence, **To Left End** and **To Right End**, both located under the **Select** menu in the **Extend Selection** submenu.

Once you have made a selection, you can copy it by going to the **Edit** menu and using the **Copy Selection** command. You can also cut the selection by using the **Cut Selection** command, also located on the **Edit** menu.

The **Paste** function will remove non-sequence characters. You can define a sequence character by going to the **Window** menu and then using the **Ambiguity Editor** from the **Ambiguity/Key Codes…** command.

For instructions on trimming ambiguous data and vectors at the time you load your sequences, see the chapters “Importing Data”, “Preparing Your Data for Assembly,, and “Customizing Sequencher and User Preferences” elsewhere in this manual.

SETTING THE BASE NUMBERING

Open the **Sequence Editor** and highlight a base. Go to the **Sequence** menu and choose **Set Base Number**. From the submenu choose either **As Base 1** or **As Base Number**. If you choose **As Base Number**, you will see a dialog where you can enter a new number for the selected base. Click **OK** to dismiss the dialog.

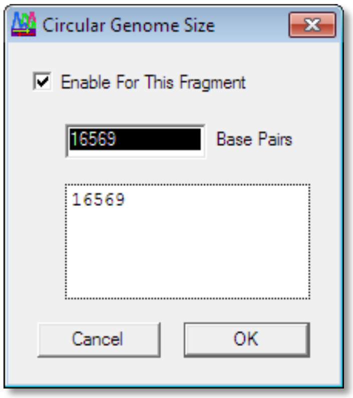
Note: If you set a base to be base 1, and this is not the first base in your sequence, every base before it will have a negative number.

SET CIRCULAR GENOME SIZE

Sometimes you will be working with circular DNA such as plasmids or even small genomes. The **Set Circular Genome Size…** command allows you to set the number of bases in your DNA circle.

Select a sequence that has already been defined as a Reference Sequence. Then go to the **Sequence** menu and choose the **Set Circular Genome Size…** command. The dialog in Figure 8-3 below will appear. Enable the circular number by clicking the checkbox. Then enter the number of base pairs in your sequence. Dismiss the dialog by clicking on the **OK** button.

Figure 8-3 The Circular Genome Size dialog



☑ Enable For This Fragment

16569 Base Pairs

16569

Note: Circular numbering can only be used in conjunction with a Reference Sequence. You must designate your sequence as a Reference Sequence before you can enable circular numbering. See Chapter 7 “The Reference Sequence” for more information.

SPLITTING A SEQUENCE

You can split a sequence into two pieces, for instance, at a known exon boundary. Put the cursor where you want to split the sequence. Go to the **Sequence** menu and choose **Split After Selection ….** **Sequencher** will ask you if you want to split the sequence after the last currently selected base. Click on the **Split** button to proceed. **Sequencher** will split the sequence into two pieces and label the new sequences for you. If your sequence was called “MyPUC”, it will be divided into two sequences called “MyPUC 5’ end” and “MyPUC 3' end.”

DUPLICATING A SEQUENCE

You can duplicate a sequence. To do this, highlight the icon, then go to the **Edit** menu and choose **Duplicate Seq Fragment**. **Sequencher** will use the name of the original sequence as a basis for the new sequence’s name. If the sequence you want to duplicate has a chromatogram associated with it, **Sequencher** will ask whether you also want to duplicate the chromatogram or duplicate the sequence without chromatograms. Click on the appropriate button to proceed.

If you wish to cancel the operation, simply press on the **Cancel** button. Chromatograms occupy a lot of memory but **Sequencher** uses a very efficient algorithm to compress trace data. However, we suggest that you only duplicate the text if you don’t need the chromatograms. (See Chapter 15 “Chromatograms” for more information.)

REVERSE AND COMPLEMENT

The data you bring into **Sequencher** may not necessarily be in a particular biologically relevant orientation. To change the orientation of the sequence, go to **View** menu and choose **Reverse & Comp** or **Sequenced Strand.** To display the original orientation, go to the **View** menu and choose **Sequenced Strand**.

Note: You will always have two copies of your data in **Sequencher**, the original data and the edited version. This duplication allows you to revert to the original data, as discussed in more detail below.

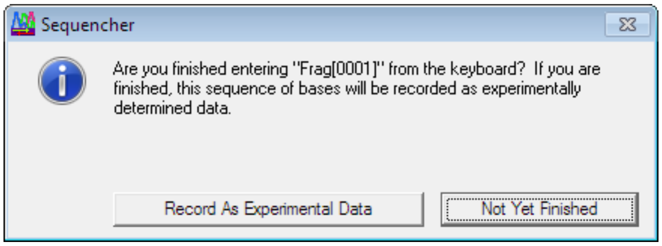
ABOUT EXPERIMENTAL DATA

CREATING A BASELINE

Sequencher notes the original base sequence when you import your data so that, after editing the sequence, you can still discard all of your edits and revert to the baseline (experimental) data. You can also do this for sequences that you enter from the keyboard.

When you close a **Sequence Editor** that has not had its data recorded as experimental data, the program explicitly asks whether the contents of the **Sequence Editor** should be recorded as experimental data (see Figure 8-4). Click on **Record As Experimental Data** if you are finished editing the sequence. If you are not finished, click on the **Not Yet Finished** button.

Figure 8-4 The Record as Experimental Data dialog



Are you finished entering "Frag[0001]" from the keyboard? If you are finished, this sequence of bases will be recorded as experimentally determined data.

Note: When a sequence becomes part of a contig, the data are automatically recorded as experimental data if they have not previously been stored in that form.

Once a version of the sequence has been recorded as experimental data, all future changes are considered to be “edits” to it. As a data quality tool, all such edits can be marked in color and displayed in **Bold Magenta** or lower-case text. This feature is invoked from the **View** menu under the **Base Edits As** command. The default is set to **Bold Magenta** but you can also choose **Not Highlighted** or **bOLD & cASE cHANGE.**

RESETTING A BASELINE

If you decide that some of the original data was incorrect (perhaps a simple typing error that has been corrected in the “edited” version of the sequence), go to the “Experimental Data” version of the sequence by clicking on the **Show Experimental** button and then click on the **New Baseline** button.

Note: Setting a new baseline for a sequence resets all of the bases to a not-yet-edited state. None of the bases that were edited will appear in **Bold Magenta** letters.

VIEWING EXPERIMENTAL DATA

After you enter a new sequence, **Sequencher** archives a copy that is “locked” (not editable). In the course of assembling your data, you may make a great number of edits to any particular sequence. This locked version enables you always to refer back to the sequence as it was originally entered. To see the Experimental Data version of the sequence, click on the **Show Experimental** button in the **Sequence Editor** button bar.

REVERTING TO EXPERIMENTAL DATA

If you edit a sequence incorrectly, you can delete all the changes to that sequence and go back to the archived baseline. To do so, open the **Sequence Editor**. Then choose **Revert To Experimental Data** from the **Sequence** menu.

If you edit a contig incorrectly, you can delete all the changes to an individual sequence and go back to the archived baseline. To do so, open the **Contig Editor** and select the sequence by clicking on its icon which can be found on the left-hand side of the **Contig Editor**. You can now remove the sequence from the contig by using the **Remove Selected Sequences…** from the **Contig** menu. Then choose **Revert To Experimental Data** from the **Sequence** menu. You can then reselect the sequences and rebuild the contig.

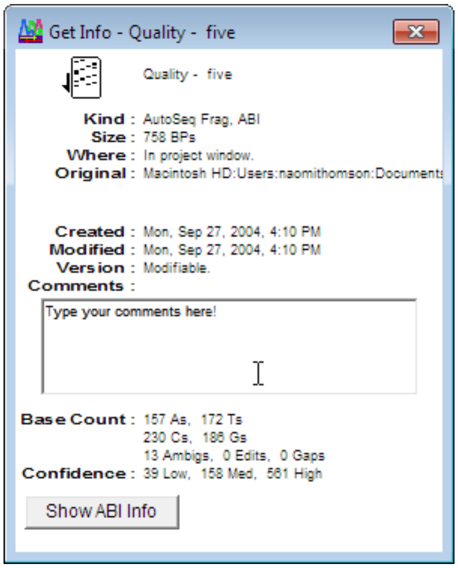
VIEWING EXPERIMENTAL DATA FOR CHROMATOGRAMS

If your data came from an automated sequencer, click the button labeled **Show Chromatogram** to view the experimental data of a sequence. This will display the original trace data. **Sequencher** allows you to scroll either vertically (the default) or horizontally. To change the scroll orientation, click the appropriate button in the left bottom corner of the sequence **Chromatogram** window. (See Chapter 15 "Chromatograms" for more information.)

VIEWING SUMMARY INFORMATION IN THE SEQUENCE EDITOR

You can bring up summary information on your data from the **Sequence Editor**. Go to the **File** menu and choose **Get Info…**. The **Get Info…** window (in Figure 8-5) provides summary information on the sequence you are currently editing.

Figure 8-5 Sequence information window



Get Info - Quality - five

Quality - five

Kind : AutoSeq Frag, ABI
Size : 758 BPs
Where : In project window.
Original : Macintosh HD:Users:naomithomson:Document

Created : Mon, Sep 27, 2004, 4:10 PM
Modified : Mon, Sep 27, 2004, 4:10 PM
Version : Modifiable.
Comments :

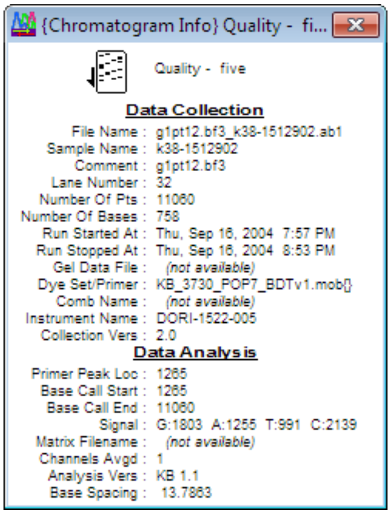
Type your comments here!

Base Count : 157 As, 172 Ts
230 Cs, 186 Gs
13 Ambigs, 0 Edits, 0 Gaps
Confidence : 39 Low, 158 Med, 561 High

If you want to annotate your data with more information than the comment box allows you to see, go to the **Edit** menu and click on **Edit Comments… Sequencher** will open a small text editing window that stores about 250 characters of text with each sequence. After you have added text with **Edit Comments**, a small "I" (for "Info") will appear at the upper left corner of the icon.

Base counts are displayed at the bottom of the window. If your sequence also includes confidence scores, you will also see a base count for each of the three confidence ranges. If you use an ABI sequencing system, click on the **Show ABI Info** button. **Sequencher** displays a window with information from the ABI data sheet, shown in Figure 8-6.

Figure 8-6 ABI data sheet info



Quality - five

Data Collection

File Name :	g1pt12.bf3_k38-1512902.ab1
Sample Name :	k38-1512902
Comment :	g1pt12.bf3
Lane Number :	32
Number Of Pts :	11060
Number Of Bases :	758
Run Started At :	Thu, Sep 16, 2004 7:57 PM
Run Stopped At :	Thu, Sep 16, 2004 8:53 PM
Gel Data File :	*(not available)*
Dye Set/Primer :	KB_3730_POP7_BDTv1.mob{}
Comb Name :	*(not available)*
Instrument Name :	DORI-1522-005
Collection Vers :	2.0

Data Analysis

Primer Peak Loc :	1265
Base Call Start :	1265
Base Call End :	11060
Signal :	G:1803 A:1255 T:991 C:2139
Matrix Filename :	*(not available)*
Channels Avgd :	1
Analysis Vers :	KB 1.1
Base Spacing :	13.7863

SELECTING BASES

Sequencher offers a number of ways to select bases. You can choose commands from the **Select** menu or you can use a keyboard combination. To select bases in the opposite orientation, press and hold the **Shift** key when using any **Select Next…** command (from the menu or with a keyboard shortcut).

FINDING AMBIGUOUS BASES

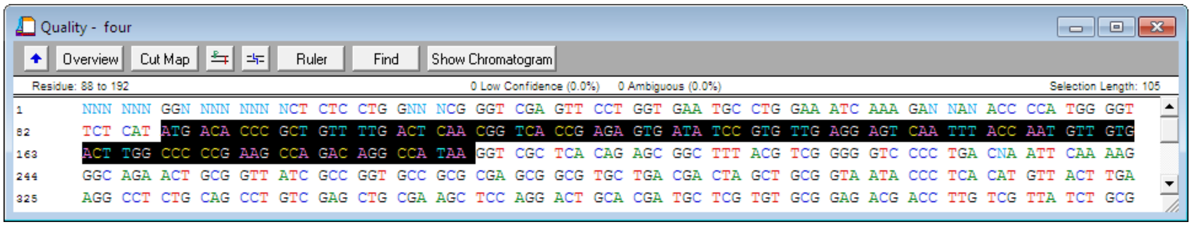
One of the most frequently executed commands is **Next Ambiguous Base** under the **Select** menu. You can use this command in the **Sequence Editor** or the **Contig Editor**. By using this command in the consensus sequence of a contig, **Sequencher** will display the next position in your contig that contains an ambiguity or disagreement. The shortcut for this command is **Ctrl+N** (Windows) or **Cmd+N** (Mac)**.** (For more on keyboard shortcuts, see Appendix 26 “Keyboard Shortcuts”.) If you use the **Shift** key with this key combination, you will go back to the previous disagreement.

Once you have used any of the **Select Next** commands, **Sequencher** will activate the space bar to execute that operation. If you use either the menu command or the shortcut key combination two consecutive times, **Sequencher** will remind you that this easier alternative is available.

FINDING OPEN READING FRAMES (ORF)

Choose **Next Met to Stop (> 0b)** to highlight the next pair of start and stop codons in one of the three forward reading frames. The number of bases shown in this command depends on whether you specified a preference for a minimum length in **User Preferences**. (For instructions on how to set a specified minimum length for open reading frames, see Chapter 23 “Customizing Sequencher and User Preferences”.) If you set preferences to highlight ORFs only when they are of some minimum length, the menu command will display the number of bases you specified.

Figure 8-7 ORF selected by Next Met to Stop command



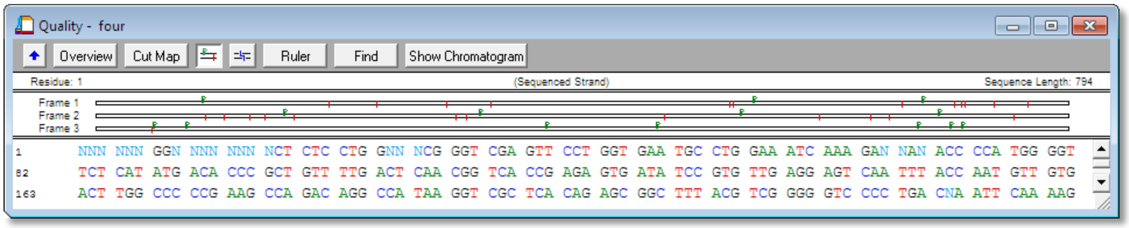
```
Residue: 88 to 192          0 Low Confidence (0.0%)   0 Ambiguous (0.0%)          Selection Length: 105
1      NNN NNN GGN NNN NNN NCT CTC CTG GNN NCG GGT CGA GTT CCT GGT GAA TGC CTG GAA ATC AAA GAN NAN ACC CCA TGG GGT
82     TCT CAT ATG ACA CCC GCT GTT TTG ACT CAA CGG TCA CCG AGA GTG ATA TCC GTG TTG AGG AGT CAA TTT ACC AAT GTT GTG
163    ACT TGG CCC CCG AAG CCA GAC AGG CCA TAA GGT CGC TCA CAG AGC GGC TTT ACG TCG GGG GTC CCC TGA CNA ATT CAA AAG
244    GGC AGA ACT GCG GTT ATC GCC GGT GCC GCG CGA GCG GCG TGC TGA CGA CTA GCT GCG GTA ATA CCC TCA CAT GTT ACT TGA
325    AGG CCT CTG CAG CCT GTC GAG CTG CGA AGC TCC AGG ACT GCA CGA TGC TCG TGT GCG GAG ACG ACC TTG TCG TTA TCT GCG
```

SEQUENCE FEATURES

LOOKING AT CODON MAPS

Click on the **Codon Map** button located in the button bar at the top of the **Sequence Editor** or go to the **View** menu and choose **Display Codon Map**. The three “outlined bars” at the top of the editor window represent the three forward frames of translation. Each green “flag” sticking partially into the top of a frame marks the position of a start codon, while each red line cutting through the frame and extending below marks the position of a stop codon. Clicking between any two markers highlights that section of the sequence.

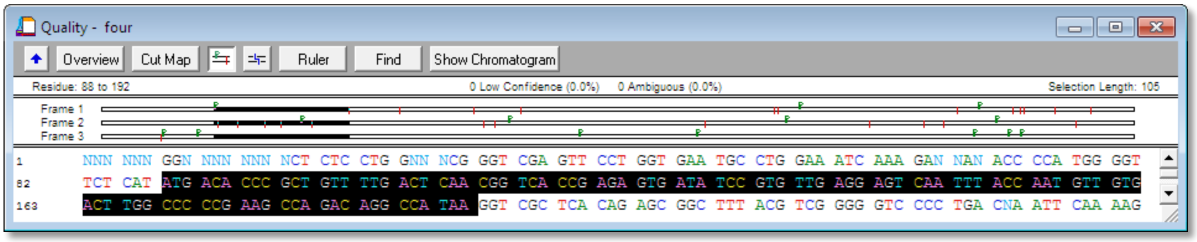
Figure 8-8 Sequence Editor with Codon Map button selected



```
Residue: 1                         (Sequenced Strand)                         Sequence Length: 794
1      NNN NNN GGN NNN NNN NCT CTC CTG GNN NCG GGT CGA GTT CCT GGT GAA TGC CTG GAA ATC AAA GAN NAN ACC CCA TGG GGT
82     TCT CAT ATG ACA CCC GCT GTT TTG ACT CAA CGG TCA CCG AGA GTG ATA TCC GTG TTG AGG AGT CAA TTT ACC AAT GTT GTG
163    ACT TGG CCC CCG AAG CCA GAC AGG CCA TAA GGT CGC TCA CAG AGC GGC TTT ACG TCG GGG GTC CCC TGA CNA ATT CAA AAG
```

In the example shown in Figure 8-9, a click between a start and a stop codon of the first reading frame has selected the corresponding bases in the editor. Note that **User Preferences** have been set to highlight ORFS of more than 100 bases. There is one such ORF in Frame 2, marked with a hash pattern. To dismiss the map, click on the **Codon Map** button again**.**

Figure 8-9 Sequence Editor with Codon Map



Residue: 88 to 192 0 Low Confidence (0.0%) 0 Ambiguous (0.0%) Selection Length: 105

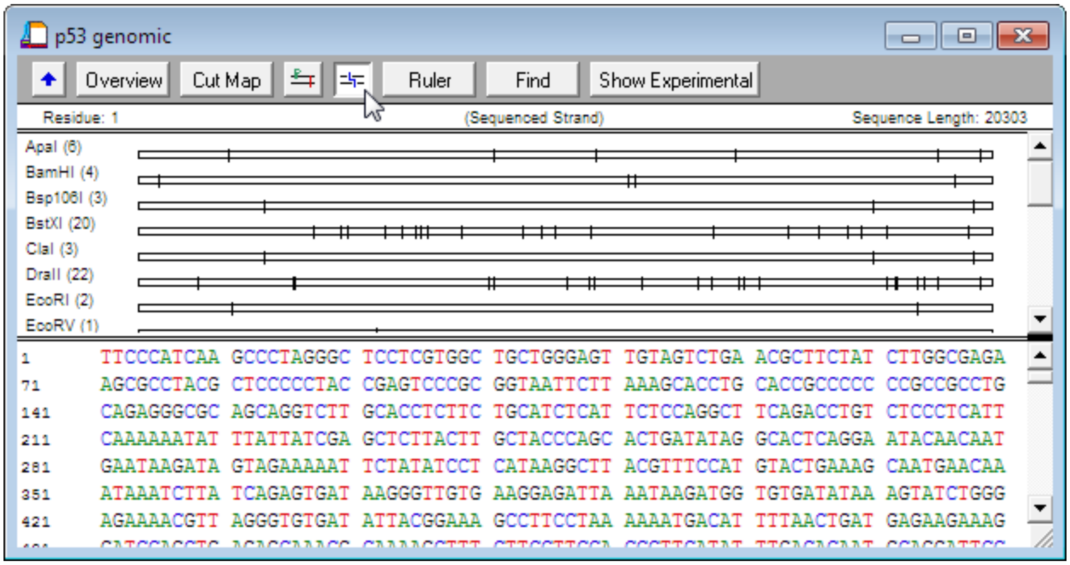
```
1     NNN NNN GGN NNN NNN NCT CTC CTG GNN NCG GGT CGA GTT CCT GGT GAA TGC CTG GAA ATC AAA GAN NAN ACC CCA TGG GGT
82    TCT CAT ATG ACA CCC GCT GTT TTG ACT CAA CGG TCA CCG AGA GTG ATA TCC GTG TTG AGG AGT CAA TTT ACC AAT GTT GTG
163   ACT TGG CCC CCG AAG CCA GAC AGG CCA TAA GGT CGC TCA CAG AGC GGC TTT ACG TCG GGG GTC CCC TGA CNA ATT CAA AAG
```

LOOKING AT RESTRICTION MAPS

You can look at a restriction map for the sequence by clicking on the **Cut Map** button at the top of the **Sequence Editor**. A restriction map replaces the sequence in the window. (See Chapter 17, “Restriction Maps,” for a general discussion of restriction maps.)

Sequencher can add a pane showing the restriction map for a sequence into the same window as the sequence itself. At the top of the **Sequence Editor** dialog is a button showing an icon for a staggered enzyme cut (shown in Figure 8-10).

Figure 8-10 Restriction Map button selected with map and sequence shown



Residue: 1 (Sequenced Strand) Sequence Length: 20303

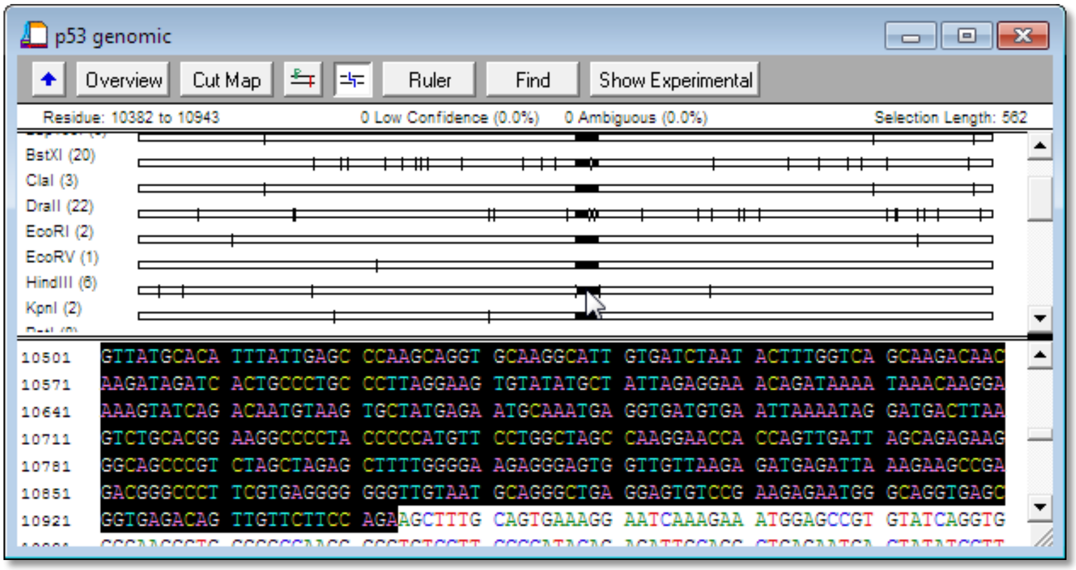
```
1     TTCCCATCAA GCCCTAGGGC TCCTCGTGGC TGCTGGGAGT TGTAGTCTGA ACGCTTCTAT CTTGGCGAGA
71    AGCGCCTACG CTCCCCCTAC CGAGTCCCGC GGTAATTCTT AAAGCACCTG CACCGCCCCC CCGCCGCCTG
141   CAGAGGGCGC AGCAGGTCTT GCACCTCTTC TGCATCTCAT TCTCCAGGCT TCAGACCTGT CTCCCTCATT
211   CAAAAAATAT TTATTATCGA GCTCTTACTT GCTACCCAGC ACTGATATAG GCACTCAGGA ATACAACAAT
281   GAATAAGATA GTAGAAAAAT TCTATATCCT CATAAGGCTT ACGTTTCCAT GTACTGAAAG CAATGAACAA
351   ATAAATCTTA TCAGAGTGAT AAGGGTTGTG AAGGAGATTA AATAAGATGG TGTGATATAA AGTATCTGGG
421   AGAAAACGTT AGGGTGTGAT ATTACGGAAA GCCTTCCTAA AAAATGACAT TTTAACTGAT GAGAAGAAAG
```

Click the button; a window with a multi-line restriction map appears. You can adjust the size of the panes by dragging the splitter bar that separates the scroll bars for each pane (see Figure

8-10). You can also invoke this command by going to the **View** menu and choosing **Display Cut Map Inset.**

Like the full-page cut map, the window restriction map is linked to the **Sequence Editor**. If you click between two cut sites, **Sequencher** highlights the resulting restriction sequence, as shown in Figure 8-11. If you click on the name of an enzyme, **Sequencher** highlights the recognition sites.

Figure 8-11 Marking restriction and recognition sites



```
Residue: 10382 to 10943     0 Low Confidence (0.0%)     0 Ambiguous (0.0%)     Selection Length: 562
```

```
10501  GTTATGCACA TTTATTGAGC CCAAGCAGGT GCAAGGCATT GTGATCTAAT ACTTTGGTCA GCAAGACAAC
10571  AAGATAGATC ACTGCCCTGC CCTTAGGAAG TGTATATGCT ATTAGAGGAA ACAGATAAAA TAAACAAGGA
10641  AAAGTATCAG ACAATGTAAG TGCTATGAGA ATGCAAATGA GGTGATGTGA ATTAAAATAG GATGACTTAA
10711  GTCTGCACGG AAGGCCCCTA CCCCCATGTT CCTGGCTAGC CAAGGAACCA CCAGTTGATT AGCAGAGAAG
10781  GGCAGCCCGT CTAGCTAGAG CTTTTGGGGA AGAGGGAGTG GTTGTTAAGA GATGAGATTA AAGAAGCCGA
10851  GACGGGCCCT TCGTGAGGGG GGGTTGTAAT GCAGGGCTGA GGAGTGTCCG AAGAGAATGG GCAGGTGAGC
10921  GGTGAGACAG TTGTTCTTCC AGAAGCTTTG CAGTGAAAGG AATCAAAGAA ATGGAGCCGT GTATCAGGTG
```

ANNOTATING SEQUENCES

You can assign a standardized GenBank **Feature Key** in addition to personal feature annotations to describe subsequences. Select the range of bases you wish to annotate, then choose **Mark Selection As Feature** from the **Sequence** menu.

If you want to assign a personal annotation, select **Sequencher** from the drop-down menu called **Feature Key:.** To give the feature a name, type your text into the **Feature Name:** box**.** You can use the default display style or you can assign a **Feature Color** and **Feature Style**, such as inverted case or underlined, from the dialog. You can also use the **Display:** radio buttons to determine whether single stranded DNA is displayed or RNA, protein translation, or the complement will be shown. (For more information, see Chapter 19 “Motifs and Features”.)

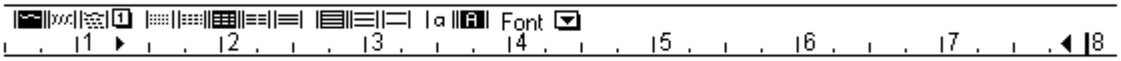
If you want to assign a GenBank feature key, make a selection from the drop-down menu called **Feature Key.** To give the feature a name, type your text into the **Feature Name:** input field**.** You can use the default display style or you can assign a **Feature Color** and **Feature Style**, such as inverted case or underlined, from the dialog. You can also use the **Display:** radio buttons to determine whether single stranded DNA is displayed or RNA, protein translation, or the complement. (For more information, see Chapter 19 “Motifs and Features”.)

CUSTOMIZING YOUR SEQUENCE VIEW

FORMATTING RULER

Formatting can be controlled with a ruler (shown in Figure 8-12) that lets you adjust several display parameters. Click on the **Ruler** button in the **Sequence Editor** or **Contig Summary** window. Alternately, you can go the **View** menu and choose **Display Format Ruler.**

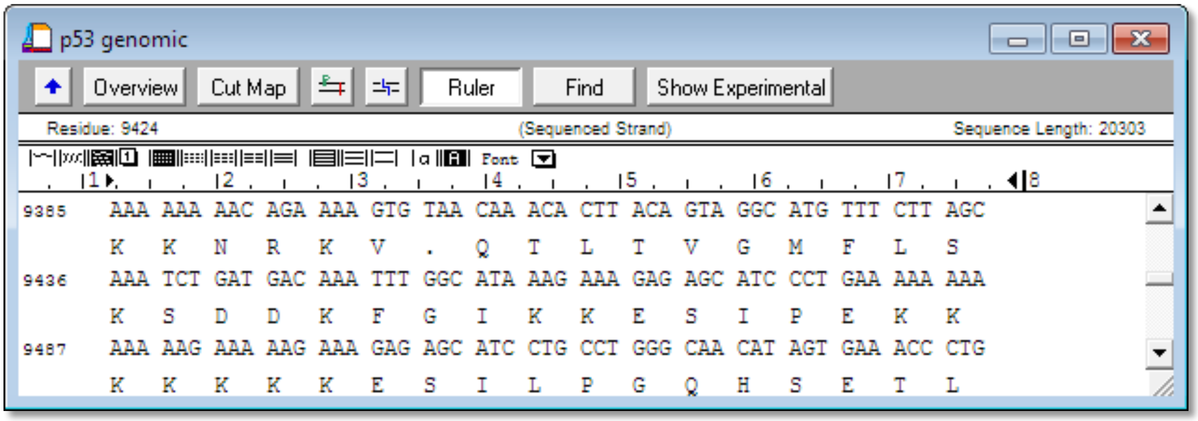
Figure 8-12 The formatting ruler



MARGINS

You can drag the margin triangles on this ruler to set the width of your sequence. Figure 8-13 shows margins that are limited to 42 bases per line.

Figure 8-13 Sequence Editor with ruler and translations



Residue: 9424 (Sequenced Strand) Sequence Length: 20303

```
9385   AAA AAA AAC AGA AAA GTG TAA CAA ACA CTT ACA GTA GGC ATG TTT CTT AGC
        K   K   N   R   K   V   .   Q   T   L   T   V   G   M   F   L   S
9436   AAA TCT GAT GAC AAA TTT GGC ATA AAG AAA GAG AGC ATC CCT GAA AAA AAA
        K   S   D   D   K   F   G   I   K   K   E   S   I   P   E   K   K
9487   AAA AAG AAA AAG AAA GAG AGC ATC CTG CCT GGG CAA CAT AGT GAA ACC CTG
        K   K   K   K   K   E   S   I   L   P   G   Q   H   S   E   T   L
```

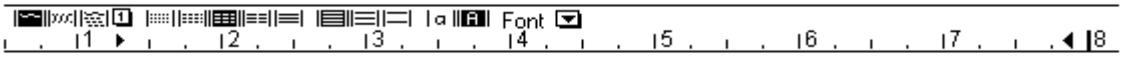
The sequence margins you set with the formatting ruler are not the same as your absolute print margins, which allow room at the left to show the first base number of a line. To change the absolute limits, go to the **File** menu and choose **Set Header & Footer…** and enter new values into the **Margin** boxes or, drag the triangular markers just under the Translation and Base Grouping icons, as shown in Figure 8-13 above.

TRANSLATION

On the left, just under the word “Residue” and above the ruler is a set of three small icons that let you specify sequence representation: single-stranded, double-stranded, or single-stranded with a protein translation.

If you select the third ‘translation’ icon, the numbered button at the right controls which frame or whether all three frames will be displayed. You can also use the translation commands in the **Translation** submenu of the **View** menu: **Single Stranded**, **Double Stranded**, **Protein 1st Frame, Protein 2nd Frame**, **Protein 3rd Frame**, **Protein All 3 Frames** (see Figure 8-14).

Figure 8-14 Translation icons on the Formatting ruler



See Chapter 23, “Customizing Sequencher and User Preferences”, for information on editing the genetic code or using one-letter abbreviations for amino acids.

BASE GROUPING

To the right of the translation display icons are the icons that let you choose blocking, the number of bases to be displayed in a single group. The available settings are **3**, **5**, **10**, **20** and **Fill-to-margin**. In Figure 8-14 above, the icon for a grouping of 10 is selected.

LINE SPACING

You can choose single, double, or triple line spacing. The three small icons that let you control the line spacing are located to the right of the base grouping icons. Figure 8-14 above shows the icon for single spacing selected.

CASE AND FONT

To the right of the line spacing icons is a pair of icons that let you choose upper or lower case for the display. In Figure 8-14 above, the icon for upper case is selected.

To the right of the upper and lower case icons is a pull down menu which lets you choose a font and size. Only fixed-width fonts are displayed, ensuring that the rows of bases will line up vertically.

VOICE VERIFICATION

Sequencher provides two forms of audio feedback to help you maintain the quality of data input when you are entering information manually from the keyboard. You can have **Sequencher** announce each base as you type it in. You can have the computer "read" the sequence aloud back to you while you look at the original source of the data. **Sequencher** stores the sounds that it uses in a sound file stored in your **Sequencher** folder.

SELECTING A VOICE FILE

First, make sure your computer sound system is enabled and that the volume is not set on mute. Then, go to the **Sequence** menu and choose **Speech**. Then go to the submenu **Select Voice File….**You will not get any audio feedback until the **Select Voice** file is loaded.

IDENTIFYING A SPEAKER

The voices Gene Codes uses to provide audio for you may be familiar, such as Walter Gilbert. To hear whose voice is being used, go to the **Sequence** menu and choose **Speech** and then the **Identify Speaker** sub-command. You can only select this command if a sound file was selected previously.

AUDIBLE KEYSTROKES

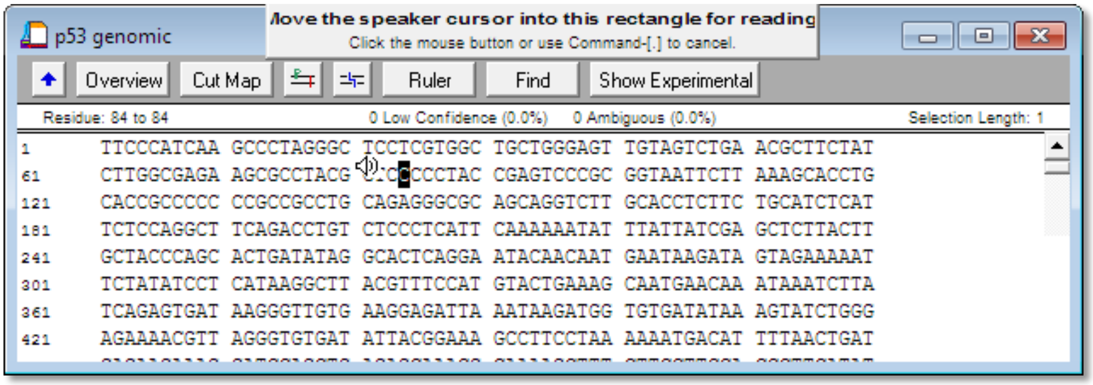
Any meaningful keystroke you type into a **Sequence Editor** dialog will be spoken aloud if you go to the **Speech** command in the **Sequence** menu and select the **Say Sequence Keystrokes** submenu. That is, if you press the "A" key, a voice will say "A," but if you press the shift key, you will not hear any voice.

READ SEQUENCE SELECTION

To check the sequence you have entered into the **Sequence Editor**, first select the bases you want the computer to read to you. Then go to the **Sequence** menu, click on **Speech** and then the **Read Sequence Selection** submenu.

Note: You must place the cursor somewhere in the small window that appears above the sequence as in Figure 8-15. To pause the reading, just move the cursor out of the window.

Figure 8-15 Speak selection cursor and instruction menu



/love the speaker cursor into this rectangle for reading
Click the mouse button or use Command-[.] to cancel.

Residue: 84 to 84   0 Low Confidence (0.0%)   0 Ambiguous (0.0%)   Selection Length: 1

```
1       TTCCCATCAA GCCCTAGGGC TCCTCGTGGC TGCTGGGAGT TGTAGTCTGA ACGCTTCTAT
61      CTTGGCGAGA AGCGCCTACG CTCCCCCTAC CGAGTCCCGC GGTAATTCTT AAAGCACCTG
121     CACCGCCCCC CCGCCGCCTG CAGAGGGCGC AGCAGGTCTT GCACCTCTTC TGCATCTCAT
181     TCTCCAGGCT TCAGACCTGT CTCCCTCATT CAAAAAATAT TTATTATCGA GCTCTTACTT
241     GCTACCCAGC ACTGATATAG GCACTCAGGA ATACAACAAT GAATAAGATA GTAGAAAAAT
301     TCTATATCCT CATAAGGCTT ACGTTTCCAT GTACTGAAAG CAATGAACAA ATAAATCTTA
361     TCAGAGTGAT AAGGGTTGTG AAGGAGATTA AATAAGATGG TGTGATATAA AGTATCTGGG
421     AGAAAACGTT AGGGTGTGAT ATTACGGAAA GCCTTCCTAA AAAATGACAT TTTAACTGAT
```

To speed up or slow down the rate at which Sequencher’s audio “reads” your data, go to the **Window** menu and choose **User Preferences...** Under the **General** section, you will find an item called **Sound**. Click on this to display a slider bar that controls the speed of read back in bases per second. To end the read back, just click your cursor anywhere on the **Sequence Editor** window.

9. SEQUENCE ASSEMBLY

In this chapter, you will learn how to set assembly conditions by using **Sequencher**'s rich variety of assembly algorithms and parameters. You will learn about automatic assembly, assembling interactively, assembling groups of sequences using sequence name information, the large gap algorithm for assembling cDNA and genomic DNA, and assembling to Reference Sequences. You will learn how to set assembly conditions for each of the different algorithms.

THE ASSEMBLY STRATEGY

Sequencher provides you with powerful options for performing your sequence assembly. In many cases, the default options will perform well. However, experimental variation can affect your data, so there will be times when you may need to change the assembly parameters. The following is a basic strategy for assembling your data.

Start with strict parameters to minimize the possibility of incorrect matches. Select the sequences you want to assemble and use the **Assemble Automatically** option. Automatic assembly relies on an exhaustive search-and-compare algorithm.

- If some sequences do not assemble, decrease the stringency of the parameters and try **Assemble Automatically** again.

- If the selected parameters become relaxed enough that the **Assemble Automatically** process may result in scientifically insupportable alignments, you should switch to interactive alignment so you can exert more control over the process.

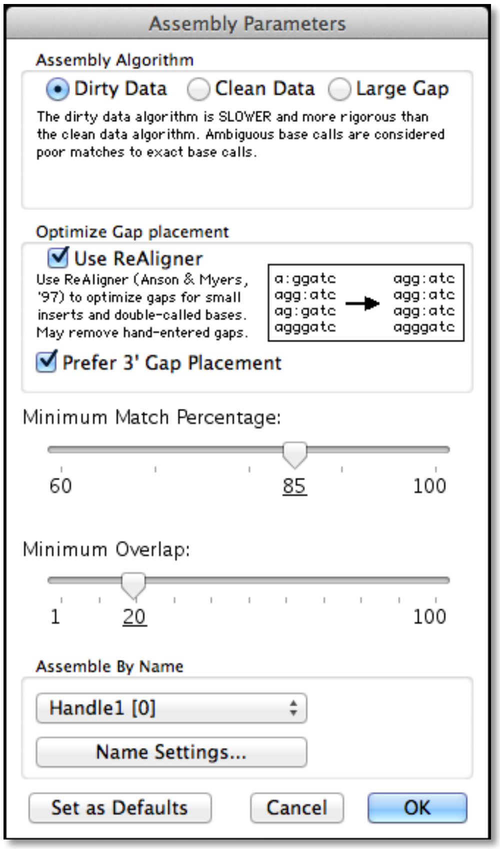
- The most frequently used options; **Assemble Automatically**, **Assemble Interactively**, and **Assemble to Reference** are available as prominent buttons in the **Project Window**. In addition, **Auto Assemble by Name** and **To Reference by Name** replace **Assemble Automatically** and **Assemble to Reference** when **Assemble by Name** is enabled. There are numerous other options under the **Assembly Parameters** and the **Assemble** menu.

SETTING THE ASSEMBLY CONDITIONS

SETTING ASSEMBLY PARAMETERS

- To change the parameter settings, select the **Project Window**. Then either click on the **Assembly Parameters** button in the **Project Window** or go to the **Assemble** menu and choose the **Assembly Parameters** command. **Sequencher** displays the dialog shown in Figure 9-1.

Figure 9-1 Assembly parameters dialog



Assembly Parameters

Assembly Algorithm

● Dirty Data ○ Clean Data ○ Large Gap

The dirty data algorithm is SLOWER and more rigorous than the clean data algorithm. Ambiguous base calls are considered poor matches to exact base calls.

Optimize Gap placement

☑ Use ReAligner

Use ReAligner (Anson & Myers, '97) to optimize gaps for small inserts and double-called bases. May remove hand-entered gaps.

```
a:ggatc          agg:atc
agg:atc   →      agg:atc
ag:gatc          agg:atc
agggatc          agggatc
```

☑ Prefer 3' Gap Placement

Minimum Match Percentage: 85

Minimum Overlap: 20

Assemble By Name

Handle1 [0]

Sequencher uses the values set in this window to control the way it assembles sequences. Once you have set the options you require click the **OK** button to dismiss the **Assembly Parameters** window.

ASSEMBLY ALGORITHMS

The algorithms in this section are devoted to working with traditional Sanger Sequencing. If you are interested in working with Next-Generation sequencing data, refer to Chapter 16 “NGS for DNA and RNA-Seq” for complete information.

To choose which assembly algorithm you want to use, click on the **Assembly Parameters** button on the **Project Window** and then click on one of the three radio buttons at the top of the window. Your choice will depend on your data.

For instance, if you click on the **Clean Data** radio button, contig assembly is faster but **Sequencher** may miss some possible matches if the ends of your sequences contain a large number of ambiguities. For best performance, make sure that you have trimmed poor quality data from your sequences.

The **Dirty Data** algorithm is somewhat slower because **Sequencher** is performing more rigorous comparisons between the sequences. You should note that a*utomated sequencers create dirty data* and **Sequencher's** algorithm is optimized to deal with that.

The **Large Gap** algorithm allows you to assemble sequences that are expected to contain insertions and deletions (gaps) of 10 or more bases long. Such gaps are typically found in evolutionary studies and sequence comparisons from different organisms. The **Large Gap** algorithm is similar to and slower than the dirty data algorithm but allows larger gaps to occur when performing the assembly. Typical examples of assemblies that require the **Large Gap** algorithm include the comparisons of a cDNA sequence to a genomic sequence and the assembly of related genes with alternative splicing.

ASSEMBLING WITH DIRTY DATA

Go to the **Assembly Parameters** dialog and select the **Dirty Data** radio button. There are two sliders for changing minimum match values. To set the *proportion* of bases that have to match, use the **Minimum Match Percentage** slider. To set the minimum number of bases that must overlap, use the **Minimum Overlap** slider.

Change the settings by positioning your cursor on the slider, holding the mouse button down, and dragging the slider left or right. The value set will automatically update as you move the slider.

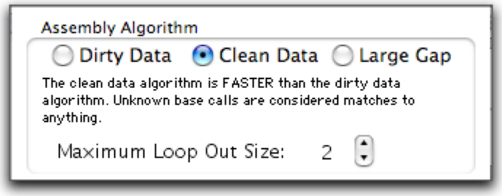
For example, if you attempt to assemble a 17-base sequence that matches perfectly to a selected 1000 base sequence with the minimum overlap set at 20 bases (**Sequencher**'s default minimum overlap), the two sequences will not assemble. You must first reduce minimum overlap to 17 bases or lower and then assemble the sequences.

ASSEMBLING WITH CLEAN DATA

When you select the **Clean Data** algorithm, an additional option called **Maximum Loop Out Size** is available. This has an elevator button for adjusting size value (Figure 9-2 below). This parameter is the largest number of *consecutively* mismatched bases that are acceptable in a potential overlap. Change the setting by positioning your cursor on the appropriate arrowhead, and holding the mouse button down. The maximum value, which can be set, is 6.

To set the *proportion* of bases that have to match, use the **Minimum Match Percentage** slider. To set the minimum number of bases that must overlap, use the **Minimum Overlap** slider.

Figure 9-2 The Clean Data algorithm with maximum Loop Out Size shown



Assembly Algorithm

○ Dirty Data ● Clean Data ○ Large Gap

The clean data algorithm is FASTER than the dirty data algorithm. Unknown base calls are considered matches to anything.

Maximum Loop Out Size: 2

ASSEMBLING WITH LARGE GAP

The **Large Gap** algorithm is specifically for data that has or might have large insertions or gaps. This might include data from evolutionary studies, different organisms or comparing cDNA to genomic DNA. When you select the **Large Gap** algorithm, there are two sliders for changing minimum match values. To set the *proportion* of bases that have to match use the **Minimum Match Percentage** slider. To set the minimum number of bases that must overlap use the **Minimum Overlap** slider.

Note: **Assemble to Reference** cannot be used with the **Large Gap** algorithm.

REFINING THE ASSEMBLY CONDITIONS

MINIMUM MATCH PERCENTAGE

The **Minimum Match Percentage** can be used with all three assembly algorithms. It is used to set the proportion of bases which must match in candidate sequences before **Sequencher** accepts the sequences as actually overlapping. The default value of **85%** can be changed by moving a slider.

If this value does not give you any overlaps, you may need to decrease the **Minimum Match Percentage** by moving the slider to the left to lower the percentage.

If you have an incorrect or unexpected overlap, you may need to increase the **Minimum Match Percentage** by moving the slider to the right (higher percentage).

MINIMUM OVERLAP

The **Minimum Overlap** can be used with all three assembly algorithms. It is used to set the minimum number of bases that must overlap before **Sequencher** accepts the sequences as actually overlapping. The default value of **20** can be changed by moving a slider. The default maximum value is **100** but this can be changed to **500** by clicking the checkbox in the **Contig User Preference** pane.

If this value does not give you any overlaps, you can decrease the **Minimum Overlap** by moving the slider to the left to indicate a smaller number of bases that must overlap**.**

If you have an incorrect or unexpected overlap, you may need to increase the **Minimum Overlap** by moving the slider to the right (larger number of bases that must overlap).

MAXIMUM LOOP OUT SIZE

The **Maximum Loop Out Size** is the largest number of consecutively mismatched bases (in this instance a gap also counts as a mismatch) that are acceptable in a potential overlap. The number of mismatching bases is set by moving an elevator button up or down. The maximum value that can be set manually is **6**. Remember that this option can be used only with the **Clean Data** algorithm.

OPTIMIZATION OF GAP PLACEMENT

ReAligner is an optional step which can be used with the **Clean Data** and **Dirty Data** algorithms. It evaluates the disposition of gaps within a contig and optimizes their placement**.** **ReAligner** facilitates editing in the consensus sequence and clearly displays the effect of insertions and deletions.

To use **ReAligner**, select the checkbox in the **Assembly Parameters** dialog. In some areas of DNA analysis, the standard is to gather the gaps to the right. If you require this, select **Prefer 3' Gap Placement**.

Note: This option cannot be used with the **Large Gap** algorithm.

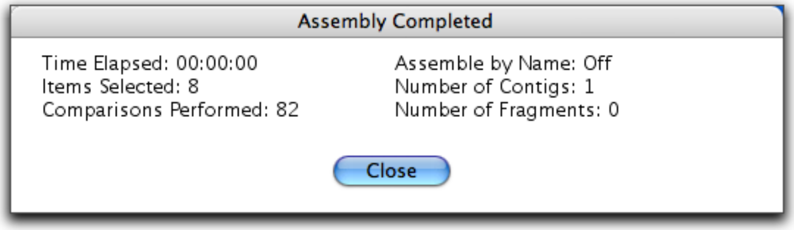
PERFORMING THE ASSEMBLY

AUTOMATIC ASSEMBLY

Bring the **Project Window** to the front and select the sequences you want to assemble. Click on the **Assemble Automatically** button or go to the **Assemble** menu and choose the **Automatically** command. **Sequencher** compares all the selected sequences, including reverse complements, and assembles the best matches that fall within the chosen assembly parameters.

When assembly is done, the advisory window appears. (See Figure 9-3.) Click **Close** to go back to the **Project Window**.

Figure 9-3 Assembly done alert



Assembly Completed

Time Elapsed: 00:00:00
Items Selected: 8
Comparisons Performed: 82

Assemble by Name: Off
Number of Contigs: 1
Number of Fragments: 0

ADDING SELECTED ITEMS TO OTHERS – INCREMENTAL BUILDING OF CONTIGS

The assembly option **Add Selected Items To Others**, which is a special case of automatic assembly, is very useful when you have a growing data set. **Sequencher** compares all the selected (new) items to the unselected (old) items. It will also compare the new items to each other. It has been designed to increase the efficiency of handling large numbers of sequences in efforts such as DNA library clustering and genome assembly. Therefore, it will not compare any unselected items to each other again.

Go to the **Assemble** menu and click on **Add Selected Items To Others**.

ASSEMBLE TO REFERENCE

The **Assemble to Reference** command allows you to assemble samples to a single Reference Sequence regardless of inconsistencies between the individual sequencess. Because it is a many-to-one comparison instead of the normal many-to-many comparison, **Assemble to Reference** is much faster than the standard **Assemble Automatically** algorithm.

To use this command, you need to have a Reference Sequence already designated. Click on your sequence and then go to the **Sequence** menu and select **Reference Sequence**. Now select the sequences you want to assemble, including the Reference Sequence. Click on the **Assemble to Reference** button or go to the **Assemble** menu and choose **Assemble to Reference**.

INTERACTIVE ASSEMBLY

The **Assemble Interactively** algorithm can be used whenever you want complete control over the assembly of a batch of sequences. The **Assemble Interactively** window provides you with detailed information regarding overlap, mismatch, and gapping for any given sequence and a set of candidate sequences. The candidate assemblies are driven by your user-defined assembly parameter settings.

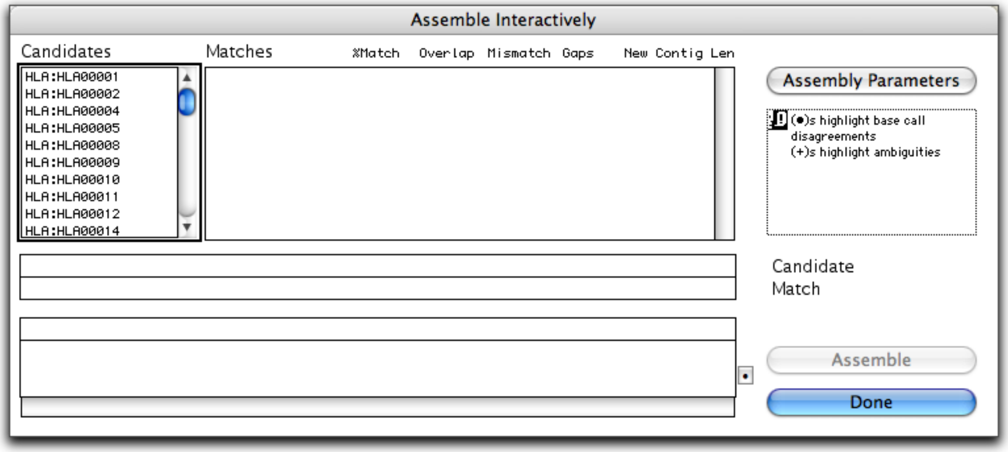
PERFORMING AN INTERACTIVE ASSEMBLY

Bring the **Project Window** to the front and select the items you want to assemble. Click on the **Assemble Interactively** button at the top of the **Project Window** or go to the **Assemble** menu and choose **Interactively…**.

Sequencher displays the **Assemble Interactively** dialog shown in Figure 9-4. The sequences and contigs you selected are displayed in the **Candidates** list. The panel on the right, showing a face in the upper left corner, is called the “agent.” It displays information about the comparisons you have requested. In the center is the **Matches** panel that displays the possible matches. For each match in the panel, **Sequencher** lists the **% Match**, the approximate number of **Overlap** bases, the number of **Mismatch** bases, the number of **Gaps**, and the **New Contig Length.**

The top line shows the Candidate sequence and the line below this shows the Match sequence. Below these two lines is the consensus sequence. There is a small button to the right of the consensus labeled with a black circle. If you click repeatedly on this button, you can display the first, second, or third reading frame or all three reading frames at once. The label on the button changes to reflect which of these choices is active.

Figure 9-4 Assemble Interactively dialog



Assemble Interactively

Candidates

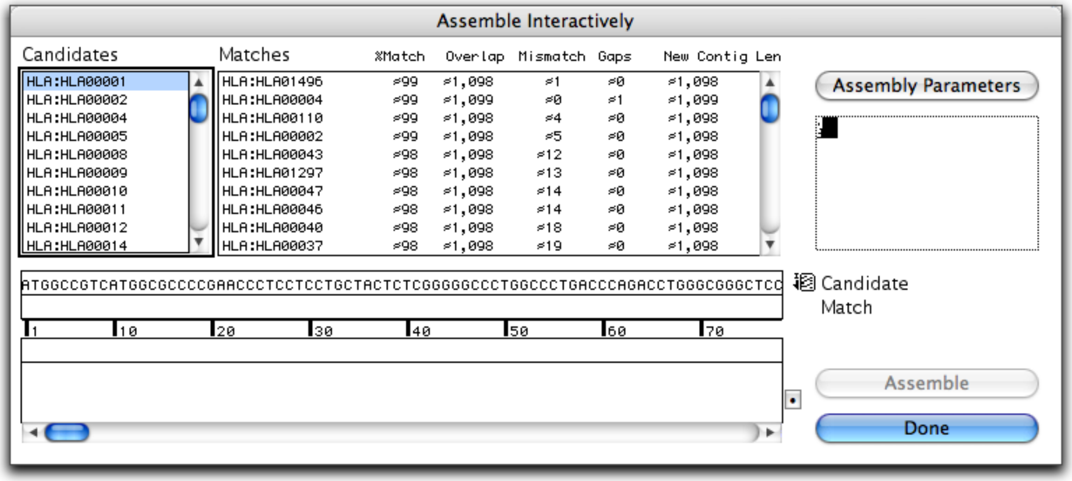
HLA:HLA00001
HLA:HLA00002
HLA:HLA00004
HLA:HLA00005
HLA:HLA00008
HLA:HLA00009
HLA:HLA00010
HLA:HLA00011
HLA:HLA00012
HLA:HLA00014

Matches	%Match	Overlap	Mismatch	Gaps	New Contig Len

(●)s highlight base call disagreements
(+)s highlight ambiguities

To have **Sequencher** compare any single sequence or contig with the other items in the **Candidates** list, click on the name of that sequence or the contig in the **Candidates** list. When the comparison is complete, the possible matches appear in the **Matches** list to the right of the **Candidates** list, as shown in Figure 9-5.

Figure 9-5 Interactive assembly showing a candidate and its matches



Assemble Interactively

Candidates

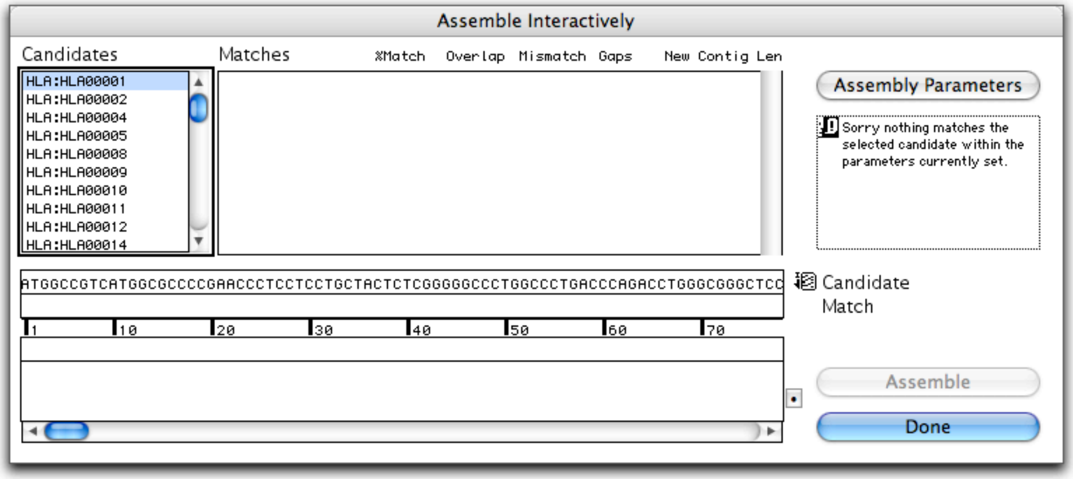
HLA:HLA00001
HLA:HLA00002
HLA:HLA00004
HLA:HLA00005
HLA:HLA00008
HLA:HLA00009
HLA:HLA00010
HLA:HLA00011
HLA:HLA00012
HLA:HLA00014

Matches	%Match	Overlap	Mismatch	Gaps	New Contig Len
HLA:HLA01496	≠99	≠1,098	≠1	≠0	≠1,098
HLA:HLA00004	≠99	≠1,099	≠0	≠1	≠1,099
HLA:HLA00110	≠99	≠1,098	≠4	≠0	≠1,098
HLA:HLA00002	≠99	≠1,098	≠5	≠0	≠1,098
HLA:HLA00043	≠98	≠1,098	≠12	≠0	≠1,098
HLA:HLA01297	≠98	≠1,098	≠13	≠0	≠1,098
HLA:HLA00047	≠98	≠1,098	≠14	≠0	≠1,098
HLA:HLA00046	≠98	≠1,098	≠14	≠0	≠1,098
HLA:HLA00040	≠98	≠1,098	≠18	≠0	≠1,098
HLA:HLA00037	≠98	≠1,098	≠19	≠0	≠1,098

```
ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCC
1        10        20        30        40        50        60        70
```

If **Sequencher** fails to find a match, as in Figure 9-6, the agent panel displays a message to that effect.

Figure 9-6 Match not found



Assemble Interactively

Candidates

HLA:HLA00001
HLA:HLA00002
HLA:HLA00004
HLA:HLA00005
HLA:HLA00008
HLA:HLA00009
HLA:HLA00010
HLA:HLA00011
HLA:HLA00012
HLA:HLA00014

Matches	%Match	Overlap	Mismatch	Gaps	New Contig Len

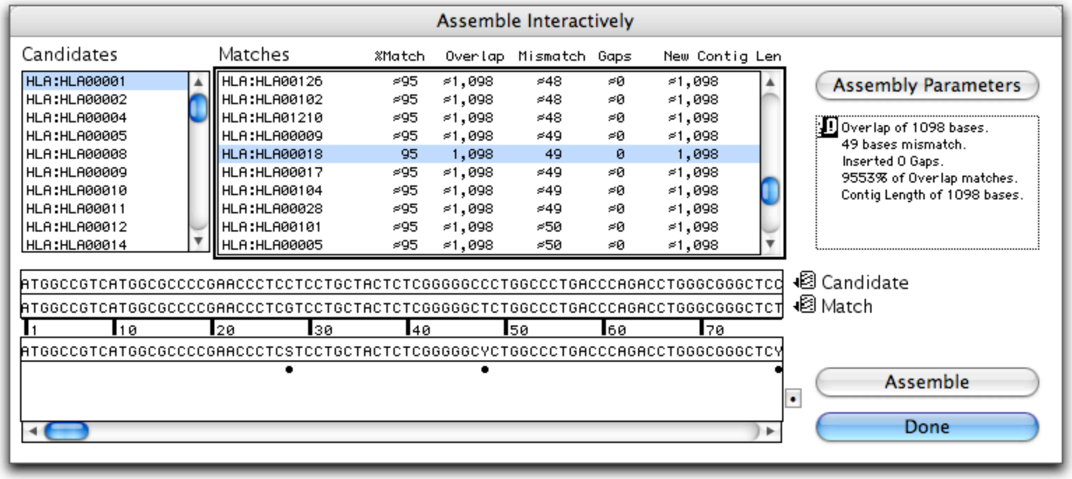
Sorry nothing matches the selected candidate within the parameters currently set.

```
ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCC
```

When you click a possible match in the **Matches** list, **Sequencher** recalculates the optimal positioning for gaps. The agent then displays a message about the selected sequences and their computed overlap.

The start of an overlap and the base consensus appear in the fields at the bottom of the interactive assembly box (Figure 9-7). You can explore the alignment further by moving the scroll bar to the right.

Figure 9-7 Interactive Assembly window showing overlap



Assemble Interactively

Candidates

HLA:HLA00001
HLA:HLA00002
HLA:HLA00004
HLA:HLA00005
HLA:HLA00008
HLA:HLA00009
HLA:HLA00010
HLA:HLA00011
HLA:HLA00012
HLA:HLA00014

Matches	%Match	Overlap	Mismatch	Gaps	New Contig Len
HLA:HLA00126	≈95	≈1,098	≈48	≈0	≈1,098
HLA:HLA00102	≈95	≈1,098	≈48	≈0	≈1,098
HLA:HLA01210	≈95	≈1,098	≈48	≈0	≈1,098
HLA:HLA00009	≈95	≈1,098	≈49	≈0	≈1,098
HLA:HLA00018	95	1,098	49	0	1,098
HLA:HLA00017	≈95	≈1,098	≈49	≈0	≈1,098
HLA:HLA00104	≈95	≈1,098	≈49	≈0	≈1,098
HLA:HLA00028	≈95	≈1,098	≈49	≈0	≈1,098
HLA:HLA00101	≈95	≈1,098	≈50	≈0	≈1,098
HLA:HLA00005	≈95	≈1,098	≈50	≈0	≈1,098

Overlap of 1098 bases.
49 bases mismatch.
Inserted 0 Gaps.
9553% of Overlap matches.
Contig Length of 1098 bases.

```
ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCC
ATGGCCGTCATGGCGCCCCGAACCCTCGTCCTGCTACTCTCGGGGGCTCTGGCCCTGACCCAGACCTGGGCGGGCTCT
ATGGCCGTCATGGCGCCCCGAACCCTCSTCCTGCTACTCTCGGGGGCYCTGGCCCTGACCCAGACCTGGGCGGGCTCY
```

If you choose to create this assembly, click on the **Assemble** button. Name the new contig in the **Set Name** dialog and click **OK**. **Sequencher** takes only a short time to assemble (or reassemble) a contig and the name of the new contig now appears at the top of the **Candidates** list. Note that as a sequence is assembled, its name is removed from the **Candidates** list.

The new contig will appear in the **Project Window** when you close the **Assemble Interactively** dialog by clicking on the **Done** button.

CHANGING PARAMETERS FOR AN INTERACTIVE ASSEMBLY

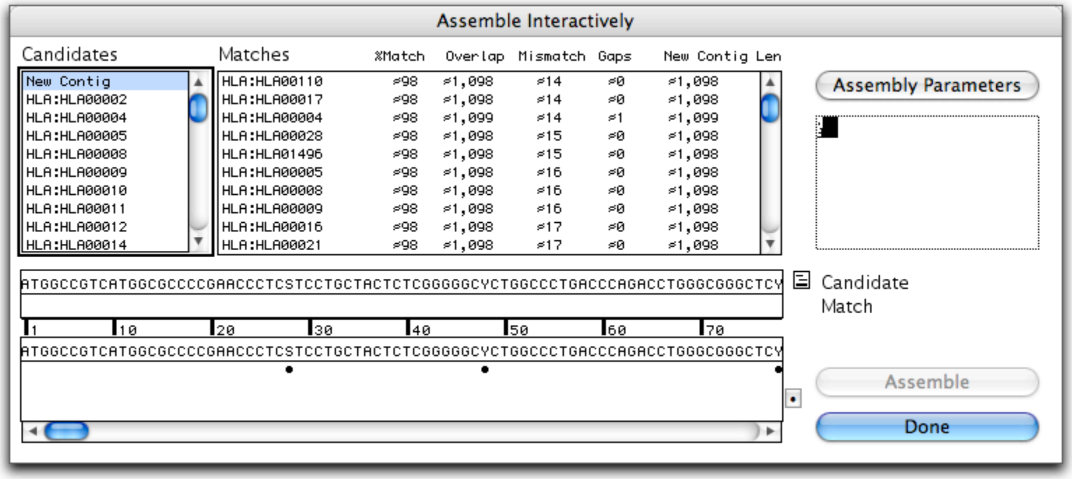
If you suspect that your sequences should form a contig but **Sequencher** cannot suggest candidate matches, you can change the parameters for assembly. To do this while using **Assemble Interactively**, click on the **Assembly Parameters** button. The **Assembly Parameters** dialog will appear. When you have changed the settings to your satisfaction, click **OK** to return to the **Assemble Interactively** window.

MINDLESSLY JOIN

Mindlessly Join allows you to put sequences together without using the conventional sets of algorithms. Examples of situations where you might wish to do this include making a plasmid or constructing an artificial cDNA sequence.

Select the sequences you want to join by dragging a box around them or **Shift**-**clicking** them.

Figure 9-8 The new contig displayed in the Assemble Interactively Window



Assemble Interactively

Candidates

New Contig
HLA:HLA00002
HLA:HLA00004
HLA:HLA00005
HLA:HLA00008
HLA:HLA00009
HLA:HLA00010
HLA:HLA00011
HLA:HLA00012
HLA:HLA00014

Matches	%Match	Overlap	Mismatch	Gaps	New Contig Len
HLA:HLA00110	≠98	≠1,098	≠14	≠0	≠1,098
HLA:HLA00017	≠98	≠1,098	≠14	≠0	≠1,098
HLA:HLA00004	≠98	≠1,099	≠14	≠1	≠1,099
HLA:HLA00028	≠98	≠1,098	≠15	≠0	≠1,098
HLA:HLA01496	≠98	≠1,098	≠15	≠0	≠1,098
HLA:HLA00005	≠98	≠1,098	≠16	≠0	≠1,098
HLA:HLA00008	≠98	≠1,098	≠16	≠0	≠1,098
HLA:HLA00009	≠98	≠1,098	≠16	≠0	≠1,098
HLA:HLA00016	≠98	≠1,098	≠17	≠0	≠1,098
HLA:HLA00021	≠98	≠1,098	≠17	≠0	≠1,098

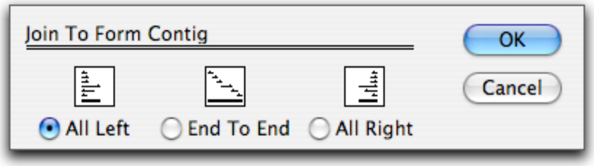
```
ATGGCCGTCATGGCGCCCCGAACCCTCSTCCTGCTACTCTCGGGGGCYCTGGCCCTGACCCAGACCTGGGCGGGCTCY
1        10        20        30        40        50        60        70
ATGGCCGTCATGGCGCCCCGAACCCTCSTCCTGCTACTCTCGGGGGCYCTGGCCCTGACCCAGACCTGGGCGGGCTCY
```

Note: They will be listed in the new contig in the order in which you selected them.

Go to the **Assemble** menu and then select **Mindlessly Join**. If **Assemble by Name** is enabled, the **Mindlessly Join** submenu item becomes **Join by Name**. **Join by Name** uses handles to group the sequences appropriately.

Sequencher will then ask if you want all the sequences aligned at the left of the contig (**All Left**), at the right of the contig (**All Right**), or joined **End to End**. Click the appropriate button. The selected items will be assembled and replaced by a contig icon in the **Project Window**.

Figure 9-9 Join by Name dialog



MUSCLE

Unlike the algorithms **Clean Data**, **Dirty Data**, and **Large Gap**, **MUSCLE** is not built into **Sequencher** but is an external tool. More information on adding **MUSCLE** to **Sequencher** can be found in the **DNA-Seq Tools Installation** pdf on the Gene Codes website **Support** page. **MUSCLE** is a multiple-sequence alignment algorithm and it will attempt to align the sequences over their entire length. This can introduce a large number of gaps into your aligned sequences, so you should be aware that it may not be the best algorithm to use for all purposes.

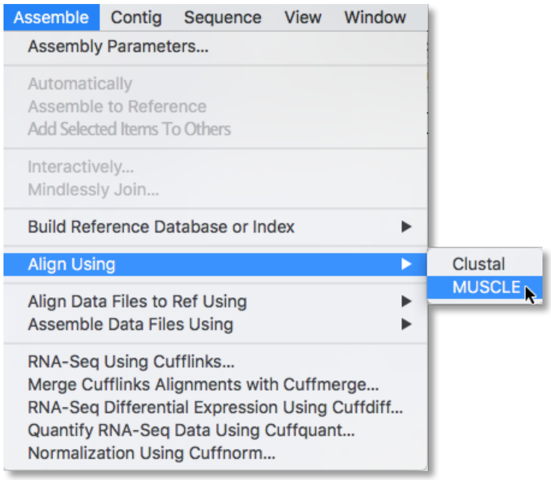
USING MUSCLE TO ALIGN SEQUENCES

Select the sequences you wish to align from the **Project Window**. Then go to the **Assemble** menu, select **Align Using**, and click on **MUSCLE** from the submenu. A new dialog appears which tells you that **MUSCLE** is running. When the alignment is completed, this dialog is dismissed.

You will see a new contig in your **Project Window** which you can analyze with **Sequencher's** other tools such as the **Variance Table**.

Note: You can read more about **MUSCLE** in this paper. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004, 5:113.

Figure 9-10 Align Using MUSCLE


Assemble
Contig
Sequence
View
Window
Assembly Parameters...
Automatically
Assemble to Reference
Add Selected Items To Others
Interactively...
Mindlessly Join...
Build Reference Database or Index
Align Using
Align Data Files to Ref Using
Assemble Data Files Using
RNA-Seq Using Cufflinks...
Merge Cufflinks Alignments with Cuffmerge...
RNA-Seq Differential Expression Using Cuffdiff...
Quantify RNA-Seq Data Using Cuffquant...
Normalization Using Cuffnorm...
Clustal
MUSCLE

10. ASSEMBLE BY NAME

Many labs apply strict sample-naming conventions to their sequences, with the name indicating something about the data. **Assemble by Name** harnesses the information from your sequence names to automate the process of selecting and naming your contigs.

In this chapter, you will learn how to use **Assemble by Name**, a powerful tool that lets you separately assemble and automatically name multiple contigs from a single selection and assembly command. For example, in just a few clicks of the mouse, **Assemble by Name** will create ninety-six appropriately named contigs from ninety-six forward and reverse sequences.

THE ASSEMBLE BY NAME STRATEGY

The names of your samples frequently contain information such as your primer, template, clone name, or sample source. **Sequencher** provides you with tools to separate the descriptive parts of your sequence name and use these to manage the assembly. These tools are the **Assembly Handles** and the **Name Delimiters**.

An **Assembly Handle** is any set of characters from your sequence name that provides information about that sequence, such as the primer or clone name. A **Name Delimiter** separates each **Assembly Handle** from the following one. Sequence names frequently have several **Assembly Handles** and **Name Delimiters**.

In its simplest form, a **Name Delimiter** may be a character such as a dash or an underscore. When you use a **Name Delimiter**, it will look like this:

handle1-handle2-handle3

or

handle1_handle2_handle3

However, a **Name Delimiter** can also be a combination of letters, numbers, and characters.

To define your **Name Delimiter**, **Sequencher** provides a set of commonly used characters as a drop-down menu in the **Assemble by Name Settings** dialog. You will see this menu if you go to **Assembly Parameters** dialog and click on the **Name Settings…** button. When your naming convention does not use simple characters as **Name Delimiters**, you can use formal descriptors known as "regular expressions" instead, which we describe later in this chapter.

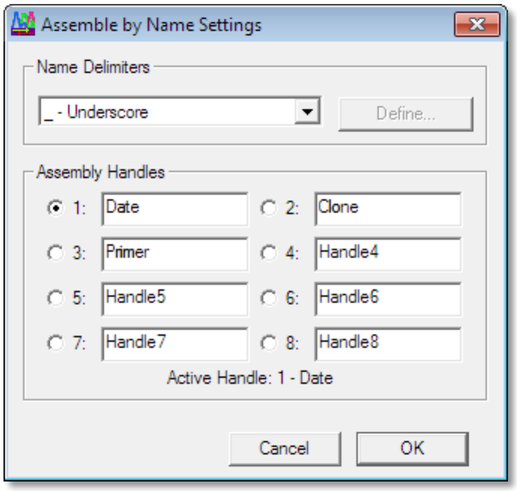
Once you have defined an **Assembly Handle**, you will be able to assemble all of the sequences in your project. Only sequences that share the same **Assembly Handle** will be analyzed for assembly into the same contig. In a clinical application, for example, several contigs might be created with each contig representing sequence data from only a single patient, provided that each sample (patient) has a unique **Assembly Handle** in its name.

SETTING THE ASSEMBLE BY NAME CONDITIONS

CONFIGURING ASSEMBLY HANDLES USING A SINGLE DELIMITER

Starting in the **Project Window**, click on the **Assembly Parameters** button. At the bottom of the **Assembly Parameters** dialog, click on the **Name Settings…** button. The **Assemble by Name Settings** dialog appears in Figure 10-1 below.

Figure 10-1 Assemble by Names Settings window



Name Delimiters: _ - Underscore

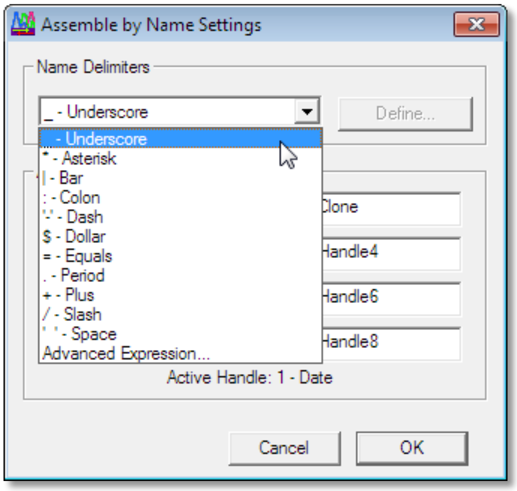
Assembly Handles

● 1:	Date	○ 2:	Clone
○ 3:	Primer	○ 4:	Handle4
○ 5:	Handle5	○ 6:	Handle6
○ 7:	Handle7	○ 8:	Handle8

Active Handle: 1 - Date

If your **Assembly Handles** are separated by a single character included in the **Name Delimiters** drop-down menu, choose that character from the list. Otherwise, follow the directions for an **Advanced Expression**.

Figure 10-2 The Name Delimiters drop-down menu



CONFIGURING ASSEMBLY HANDLES USING ADVANCED EXPRESSIONS

You will need to create an Assembly Handle using a regular expression if your Assembly Handle is not separated by one of the characters in the **Name Delimiters** drop-down menu.

WHAT IS A REGULAR EXPRESSION?

A regular expression is a way of describing a text pattern using letters, numbers, and special characters. These expressions follow certain syntactical rules.

Starting in the **Project Window**, click on the **Assembly Parameters** button. At the bottom of the **Assembly Parameters** dialog, click on the **Name Settings**... button. The **Assemble by Name Settings** dialog appears.

Choose **Advanced Expression**... from the **Name Delimiters** drop-down menu. You will notice that the **Define**... button is now enabled. Click on the **Define**... button.

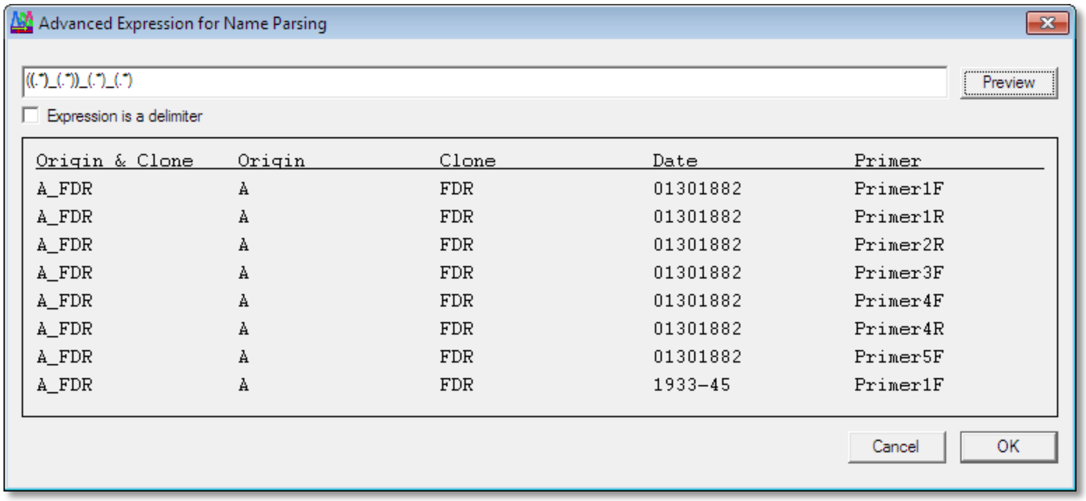
The **Advanced Expression for Name Parsing** dialog appears. At the top of the dialog is an input field where you can type in a regular expression. **Assemble by Name** uses regular expressions in two ways. When the **Expression is a delimiter** box is checked, the regular expression you type in lets you define a delimiter option other than one available from the drop-down menu.

If the **Expression is a delimiter** box is *not* checked, your regular expression must completely define each **Assembly Handle** and **Name Delimiter** in your sequence name. The regular expression you write will not work unless it describes the entire name of your sequences. When you type the regular expression into the text field, it should be in the form:

(**Assembly Handle1**)**Name Delimiter**(**Assembly Handle2**)

In the example in Figure 10-3 below, a regular expression is used to combine the first two parts of the sequence name, Origin and Clone, to create a new Assembly Handle, Origin&Clone. The **Preview** button displays the results of your regular expression. If you are satisfied, click on the **OK** button.

Figure 10-3 Advanced Expression for Name Parsing dialog



Advanced Expression for Name Parsing

((.*)_(.*))_(.*)_(.*)

□ Expression is a delimiter

Origin & Clone	Origin	Clone	Date	Primer
A_FDR	A	FDR	01301882	Primer1F
A_FDR	A	FDR	01301882	Primer1R
A_FDR	A	FDR	01301882	Primer2R
A_FDR	A	FDR	01301882	Primer3F
A_FDR	A	FDR	01301882	Primer4F
A_FDR	A	FDR	01301882	Primer4R
A_FDR	A	FDR	01301882	Primer5F
A_FDR	A	FDR	1933-45	Primer1F

For more details on regular expressions, see Appendix 27 “Advanced Expressions” and the tutorials in the Tutorials folder in your **Sequencher** installation folder.

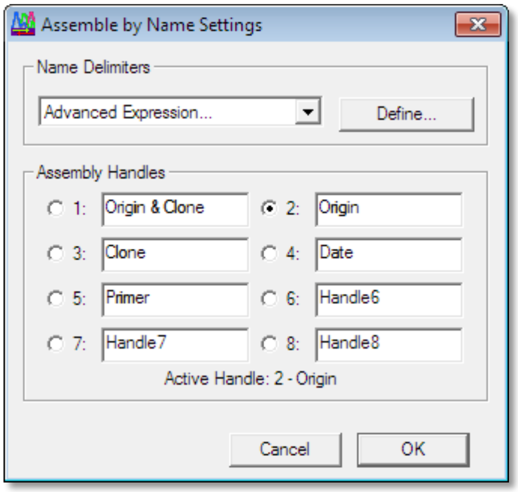
SETTING YOUR ASSEMBLY HANDLE NAMES

Starting in the **Project Window**, click on the **Assembly Parameters** button. At the bottom of the **Assembly Parameters** dialog, click on the **Name Settings…** button. The **Assemble by Name Settings** dialog appears as in Figure 10-4.

You can describe up to eight **Assembly Handles** of which only one is active at any one time.

For each **Assembly Handle**, you can use the default name **Sequencher** automatically gives your contigs or you can type a descriptive title into the input field. Click on the radio button to the left of the handle you wish to activate. **Sequencher** displays the number and the name of the **Active Handle** at the bottom of this dialog. Click on the **OK** button to dismiss this dialog and return to the **Assembly Parameters** dialog.

Figure 10-4 The Assemble by Name Setting window



Name Delimiters

Advanced Expression...

Assembly Handles

○ 1: Origin & Clone
● 2: Origin
○ 3: Clone
○ 4: Date
○ 5: Primer
○ 6: Handle6
○ 7: Handle7
○ 8: Handle8

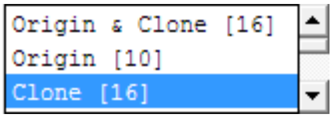
Active Handle: 2 - Origin

CHOOSING A NEW ASSEMBLY HANDLE

There are two ways to choose an **Assembly Handle**. If you are in the **Assembly Parameters** dialog, you can change the handle by choosing a new one from the drop-down menu.

Note: The probable number of contigs that can be formed using this handle is indicated in square brackets to the right of the handle.

Figure 10-5 Assembly Handles drop-down menu

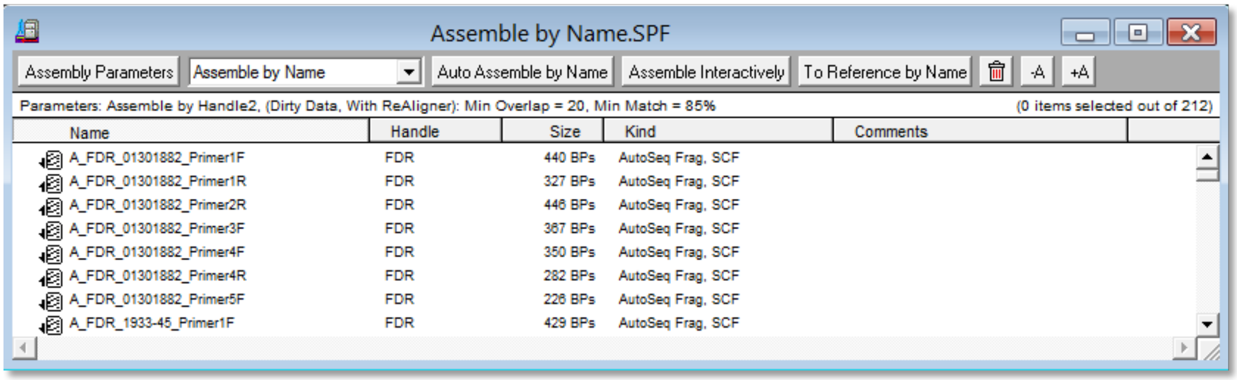


You can also click on the **Name Settings…** button and then click on the radio button next to your preferred handle.

ENABLING ASSEMBLE BY NAME

You enable **Assemble by Name** from the **Project Window** by clicking on the **Assembly** Modes drop-down menu in the button bar and choosing **Assemble by Name**. If you are in the List View, you will notice a new column titled **Handle**, which lists the active handle for each sequence. You will also see that the text of the assembly command buttons has changed to **Auto Assemble by Name** and **To Reference by Name**. The new assembly status is also reflected in the current parameters above the column headers.

Figure 10-6 The Project Window with Assemble by Name enabled



Parameters: Assemble by Handle2, (Dirty Data, With ReAligner): Min Overlap = 20, Min Match = 85% (0 items selected out of 212)

Name	Handle	Size	Kind	Comments
A_FDR_01301882_Primer1F	FDR	440 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer1R	FDR	327 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer2R	FDR	446 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer3F	FDR	367 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer4F	FDR	350 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer4R	FDR	282 BPs	AutoSeq Frag, SCF	
A_FDR_01301882_Primer5F	FDR	226 BPs	AutoSeq Frag, SCF	
A_FDR_1933-45_Primer1F	FDR	429 BPs	AutoSeq Frag, SCF	

SETTING ASSEMBLY PARAMETERS

Sequencher uses the values set in the **Assembly Parameters** dialog to control the way it assembles sequences. Once you have set your required options, you can store these as your default **Assembly Parameters**. Click on the **Set as Defaults** button at the bottom of the dialog and **Sequencher** will use your preferred **Assembly Parameters** whenever it launches. If you just click **OK** to dismiss the **Assembly Parameters** dialog, **Sequencher** will use your new parameters only with this project or session of **Sequencher**.

PERFORMING AN ASSEMBLY WITH ASSEMBLE BY NAME

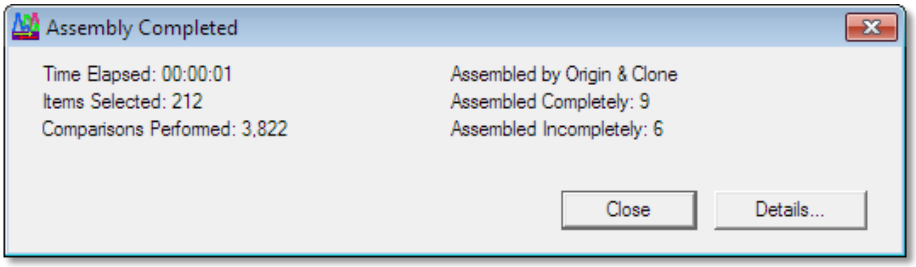
Once you have chosen your Name Delimiters and Assembly Handles, you are nearly ready to perform your assembly. In addition to the standard **Auto Assemble by Name** command, you can use the **Assemble by Name** function to assemble your sequences with a Reference Sequence or to Mindlessly Join your sequences.

Under the **Assemble** menu, the following menu items will change if you have enabled **Assemble by Name**. The **Mindlessly Join** command becomes **Join by Name**…. Similarly, the **Assemble to Reference** button will now be labeled **To Reference by Name**.

Before you use the **To Reference by Name** command, you will need to define a Reference Sequence. You will also need to choose and set your assembly algorithm and any associated parameters. (For more details, see Chapter 9 “Sequence assembly”.)

Starting from the **Project Window**, go to the **Select** menu and click on the **Select All** command to select all the sequences in your project. Click on the **Auto Assemble by Name** button, or the **To Reference by Name** button if you have a Reference Sequence. The **Assembly Preview** window will appear. This window contains the names of the contigs that may be generated using your chosen handle and the probable number of sequences in each contig. Click on the **Assemble** button to continue with the assembly, or click on the **Cancel** button. **Sequencher** displays the **Assembly Completed** dialog when the assembly has finished assembling your sequences.

Figure 10-7 The Assembly Completed dialog



Assembly Completed

Time Elapsed: 00:00:01
Items Selected: 212
Comparisons Performed: 3,822

Assembled by Origin & Clone
Assembled Completely: 9
Assembled Incompletely: 6

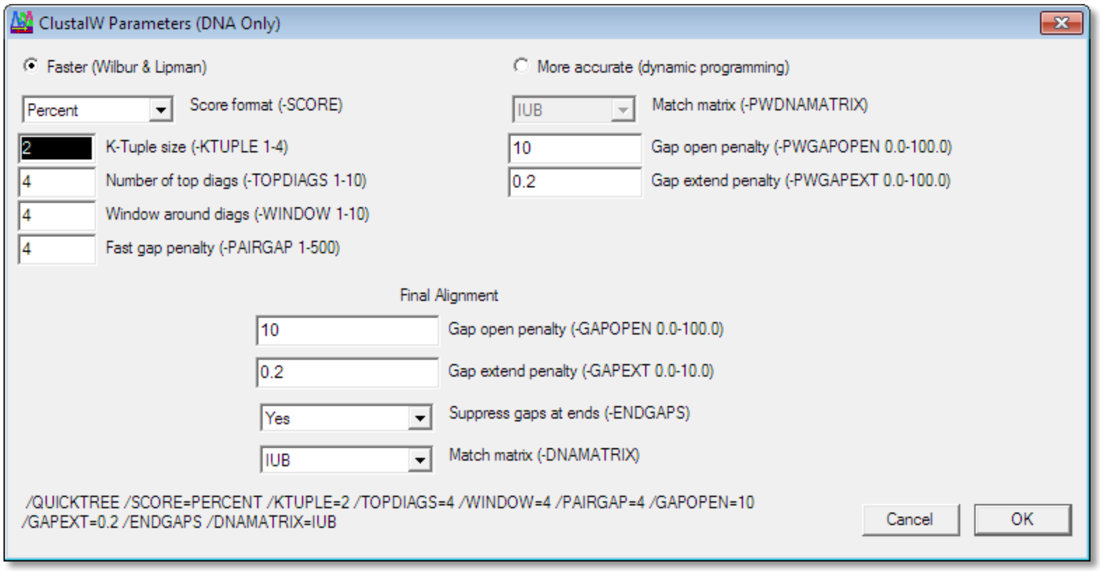
This dialog gives a summary of the assembly process. If this information is sufficient, click the **Close** button. This will dismiss the **Assembly Completed** window. If you require more detailed information, click on the **Details…** button. You will see a report that contains further information about the assembly. You can save a text copy of this report by clicking on the **Save As** button. Dismiss the dialog by clicking on the **Close** button.

Note: The name of each contig formed using **Assemble by Name** will be based on the **Assembly Handle** used.

ALIGNING SEQUENCES WITH CLUSTAL

Unlike the algorithms **Clean Data**, **Dirty Data**, and **Large Gap**, **Clustal** is not built into **Sequencher** but is an external tool. More information on adding **Clustal** to **Sequencher** can be found in the **DNA-Seq Tools Installation** pdf on the Gene Codes website **Support** page.

Figure 10-8 Clustal Options dialog



ClustalW Parameters (DNA Only)

● Faster (Wilbur & Lipman)

Percent — Score format (-SCORE)
2 — K-Tuple size (-KTUPLE 1-4)
4 — Number of top diags (-TOPDIAGS 1-10)
4 — Window around diags (-WINDOW 1-10)
4 — Fast gap penalty (-PAIRGAP 1-500)

○ More accurate (dynamic programming)

IUB — Match matrix (-PWDNAMATRIX)
10 — Gap open penalty (-PWGAPOPEN 0.0-100.0)
0.2 — Gap extend penalty (-PWGAPEXT 0.0-100.0)

Final Alignment

10 — Gap open penalty (-GAPOPEN 0.0-100.0)
0.2 — Gap extend penalty (-GAPEXT 0.0-10.0)
Yes — Suppress gaps at ends (-ENDGAPS)
IUB — Match matrix (-DNAMATRIX)

/QUICKTREE /SCORE=PERCENT /KTUPLE=2 /TOPDIAGS=4 /WINDOW=4 /PAIRGAP=4 /GAPOPEN=10
/GAPEXT=0.2 /ENDGAPS /DNAMATRIX=IUB

USING CLUSTAL TO ALIGN SEQUENCES

Select the sequences you wish to align from the **Project Window**. Then go to the **Assemble** menu, select **Align Using**, and click on **Clustal** from the submenu.

Click on either the **Wilbur & Lipman** radio button or the **dynamic programming** radio button. Now set any other parameters by changing the values in the input fields or choosing from the appropriate drop-down menu. At the bottom of the dialog, you will see a series of values and parameter names, these are the settings that will be sent to **Clustal**. If you are not sure about the meaning or effect of any parameter, then use the default value.

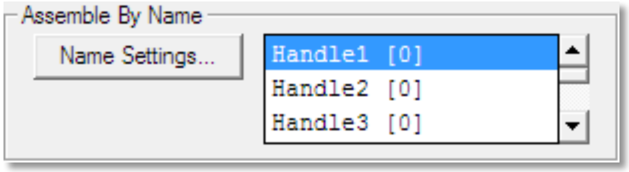
To start the alignment, click on the **OK** button. A new window appears which asks you to be patient while the alignment takes place. The results will appear as a single contig in your **Project Window**.

USING CLUSTAL WITH ASSEMBLE BY NAME

You can also use **Clustal** with **Assemble by Name**. Typically you would use **Assemble by Name** when you have multiple sequences from different sources, taking this approach will save you time but works best when you have a well standardised naming scheme for your sequences.

Click in the **Assembly Mode** drop-down menu on the button bar and choose **Assemble by Name**. Then click on the **Assembly Parameters** button on the button bar at the top of the **Project Window** and select the **Name Settings…** button.

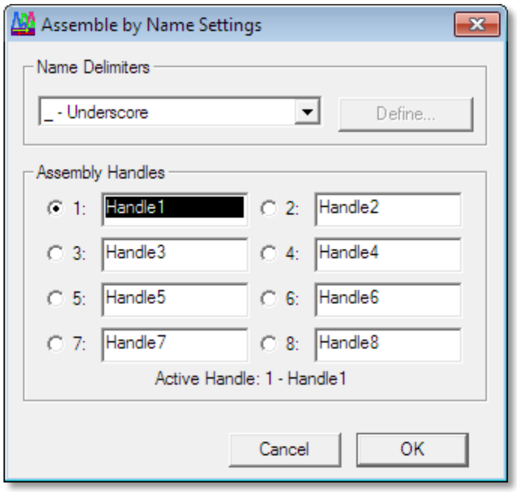
Figure 10-9 Assemble by Name and Name Settings…



Handle1 [0]
Handle2 [0]
Handle3 [0]

Next you will need to set your **Assemble by Name** parameters by choosing the **Name Delimiter** from the drop-down menu and the **Assembly Handle** by clicking on the appropriate radio button that represents the portion of the sequence name to be used for assembling the contigs.

Figure 10-10 Assembly Handles with Name Delimiter chosen


Assemble by Name Settings

Name Delimiters

_ - Underscore

Assembly Handles

● 1: Handle1 ○ 2: Handle2

○ 3: Handle3 ○ 4: Handle4

○ 5: Handle5 ○ 6: Handle6

○ 7: Handle7 ○ 8: Handle8

Active Handle: 1 - Handle1

Once these have been set and the **Assemble by Name** function has been enabled, you can then proceed with the alignment.

Select the sequences you wish to align from the **Project Window**. Then go to the **Assemble** menu, select **Align by Name Using**, and click on **Clustal** from the submenu. A new window appears which asks you to be patient while the alignment takes place. When the alignment is completed, you will see the **Alignment Completed** dialog. Click on the **OK** button to dismiss this window.

You will see new contigs appear in your **Project Window** which you can analyse with **Sequencher's** other tools such as the **Variance Table**.

11. THE CONTIG EDITOR

In this chapter, you will learn about the overview and bases view of the **Contig Editor**. You will learn how to use these options to explore your contig. We will discuss the steps to take before you start editing: base numbering, setting a consensus calculation, and assigning a Reference Sequence.

CONTIG EDITING

DNA sequencing is an error-prone process, and so the base calls from an automated DNA sequencer may not correctly represent the sample being sequenced. This means that when working with automated DNA sequencing data, it is usually necessary to check and edit your contigs.

Sequencher provides a powerful interface between the user and the data so that you can analyze and edit sequences and contigs according to your own specifications. You can move directly from one ambiguity or disagreement to the next to analyze and override base calling errors, eliminating discrepancies and ambiguities. **Sequencher** allows you to choose from a number of consensus calculations and continually recalculates the consensus of the contig as you edit the sequences.

OVERVIEW OF THE CONTIG

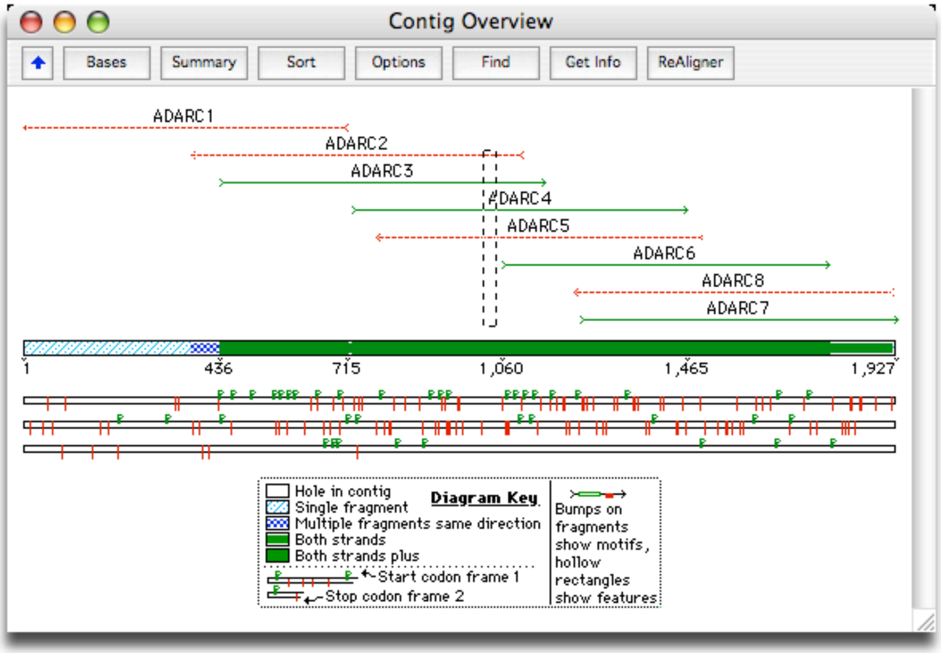
THE CONTIG OVERVIEW WINDOW

When you first open a contig by double-clicking on it, you will see an overview that shows how the sequences have been assembled (Figure 11-1). There is also a button bar at the top of the **Overview** window.

The schematic provides you with a wealth of useful information enabling you to assess and manage your data. A horizontal line represents each sequence in the contig. The line will be green and solid if the sequence is in the forward orientation and red and dotted if it is in the reverse orientation. These orientations may change as more data is added to the contig, unless you have used a Reference Sequence in your assembly. In this case, the Reference Sequence determines the contig orientation. Below the sequences, you will see the coverage map and below the coverage map you will see the start and stop map. Notice the diagram key at the base of the window.

If your sequence(s) had a GenBank style feature table, then the features would automatically be applied on import into **Sequencher**. You would see these features marked in the **Overview**.

Figure 11-1 Contig Overview



SELECTION MARQUEE

In the **Overview**, you will notice a dashed rectangle. This animated marquee (sometimes called the marching ants) highlights the section of the contig displayed when you switch to the **Bases View**. You can change the location you want to view by dragging the marquee to a new location. As you move the cursor over the marquee, you will see it change to a hand icon. You can now hold down the mouse button and drag the marquee to the desired location.

CHECKING COVERAGE IN THE OVERVIEW

The bar across the bottom of the **Overview** shows various levels of sequence coverage indicated by different graphic patterns. For instance, a wide solid green line indicates that the sequence coverage comprises both strands. A region with no color or pattern indicates a hole in the coverage. The full array of patterns indicating coverage are shown at the bottom of the **Overview** in the figure above.

In the example in the figure above, the **Overview** shows the contig starting out with a single, unsupported sequence. Then a second sequence starts to overlap with the first. Next there is an area with "full" coverage: both strands are sequenced with additional confirming sequences. At the extreme right end of the contig, there are just two sequences in opposite directions.

OVERVIEW FEATURES

BASES

To get from the **Overview** to the view used for editing, click on the **Bases** button. You can also go to the **View** menu and choose **Bases** from the **Bases**, **Map**, **Overview**, … submenu. **Sequencher** displays the **Contig Editing** window.

SUMMARY

You can press the **Summary** button to view a compact report of the contig. To see this report, go to the **View** menu, choose the **Bases, Map, Overview, …** submenu, and click on the **Summary Report** option. The features of the **Summary Report** will be dealt with in more detail later (see Chapter 12 “Editing Contigs” for more information).

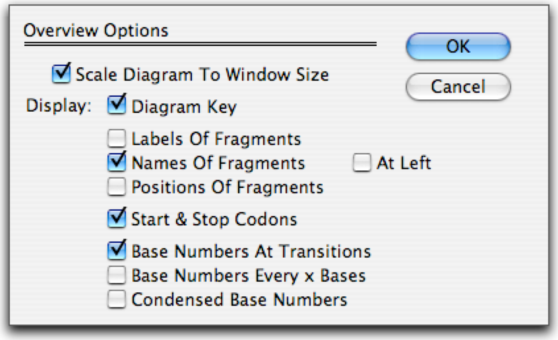
SORT

You can sort the sequences in a contig vertically. Click on the **Sort** button to display a dialog of options. Choose an option by clicking on its radio button and then pressing **OK.** You may also use the sorting options in the **View** menu under the **Sort/Cleanup** submenu.

OVERVIEW OPTIONS

To choose the display options, click on the **Options** button or go to the **View** menu and click on View **Options**. An **Overview Options** dialog will appear where you can set your choices.

Figure 11-2 Contig Overview options dialog



Overview Options

☑ Scale Diagram To Window Size

Display: ☑ Diagram Key

□ Labels Of Fragments

☑ Names Of Fragments □ At Left

□ Positions Of Fragments

☑ Start & Stop Codons

☑ Base Numbers At Transitions

□ Base Numbers Every x Bases

□ Condensed Base Numbers

SCALE DIAGRAM TO WINDOW SIZE

Selecting **Scale Diagram To Window Size** shrinks the whole overview to fit in the window, even if the names of the individual sequences become unreadable. When you turn this option off, you make the names of the sequences readable, but you may have to scroll the window to see all the sequences in the diagram.

DIAGRAM KEY

Clicking this checkbox displays or hides a key that explains the **Overview**.

LABELS OF FRAGMENTS

You must have already defined and applied some labels to your sequences with **User Preferences** (see Chapter 23, “Customizing Sequencher and User Preferences”) and the **Label** command. Clicking **Labels of Fragments** will display or hide any fragment label names.

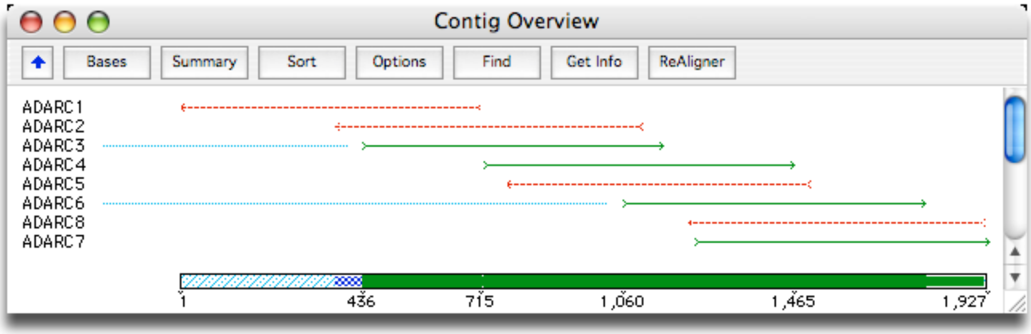
NAMES OF FRAGMENTS

Clicking this checkbox displays or hides the sequence names in the **Overview.** You may have this and the previous item checked at the same time.

NAMES OF FRAGMENTS, AT LEFT

Clicking the **At Left** box in conjunction with **Names of Fragments** lists the fragment names down the left side of the window instead of on top of each arrow (as shown in Figure 11-3). If, for example, the sequences are also sorted by position, a faint blue line will connect the name with its fragment arrow.

Figure 11-3 Names Of Fragments At Left



POSITIONS OF FRAGMENTS

Clicking **Positions of Fragments** displays the numerical range positions of the sequences in the **Overview**. If you do not wish to see this information, leave the box unchecked.

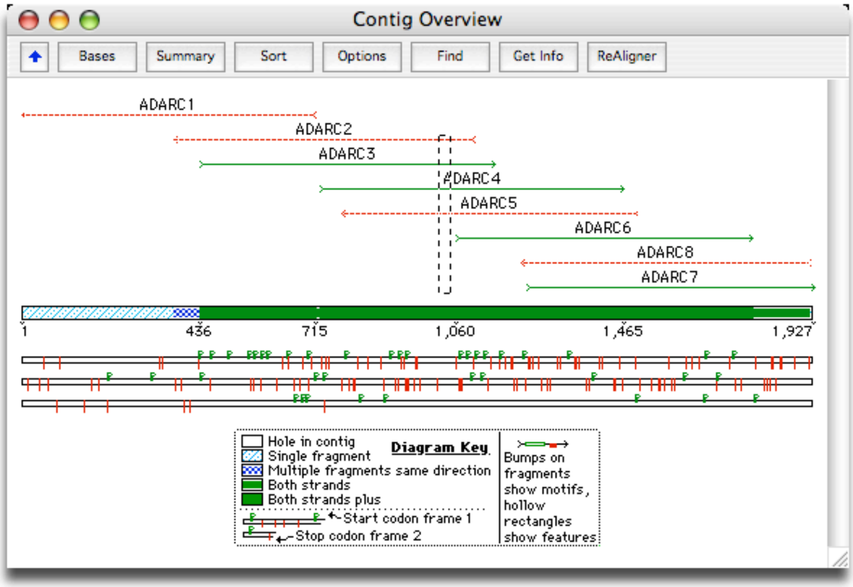
START & STOP CODONS

You can display a start and stop **Codon Map** showing all the start and stop codons in the three reading frames of one strand. Start codons are indicated by green flags and stop codons are denoted by vertical red lines. To display this map, check the **Start & Stop Codons** box. The map will be positioned at the bottom in the **Overview**.

In Figure 11-4, the three-frame map of start and stop codons is located just above the diagram key.

To see start and stop codons in the other strand, go the **View** menu and choose **Reverse & Comp**. This will reverse and complement your sequences. Choosing **Reverse & Comp** again will restore your contig's original orientation.

Figure 11-4 Start and Stop Codon Map



BASE NUMBERS AT TRANSITIONS

Clicking **Base Numbers at Transitions** displays or hides the numbers showing approximate positions of the coverage transitions.

BASE NUMBERS AT EVERY X BASES

Selecting Base Numbers Every x Bases shows the base numbers at constant intervals.

CONDENSED BASE NUMBERS

Selecting the Condensed Base Numbers option changes the font of the base numbers to a compressed version. This option can be used with either Base Numbers at Transitions or Base Numbers at Every x Bases.

CONTIG COVERAGE GRAPH

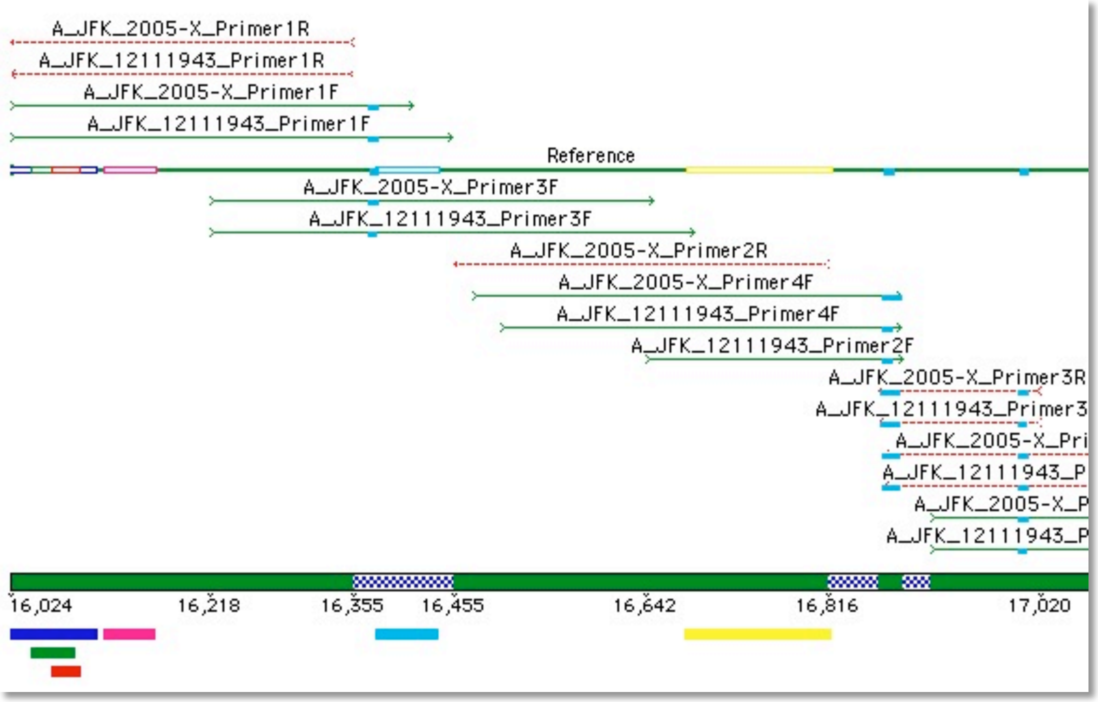
As well as the standard **Overview** you can see a representation of your contig in graph form. This is especially useful when you are dealing with large numbers of sequences. To see the **Coverage Graph**, go to the **View** menu and choose **Bases, Map, Overview, …** and select the **Coverage Graph** command.

GRAPHIC FEATURE MAPS

When you annotate your sequence with features or if you have imported a sequence containing a GenBank Feature Table, **Sequencher** will present a graphical representation of the features under the display of overviews and restriction maps.

The graphical feature map has two parts. Above the sequences (red/green lines) the features are presented on one line. Below the coverage map the feature may be presented over several lines. You can see this at the five prime end of the sequence in the image below, where there are three features in blue, red and green which overlap. In the bottom map the features are displayed on separate lines. If you allow the mouse cursor to stay over one of these features a tool tip appears, containing information about the feature.

Figure 11-5 Graphic Feature Maps on Overview



GETTING MORE INFORMATION

FIND

You can search for a specific subsequence in your contig. Fragments, which contain the subsequence, will appear with a heavier line in the **Overview**.

You can choose to find the first occurrence by clicking on the **Find** button. You can search for subsequent occurrences using the **Find Next** button. Three radio buttons allow you to control the specificity of the search by choosing **Exact Matches, Specified Bases** (which would allow you to use ambiguity codes such as W to find T or A), or **Any Ambiguous Base**.

Use the **Any Ambiguous Base** command to find any degenerate DNA that may match. There is also a drop-down menu from which you may choose restriction enzyme sequences.

If you have used a Reference Sequence and want to include it in the search, select the **Include Reference Sequence in Find** checkbox. Close the dialog by pressing **Done.**

GET INFO

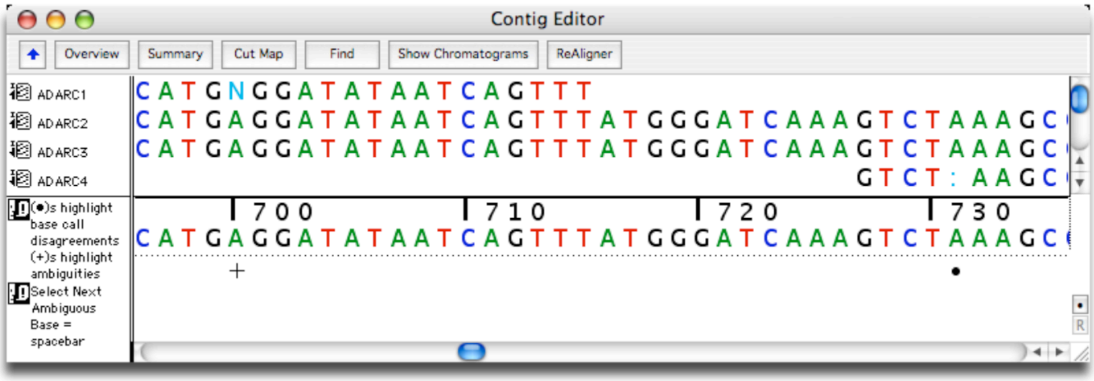
Summary information about the open contig can be obtained by pressing the **Get Info** button or go to the **File** menu and click on the **Get Info** command.

THE BASES VIEW

WORKING IN THE BASES VIEW

To get from the **Overview** to the view used for editing a contig, click on the **Bases** button. Alternatively, you can go to the **View** menu and then go to the **Bases, Map, Overview…** submenu and choose **Bases**. **Sequencher** displays the Contig Editing window shown in Figure 11-6.

Figure 11-6 The Contig Editing window


Contig Editor
Overview
Summary
Cut Map
Find
Show Chromatograms
ReAligner
(●)s highlight base call disagreements
(+)s highlight ambiguities
Select Next Ambiguous Base = spacebar
R

```
ADARC1  CATGNGGATATAATCAGTTT
ADARC2  CATGAGGATATAATCAGTTTATGGGATCAAAGTCTAAAGC
ADARC3  CATGAGGATATAATCAGTTTATGGGATCAAAGTCTAAAGC
ADARC4                                 GTCT:AAGC
           700       710       720       730
        CATGAGGATATAATCAGTTTATGGGATCAAAGTCTAAAGC
            +                                ●
```

To return to the **Overview**, click on the **Overview** button or go to the **Bases**, **Map**, **Overview**… submenu and choose **Overview**.

The labels in Figure 11-7 identify the different parts of the **Contig Editor**.

Figure 11-7 Identifying parts of the contig editor

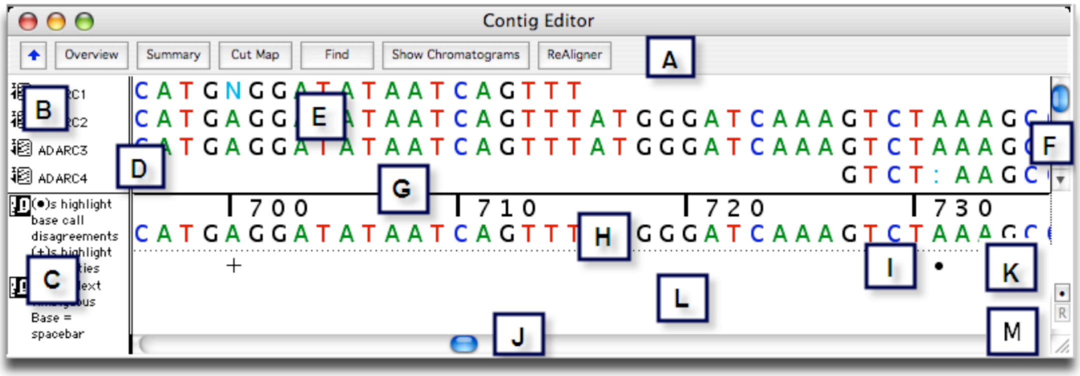


Table 11-1 Parts of the Contig Editor

Label	Name	Function
A	**Button Bar**	Contains several buttons you click to perform frequently-used functions.
B	**Sequence list**	Shows which sequences make up the contig you are editing.
C	**Agent**	Describes current selection and the current status of the spacebar shortcut.
D	**Splitter bar**	Controls the width of list/message area and the editing area; dragging the splitter bar to the left or right changes the size of these items.
E	**Sequence Bases**	Displays the sequence bases aligned.
F	**Sequence Scroll**	Scrolls up and down through the sequence bases in the contig.
G	**Base Numbers**	Shows which bases you are viewing, numbered from the beginning of the contig.

Label	Name	Function
H	**Consensus line**	Shows the calculated consensus of the sequences that were aligned to form this contig.
I	**Ambiguities**	Displays a plus (+) below any ambiguous base calls in the contig line and a bullet (•) below any base calls with disagreeing bases above them. This area also displays protein translations with a bullet (•).
J	**Base Scroll**	Scroll left and right through the bases of the contig.
K	**Translation Button**	This button toggles between the display of ambiguities and protein translations.
L	**Notation Line**	Displays translation of Reference Sequence and user text notes

VIEWING SEQUENCE NAMES

If the names of your sequences are very long, you may want to move the splitter (item D in the key) in the **Contig Editor** dialog.

BEFORE YOU START TO EDIT

SETTING THE BASE NUMBERING

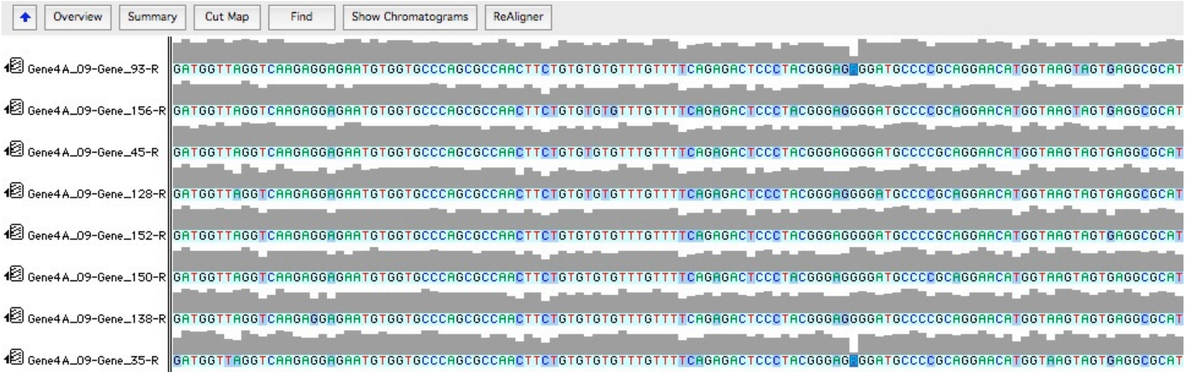
You have control over the base numbering of the consensus. For instance, you can select any base in the consensus and make it the “origin” base (base #1). *All the bases before that position will have negative numbers.* To set the numbering, select a base, then go to the **Sequence** menu and choose **Set Base Number** with an appropriate suboption.

Note: If you have already used a Reference Sequence in your assembly, this will set the base numbering.

CONFIDENCE HISTOGRAMS

Many **Sequencher** users have data which contains confidence scores. Phred scores have a range with a maximum of **60**, other systems have a dynamic range which goes up to **100**. To see the histograms which are displayed above the bases in the **Contig Editor**, choose one of the following from the **View** menu. For Phred scores, choose **Confidence Histogram (max 60)**. For other systems, choose **Confidence Histogram (max 100)**.

Figure 11-8 Confidence histograms in the Contig Editor



```
Gene4A_09-Gene_93-R  GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGRGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_156-R GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_45-R  GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_128-R GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_152-R GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_150-R GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_138-R GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGGGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
Gene4A_09-Gene_35-R  GATGGTTAGGTCAAGAGGAGAATGTGGTGCCCAGCGCCAACTTCTGTGTGTGTTTGTTTTCAGAGACTCCCTACGGGAGRGGATGCCCCGCAGGAACATGGTAAGTAGTGAGGCGCAT
```

CONSENSUS CALCULATION

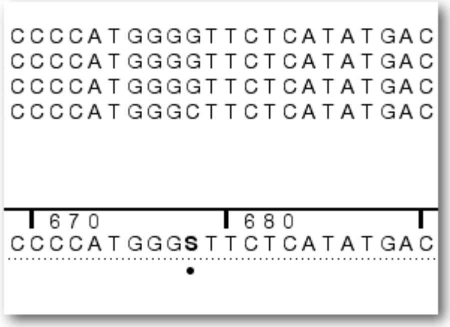
Under the **Contig** menu, there are four choices for calculating the consensus of your sequences: **Consensus Inclusively**, **Consensus by Plurality**, **Consensus to Forensic Standards**, and **Consensus by Confidence**. The consensus is displayed in the bottom line of the **Contig Editor** and is continuously updated as you edit. Thus, the consensus will alter after *any* change you make, whether that is altering bases or adding new sequences to your contig. Bullets under the consensus highlight discrepancies and pluses indicate ambiguities.

If you choose a consensus calculation for a project and then save and close the project, the consensus choice will be set as the default calculation for that project.

CONSENSUS INCLUSIVELY

With **Consensus Inclusively**, **Sequencher** determines the base in the consensus line by the smallest category of ambiguity that covers all available data. In the example in Figure 11-9, a column with one sequence containing a G and the other containing a C is shown in the consensus line as an S (the IUPAC code for C or G). This consensus is particularly useful in any work involving mutations.

Figure 11-9 Consensus inclusively example



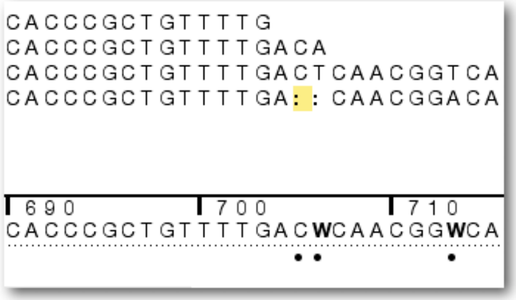
```
CCCCATGGGGTTCTCATATGAC
CCCCATGGGGTTCTCATATGAC
CCCCATGGGGTTCTCATATGAC
CCCCATGGGCTTCTCATATGAC

 670       680
CCCCATGGGSTTCTCATATGAC
         •
```

CONSENSUS BY PLURALITY

Consensus By Plurality is **Sequencher**'s default setting. With this consensus calculation, **Sequencher** determines the consensus line bases by majority rule. If there is no clear majority, **Sequencher** chooses the smallest category that fits the available data.

In the example in Figure 11-10 below, a column with two sequences that contains a T and an A is still shown in the consensus line as a W, because there is no clear majority. However, a column with three sequences that contains a *gap* in one sequence is shown in the consensus line as an ambiguity code in the consensus line only if the other two sequences do not agree.

Figure 11-10 Consensus by plurality example



```
CACCCGCTGTTTTG
CACCCGCTGTTTTGACA
CACCCGCTGTTTTGACTCAACGGTCA
CACCCGCTGTTTTGA::CAACGGACA

 690       700       710
CACCCGCTGTTTTGACWCAACGGWCA
               ••      •
```

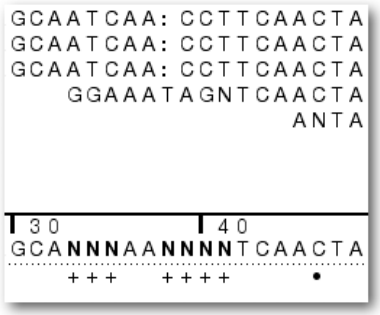
CONSENSUS TO FORENSIC STANDARDS

In certain applications, such as forensic mtDNA sequencing, there is an established standard reference for human identification. This calculation is based on the coverage of sequences at a given position and taken from work at the U.S. Armed Forces DNA Identification Laboratory (AFDIL) in Rockville, Maryland.

All positions with a coverage of only one sequence are marked as N (ambiguous).

- All IUPAC codes apart from ACG and T are treated as N.
- If any positions disagree where the coverage is between two and four, this position is marked as N. If the coverage is between five and seven, any position with a disagreement will be marked as ambiguous.
- If there are two or more disagreeing bases at a position, this will be called an N regardless of the coverage.

Figure 11-11 Consensus to Forensic Standards example



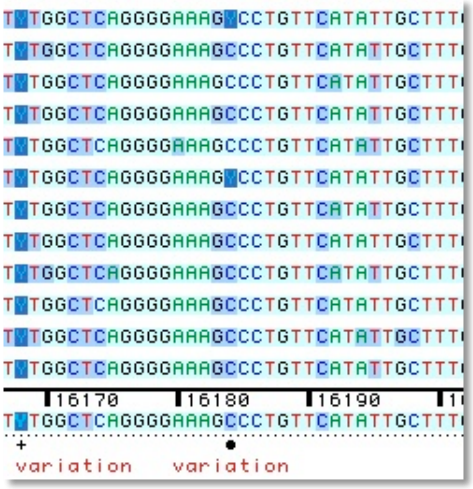
```
GCAATCAA:CCTTCAACTA
GCAATCAA:CCTTCAACTA
GCAATCAA:CCTTCAACTA
   GGAAATAGNTCAACTA
               ANTA

 30         40
GCANNNAANNNNTCAACTA
...................
   +++  ++++    •
```

CONSENSUS BY CONFIDENCE

While **Consensus by Plurality** and **Inclusively** are good approximate calculations for creating a snapshot of the consensus, they do not contain any information about the quality of the bases chosen. With **Consensus by Confidence**, you can take advantage of the confidence scores which your sequencer provides and which are imported with your data. With this method, **Sequencher** looks at each column of data and calculates the consensus using the underlying confidence data.

Additionally, if you have enabled **Display Base Confidences** from the **View** menu, you will see light, medium, or dark blue backgrounds behind the reads and the consensus line.

Figure 11-12 Consensus by Confidence example



```
TYTGGCTCAGGGGAAAGYCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGYCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
    16170     16180     16190     1
TYTGGCTCAGGGGAAAGCCCTGTTCATATTGCTTT
 +               ●
 variation     variation
```

ASSIGNING A REFERENCE SEQUENCE

To designate a sequence as a Reference Sequence, select the sequence icon in the **Project Window** and check the **Reference Sequence** item in the **Sequence** menu. From now on, that sequence icon will include a small letter "R" to remind you that it has been marked as a Reference Sequence (see Chapter 7 “The Reference Sequence” for more information).

12. EDITING CONTIGS

In this chapter, we explain how to find bases such as ambiguities and low-confidence bases. We discuss how to perform your edits, move and delete bases and sequences, insert gaps or bases, and create a new sequence from a consensus. You will learn how to compare sequences to highlight differences and how to create customized reports.

FINDING BASES WHICH NEED ATTENTION

You are likely to see a number of issues in your data. Some will need to be examined and perhaps edited. These issues might be indicated as an N, as a disagreement, as a gap, or as having a low confidence rating. You can use one of the **Next** commands from the **Select** menu in the consensus sequence of a contig to find any of these problem areas.

Once you have used any of the **Next** commands, **Sequencher** will activate the space bar to execute that operation. If you use either the menu command or the shortcut key combination two consecutive times, **Sequencher** will remind you that this easier alternative is available.

VIEWING AN INDIVIDUAL SEQUENCE FROM THE CONTIG

To look at the data for one of the sequences in your contig, double-click the sample name in the sequence list. **Sequencher** opens a locked sequence viewer containing the data for that sequence.

You can also select a sequence in the **Contig Editor** and then go to the **File** menu and choose **Open Window**.

FINDING DISAGREEING AMBIGUITIES

You can find ambiguities quickly by typing **Ctrl+N** (Windows) or **Cmd+N** (Mac) or by clicking in the consensus line and going to the **Select** menu and choosing **Next Ambiguous Base**. This will find data that is indicated as an N or as a disagreement or gap. The **Next Ambiguous Base** command will also find positions in the contig where the only contribution to the consensus is from a base with a low confidence score. This type of base will be marked with a + (plus) or a • (bullet) below the consensus line.

Each time you press **Ctrl+N** (Windows) or **Cmd+N** (Mac), **Sequencher** moves the cursor to the next ambiguity in the consensus, starting *after* the currently selected character. After establishing this function, you can move to the next ambiguous base by hitting the spacebar or going to the **Select** menu and using the **Repeat Select** command. If you use the **Shift** key in conjunction with a **Next Ambiguous Base** key combination, the cursor will move in the reverse direction.

FINDING DISAGREEING BASES

You can find disagreeing bases quickly by typing **Ctrl+D** (Windows) or **Cmd+D** (Mac) or clicking in the contig line, going to the **Select** menu, and then choosing **Next Contig Disagree**. This type of base will be marked with a • (bullet) below the consensus line.

Sequencher moves the cursor to the next disagreement in the consensus, starting *after* the currently selected character and ignoring ambiguities.

FINDING LOW CONFIDENCE BASES

You can quickly locate low confidence bases by typing **Ctrl+L** (Windows) or **Cmd+L** (Mac) or by going to the **Select** menu and choosing **Select Next Low Confidence Base**. You can set the ranges for Low, Medium, and High confidence in the **User Preferences** (see Chapter 23 “Customizing Sequencher and User Preferences”).

Note: You can use the **Next Low Confidence Base** command to help direct your editing. Under the **Window** menu, click on **User Preferences…**, select the **Confidence** option under **General** settings , and set the **Low** Confidence range appropriately for the value range you wish to find. Go to the **Select** menu and click on **Next Low Confidence Base** to move your cursor directly to the bases that require attention.

FINDING EDITED BASES

You can quickly review edits already made to a contig by typing **Ctrl+E** (Windows) or **Cmd+E** (Mac) or by going to the **Select** menu and choosing **Next Edited Base.**

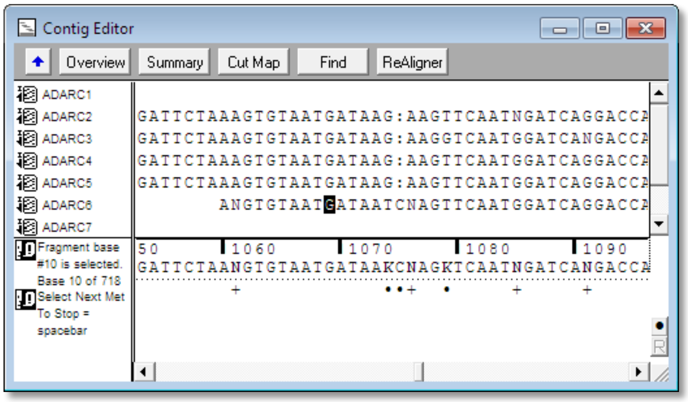
Note: This function will not find deletions.

PERFORMING EDITS

THE EDIT COMMAND

To edit a base, first set the **Edit** mode you wish to use, then select the base by clicking on it (see Figure 12-1). Now type the new character over the old. If you accidentally type the wrong character, go to the **Edit** menu and choose **Undo** or just type over the mistake. **Sequencher** will recalculate the consensus line automatically. If you resolve an ambiguity at a particular position, the ambiguity marker below the consensus line disappears (see Figure 12-1).

Figure 12-1 A base is selected in the contig editor



```
ADARC1
ADARC2   GATTCTAAAGTGTAATGATAAG:AAGTTCAATNGATCAGGACCA
ADARC3   GATTCTAAAGTGTAATGATAAG:AAGGTCAATGGATCANGACCA
ADARC4   GATTCTAAAGTGTAATGATAAG:AAGTTCAATGGATCAGGACCA
ADARC5   GATTCTAAAGTGTAATGATAAG:AAGTTCAATGGATCAGGACCA
ADARC6          ANGTGTAATGATAATCNAGTTCAATGGATCAGGACCA
ADARC7
Fragment base   50      1060      1070      1080      1090
#10 is selected. GATTCTAANGTGTAATGATAAKCNAGKTCAATNGATCANGACCA
Base 10 of 718  ........................................
Select Next Met          +            ••+  •     +     +
To Stop =
spacebar
```

EDITING MODES

The **Edit** command allows you to replace/overstrike, insert, or erase a base call. The specific option you choose from the **Edit** menu will create that behavior. When you use **Replace When Editing**, the base you have just edited will remain highlighted. When you use **Overstrike When Editing**, the cursor will move one base to the right and this base will then be highlighted. **Insert When Editing** allows you to insert bases before the currently selected base, pushing all of the following bases forward. This also allows you to add bases beyond the end of a sequence. **Sequencher** uses **Replace When Editing** as its default setting.

Note: If you insert or delete a base from any one sequence, the sequence may no longer align with other sequences after the point where you made your change. The consensus line will reflect this by showing lots of ambiguities in the form of bullets •. If you have made only one change, go to the **Edit** menu and the **Undo** command will remove it.

VIEWING BASE EDITS

To assess the quality of your data, you need to be able to see how much modification has already been done to your contig as you edit it.

When a base is edited, **Sequencher** changes its appearance by displaying it in magenta and bold-faced text. If you want edits to stand out even more (for instance in a black and white printout), you can also invert the case of the characters (e.g. change all edits from uppercase to lowercase text).

To see the options for displaying edits, go to the **View** menu and click on the **Base Edits As…** submenu to choose from **Not Highlighted**, **Bold Magenta**, or **bOLD and cASE cHANGE**.

COLLECT GAPS

The **Collect Gaps** command allows you to select a range in a sequence or contig consensus and move all of the gaps within this selection to either the 5’ or the 3’ end of the selection.

If you wish to collect the gaps within a single sequence, you must select the range within that sequence. Go to the **Sequence** menu and choose the **Collect Gaps** command. To move the gaps to the 5’ end, select **Left** from the submenu. To move gaps to the 3’ end of the selection, choose **Right** from the submenu.

If you wish to collect the gaps within a region of a single contig you must select the range from the consensus line. Go to the **Sequence** menu and choose the **Collect Gaps** command. To move the gaps to the 5’ end, select **Left** from the submenu. To move gaps to the 3’ end of the selection, choose **Right** from the submenu.

Note: The **Collect Gaps** command may have the effect of increasing the number of ambiguities. Check the consensus line to see if the number of ambiguities has increased or decreased. You can go to the **Edit** menu and use the **Undo** command to reverse the **Collect Gaps** command.

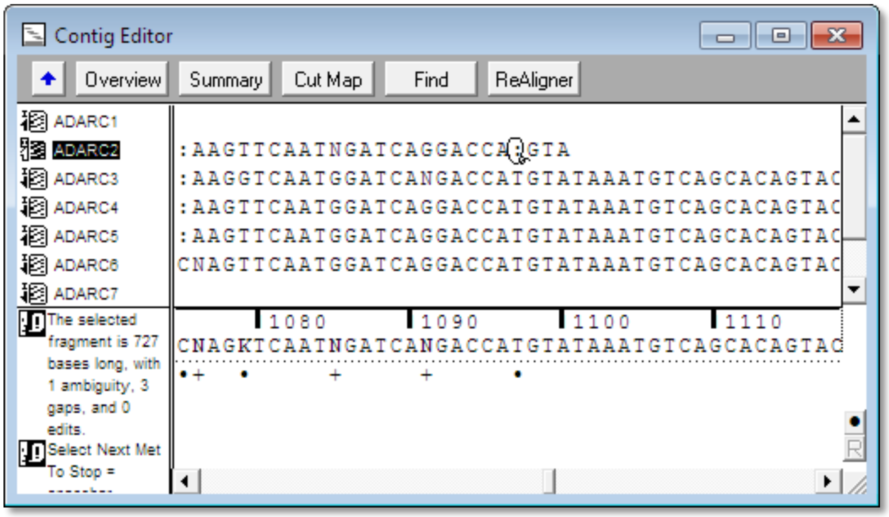
MOVING BASES AND GAPS

You can move a selection within a single sequence using the lasso tool. The selection can be a single base or gap or a number of bases or gaps.

Begin by making your selection by dragging the cursor across the bases or gaps you wish to move. The region will be highlighted. Then hold down the **Alt** key and your cursor will turn into a little “lasso” tool. Hold down the **Alt** key and click on the highlighted selection you want to move. Then drag it to its new location. The selection will change to **Bold Magenta** text in its new position.

Note If you have used the **Large Gap** algorithm to assemble sequences you can double click anywhere within the gap region to select it.

Figure 12-2 The lasso tool around a base to be moved



```
ADARC1
ADARC2   :AAGTTCAATNGATCAGGACCA:GTA
ADARC3   :AAGGTCAATGGATCANGACCATGTATAAATGTCAGCACAGTAC
ADARC4   :AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTAC
ADARC5   :AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTAC
ADARC6   CNAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTAC
ADARC7
              1080      1090      1100      1110
         CNAGKTCAATNGATCANGACCATGTATAAATGTCAGCACAGTAC
         •+   •     +     +     •
```

The selected fragment is 727 bases long, with 1 ambiguity, 3 gaps, and 0 edits.

Select Next Met To Stop =

USING REALIGNER TO CLEAN UP CONTIGS

The **ReAligner** button on the **Contig Editor** button bar lets you apply the **ReAligner** algorithm to optimize gap placement in your sequences without first dissolving your contig. This is especially useful when you are left with gaps and unaligned bases after removing one or more sequences from a contig.

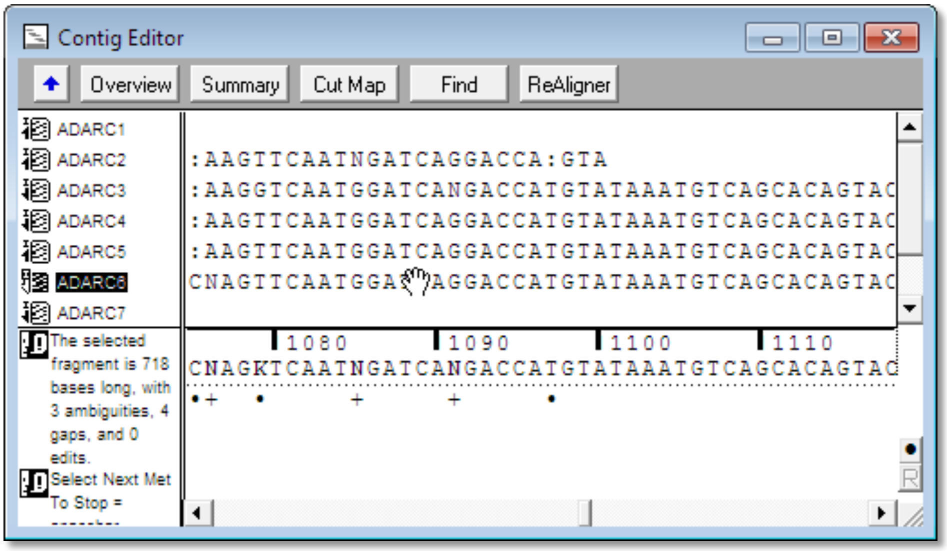
Click on the **ReAligner** button. Once the ReAlignment has finished, a new window called **ReAlignment Complete** will appear. This window displays a summary of the disagreements, gaps, and ambiguities before and after the ReAlignment.

Note: ReAlignment is not the same as reassembly. If you wish to reassemble your sequences, you must first dissolve your contig. (See Chapter 9 “Sequence Assembly” for more information.)

MOVING SEQUENCES

If you hold down the **Ctrl** (Windows) or **Cmd** (Mac) key, the “grabber” tool is invoked. It resembles a small hand. You can use the grabber tool to drag an entire sequence to the left or right. This method of moving entire sequences also works in the **Overview**. The grabber tool is shown in Figure 12-3. You may notice that some or all of your sequence no longer aligns and that a large number of bullets (•) appear in the consensus line. If you decide that the new position is not correct, you can use the **Undo** command to restore it to its previous place.

Figure 12-3 Grabber tool



```
ADARC1
ADARC2   :AAGTTCAATNGATCAGGACCA:GTA
ADARC3   :AAGGTCAATGGATCANGACCATGTATAAATGTCAGCACAGTAC
ADARC4   :AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTAC
ADARC5   :AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTAC
ADARC6   CNAGTTCAATGGA AGGACCATGTATAAATGTCAGCACAGTAC
ADARC7
                  1080      1090      1100      1110
         CNAGKTCAATNGATCANGACCATGTATAAATGTCAGCACAGTAC
         •+   •      +       +       •
```

The selected fragment is 718 bases long, with 3 ambiguities, 4 gaps, and 0 edits.

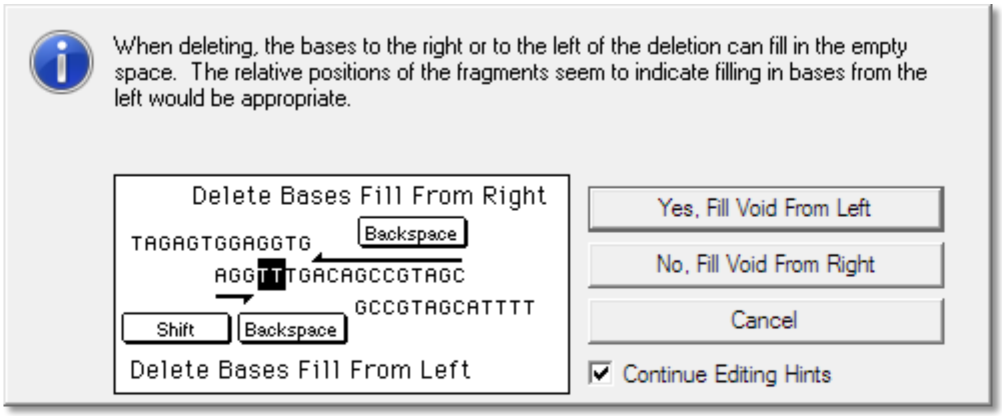
Select Next Met To Stop =

Note: Moving sequences in a low coverage region could leave a hole in the middle of the contig. This is not recommended since the hole may cause the contig to align inappropriately with other data. **Sequencher** will warn you if this about to happen.

DELETING BASES

You can delete bases in the middle of a sequence by simply highlighting them and then using the **Delete** key. **Sequencher** asks if you want to fill the gap with bases moved from either the left or the right of the void you have created. Click on either **Yes, Fill Void From Left**, or **No, Fill Void From Right** button. (Figure 12-4) **Sequencher** then moves the rest of the bases accordingly. Click on **Cancel** if you do not want to delete anything. You can also perform the deletion by going to the **Sequence** menu, using the **Delete Bases** command, and then choosing either **Fill Void From Left** or **Fill Void From Right** from the submenu.

Figure 12-4 Fill Void dialog



When deleting, the bases to the right or to the left of the deletion can fill in the empty space. The relative positions of the fragments seem to indicate filling in bases from the left would be appropriate.

DELETING BASES FROM THE 5’ OR 3’ ENDS

Although you may have automatically trimmed your sequences for poor quality data when you imported them (see Chapter 6 “Preparing your data for assembly”), you may still need to delete a few 5’ bases to correct any mistakes made during the amplification process.

Select the bases you want to delete. If you are at the 5' (left end), then click the **Forward Delete** key. **Sequencher** will delete those bases without moving the entire sequence to the left. Similarly, if you choose bases on the 3‘ (right) end, use the **Backspace** key for Windows or the **Delete** key on Mac and **Sequencher** will automatically delete those bases without realigning your sequence.

Note: When editing from the consensus line, you do not have to worry about the bases shifting to the left and right.

INSERTING GAPS OR BASES INTO A CONTIG

If you want to insert gaps, go to the **Sequence** menu and click on **Insert Gaps & Move Bases**, then choose **Right** or **Left** from the submenu. You can also create gaps by selecting a number of bases equivalent to the gap size you want to create. Press the **Tab** key if you want the selected bases to move to the right. To move to the left, hold down the **Alt** (Windows) or **Option** (Mac) key when you press **Tab**. The space created will be filled with gap marks (small colon marks) that will appear without disrupting the rest of your contig.

Once your gaps are in place, you can either type the bases you want into the gap using **Overstrike When Editing** mode or use the lasso and grabber tools to move existing bases into the new position.

MAKING EDITS FROM THE CONSENSUS LINE

You can simultaneously edit all the bases in the column above the position you choose in the consensus line, instead of individually changing each one. To do this, select a base in the consensus line and enter the character that represents the edit you want to make in that column. If you currently have a base selected in a sequence, go to the **Select** menu and choose **Contig Column** (**Ctrl+K** (Windows) or **Cmd+K** (Mac))**.**

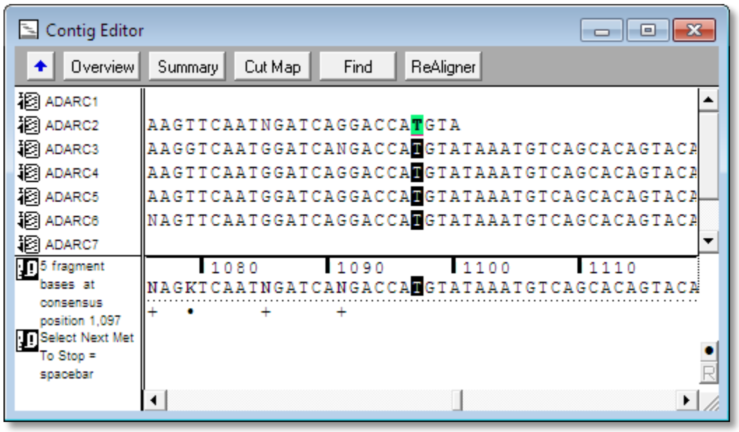
Any bases in the sequences above the contig column you select that do not match the character you type *will be changed* to match the character you type in the consensus line. Your edits will appear according to the preferences you set in the **View** menu. Generally, this will be magenta and bold-text.

You can also delete bases from the consensus line. You cannot use the **Undo** command to reverse this action if your contig contains a Reference Sequence.

Note: Remember that changing bases from the consensus line edits only the sequences, after which the consensus is recalculated and displayed. Your action does not change the consensus directly. Changes in the consensus result from matching sequences with each other and evaluating them according to the type of consensus calculation you have chosen (e.g. by plurality, inclusively, forensic, confidence).

Further note that this type of editing does not change Reference Sequences.

Figure 12-5 The base is changed



```
ADARC1
ADARC2   AAGTTCAATNGATCAGGACCATGTA
ADARC3   AAGGTCAATGGATCANGACCATGTATAAATGTCAGCACAGTACA
ADARC4   AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTACA
ADARC5   AAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTACA
ADARC6   NAGTTCAATGGATCAGGACCATGTATAAATGTCAGCACAGTACA
ADARC7
              1080      1090      1100      1110
         NAGKTCAATNGATCANGACCATGTATAAATGTCAGCACAGTACA
         +  •      +     +
```

5 fragment bases at consensus position 1,097

Select Next Met To Stop = spacebar

SPLIT AFTER SELECTION…

You can split a sequence into two pieces. Put the cursor where you want to split the sequence. Go to the **Sequence** menu and choose **Split After Selection….**

Sequencher will ask you if you want to split the sequence after the last currently selected base. Click on the **Split** button to proceed. **Sequencher** will then split the sequence in two and label the new sequences for you.

You can also apply the **Split After Selection…** command to an entire contig by simply making your selection in the consensus sequence, and then choosing **Split After Selection…** from the **Sequence** menu.

AUTO MATCH

Sometimes you may wish to edit a number of sequences simultaneously, perhaps to match a high quality segment of sequence. **Sequencher** allows you to perform this type of mass edit on a range of bases using the **Auto Match** command.

First select the base or range of bases to which the other sequences should match, then go to the **Edit** menu and choose **Auto Match**. This command will edit (and highlight) every overlapping base so that it will agree with the selected sequence. To reverse this procedure, go to the **Edit** menu and click on **Undo**.

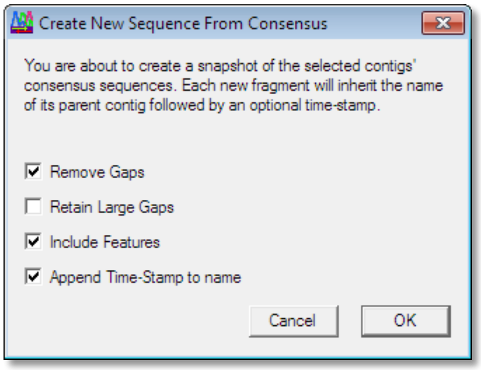
CREATE NEW SEQUENCE FROM A CONSENSUS

You can create a new sequence from the consensus sequence by going to the **Contig** menu and choosing **Create New Seq From Consensus...** In addition to creating a new sequence, you can use this feature to track the changes made to your contig over time.

Note: If you have used the **Consensus by Confidence** calculation, then the newly created consensus will also contain the confidence scores calculated by this consensus method.

Sequencher displays a dialog that tells you that you are about to create a snapshot consensus of the selected contig. This dialog allows you to control some basic features of the new sequence. Figure 12-6 shows the options dialog for this command.

Figure 12-6 Create New Sequence From Consensus options dialog



Create New Sequence From Consensus

You are about to create a snapshot of the selected contigs' consensus sequences. Each new fragment will inherit the name of its parent contig followed by an optional time-stamp.

☑ Remove Gaps

□ Retain Large Gaps

☑ Include Features

☑ Append Time-Stamp to name

Check or uncheck the boxes to indicate how you want the new sequence to appear. If the default setting **Remove Gaps** is checked, then **Sequencher** will remove all gaps (large and small) when it creates the new sequence.

If you want to keep large gaps independently from other gaps, then select the checkbox called **Retain Large Gaps**. In **Sequencher**, a large gap consists of 10 or more consecutive bases. Such a gap might appear if, for example, you have used the **Large Gap** algorithm to assemble genomic and cDNA. The **Retain Large Gaps** option is not available if **Remove Gaps** is checked.

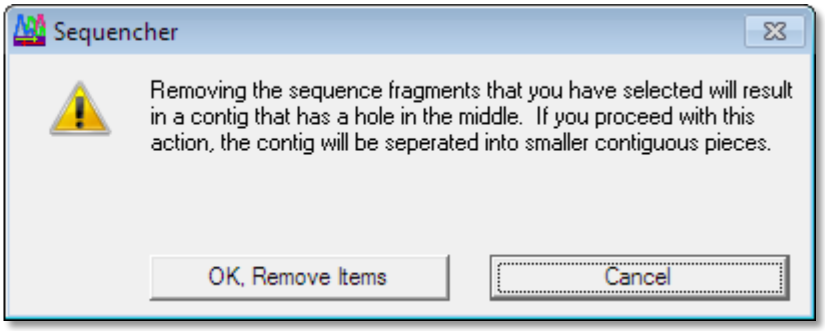
Note: To see the new sequence, go to the **Project Window**. Remember this is a snapshot of the consensus at the time of its creation. Future edits will not be marked.

REMOVING SEQUENCES FROM A CONTIG

To remove sequences from a contig, click on the name of the sequence you wish to remove from the list at the left side of the **Contig Editor** window or in the **Overview** window. To remove more than one sequence at a time, hold down the **Shift** key while you click the names. Next, go to the **Contig** menu and choose **Remove Selected Sequences… Sequencher** will ask you if you are sure you wish to remove the sequence(s). Click on the **Remove Fragments** button to delete the sequences from the contig. The sequences you just removed from the contig are moved out of the contig and back into the **Project Window**. They will also stay selected.

Note: This action may break a contig into multiple pieces. If this is about to happen, **Sequencher** will display a dialog (Figure 12-7) to ask if this is really what you intend to do. Press the **Cancel** button if you do not wish to proceed. Any new contigs generated in this way are named by appending an alphabetic character to the name of the original contig.

Figure 12-7 Warning dialog when about to make a hole in a contig



Removing the sequence fragments that you have selected will result in a contig that has a hole in the middle. If you proceed with this action, the contig will be seperated into smaller contiguous pieces.

DISSOLVING A CONTIG

To dissolve a contig, first select the contig's icon from the project. Go to the **Contig** menu and choose **Dissolve Contig…**. You can also perform this command while you are in the **Contig Editor**.

Sequences from the dissolved contig are shown highlighted in the project. It is important to remember that all edits made to a sequence while in the contig will be retained. If you want to restore the original sequences, you must go to the **Sequence** menu and click on **Revert to Experimental Data**.

13. THE VARIANCE TABLE

In this chapter, we explain how to use the **Variance Table**, an interactive display of the differences among your sequences or contigs. We go on to discuss how you find differences and work with your trace and confidence data. The **Variance Table** is useful for SNP analysis, mutation detection, and clone checking.

THE VARIANCE TABLE STRATEGY

WHAT IS A VARIANCE TABLE

In its simplest form, a **Variance Table** compares and displays the differences between two sequences. The data in the **Variance Table** are dynamically linked to the data in the underlying contigs.

You can compare some or all of the sequences in a contig to the **Consensus** sequence, the **Reference Sequence**, or the **Top Sequence** of that contig (see Figure 13-1 below). This generates a **Variance Table** which will display the differences between the exemplar sequence and your chosen sequence(s). The results may be from just one contig, but can represent an assembly of hundreds or even thousands of sequences. If you are performing activities such as de novo sequencing, clone checking, or resequencing, you may want to use this form of the **Variance Table**.

THE VARIANCE TABLE AND CONSENSUS SEQUENCES

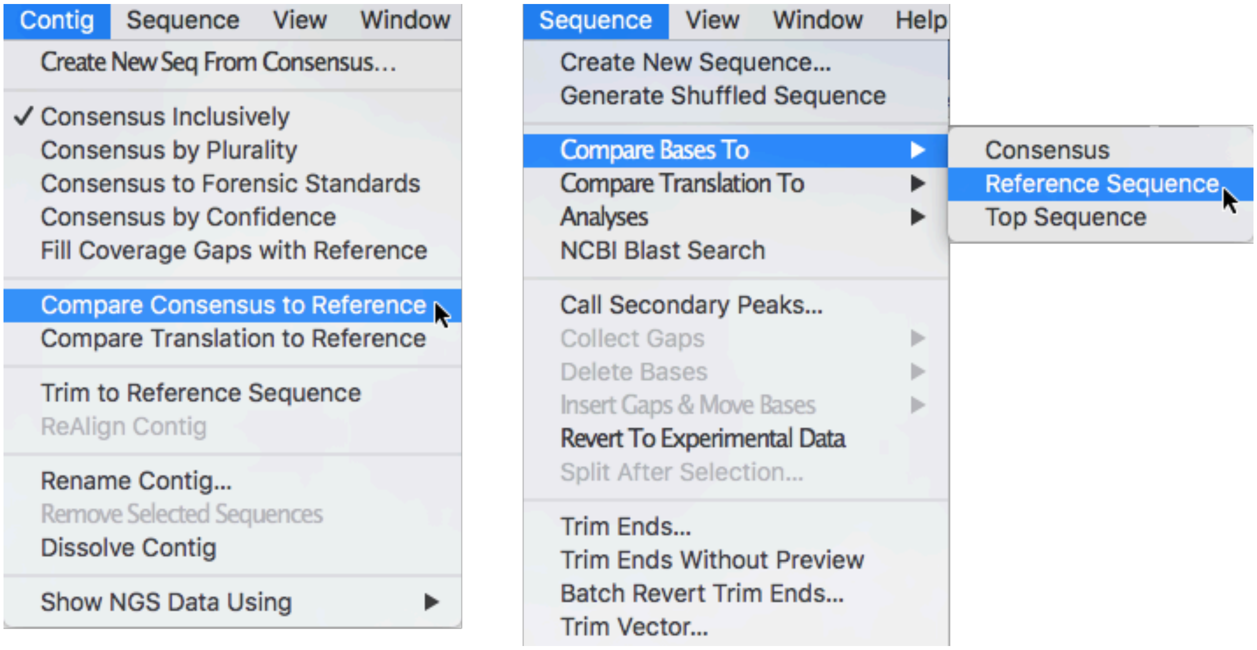
You can also create a **Variance Table** that summarizes all of the differences between selected contig consensus sequences (see Figure 13-1 below) in your project and a common Reference Sequence (consensus **Variance Table**). When you are working with multiple samples from multiple sources and have used **Assemble by Name** with a Reference Sequence, you should use this **Variance Table**. If you are performing activities such as comparative sequencing, clone checking, or SNP analysis, you probably want to use this form of the **Variance Table**.

THE TRANSLATED VARIANCE TABLE

If you are interested in focusing on the differences between sequences or contigs at the amino acid level, you should use the **Translated Variance Table**. This table differs from the **Variance Table** by displaying both codons and their associated amino acid residues. Each row in the **Translated Variance Table** summarizes the variations among all of the selected sequence translations at a given amino acid position *relative to the exemplar.*

You can also create a **Translated Variance Table** that summarizes the differences between the translations of selected contig consensus sequences in your project and the translation of a common Reference Sequence (Translated consensus Variance Table). When you are working with multiple samples from multiple sources, and have used **Assemble by Name** with a Reference Sequence, you should use this form of the **Translated Variance Table**.

Figure 13-1 Variance Table menu commands



THE STRUCTURE OF THE VARIANCE TABLE

Each row in a **Variance Table** summarizes the variation among all of the selected sequences at a given base position *relative to the exemplar* (see Figure 13-2). The exemplar can be a Reference Sequence, a consensus sequence, or any other sequence you choose. In the consensus **Variance Table**, the exemplar is limited to the Reference Sequence.

The range of bases included in a **Variance Table** is called the **Comparison Range**. The length of the exemplar determines the default **Comparison Range** and sets the base numbering.

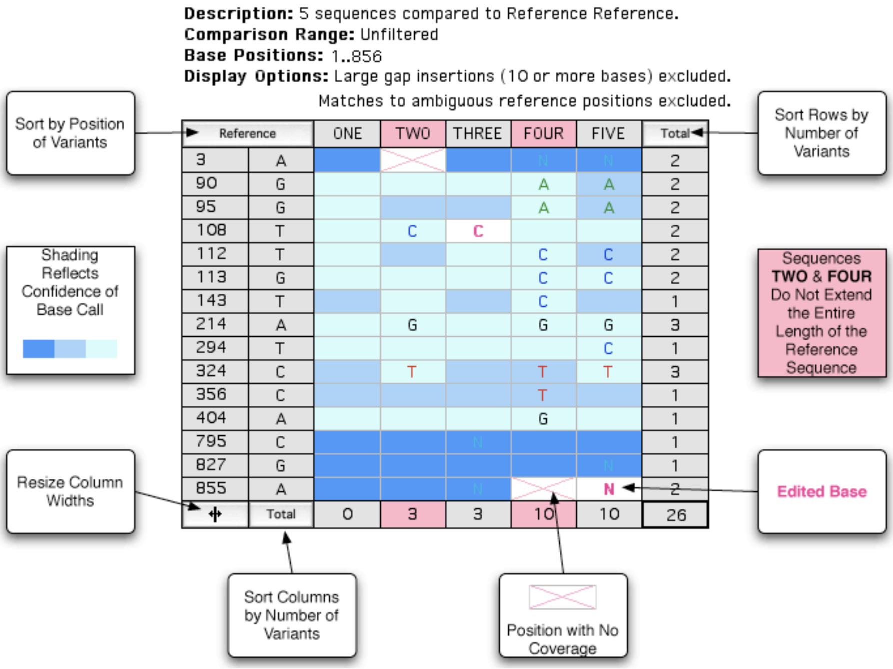
The extent to which the exemplar matches any sample sequence is called the **Coverage Range**. The **Coverage Range** can be complete (full range) or incomplete. If the range is full, the entire length of the sample sequence has been compared to the entire length of the exemplar sequence.

If the range is listed as incomplete, the display of the sample sequence comparison runs along only part of the exemplar sequence. If a sequence extends the full length of the **Comparison Range**, its column header will be shaded in gray. If a sequence in the table does not extend the full length of the **Comparison Range**, its column header will be shaded in pink.

The first column of the **Variance Table** displays the position and the second column displays the base call of the exemplar sequence, but *only at positions that differ*. The additional columns in the **Variance Table** are created by the sequences or contigs you select for comparison.

You can view or edit the data in any version of the **Variance Table** by double clicking on any of the cells in that table. In **Review** mode, when you type a new base in the **Variance Table**, you edit the underlying sequence or contig automatically.

Figure 13-2 Components of Variance Table



Description: 5 sequences compared to Reference Reference.
Comparison Range: Unfiltered
Base Positions: 1..856
Display Options: Large gap insertions (10 or more bases) excluded.
Matches to ambiguous reference positions excluded.

► Reference		ONE	TWO	THREE	FOUR	FIVE	Total ◄
3	A		☒		N	N	2
90	G				A	A	2
95	G				A	A	2
108	T		C	C			2
112	T				C	C	2
113	G				C	C	2
143	T				C		1
214	A		G		G	G	3
294	T					C	1
324	C		T		T	T	3
356	C				T		1
404	A				G		1
795	C			N			1
827	G					N	1
855	A			N	☒	N	2
✢	Total	0	3	3	10	10	26

Sort by Position of Variants
Shading Reflects Confidence of Base Call
Resize Column Widths
Sort Columns by Number of Variants
Position with No Coverage
Sort Rows by Number of Variants
Sequences TWO & FOUR Do Not Extend the Entire Length of the Reference Sequence
Edited Base

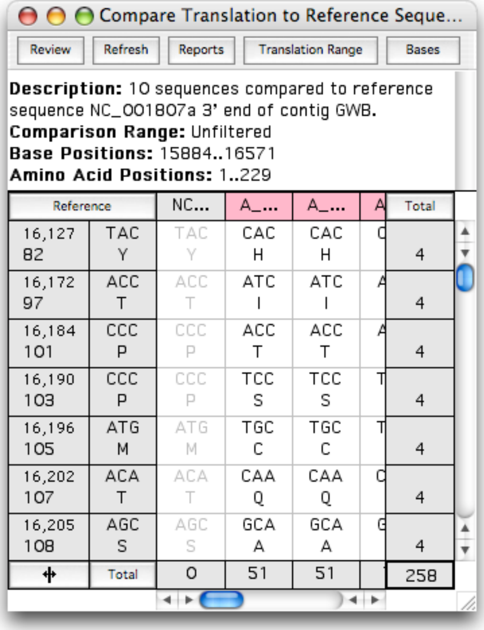
STRUCTURE OF THE TRANSLATED VARIANCE TABLE

The structure of the **Translated Variance Table** follows the format of the **Variance Table** but its display consists of an amino acid and its codon in each cell (see Figure 13-3).

When you see a cell in the table with an amino acid and its codon written in black typeface, this indicates a difference between your exemplar and your sample. You may also see some cells in the table where the amino acid and codon are displayed in light gray. Amino acids in these cells match the exemplar but use either an identical or synonymous codon.

The numbers that appear in the first column of the **Translated Variance Table** refer to the position of the first base of the codon (in the upper location) and the position of the amino acid (in the lower one).

Figure 13-3 The Translated Variance Table



Description: 10 sequences compared to reference sequence NC_001807a 3' end of contig GWB.
Comparison Range: Unfiltered
Base Positions: 15884..16571
Amino Acid Positions: 1..229

Reference		NC...	A_...	A_...	A	Total
16,127 82	TAC Y	TAC Y	CAC H	CAC H	C	4
16,172 97	ACC T	ACC T	ATC I	ATC I	A	4
16,184 101	CCC P	CCC P	ACC T	ACC T	A	4
16,190 103	CCC P	CCC P	TCC S	TCC S	T	4
16,196 105	ATG M	ATG M	TGC C	TGC C	T	4
16,202 107	ACA T	ACA T	CAA Q	CAA Q	C	4
16,205 108	AGC S	AGC S	GCA A	GCA A	G	4
✣	Total	0	51	51		258

Note: When **Sequencher** encounters a gap, it skips to the next available base to include it as a component of the codon. If an N is in the first or second position of the codon, **Sequencher** displays a question mark (**?)** instead of an amino acid. If the N is in the third position of the codon, a residue may be displayed. This will depend on the genetic code redundancy.

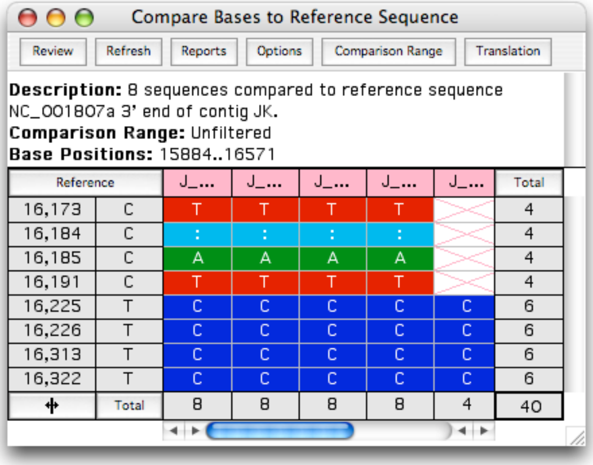
CREATING A VARIANCE TABLE

VARIANCE TABLE FOR SEQUENCES IN THE SAME CONTIG

When you are ready to begin creating your **Variance Table**, either select individual sequences from within a contig or go to the **Project Window** and select a contig by clicking on its icon. This will select all of the contig's sequences.

Then go to the **Sequence** menu and choose **Compare Bases To** and either **Consensus**, **Reference Sequence**, or **Top Sequence** from the submenu. The **Variance Table** appears in a new window (Figure 13-4).

Figure 13-4 Variance Table Created with Compare Bases To Reference Sequence



Description: 8 sequences compared to reference sequence NC_001807a 3' end of contig JK.
Comparison Range: Unfiltered
Base Positions: 15884..16571

Reference		J_...	J_...	J_...	J_...	J_...	Total
16,173	C	T	T	T	T	✕	4
16,184	C	:	:	:	:	✕	4
16,185	C	A	A	A	A	✕	4
16,191	C	T	T	T	T	✕	4
16,225	T	C	C	C	C	C	6
16,226	T	C	C	C	C	C	6
16,313	T	C	C	C	C	C	6
16,322	T	C	C	C	C	C	6
✥	Total	8	8	8	8	4	40

THE CONSENSUS VARIANCE TABLE

When you want to compare a consensus sequence to a Reference Sequence, you use the **Contig** menu's **Compare Consensus to Reference** command.

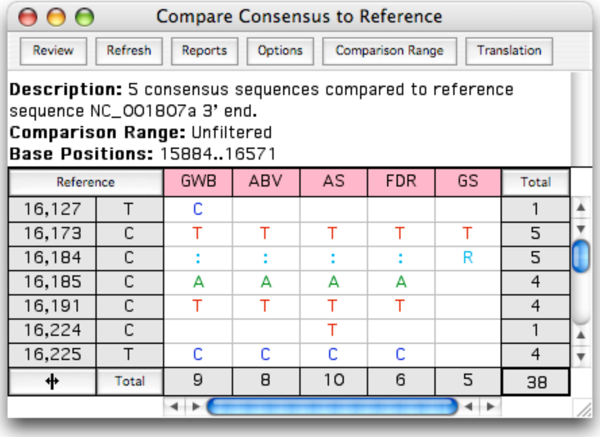
You can also use the **Compare Consensus to Reference** command to compare the consensus sequences from multiple contigs that share the same Reference Sequence. These contigs are the result of using the **Assemble By Name** command with a Reference Sequence.

For each of these operations, go the **Project Window** and select one or more contigs that share a common Reference Sequence. Then, from the **Contig** menu, choose the **Compare Consensus to Reference** command. The **Variance Table** appears in a new window (see Figure 13-5).

Each cell in the resulting Variance Table will display a consensus base when that base differs from the Reference Sequence. Pink shading in the header and footer cells of the **Variance Table** highlight contigs where the consensus sequence does not cover the span of the Reference Sequence. The pink X's mark specific cells where coverage is missing (see Figure 13-5). It is possible for a sequence to have partial coverage, a pink header and footer, but not have any pink X's in the column.

Note: When comparing multiple consensus sequences, the **View** menu's **Display Base Confidences** option is not relevant, because confidence values exist only in sequences and not in consensus sequences.

Figure 13-5 Compare Consensus to Reference Variance Table



Description: 5 consensus sequences compared to reference sequence NC_001807a 3' end.
Comparison Range: Unfiltered
Base Positions: 15884..16571

Reference		GWB	ABV	AS	FDR	GS	Total
16,127	T	C					1
16,173	C	T	T	T	T	T	5
16,184	C	:	:	:	:	R	5
16,185	C	A	A	A	A		4
16,191	C	T	T	T	T		4
16,224	C			T			1
16,225	T	C	C	C	C		4
✣	Total	9	8	10	6	5	38

For more information about **Assemble by Name** and **Reference Sequences**, see Chapter 10 “Assemble by Name” and Chapter 7 “The Reference Sequence”.

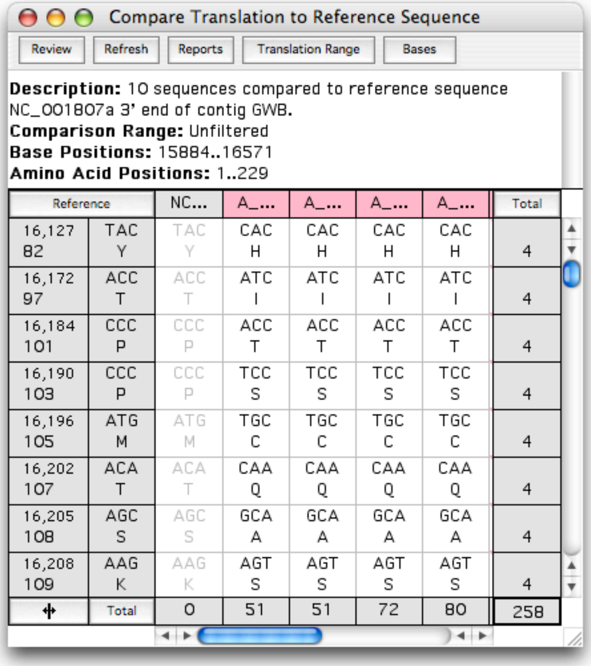
TRANSLATED VARIANCE TABLE FOR SEQUENCES IN THE SAME CONTIG

When you are ready to begin creating your **Translated Variance Table**, either select individual sequences from within a contig or go to the **Project Window** and select a contig by clicking on its icon. This will select all of the contig’s sequences.

Then go to the **Sequence** menu and choose **Compare Translation To** and either **Consensus**, **Reference Sequence**, or **Top Sequence** from the submenu. The **Translated Variance Table** appears in a new window (see Figure 13-6).

The **Translated Variance Table** header summarizes the table’s contents including a Description, the Base Positions, and the Amino Acid Positions. If a sequence in the table does not extend the full length of the Comparison Range, its column header will be shaded in pink. The pink X’s mark cells that are not covered by sequence data, indicating that the status of the base at that position is unknown.

Figure 13-6 Translated Variance Table for sequences in the same contig



Description: 10 sequences compared to reference sequence NC_001807a 3' end of contig GWB.
Comparison Range: Unfiltered
Base Positions: 15884..16571
Amino Acid Positions: 1..229

Reference		NC...	A_...	A_...	A_...	A_...	Total
16,127 82	TAC Y	TAC Y	CAC H	CAC H	CAC H	CAC H	4
16,172 97	ACC T	ACC T	ATC I	ATC I	ATC I	ATC I	4
16,184 101	CCC P	CCC P	ACC T	ACC T	ACC T	ACC T	4
16,190 103	CCC P	CCC P	TCC S	TCC S	TCC S	TCC S	4
16,196 105	ATG M	ATG M	TGC C	TGC C	TGC C	TGC C	4
16,202 107	ACA T	ACA T	CAA Q	CAA Q	CAA Q	CAA Q	4
16,205 108	AGC S	AGC S	GCA A	GCA A	GCA A	GCA A	4
16,208 109	AAG K	AAG K	AGT S	AGT S	AGT S	AGT S	4
✢	Total	0	51	51	72	80	258

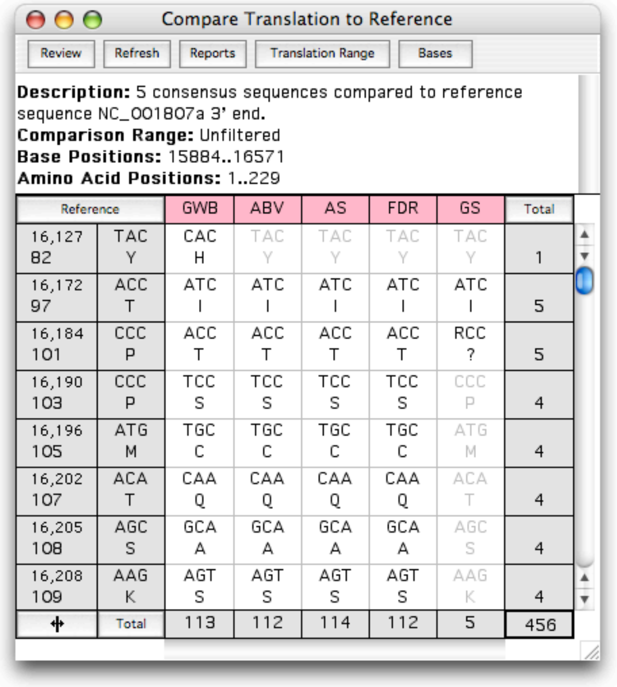
THE TRANSLATED CONSENSUS VARIANCE TABLE

Sequencher allows you to compare the translation of a consensus sequence to the translation of a Reference Sequence. Go to the **Contig** menu and click on **Compare Translation to Reference.**

You can also use the **Compare Translation to Reference** command to compare the translations of consensus sequences from multiple contigs to the translation of their common Reference Sequence. These contigs are formed when you use **Assemble By Name** with a Reference Sequence.

For both of these operations, select one or more contigs in the **Project Window** that share a common Reference Sequence. Then, from the **Contig** menu, choose **Compare Translation to Reference.** The Translated consensus Variance Table appears in a new window (see Figure 13-7).

Figure 13-7 Translated variance Table for consensus sequences from different contigs



Description: 5 consensus sequences compared to reference sequence NC_001807a 3' end.
Comparison Range: Unfiltered
Base Positions: 15884..16571
Amino Acid Positions: 1..229

Reference		GWB	ABV	AS	FDR	GS	Total
16,127 82	TAC Y	CAC H	TAC Y	TAC Y	TAC Y	TAC Y	1
16,172 97	ACC T	ATC I	ATC I	ATC I	ATC I	ATC I	5
16,184 101	CCC P	ACC T	ACC T	ACC T	ACC T	RCC ?	5
16,190 103	CCC P	TCC S	TCC S	TCC S	TCC S	CCC P	4
16,196 105	ATG M	TGC C	TGC C	TGC C	TGC C	ATG M	4
16,202 107	ACA T	CAA Q	CAA Q	CAA Q	CAA Q	ACA T	4
16,205 108	AGC S	GCA A	GCA A	GCA A	GCA A	AGC S	4
16,208 109	AAG K	AGT S	AGT S	AGT S	AGT S	AAG K	4
	Total	113	112	114	112	5	456

SETTING THE VARIANCE TABLE CONDITIONS

RESTRICTING THE COMPARISON RANGE

You can restrict the Comparison Range by Feature Key if you have any Features in your exemplar. Construct a Variance Table then go to the button bar and click on the **Comparison Range** button. The **Comparison Range** dialog appears (see Figure 13-8).

In the default setting the **Unfiltered** radio button is checked. Leave this radio button checked if you decide to leave your table unfiltered.

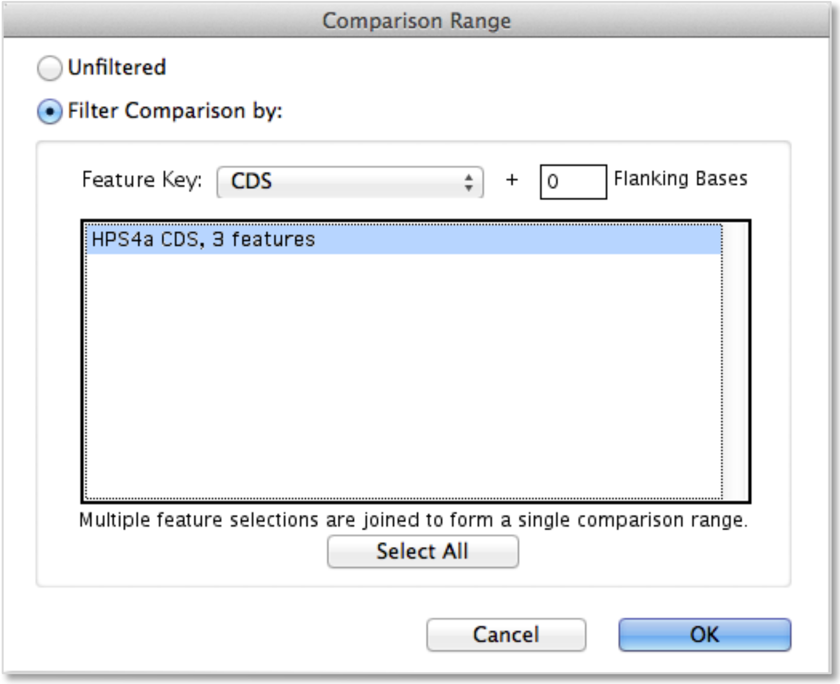
If you want to restrict the Comparison Range, click the **Filter Comparison by:** radio button. Once you have clicked on that button, the **Feature Key:** drop-down menu becomes active. You can now choose which Filter Key to use by making a selection from this menu.

Once you have selected a Feature Key, you will see a list of features annotated with that key in the pane located below the **Feature Key:** drop-down menu. You can select a single feature, a continuous range or a discontinuous range of features. You can also select all of the features within this pane by clicking on the **Select All** button.

You can widen your Comparison Range by typing a number into the **Flanking Bases** box. For example, by typing the number ten in this box, the Comparison Range will include 10 flanking bases on either side of the chosen Feature Key.

Once you have made your selection, click on the **OK** button. The Variance Table will be rebuilt automatically.

Figure 13-8 Comparison Range dialog



Comparison Range

○ Unfiltered

● Filter Comparison by:

Feature Key: CDS + 0 Flanking Bases

HPS4a CDS, 3 features

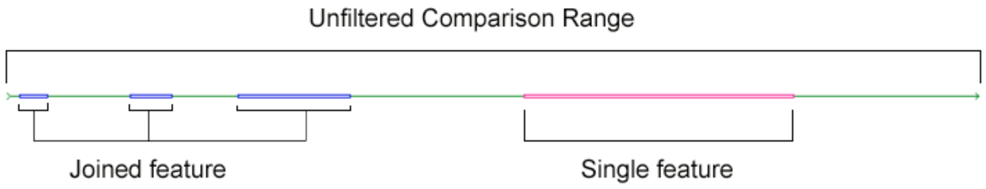
Multiple feature selections are joined to form a single comparison range.

RESTRICTING THE TRANSLATION RANGE

Reference Sequences are often marked up with GenBank-style features. (See Chapter 19 “Motifs and Features” for further information.)

You can reduce the scope of your Translated Variance Table to a single feature. Some features, such as mRNA or CDS, may be composed of several elements. These elements are concatenated and treated as a single feature. Translation of the DNA sequence starts in the first frame, with the first base in the feature range. Figure 13-9 illustrates an unfiltered Comparison Range, a joined feature (blue), and a single feature (pink).

Figure 13-9 Restricting the Comparison Range using Features



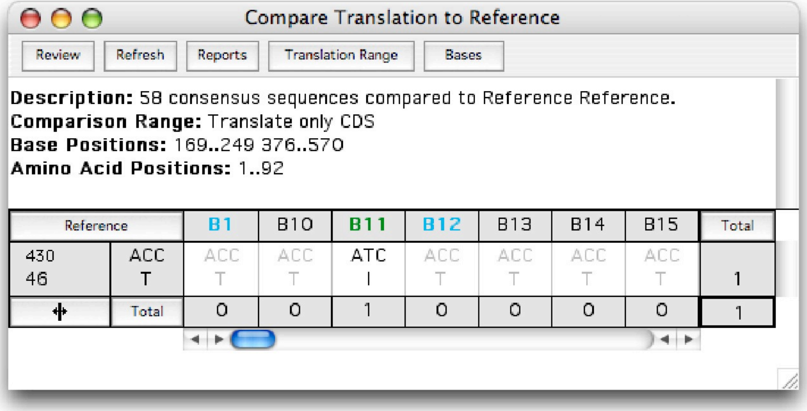
View your Translated Variance Table and go to the button bar and click on the **Translation Range** button. The **Translation Range** dialog appears. You will see that the **Translate Entire Sequence** radio button is checked. This is your default. Leave this radio button checked if you decide to leave your table unfiltered.

If you want to restrict the Comparison Range, click on the **Filter Translation** by: radio button. Once you have clicked the **Filter Translation by**: radio button the **Feature** Key: drop-down menu becomes active. You can now choose which Filter Key to use by making a selection from this menu.

Once you have selected a Feature Key, you will see a list of features annotated with that key in the pane located below the **Feature Key:** drop-down menu. You can select a single feature, a continuous range or a discontinuous range of features. You can also select all the features within this pane by clicking on the **Select All** button.

Once you have made your selection click the **OK** button. The Translated Variance Table will be rebuilt automatically (see Figure 13-10).

Figure 13-10 Translated Variance Table restricted by Translation Range



Description: 58 consensus sequences compared to Reference Reference.
Comparison Range: Translate only CDS
Base Positions: 169..249 376..570
Amino Acid Positions: 1..92

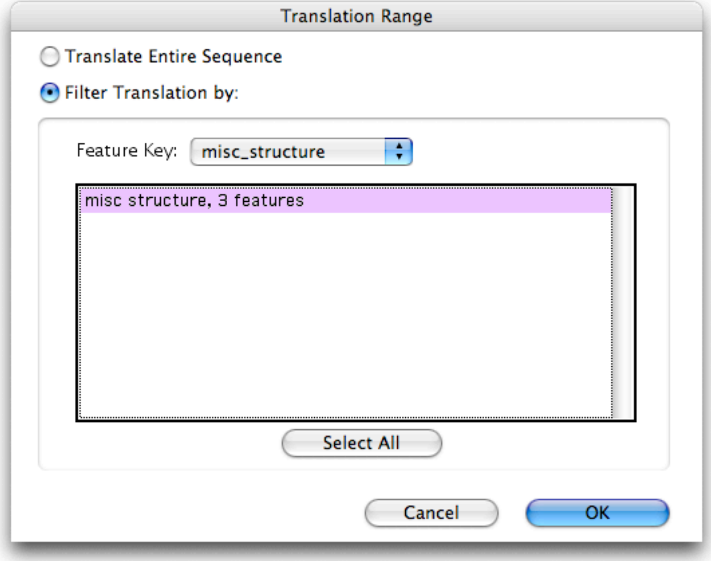
Reference		B1	B10	B11	B12	B13	B14	B15	Total
430 46	ACC T	ACC T	ACC T	ATC I	ACC T	ACC T	ACC T	ACC T	1
✣	Total	0	0	1	0	0	0	0	1

The table header indicates how the Comparison Range was restricted. For some Feature Keys, the actual feature is composed of several components joined together (see Figure 13-9). The Base Positions indicate the exact range of bases used for each component.

Features which share the same Feature Key, such as CDS, and which are joined (the Feature has a Join qualifier), will appear in the **Feature Listing** as CDS [01], CDS [02]. The same Feature will appear in the Translation Range dialog as a single item (see Figure 13-11).

See Chapter 19 "Motifs and Features" and Appendix 30 "Feature Keys and Qualifiers" for more information.

Figure 13-11 Feature Keys with more than one component


Translation Range

○ Translate Entire Sequence

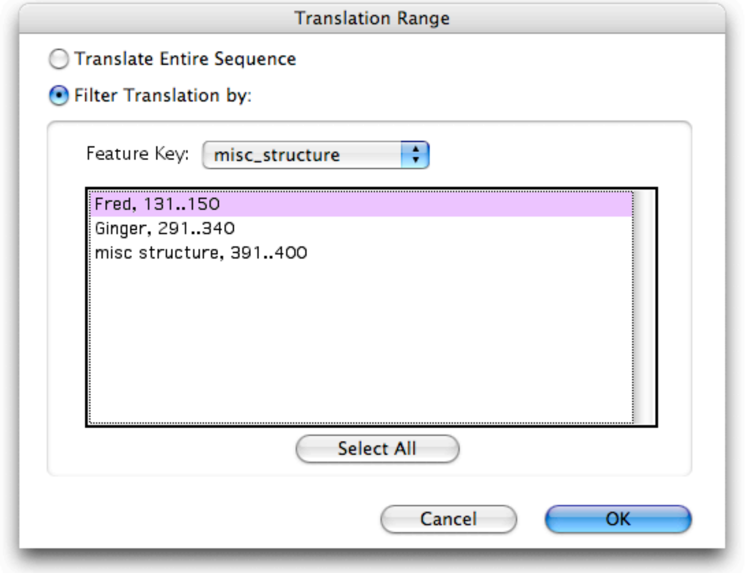
● Filter Translation by:

Feature Key: misc_structure

misc structure, 3 features

If you prefer to have each component of a Joined Feature as a separate item, you must give them different names. You can do this by editing the Feature using the **Edit Features…** command (see figure below).

Figure 13-12 Appearance of features once they have been renamed



Translation Range

○ Translate Entire Sequence

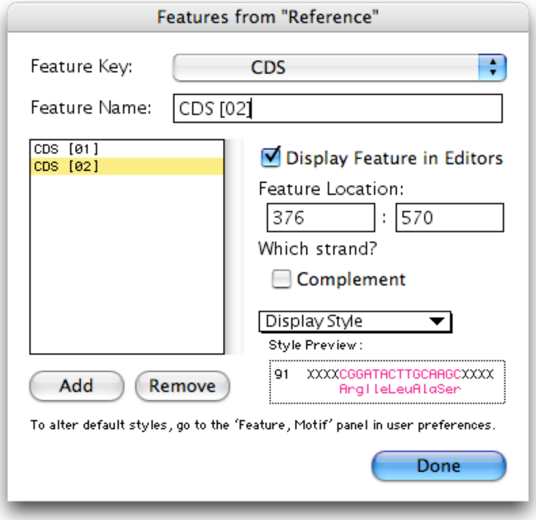
● Filter Translation by:

Feature Key: misc_structure

Fred, 131..150
Ginger, 291..340
misc structure, 391..400

You can also create your own joined features. Use the **Edit Features…** command to create a series of features which share the same name followed by a bracketed number, CDS [01], CDS [02] (see Figure 13-13).

Figure 13-13 Creating a Joined Feature



Features from "Reference"

Feature Key: CDS

Feature Name: CDS [02]

CDS [01]
CDS [02]

☑ Display Feature in Editors

Feature Location: 376 : 570

Which strand?

□ Complement

Display Style

Style Preview:

```
91   XXXXCGGATACTTGCAAGCXXXX
         ArgIleLeuAlaSer
```

To alter default styles, go to the 'Feature, Motif' panel in user preferences.

SETTING THE VARIANCE TABLE DISPLAY OPTIONS

Throughout **Sequencher**, you can customize the display of your sequences using the commands from the **View** menu. When creating a Variance Table, you can also specify a number of display options (see Table 13-1).

Table 13-1 View Options

	Variance Table	Consensus Variance Table	Translated Variance Table	Translated Consensus Variance Table
Base Ambiguities	No	No	No	No
Base Edits	Yes	No	No	No
Display Color Bases	Yes	Yes	No	No
Display Features	Yes	Yes	No	No
Display Motifs	No	No	No	No
Colors As Backgrounds	Yes	Yes	No	No
Display Base Confidences	Yes	No	No	No
Labels	Yes	Yes	Yes	Yes

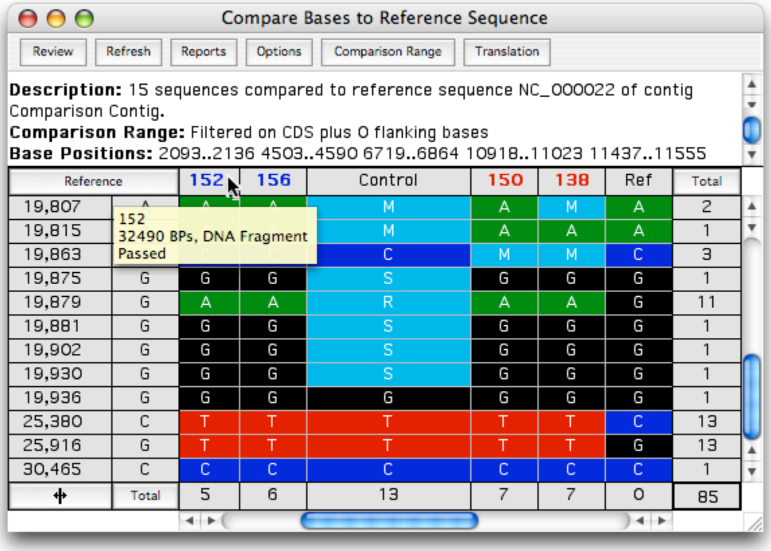
To highlight variants in your data, go to the **View** menu and click on **Display Base Confidences.** This will add a background color to the cells in your Variance Table to indicate that there is an associated confidence or quality score range for the variant. Variants that require additional checking because they have a low confidence or quality score will be highlighted with dark blue. Those with medium or high confidence scores will be highlighted with a medium or pale blue background.

If you click on **Display Color Bases** and **Display Colors as Backgrounds**, it can help you to see patterns in your data (see Figure 13-14). You can use **Base Ambiguities As** to distinguish ambiguous or edited bases. You can also toggle the display of features from the View menu.

You can use the **Label** command to mark sequences in the table. Set your preferences using the **Label & Name** preferences pane. Use the boxes grouped under **Available Labels** in the preference pane to choose label colors and descriptions for marking individual sequences. Select the sequence you want to label by clicking its column header. Go to the **Edit** menu and choose the **Label** submenu. The display name of the sequence will change to match the color of the **Label.** (See Figure 13-4.) Notice how the tool tip provides the name of the sample, its label, and other attributes.

(Refer to Chapter 23 “Customizing **Sequencher** and User Preferences” for more detailed information.)

Figure 13-14 The Variance Table Display Options



Description: 15 sequences compared to reference sequence NC_000022 of contig Comparison Contig.
Comparison Range: Filtered on CDS plus 0 flanking bases
Base Positions: 2093..2136 4503..4590 6719..6864 10918..11023 11437..11555

Reference		152	156	Control	150	138	Ref	Total
19,807	A	A	A	M	A	M	A	2
19,815	[illegible]	[illegible]	A	M	A	A	A	1
19,863	[illegible]	C	C	C	M	M	C	3
19,875	G	G	G	S	G	G	G	1
19,879	G	A	A	R	A	A	G	11
19,881	G	G	G	S	G	G	G	1
19,902	G	G	G	S	G	G	G	1
19,930	G	G	G	S	G	G	G	1
19,936	G	G	G	G	G	G	G	1
25,380	C	T	T	T	T	T	C	13
25,916	G	T	T	T	T	T	G	13
30,465	C	C	C	C	C	C	C	1
✥	Total	5	6	13	7	7	0	85

152
32490 BPs, DNA Fragment
Passed

SETTING THE CONSENSUS CALCULATION FOR A VARIANCE TABLE

You can generate a Variance Table where the exemplar is a consensus sequence rather than the Reference Sequence by using the **Compare Bases To** command from the **Sequence** menu. You generate a consensus Variance Table using the **Compare Consensus to Reference** command from the **Contig** menu.

In either case, the method you choose to calculate your consensus can affect your results. Generally, using the command **Consensus Inclusively** will report more variants than **Consensus by Plurality** while **Consensus by Confidence** will use a method where the bases are calculated to be the best based on confidence scores.

Sequencher recalculates the consensus on the fly, so if you change the methods you are using to compute your consensus, you can reconstruct the Variance Table quickly. Just click on the **Refresh** button in the button bar at the top of the Variance Table to display the results of the new consensus calculation.

For more information on consensus calculations, see Chapter 11 “The Contig Editor”.

VARIANCE TABLE OPTIONS

If you need to change User Preferences for your Variance Table, click on the **Options** button on the Variance Table button bar. The **User Preferences** dialog opens in a new window. (See Chapter 23 “Customizing Sequencher and User Preferences” for more detailed information.)

WORKING WITH THE VARIANCE TABLE

NAVIGATING A VARIANCE TABLE

To navigate through any form of Variance Table, use the arrow keys to move across the rows or down the columns.

When you are working with any Variance Table and only want to focus on differences, use an arrow key with the **Alt key**. This will skip any cells that match the comparison sequence. If you try to move the cursor beyond the edge of the table, you will be alerted by a beep.

Note: If you are using the **Compare Consensus to Reference** command, you will see the consensus base(s) for each contig you have selected in the same position. If you are using the **Compare Translation to Reference** command, you will be viewing the consensus codon for each contig you have selected in the same position.

SORTING THE VARIANCE TABLE

The Variance Table’s **Sort** functions are useful when, for instance, you need to segregate clone sequences that perfectly match your Reference Sequence or cluster sequences of like alleles. The Variance Table simultaneously displays differences and calculates summary information about your selections. The Table lists the total number of differences *for each sequence* at the bottom of that sequence’s column. The Variance Table also displays the total number of differences *for each position* to the right of each row. There are two **Total** buttons, at the bottom left and top right corners of the Variance Table. These act as **Sort** buttons**.**

The **Total** button, on the bottom left of the table, sorts the sample columns so that the samples with the greatest number of differences move to the left. Click **Total** again, and the columns sort in the opposite order.

There are two ways to define the sort order for the rows. You can sort by total number of differences or you can sort by 5’-3’order. In the default, unsorted display of the Variance Table, rows are sorted by base position in ascending (5'- 3') order. When you click on **Total** in the top right corner of the Variance Table, the rows are rearranged so that those with the greatest number of differences sort to the top. Click the **Total** button again to reverse the order.

To restore the order by base position, click on the button in the top left corner of the Variance Table. Click here again to have the rows sort by position in descending (3’-5’) order.

Note: The text on the sort button is context sensitive and will change depending on the **Compare Bases To** command used to generate the Variance Table.

REMOVING COLUMNS FROM THE VARIANCE TABLE

As you start to work with your Variance Table, you may decide that some sequences are no longer of interest. The **Edit** menu's **Remove From Table** command allows you to delete sequences selectively from your Variance Table.

To select a continuous range of sequences, choose the first sequence by clicking on its column header and then, holding down the **Shift** key, select the last sequence in the same fashion. To select a discontinuous list of sequences, hold down the **Ctrl** (Windows) or **Cmd** (Mac) key while clicking on the sequences. Go to the **Edit** menu and choose the **Remove From Table** command.

Note: The **Remove From Table** command does not remove sequences from the underlying contig.

REVIEW MODE

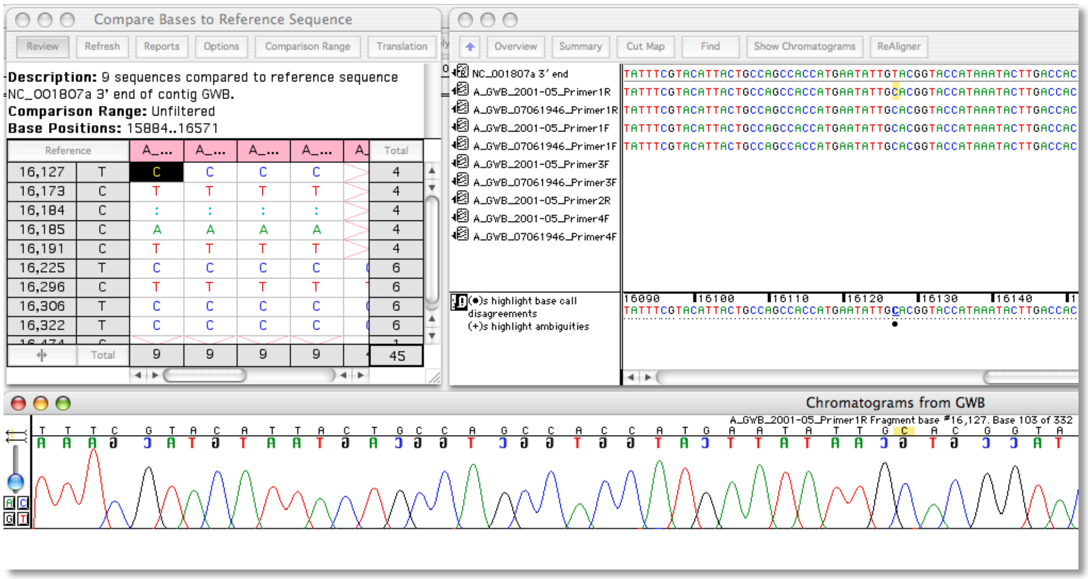
In **Review** mode, the Variance Table lets you view the data supporting each cell (see Figure 13-15). You can access **Review** mode by clicking on the **Review** button in the Variance Table button bar or by double-clicking on any cell in the table. **Sequencher** will open both the **Contig** and **Chromatogram** editors with the associated base or column of bases selected.

You can specify the layout of these three Review windows, the **Variance Table**, the **Contig Editor**, and the **Contig Chromatogram Editor**, and then set the new organization as your default layout. After you have arranged your windows, go to the **Window** menu and choose the **Remember Window Layout** command, then select **Single Variance Table Review** from the submenu**.**

You can edit your data directly in the Variance Table in **Review Mode**. Select the cell to be edited in the table, then type in your desired change. The change will be reflected in the **Variance Table** and the **Contig Editor** immediately.

Note: You can create features in the Variance Table by going to the **Sequence** menu and clicking on the **Mark Selection as Feature** command.

Figure 13-15 Comparing sequences in Review Mode



Description: 9 sequences compared to reference sequence NC_001807a 3' end of contig GWB.
Comparison Range: Unfiltered
Base Positions: 15884..16571

Reference		A_...	A_...	A_...	A_...	A_	Total
16,127	T	C	C	C	C		4
16,173	C	T	T	T	T		4
16,184	C	:	:	:	:		4
16,185	C	A	A	A	A		4
16,191	C	T	T	T	T		4
16,225	T	C	C	C	C	C	6
16,296	C	T	T	T	T	T	6
16,306	T	C	C	C	C	C	6
16,322	T	C	C	C	C	C	6
16,474	C						1
	Total	9	9	9	9	4	45

```
NC_001807a 3' end            TATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer1R       TATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_07061946_Primer1R      TATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer1F       TATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_07061946_Primer1F      TATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer3F
A_GWB_07061946_Primer3F
A_GWB_2001-05_Primer2R
A_GWB_2001-05_Primer4F
A_GWB_07061946_Primer4F

(●)s highlight base call     16090    16100    16110    16120    16130    16140
disagreements                TATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
(+)s highlight ambiguities
```

A_GWB_2001-05_Primer1R Fragment base #16,127. Base 103 of 332

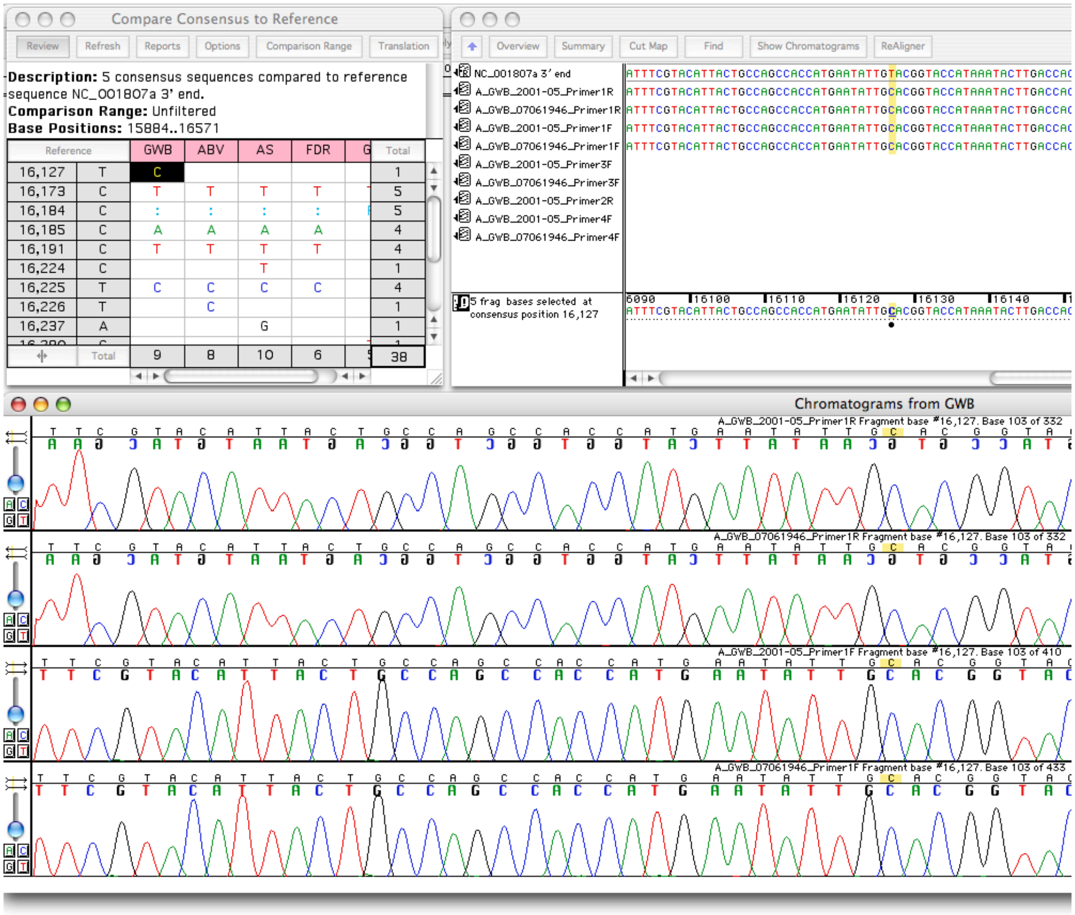
THE REVIEW MODE AND COMPARE CONSENSUS TO REFERENCE

The **Contig** menu's **Compare Consensus to Reference** command creates a Variance Table that reports differences from multiple contigs (see Figure 13-16). The **Review** mode lets you navigate efficiently between a difference in one contig and the equivalent position in the next. Each column in the Variance Table contains the differences from separate contigs. As you change your selection across the Variance Table rows, Sequencher closes the **Contig Editor** and the **Contig Chromatogram Editor** from your previous selection. It will then open the editors corresponding to your current selection.

Just as you can edit contigs from the consensus line in a **Contig** E**ditor**, you can also edit contigs from a selection in the Variance Table **Review** mode. Your selection in the Variance Table corresponds to a selection in the consensus of the underlying contig.

Note: Remember that when you type over a base in the Variance Table, the underlying data in your contig will also change.

Figure 13-16 Comparing contigs in Review Mode



Compare Consensus to Reference

Description: 5 consensus sequences compared to reference sequence NC_001807a 3' end.
Comparison Range: Unfiltered
Base Positions: 15884..16571

Reference		GWB	ABV	AS	FDR	G	Total
16,127	T	C					1
16,173	C	T	T	T	T	T	5
16,184	C	:	:	:	:	F	5
16,185	C	A	A	A	A		4
16,191	C	T	T	T	T		4
16,224	C			T			1
16,225	T	C	C	C	C		4
16,226	T		C				1
16,237	A			G			1
	Total	9	8	10	6	5	38

```
NC_001807a 3' end              ATTTCGTACATTACTGCCAGCCACCATGAATATTGTACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer1R         ATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_07061946_Primer1R        ATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer1F         ATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_07061946_Primer1F        ATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
A_GWB_2001-05_Primer3F
A_GWB_07061946_Primer3F
A_GWB_2001-05_Primer2R
A_GWB_2001-05_Primer4F
A_GWB_07061946_Primer4F

5 frag bases selected at       6090     16100     16110     16120     16130     16140
consensus position 16,127      ATTTCGTACATTACTGCCAGCCACCATGAATATTGCACGGTACCATAAATACTTGACCAC
```

Chromatograms from GWB

A_GWB_2001-05_Primer1R Fragment base #16,127. Base 103 of 332

A_GWB_07061946_Primer1R Fragment base #16,127. Base 103 of 332

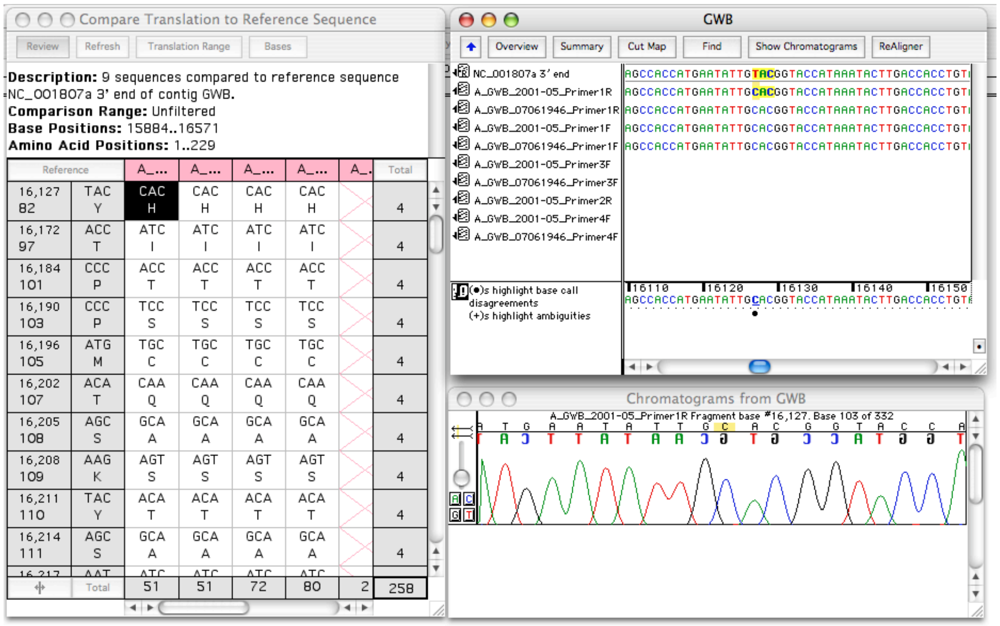
A_GWB_2001-05_Primer1F Fragment base #16,127. Base 103 of 410

A_GWB_07061946_Primer1F Fragment base #16,127. Base 103 of 433

THE REVIEW MODE AND THE TRANSLATED VARIANCE TABLE

The Translated Variance Table also has a **Review** button in its button bar. When you click on this button, you will see the associated **Contig Editor** and **Chromatograms** (see Figure 13-17). Once you have arranged the individual windows to your satisfaction, you can save the layout by going to the **Window** menu and choosing **Remember Window Layout.** Then click on **Single Table Variance Review** from the submenu. When you click on an individual cell in the table, its codon is highlighted in the **Contig Editor** and the first base is selected. You will notice that the equivalent codon in the exemplar sequence is also highlighted.

Figure 13-17 The Translated Variance Table in Review Mode



Description: 9 sequences compared to reference sequence NC_001807a 3' end of contig GWB.
Comparison Range: Unfiltered
Base Positions: 15884..16571
Amino Acid Positions: 1..229

Reference		A_...	A_...	A_...	A_...	A_.	Total
16,127 82	TAC Y	CAC H	CAC H	CAC H	CAC H		4
16,172 97	ACC T	ATC I	ATC I	ATC I	ATC I		4
16,184 101	CCC P	ACC T	ACC T	ACC T	ACC T		4
16,190 103	CCC P	TCC S	TCC S	TCC S	TCC S		4
16,196 105	ATG M	TGC C	TGC C	TGC C	TGC C		4
16,202 107	ACA T	CAA Q	CAA Q	CAA Q	CAA Q		4
16,205 108	AGC S	GCA A	GCA A	GCA A	GCA A		4
16,208 109	AAG K	AGT S	AGT S	AGT S	AGT S		4
16,211 110	TAC Y	ACA T	ACA T	ACA T	ACA T		4
16,214 111	AGC S	GCA A	GCA A	GCA A	GCA A		4
16,217	AAT	ATC	ATC	ATC	ATC		
	Total	51	51	72	80	2	258

NC_001807a 3' end AGCCACCATGAATATTGTACGGTACCATAAATACTTGACCACCTGT
A_GWB_2001-05_Primer1R AGCCACCATGAATATTGCACGGTACCATAAATACTTGACCACCTGT
A_GWB_07061946_Primer1R AGCCACCATGAATATTGCACGGTACCATAAATACTTGACCACCTGT
A_GWB_2001-05_Primer1F AGCCACCATGAATATTGCACGGTACCATAAATACTTGACCACCTGT
A_GWB_07061946_Primer1F AGCCACCATGAATATTGCACGGTACCATAAATACTTGACCACCTGT
A_GWB_2001-05_Primer3F
A_GWB_07061946_Primer3F
A_GWB_2001-05_Primer2R
A_GWB_2001-05_Primer4F
A_GWB_07061946_Primer4F

(●)s highlight base call disagreements
(+)s highlight ambiguities

16110 16120 16130 16140 16150
AGCCACCATGAATATTGCACGGTACCATAAATACTTGACCACCTGT

A_GWB_2001-05_Primer1R Fragment base #16,127. Base 103 of 332

Note: You cannot edit your sequences directly from the Translated Variance Table (although you can do so in the ordinary Variance Table). Any edits to the Translated Variance Table must be made in the **Contig Editor**. In order to see the results of your changes in the Translated Variance Table, click on the **Refresh** button in the button bar.

THE REVIEW MODE AND SISTER TABLES

You can generate a "Sister Table" when you are working in an Individual Variance Table. If you have generated a Variance Table, you can see the Translated Variance Table for the same data set by going to the button bar and clicking on the **Translation** button (see Figure 13-18).

Similarly, if you have already generated a Translated Variance Table, you can see the Variance Table for the DNA sequences of the same data set by going to the button bar and clicking on the **Bases** button.

If you are in **Review** mode and have generated a Sister Table, you can arrange all four windows and save the layout by going to the **Window** menu and choosing **Remember Window Layout**. Then click on **Double Table Variance Review** from the submenu.

VARIANCE TABLE REPORTS

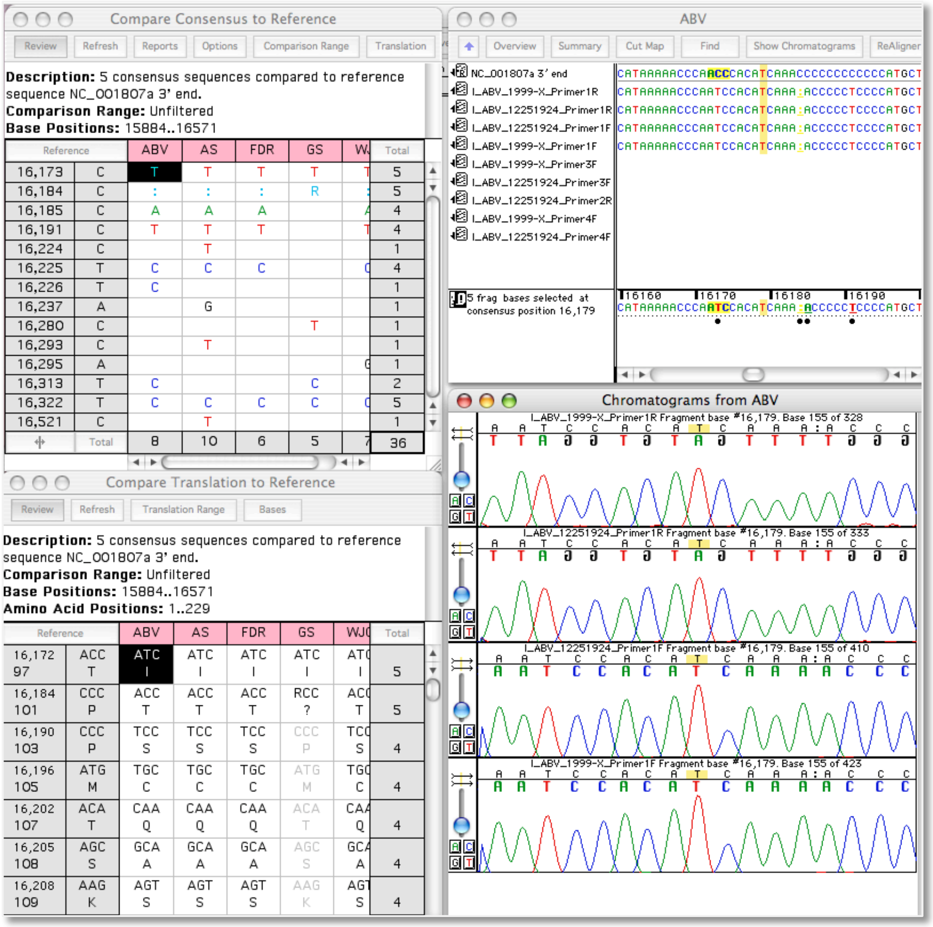
The format of each Variance Table report is broadly similar. The report starts with a Common Header. This header includes the date, the project name, the type of table, the Comparison

Range, base positions, and options. This is followed by a header containing more pertinent information, and finally the body of the report.

VARIANCE TABLE REPORT

The **Variance Table Report** consists of three parts: the Common Header, the Variance Table, and the Comparison Range Coverage table. This table lists the name of each sequence, the type of coverage (complete or incomplete), and the range of bases in the comparison (see Figure 13-19).

Figure 13-18 Review Mode showing Sister Tables



Description: 5 consensus sequences compared to reference sequence NC_001807a 3' end.
Comparison Range: Unfiltered
Base Positions: 15884..16571

Reference		ABV	AS	FDR	GS	WJ	Total
16,173	C	T	T	T	T	T	5
16,184	C	:	:	:	R	:	5
16,185	C	A	A	A		A	4
16,191	C	T	T	T		T	4
16,224	C		T				1
16,225	T	C	C	C		C	4
16,226	T	C					1
16,237	A		G				1
16,280	C				T		1
16,293	C		T				1
16,295	A					G	1
16,313	T	C			C		2
16,322	T	C	C	C	C	C	5
16,521	C		T				1
	Total	8	10	6	5	7	36

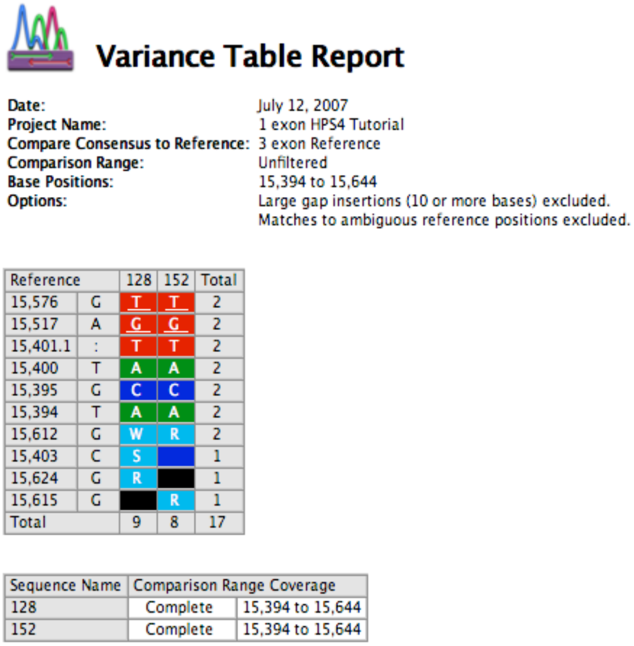
Description: 5 consensus sequences compared to reference sequence NC_001807a 3' end.
Comparison Range: Unfiltered
Base Positions: 15884..16571
Amino Acid Positions: 1..229

Reference		ABV	AS	FDR	GS	WJ(	Total
16,172 97	ACC T	ATC I	ATC I	ATC I	ATC I	ATC I	5
16,184 101	CCC P	ACC T	ACC T	ACC T	RCC ?	ACC T	5
16,190 103	CCC P	TCC S	TCC S	TCC S	CCC P	TCC S	4
16,196 105	ATG M	TGC C	TGC C	TGC C	ATG M	TGC C	4
16,202 107	ACA T	CAA Q	CAA Q	CAA Q	ACA T	CAA Q	4
16,205 108	AGC S	GCA A	GCA A	GCA A	AGC S	GCA A	4
16,208 109	AAG K	AGT S	AGT S	AGT S	AAG K	AGT S	4

```
NC_001807a 3' end              CATAAAAACCCAACCCACATCAAACCCCCCCCCCCCATGCT
I_ABV_1999-X_Primer1R          CATAAAAACCCAATCCACATCAAA:ACCCCCTCCCCATGCT
I_ABV_12251924_Primer1R        CATAAAAACCCAATCCACATCAAA:ACCCCCTCCCCATGCT
I_ABV_12251924_Primer1F        CATAAAAACCCAATCCACATCAAA:ACCCCCTCCCCATGCT
I_ABV_1999-X_Primer1F          CATAAAAACCCAATCCACATCAAA:ACCCCCTCCCCATGCT
I_ABV_1999-X_Primer3F
I_ABV_12251924_Primer3F
I_ABV_12251924_Primer2R
I_ABV_1999-X_Primer4F
I_ABV_12251924_Primer4F

5 frag bases selected at       16160     16170     16180     16190
consensus position 16,179      CATAAAAACCCAATCCACATCAAA:ACCCCCTCCCCATGCT
```

Chromatograms from ABV
I_ABV_1999-X_Primer1R Fragment base #16,179. Base 155 of 328
I_ABV_12251924_Primer1R Fragment base #16,179. Base 155 of 333
I_ABV_12251924_Primer1F Fragment base #16,179. Base 155 of 410
I_ABV_1999-X_Primer1F Fragment base #16,179. Base 155 of 423

Figure 13-19 Extract from Variance Table Report



Variance Table Report

Date:	July 12, 2007
Project Name:	1 exon HPS4 Tutorial
Compare Consensus to Reference:	3 exon Reference
Comparison Range:	Unfiltered
Base Positions:	15,394 to 15,644
Options:	Large gap insertions (10 or more bases) excluded. Matches to ambiguous reference positions excluded.

Reference		128	152	Total
15,576	G	T	T	2
15,517	A	G	G	2
15,401.1	:	T	T	2
15,400	T	A	A	2
15,395	G	C	C	2
15,394	T	A	A	2
15,612	G	W	R	2
15,403	C	S	🟦	1
15,624	G	R	⬛	1
15,615	G	⬛	R	1
Total		9	8	17

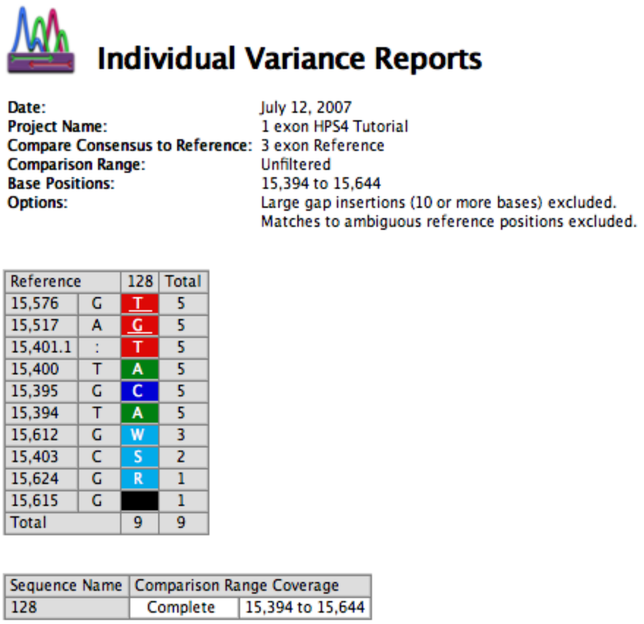
Sequence Name	Comparison Range Coverage	
128	Complete	15,394 to 15,644
152	Complete	15,394 to 15,644

INDIVIDUAL VARIANCE REPORTS

You can create a separate report for each sample in the table. The **Individual Variance Report** lists the variants in a single column. This table is followed by the Comparison Range Coverage table which lists the name of each sequence, whether coverage is complete or incomplete, and the range of bases in the comparison (see Figure 13-20).

Note: This can produce a very long report.

Figure 13-20 Extract from Individual Variance Table Report



Individual Variance Reports

Date:	July 12, 2007
Project Name:	1 exon HPS4 Tutorial
Compare Consensus to Reference:	3 exon Reference
Comparison Range:	Unfiltered
Base Positions:	15,394 to 15,644
Options:	Large gap insertions (10 or more bases) excluded. Matches to ambiguous reference positions excluded.

Reference		128	Total
15,576	G	T	5
15,517	A	G	5
15,401.1	:	T	5
15,400	T	A	5
15,395	G	C	5
15,394	T	A	5
15,612	G	W	3
15,403	C	S	2
15,624	G	R	1
15,615	G	■	1
Total		9	9

Sequence Name	Comparison Range Coverage	
128	Complete	15,394 to 15,644

VARIANCE DETAIL REPORT

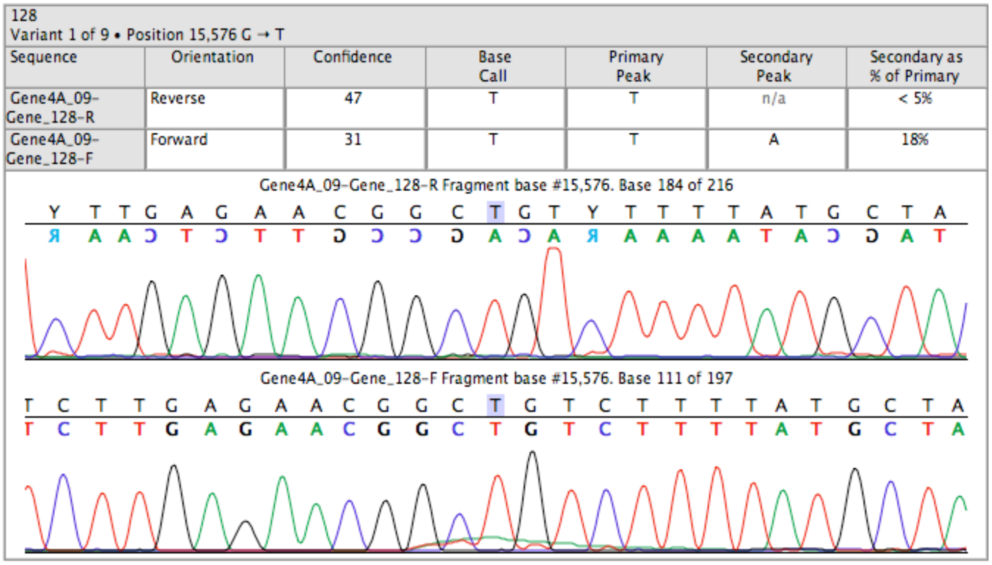
As its name indicates, the **Variance Detail Report** contains comprehensive information on the variants in a Variance Table. Below the Common Header, you will see detailed information for each of the described variants.

The information for each variant consists of the sample name and comparison range for the parent sequence. The Detail Table is divided into two parts. The first part contains the sequence name, orientation, confidence score (where available), base call, primary peak base call, secondary peak base call, and the height of the secondary peak as a percentage of the primary peak. The second part of the table contains a tracelet, which is an extract of the sequence trace of the variant combined with some flanking bases (see Figure 13-21). It only appears in the report if the sequence has associated chromatogram data.

Note: If the Detail Table displays a gap for a specific base position, this will still be given a confidence value. The confidence value will be equal to the average confidence ratings of the bases flanking the gap. **Sequencher** will append an asterisk to the confidence value and a footnote will explain how the value was derived.

When the report contains a great deal of data, it may be useful to use the Chromatogram Scale function which can be accessed by clicking on the Report's **Option** button.

Figure 13-21 Extract from Variance Detail Report



128
Variant 1 of 9 • Position 15,576 G → T

Sequence	Orientation	Confidence	Base Call	Primary Peak	Secondary Peak	Secondary as % of Primary
Gene4A_09-Gene_128-R	Reverse	47	T	T	n/a	< 5%
Gene4A_09-Gene_128-F	Forward	31	T	T	A	18%

Gene4A_09-Gene_128-R Fragment base #15,576. Base 184 of 216

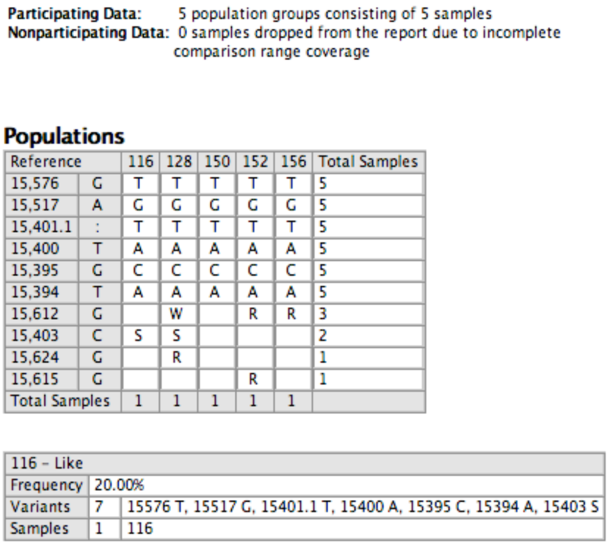
Y T T G A G A A C G G C T G T Y T T T T A T G C T A
Я A A Ɔ T Ɔ T T Ә Ɔ Ɔ Ә A Ɔ A Я A A A A T A Ɔ Ә A T

Gene4A_09-Gene_128-F Fragment base #15,576. Base 111 of 197

T C T T G A G A A C G G C T G T C T T T T A T G C T A
T C T T G A G A A C G G C T G T C T T T T A T G C T A

Note: If the Detail Table displays a gap for a specific base position, this will still be given a confidence value. The confidence value will be equal to the average confidence ratings of the bases flanking the gap. **Sequencher** will append an asterisk to the confidence value and a footnote will explain how the value was derived.

Figure 13-22 Extract from a Population Report



Participating Data: 5 population groups consisting of 5 samples
Nonparticipating Data: 0 samples dropped from the report due to incomplete comparison range coverage

Populations

Reference		116	128	150	152	156	Total Samples
15,576	G	T	T	T	T	T	5
15,517	A	G	G	G	G	G	5
15,401.1	:	T	T	T	T	T	5
15,400	T	A	A	A	A	A	5
15,395	G	C	C	C	C	C	5
15,394	T	A	A	A	A	A	5
15,612	G		W		R	R	3
15,403	C	S	S				2
15,624	G		R				1
15,615	G				R		1
Total Samples		1	1	1	1	1	

116 - Like		
Frequency	20.00%	
Variants	7	15576 T, 15517 G, 15401.1 T, 15400 A, 15395 C, 15394 A, 15403 S
Samples	1	116

WORKING WITH VARIANCE TABLE REPORTS

You now have a number of ways to present the data from your Variance Table. You can use the table's **Reports** function to print your data or to export it. The options vary depending on the data you have chosen to use in your report.

PRINTING A VARIANCE TABLE REPORT

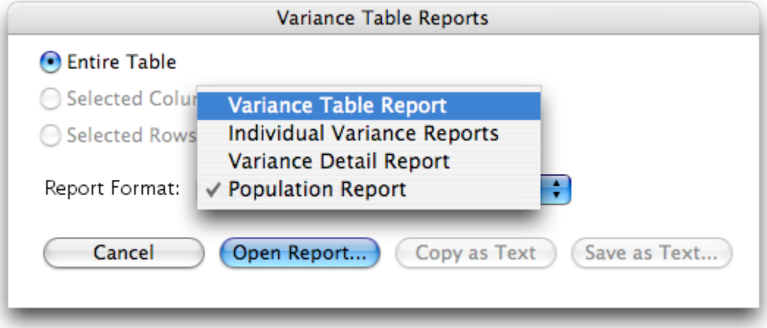
You can print the data from an entire table, selected columns, or selected rows. The type of report you can generate will depend on this initial selection (see Table 13-2).

Table 13-2 Variance Table Print Options

	Variance Table Report	**Individual Variance Table Reports**	**Variance Detail Report**	**Population Report**
Entire Table	Yes	Yes	Yes	Yes
Selected	Yes	Yes	Yes	Yes
Selected Rows	Yes	n/a	Yes	n/a

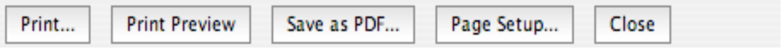
To print all the data in a table, click on the **Reports** button. The **Variance Table Reports** dialog appears (see Figure 13-23). A radio button called **Entire Table** will be checked. Below that you will see a drop-down menu called **Report Format:**. Choose **Variance Table Report**, **Individual Variance Table Reports**, **Variance Detail Report**, or **Population Report**. Click on the **Open Report…** button. The **Sequencher Report Viewer** opens in a new window.

Figure 13-23 Variance Table Reports drop-down menu



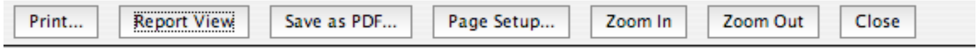
Use the **Print Preview** button from the button bar at the top of the **Report Viewer** to preview your report (see Figure 13-24).

Figure 13-24 Reports Viewer button bar



The **Report View** is replaced by the **Preview View**. Note the **Zoom In** and **Zoom Out** buttons on the button bar (see Figure 13-25). Click on the **Report View** button to return to the previous view.

Figure 13-25 Print Preview button bar



Use the **Page Setup** button to choose paper size. Click on the **Print** button to send the report to a printer. You can also save the report in PDF format by clicking on the **Save as PDF…** button. If you wish to dismiss the window without performing any of these actions, click on the **Close** button.

It is important to note that you can reduce the size of the tracelets within the **Variance Detail Report**, this can decrease the number of pages you need to print out for your report. Having chosen **Variance Detail Report** as your report format, click on the **Options…** button on the button bar. Now either type in the scaling value you wish to use or use the elevator buttons. You can also limit the number of chromatograms displayed to **None** or a specific number by clicking on the **Display** radio button and inserting a number.

PRINTING SELECTED DATA FROM A VARIANCE TABLE

You can print selected data from your table. You have several options. You can remove sample sequences from the table using the **Remove from Table** command. You can print selected variants by choosing selected rows and you can print selected samples by choosing selected columns. You may already have restricted the data in the table by using the **Comparison Range** command.

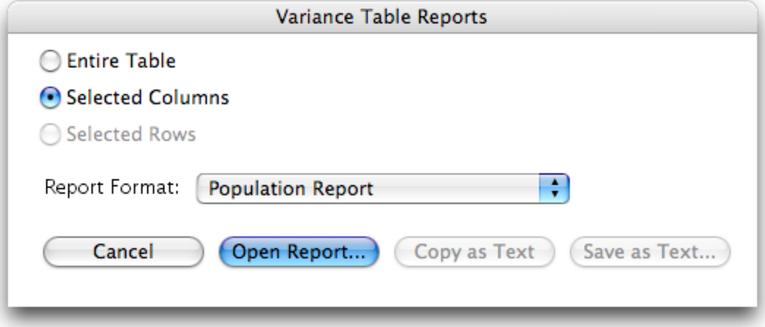
If you want to remove sample sequences from your table, you can either select a continuous range of sequences or a discontinuous range of sequences. For a continuous range of sequences, select the first sequence in your range by clicking on its column header, and then, holding down the **Shift** key, select the last sequence in the same fashion. To select a discontinuous list of sequences, hold down the **Ctrl** (Windows) or **Cmd** (Mac) key while clicking on the sequences. Then go to the **Edit** menu and choose the **Remove From Table** command.

Note: This does not remove sequences from the underlying contig.

If you want to print selected samples from your table, you can choose a continuous or a discontinuous range of sequences by clicking in the column header as described previously.

Click on the **Reports** button. The **Variance Table Reports** dialog appears (see Figure 13-26). A radio button called **Selected Columns** will be checked. Below that you will see a drop-down menu called **Report Format:**. Choose **Variance Table Report**, **Individual Variance Table Reports**, **Variance Detail Report**, or **Population Report**. Click on the **Open Report…** button. The **Sequencher Report Viewer** opens in a new window.

Figure 13-26 Variance Table Reports dialog



Variance Table Reports

○ Entire Table

● Selected Columns

○ Selected Rows

Report Format: Population Report

Use the **Print Preview** button from the button bar at the top of the **Report Viewer** to preview your report. Click on the **Print** button to send the report to a printer.

If you want to print selected variants from your table, you can choose a continuous or a discontinuous range of base positions by clicking the equivalent exemplar base position as described previously.

Click on the **Reports** button. The **Variance Table Reports** dialog appears (see Figure 13-26). A radio button called **Selected Rows** will be checked. Below that you will see a drop-down menu called **Report Format:**. Choose either **Variance Table Report** or **Variance Detail Report**. Click on the **Open Report…** button. The **Sequencher Report Viewer** opens in a new window.

Use the **Print Preview** button from the button bar at the top of the **Report Viewer** to preview your report. Click on the **Print** button to send the report to a printer.

Note: Not all reports are available when you print selected rows.

EXPORTING A VARIANCE TABLE REPORT

You can export the data from an entire table, selected columns, or selected rows. The type of report you can generate will depend on this selection (see Table 13-3).

Table 13-3 Variance Table Report export options

	Variance Table Report	Individual Variance Table Reports	Variance Detail Report	Population Report
Entire Table	Yes	Yes	No	No
Selected Columns	Yes	Yes	No	No
Selected	Yes	n/a	No	n/a

You can remove sample sequences from your table before you export your data. Select a continuous range of sequences. Choose the first sequence by clicking on its column header. Then, holding down the **Shift** key, select the last sequence in the same fashion. To select a discontinuous list of sequences, hold down the **Ctrl** (Windows) or **Cmd** (Mac) key while clicking on the sequences. Go to the **Edit** menu and choose the **Remove From Table** command.

Note: This does *not* remove sequences from the underlying contig. You can then export the remaining data as normal.

To export all the data in the table, click on the **Reports** button. The **Variance Table Reports** dialog appears. A radio button called **Entire Table** will be checked. Below that you will see a drop-down menu called **Report Format:**. Choose either **Variance Table Report** or **Individual Variance Table Reports**. If you choose **Variance Table**, **Sequencher** will generate a single table. If you choose **Individual Variance Table Reports**, each sample will appear in a separate report.

There are four buttons below the **Report Format** drop-down menu, **Cancel**, **Reports**, **Copy as Text**, and **Save as Text... Save as Text…** is the default setting. If you click on the **Save as Text…** button, **Sequencher** presents you with a dialog to assign the name and location of your export. The report is saved in a tabbed text format.

Note: You can select and then remove columns from your table using the **Remove From Table** command before you export your data.

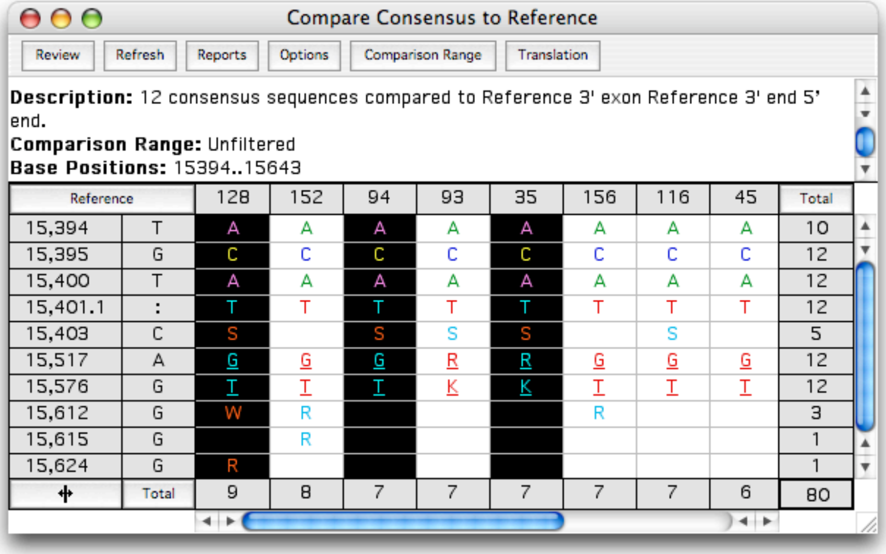
EXPORTING SELECTED SAMPLE SEQUENCES

Each column in the table represents a sequence or consensus. If you need to export specific columns in the table, make your selection by using **Shift+click** for a continuous range or **Ctrl+click** (Windows) or **Cmd+click** (Mac) for a discontinuous range (see Figure 13-27). Then click on the **Reports** button. The **Variance Table Reports** dialog appears, the **Selected Columns** radio button is checked. Choose a **Report Format** from the drop-down menu.

Select **Individual Variance Reports** if you want to save each column (sample) as a separate file. Select **Variance Table Report** if you want the selected columns in one table. Click on the **Save as Text…** button, **Sequencher** presents you with a dialog to assign the name and location of your export.

You can copy your Variance Table into another document or application that accepts tabbed text. To paste the contents of your **Variance Table Report** in the document of your choice, click on the **Copy as Text** button. The data can now be pasted into its new location.

Figure 13-27 Variance Table with discontinuous range of columns selected



Description: 12 consensus sequences compared to Reference 3' exon Reference 3' end 5' end.
Comparison Range: Unfiltered
Base Positions: 15394..15643

Reference		128	152	94	93	35	156	116	45	Total
15,394	T	A	A	A	A	A	A	A	A	10
15,395	G	C	C	C	C	C	C	C	C	12
15,400	T	A	A	A	A	A	A	A	A	12
15,401.1	:	T	T	T	T	T	T	T	T	12
15,403	C	S		S	S	S		S		5
15,517	A	G	G	G	R	R	G	G	G	12
15,576	G	T	T	T	K	K	T	T	T	12
15,612	G	W	R				R			3
15,615	G		R							1
15,624	G	R								1
✢	Total	9	8	7	7	7	7	7	6	80

EXPORTING DATA FOR SELECTED VARIANTS

Each row in the table represents a variant. If you need to export selected rows, make your selection by using **Shift+click** for a continuous range or **Ctrl+click** (Windows) or **Cmd+click** (Mac) for a discontinuous range. Then click on the **Reports** button. The **Variance Table Reports** dialog appears. The **Selected Rows** radio button is checked. There are four buttons, **Cancel, Open Report…, Copy as Text**, and **Save as Text...**. **Save as Text…** is the default setting.

The chosen rows are saved as one table. The options in the **Reports Format:** drop-down menu are limited to **Variance Table Report** and **Variance Detail Report**. Click on the **Save as Text…** button. You will see a dialog prompting you to assign the name and location of your export. You can type a name for the table in this window. Click on the **Save** button to save your table as a tabbed text file.

Although the **Selected Rows** radio button is checked, you can change your selection to include all rows by clicking on the **Entire Table** radio button.

You can copy your Variance Table into another document or application that accepts tabbed text. To paste the contents of your **Variance Table Report** in the document of your choice, click on the **Copy as Text** button. The data can now be pasted into its new location.

TRANSLATED VARIANCE TABLE REPORT

TRANSLATED VARIANCE TABLE REPORT

The **Translated Variance Table Report** consists of three parts: the header information, the Comparison Range Coverage, and the Translated Variance Table.

The Translated Variance Table contains the name of the sample sequence, the variant codon, the amino acid residue, and the type of coverage (whether the coverage is complete or incomplete) for each variant in the table. The Comparison Range Coverage table lists the name of each sequence and the range of bases in the comparison.

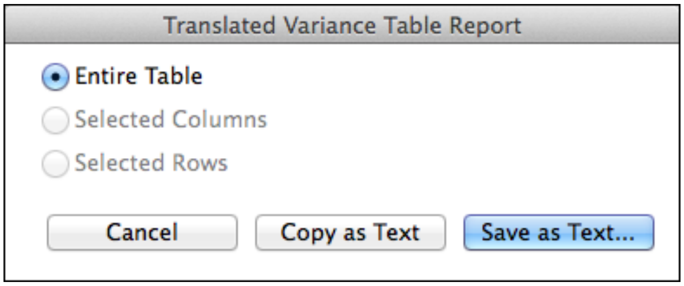
WORKING WITH A TRANSLATED VARIANCE TABLE REPORT

There are a number of ways to present the data from a Translated Variance Table. You can export the data from an entire table, selected columns, or selected rows. You can also remove data from the table before you export it.

EXPORTING A TRANSLATED VARIANCE TABLE REPORT

To export all the data in the table, click on the **Reports** button. The **Translated Variance Table Reports** dialog appears (see Figure 13-28). A radio button called **Entire Table** will be checked.

Figure 13-28 Translated Variance Table Reports dialog



You will see three enabled buttons in the **Translate Variance Table Reports** dialog, **Cancel**, **Copy as Text**, and **Save as Text...**

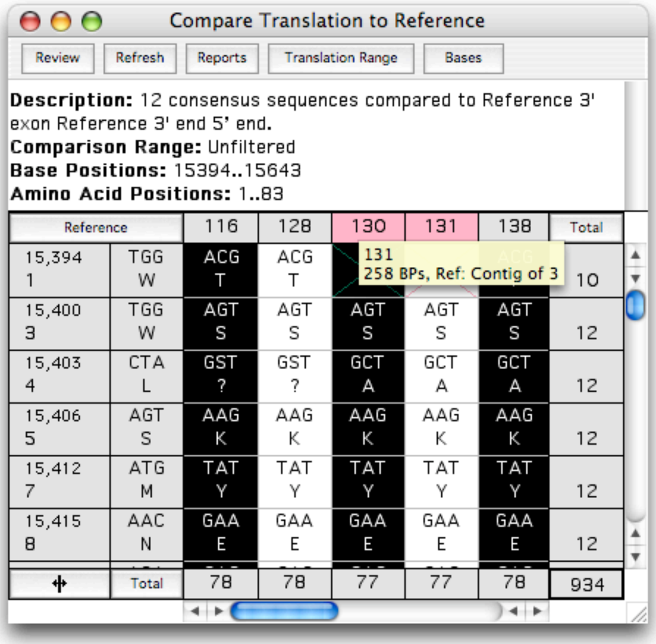
Save as Text… is the default setting. If you click on the **Save as Text…** button, **Sequencher** presents you with a dialog to assign the name and location of your export. The report is saved in a tabbed text format.

If you prefer to copy your Translated Variance Table to another document or application that accepts tabbed text, click on the **Copy as Text** button. The data can now be pasted into its new location.

EXPORTING SELECTED SAMPLE SEQUENCES FROM A TRANSLATED VARIANCE TABLE

Each column in the table represents a sequence or consensus. If you need to export specific columns in the table, make your selection by using **Shift+click** for a continuous range or **Ctrl+click** (Windows) or **Cmd+click** (Mac) for a discontinuous range (see Figure 13-29). Then click on the **Reports** button. The **Translated Variance Table Reports** dialog appears (see figure above). A radio button called **Selected Columns** will be checked.

Figure 13-29 Discontinuous selection of columns



Description: 12 consensus sequences compared to Reference 3' exon Reference 3' end 5' end.
Comparison Range: Unfiltered
Base Positions: 15394..15643
Amino Acid Positions: 1..83

Reference		116	128	130	131	138	Total
15,394 1	TGG W	ACG T	ACG T			ACG	10
15,400 3	TGG W	AGT S	AGT S	AGT S	AGT S	AGT S	12
15,403 4	CTA L	GST ?	GST ?	GCT A	GCT A	GCT A	12
15,406 5	AGT S	AAG K	AAG K	AAG K	AAG K	AAG K	12
15,412 7	ATG M	TAT Y	TAT Y	TAT Y	TAT Y	TAT Y	12
15,415 8	AAC N	GAA E	GAA E	GAA E	GAA E	GAA E	12
	Total	78	78	77	77	78	934

131
258 BPs, Ref: Contig of 3

Choose one of the three enabled buttons in the **Translated Variance Table Reports** dialog, **Cancel**, **Copy as Text**, and **Save as Text... Save as Text…** is the default setting. If you click on the **Save as Text…** button, **Sequencher** presents you with a dialog to assign the name and location of your export. The report is saved in a tabbed text format.

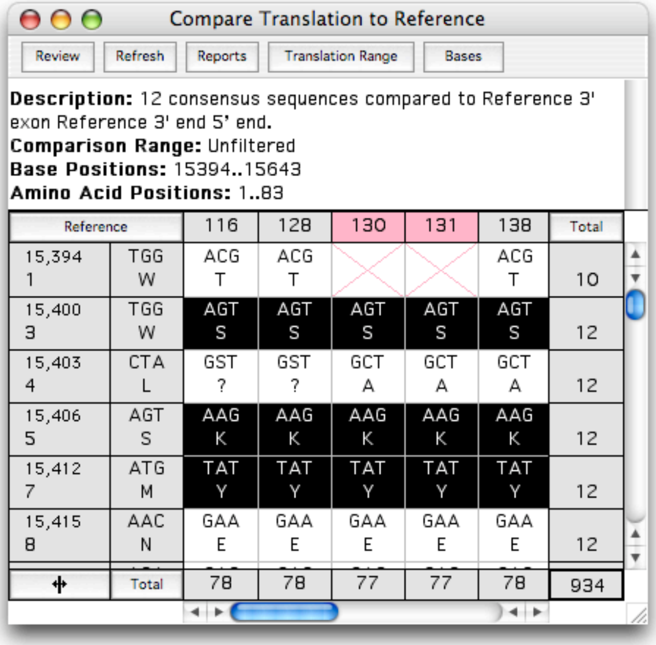
Although the **Selected Columns** radio button is checked, you can change your selection to include all rows by clicking on the **Entire Table** radio button.

You can copy your Translated Variance Table into another document or application that accepts tabbed text. To copy the contents of your Translated Variance Table Report, click on the **Copy as Text** button. The data can now be pasted into its new location.

EXPORTING DATA FOR SELECTED VARIANTS

Each row in the table represents a variant. If you need to export selected rows, make your selection by using **Shift+click** for a continuous range or **Ctrl+click** (Windows) or **Cmd+click** (Mac) for a discontinuous range (see Figure 13-30). Then click on the **Reports** button.

Figure 13-30 Translated Variance Table - selected variants

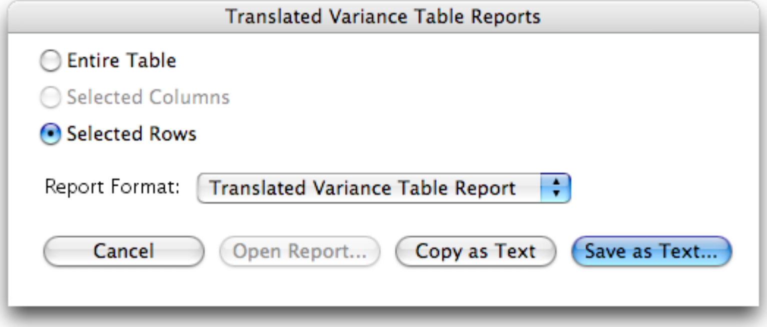


Description: 12 consensus sequences compared to Reference 3' exon Reference 3' end 5' end.
Comparison Range: Unfiltered
Base Positions: 15394..15643
Amino Acid Positions: 1..83

Reference		116	128	130	131	138	Total
15,394 1	TGG W	ACG T	ACG T			ACG T	10
15,400 3	TGG W	AGT S	AGT S	AGT S	AGT S	AGT S	12
15,403 4	CTA L	GST ?	GST ?	GCT A	GCT A	GCT A	12
15,406 5	AGT S	AAG K	AAG K	AAG K	AAG K	AAG K	12
15,412 7	ATG M	TAT Y	TAT Y	TAT Y	TAT Y	TAT Y	12
15,415 8	AAC N	GAA E	GAA E	GAA E	GAA E	GAA E	12
✣	Total	78	78	77	77	78	934

The **Translated Variance Table Reports** dialog appears (see Figure 13-31). Although the **Selected Rows** radio button is checked, you can change your selection to include all rows by clicking on the **Entire Table** radio button.

Figure 13-31 Translated Variance Table Reports dialog - Selected Rows



Translated Variance Table Reports

○ Entire Table

○ Selected Columns

● Selected Rows

Report Format: Translated Variance Table Report

Choose one of three buttons in the **Translated Variance Table Reports** dialog, **Cancel, Copy as Text**, and **Save as Text..**. **Save as Text…** is the default setting. The rows you have selected will be saved in one table. Click on the **Save as Text…** button. A dialog to assign the name and location of your export appears. You can type a name for the table in this window. Click on the **Save** button to save your table as a tabbed text file.

You can copy your Translated Variance Table into another document or application that accepts tabbed text. To paste the contents of your Translated Variance Table Report, click on the **Copy as Text** button.

REMOVING DATA FROM A TABLE BEFORE EXPORT

If you need to reduce the amount of data in your table, you can remove selected sample sequences.

Make your selection by using **Shift+click** for a continuous range or **Ctrl+click** (Windows) or **Cmd+click** (Mac) for a discontinuous range.

Go to the **Edit** menu and choose the **Remove From Table** command. Your selection will be removed. You can then export the remaining data as usual.

Note: This command does *not* remove sequences from the underlying contig.

14. THE SUMMARY REPORT

In this chapter, we explain how to use the **Summary Report**. This is a static report that has a number of options making it a useful addition to the Variance Table for mutation hunting, verifying re-sequenced fragments, comparisons to a Reference Sequence, or just as a report for your lab notebook.

THE SUMMARY REPORT VIEW

VIEW BY SUMMARY

You can display this report by going to the **Contig Editor** and clicking on the **Summary** button in either the **Bases View** window or the **Overview**…. Alternatively, you can go to the **View** menu, choose Bases, Map, Overview…, and then click on **Summary Report**.

You will be presented with a new window which shows the assembled sequences within your chosen contig. There are five buttons arranged on a button bar (see below).

Table 14-1 Button bar

Button	Description
Bases	This button returns you to the **Contig Editor**.
Overview	This button returns you to the **Contig Overview**.
Ruler	This button brings up a simple text editor-like ruler.
Options	This button allows you to control which information will be displayed.
Find	This button brings up a window into which you can type a substring and perform a search (see Chapter 20 “Finding Items”).

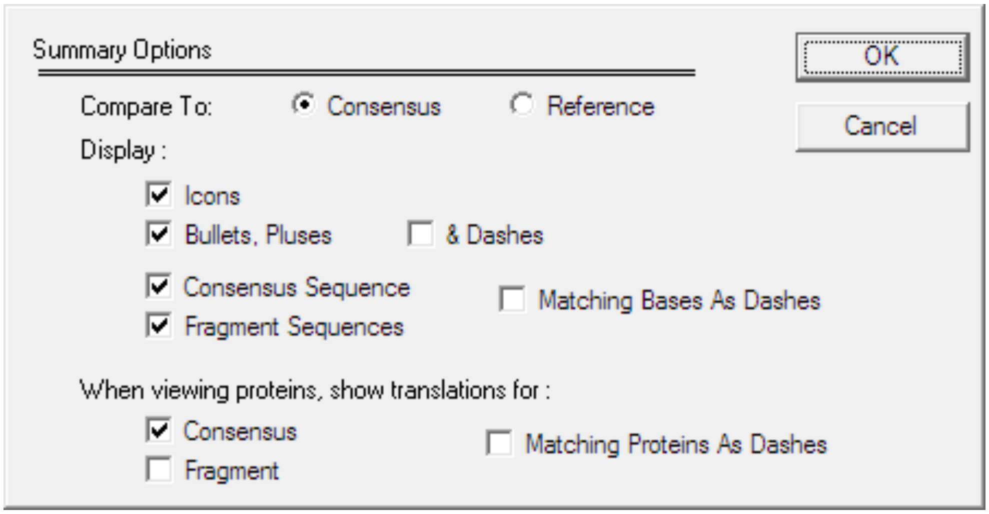
SUMMARY REPORT DISPLAY OPTIONS

Sequencher has a default format for the **Summary Report** but you can change it to display different types of information. Click on the **Options** button or go to the **View** menu and choose **View Options…**

The **Summary Options** are divided into those that affect the sequence fragments and those that affect the display of protein translations. You can change the consensus calculation while in this view. Any change you make will be reflected immediately in the report.

Sequencher displays the window shown in Figure 14-1 Summary Report Options.

Figure 14-1 Summary Report Options



Summary Options

Compare To: ● Consensus ○ Reference

Display :

☑ Icons

☑ Bullets, Pluses ☐ & Dashes

☑ Consensus Sequence ☐ Matching Bases As Dashes

☑ Fragment Sequences

When viewing proteins, show translations for :

☑ Consensus ☐ Matching Proteins As Dashes

☐ Fragment

COMPARE TO CONSENSUS OR REFERENCE SEQUENCE

If you have a Reference Sequence in your contig, you can choose whether to display a **Summary Report** compared to the consensus sequence or to the assembled Reference Sequence. The default setting is **Compare To: Consensus** but you can change this by clicking on the **Reference** radio button. Some of the **Display** options are context sensitive and will change to reflect your choice.

The **Reference** radio button will be grayed out when the selected contig does not contain a Reference Sequence.

ICONS

If you check the box for the **Icons** display, you will see the sequence fragment icons positioned to the left of the sequence name. Unchecking this option will hide the icons.

Note: When you have a Reference Sequence in your contig, its icon is distinguished by the addition of an "R".

BULLETS, PLUSES AND DASHES

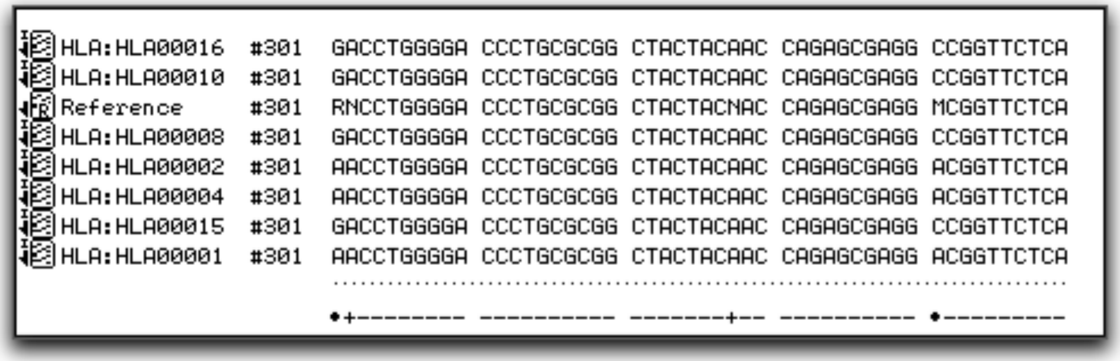
Clicking the **Bullets, Pluses** boxes highlights disagreements and ambiguities in a line below the consensus. If you also click on **& Dashes**, you are marking positions where the bases in individual sequences match so that the disagreements are more visible.

Note: You can display Bullets, Pluses without the & Dashes but you cannot display & Dashes without the Bullets, Pluses. However, see Matching Bases As Dashes below.

CONSENSUS SEQUENCE

You can choose to hide the consensus sequence by clicking on the **Consensus Sequence** box. Figure 14-2 shows a summary line without the consensus line but with **Bullets, Pluses** and **& Dashes** all checked.

Figure 14-2 Summary Report with no consensus line, with bullets, pluses and dashes



```
HLA:HLA00016   #301   GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG CCGGTTCTCA
HLA:HLA00010   #301   GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG CCGGTTCTCA
Reference      #301   RNCCTGGGGA CCCTGCGCGG CTACTACNAC CAGAGCGAGG MCGGTTCTCA
HLA:HLA00008   #301   GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG CCGGTTCTCA
HLA:HLA00002   #301   AACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACGGTTCTCA
HLA:HLA00004   #301   AACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACGGTTCTCA
HLA:HLA00015   #301   GACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG CCGGTTCTCA
HLA:HLA00001   #301   AACCTGGGGA CCCTGCGCGG CTACTACAAC CAGAGCGAGG ACGGTTCTCA
                      ......................................................

                      •+-------- ---------- -------+-- ---------- •---------
```

FRAGMENT SEQUENCES

You can choose to hide the fragment sequence by clicking on the **Fragment Sequences** box. This gives you a very compressed report that can be useful when viewing protein translations. Figure 14-2 above shows a summary line without the consensus line, but with **Bullets, Pluses** and **& Dashes** all checked.

MATCHING BASES AS DASHES

Matching Bases As Dashes is a very powerful option when you are interested in any kind of mutation detection. When you check this option, all the bases that match each other are converted to dashes.

If you have selected **Display Color Bases**, the dashes will be marked in the color of the original base call. If you also go to the **View** menu and chose **Colors as Background,** each dash will appear embedded in a small color block. Bases that disagree are displayed as characters and as such stand out in stark relief. With this option, you can easily see mutation hot spots or clusters and potential SNPs.

PROTEIN TRANSLATIONS AND THE SUMMARY REPORT

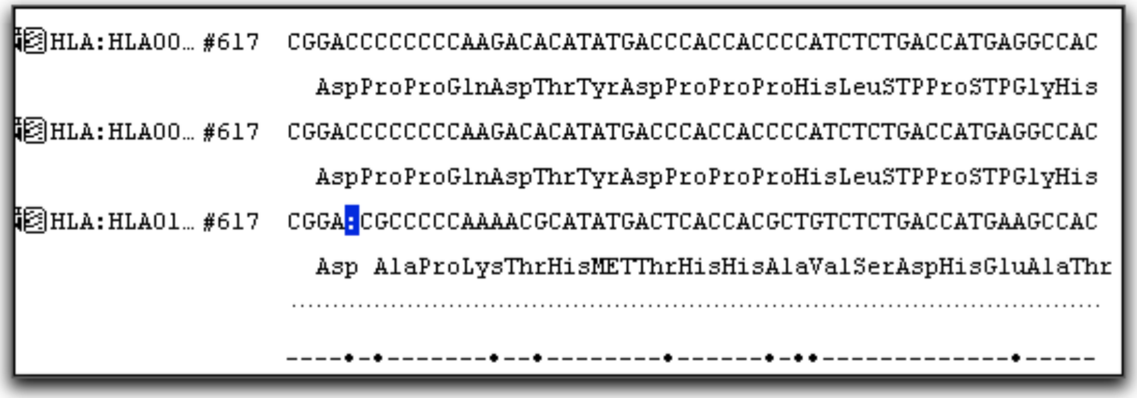
Sequencher can show protein translations for individual sequences and mark matching amino acids with dashes. This helps you see whether a particular discrepancy in an alignment has functional implications or whether it may simply be a silent mutation. If you combine options, you can turn off the DNA display altogether and just focus on your translations. Alternately, you can use a mixture of options to tailor the report to your specific needs.

ENABLING THE PROTEIN DISPLAY

To enable the protein display in the summary view, go to the **View** menu and click on **Translation**. You can then choose from the following list of menu suboptions: **Single Stranded, Double Stranded, Protein 1st Frame, Protein 2nd Frame, Protein 3rd Frame** and**, Protein All 3 Frames**.

To see the protein translations for the opposite strand, go the **View** menu and select **Reverse & Comp**. Then select the desired option. Where the assembly algorithm has inserted a gap in a sequence, **Sequencher** will translate the codon correctly.

Figure 14-3 Insert Summary Report displaying translation across a sequence gap



```
HLA:HLA00… #617  CGGACCCCCCCAAGACACATATGACCCACCACCCCATCTCTGACCATGAGGCCAC
                   AspProProGlnAspThrTyrAspProProProHisLeuSTPProSTPGlyHis
HLA:HLA00… #617  CGGACCCCCCCAAGACACATATGACCCACCACCCCATCTCTGACCATGAGGCCAC
                   AspProProGlnAspThrTyrAspProProProHisLeuSTPProSTPGlyHis
HLA:HLA01… #617  CGGA:CGCCCCCAAAACGCATATGACTCACCACGCTGTCTCTGACCATGAAGCCAC
                   Asp AlaProLysThrHisMETThrHisHisAlaValSerAspHisGluAlaThr
                 ........................................................

                 ----•-•--------•--•---------•-------•-••--------------•-----
```

PROTEIN TRANSLATIONS FOR THE CONSENSUS OR REFERENCE SEQUENCE

This option is context sensitive. If you do not have a Reference Sequence or you have selected the **Consensus** radio button, you will be able to show protein translations for the consensus. Click on the **Consensus** checkbox at the bottom of the **Summary Options** dialog.

If you have a Reference Sequence in your contig and you have selected the **Reference** radio button, you will be able to show translations for the Reference Sequence. Click on the **Reference** checkbox at the bottom of the **Summary Options** dialog.

PROTEIN TRANSLATIONS FOR FRAGMENTS

To show protein translations for fragments, click on the **Fragment** checkbox in the **Summary Options** dialog.

MATCHING PROTEINS AS DASHES

To show Matching Proteins As Dashes, click on that checkbox in the **Summary Options** dialog.

USING THE FORMATTING RULER

You can use the formatting ruler to change the blocking (for example, to groups of 3s instead of 10s) or to show protein translations. (See Chapter 8 “The Sequence Editor” for more information on the formatting ruler.)

Click on the **Ruler** button in the button bar or go to the **View** menu and choose **Display Format Ruler**. You can use the smaller font options in conjunction with the landscape page format to obtain full-width displays of your assembled sequences for use in posters and figures.

REVERSE AND COMPLEMENT A CONTIG

You can change the orientation of the contig while displaying any of its views by going to the **View** menu and choosing **Reverse & Comp.** You can display the contig’s original orientation by choosing **Reverse & Comp** again.

15. CHROMATOGRAMS

In this chapter, you will learn how to use the chromatograms from an automated sequencer to help you when editing your contigs. We discuss how **Sequencher** displays traces and secondary peaks to find heterozygotes and how you can edit bases from traces or revert to experimental data.

WORKING WITH AUTOMATED SEQUENCER DATA

When you import a trace file from an automated sequencer, **Sequencher** imports the chromatograms and the confidence values if these are available.

Note: If you import a text file containing just ASCII characters, you will not have imported traces. To import traces, you must import the trace file. Text files from an automated sequencer will be only a couple of kilobytes in size, whereas trace files will be well over 50 kilobytes.

When you copy a **Sequencher** project file from one machine to another, you don't need to copy the trace files separately because **Sequencher** stores the trace information along with everything else in the project file. **Sequencher** compresses the imported traces to conserve disk space, so don't be alarmed if a project file containing several sequences from an automated sequencer is smaller than any one of the trace files.

VIEWING CHROMATOGRAMS FROM A SEQUENCE EDITOR

Sequencher's chromatogram display is tightly integrated with its other editors. You can view the entire original trace for a sequence by opening the **Sequence Editor** for that sequence. Click on the **Show Chromatogram** button on the button bar or go to the **Window** menu and choose **Chromatogram.** You can scroll either vertically (the default) or horizontally. To change orientation, click the appropriate button in the left bottom corner of the sequence chromatogram window.

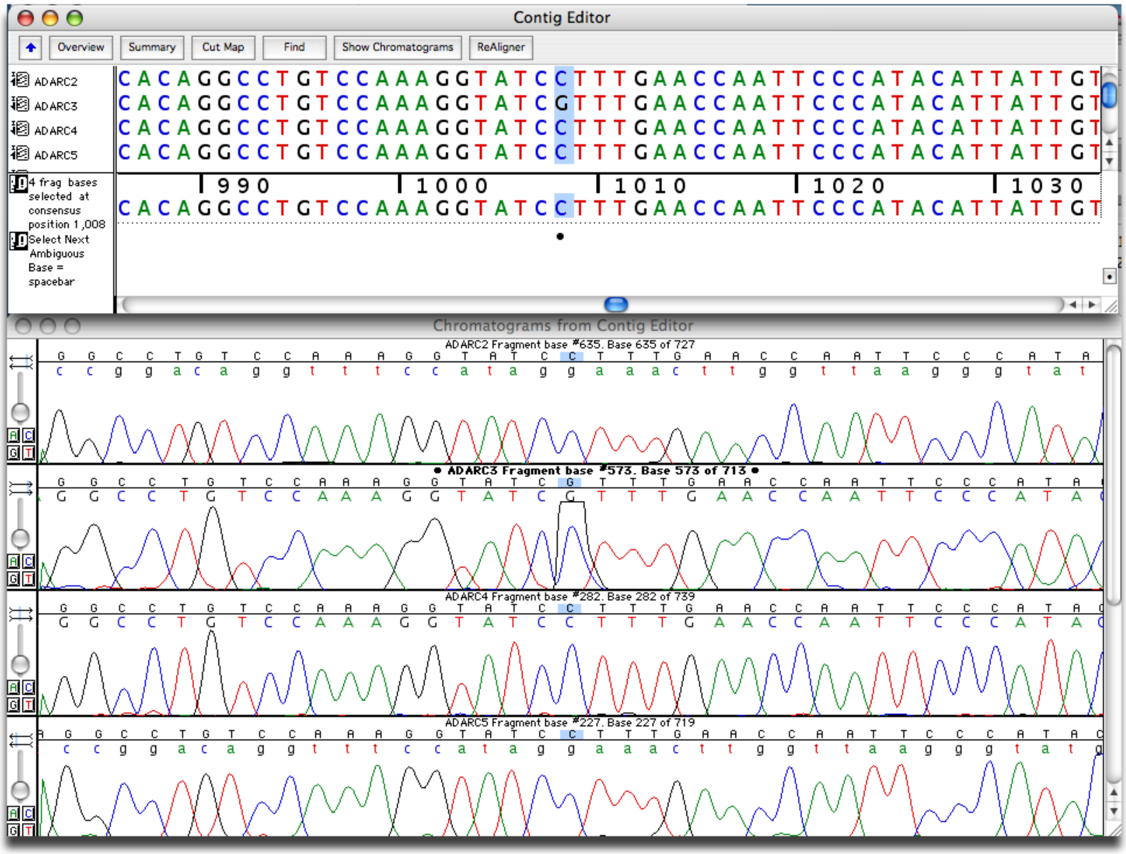
Note: Remember that, even if a sequence is incorporated into a contig, you can still open its editor by double-clicking the sequence name in the list at the left of the **Contig Editor**.

VIEWING CHROMATOGRAMS FROM A CONTIG EDITOR

The **Sequencher Contig Editor** allows you to view simultaneously all chromatogram data relevant to a consensus base call. If you are at the 5' end of the contig, **Sequencher** selects the most 5' base in the consensus and displays all the associated traces. If you are elsewhere in the contig, **Sequencher** selects the center-most consensus base in the contig display and displays the associated traces. This allows you to check the signal strength of any peak in conjunction with its base call and edit your data accordingly.

Select a contig column by clicking on its base in the consensus line (Figure 15-1). Then click on the **Show Chromatograms** button or go to the **Window** menu and choose **Chromatogram. Sequencher** opens a window to display the traces (Figure 15-1). The bases you highlighted in the **Contig Editor** will also be highlighted in the trace windows.

Figure 15-1 Selected column chromatograms


Contig Editor
Overview
Summary
Cut Map
Find
Show Chromatograms
ReAligner
ADARC2
ADARC3
ADARC4
ADARC5
CACAGGCCTGTCCAAAGGTATCCTTTGAACCAATTCCCATACATTATTGT
CACAGGCCTGTCCAAAGGTATCGTTTGAACCAATTCCCATACATTATTGT
CACAGGCCTGTCCAAAGGTATCCTTTGAACCAATTCCCATACATTATTGT
CACAGGCCTGTCCAAAGGTATCCTTTGAACCAATTCCCATACATTATTGT
4 frag bases selected at consensus position 1,008
Select Next Ambiguous Base = spacebar
990
1000
1010
1020
1030
CACAGGCCTGTCCAAAGGTATCCTTTGAACCAATTCCCATACATTATTGT
Chromatograms from Contig Editor
ADARC2 Fragment base #635, Base 635 of 727
ADARC3 Fragment base #573, Base 573 of 713
ADARC4 Fragment base #282, Base 282 of 739
ADARC5 Fragment base #227, Base 227 of 719
A C
G T

When you click the **Show Chromatograms** button in a region without trace data, you will see a “No Chromatogram Data” message in the **Contig Chromatogram** window.

When you move the cursor back into a region where there are sequences with chromatogram data, the traces will reappear.

UNDERSTANDING THE TRACE DISPLAY

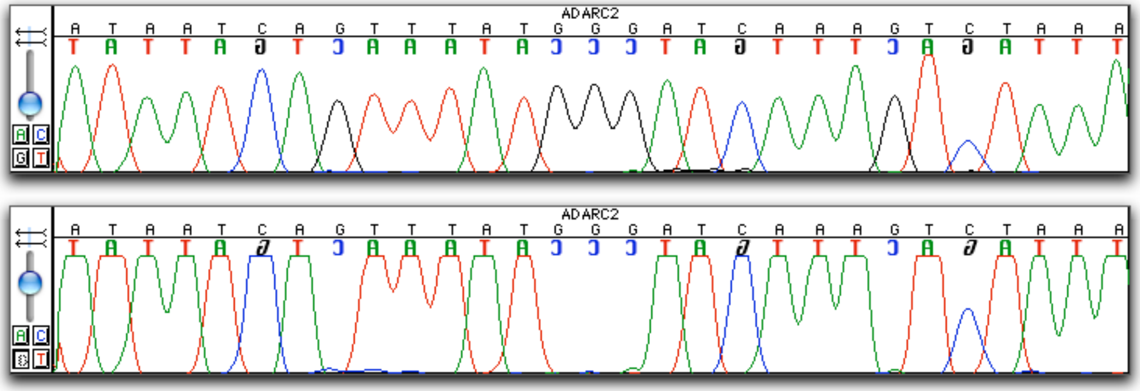
Above the traces are two lines of bases. The upper line displays the sequence as it currently appears in the editor, including any edits you may have made. The lower line displays the base calls as originally imported, including the trimmed data. The arrows to the left show the orientation and relative lengths of the trimmed and untrimmed data.

Below the arrows is a slider for changing the height of the trace. Below the slider are four buttons that let you hide the lanes for any or all of the traces. Click on the button to hide any lane. Click it again to display the lane. Figure 15-2 shows a contig with adjusted scales and hidden lanes.

Scrolling through a single chromatogram is simple. Position your cursor on the small rectangular slider button to the right of the trace. With the mouse button held down, you can move the slider down or up, depending on your current position within the sequence. Alternately, you can use the left arrow and right arrow keys to step through the sequence trace base by base.

When you are working in a contig, if you highlight a base in the consensus line you can use the arrow keys to move through the consensus sequence. Note that, as you do so, all the contig chromatograms will move left or right, in step with the highlighted base.

Figure 15-2 Trace with hidden lanes and adjusted scales


ADARC2
A T A A T C A G T T T A T G G G A T C A A A G T C T A A A
ADARC2
A T A A T C A G T T T A T G G G A T C A A A G T C T A A A
A C
G T

You can specify the settings for trace display by going to the **Window** menu and choosing **User Preferences** and selecting C**hromatogram**. You can specify the height of the trace, the width from peak to peak, the way the lanes are identified, and various screen display attributes. (See Chapter 23 on “Customizing Sequencher and User Preferences” for more information.)

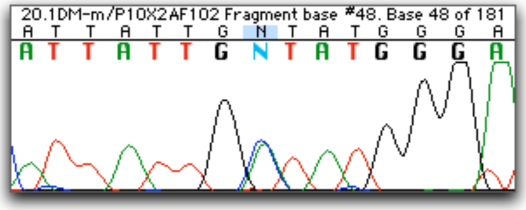
If you select the **Contig Chromatogram** section, you can specify how many columns of traces you want to view at once, whether **Sequencher** should scroll to the selected column when you make a new selection in the contig, and where the window should appear.

Note: If you hold down the **Ctrl** (Windows) or **Command** (Mac) key while changing the scale, the new scale will apply to all the traces in that window. If you hold down the **Ctrl** (Windows) or **Command** (Mac) key while clicking any of the buttons to hide a lane, the lane(s) will be hidden for all the traces in the window. (See Figure 15-2 above.)

DISPLAY SECONDARY PEAK

You can identify heterozygotes by identifying the second-highest peak beneath the primary peak. If you have a particular candidate for a heterozygote, first open the chromatogram for the individual sequence then click on the base call(s) as shown in Figure 15-3.

Figure 15-3 Trace with base selected



20.1DM-m/P10X2AF102 Fragment base #48. Base 48 of 181

A T T A T T G N T A T G G G A

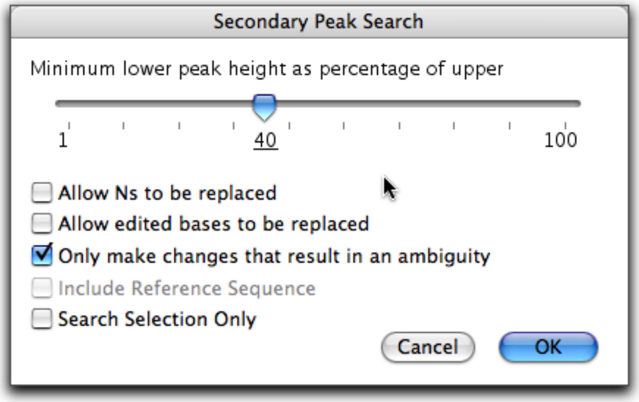
A T T A T T G N T A T G G G A

Go to the **Sequence** menu and choose **Call Secondary Peaks…**. A dialog lets you specify how the secondary peak should be called (Figure 15-4).

The slider lets you specify how significant the second-highest peak must be to generate a change. In Figure 15-4 above, the slider is set so the secondary peak has to be at least 75% as high as the highest peak.

If you want Ns to be replaced, then select the Allow Ns to be replaced checkbox.

Figure 15-4 Secondary peak parameters



Secondary Peak Search

Minimum lower peak height as percentage of upper

1 40 100

☐ Allow Ns to be replaced

☐ Allow edited bases to be replaced

☑ Only make changes that result in an ambiguity

☐ Include Reference Sequence

☐ Search Selection Only

Cancel OK

If you've already edited bases by hand and don't want them changed, you must *deselect* the checkbox called **Allow edited bases to be replaced**.

If you select **Only make changes that result in an ambiguity**, **Sequencher** may change an ambiguous base call to an A, C, G, or T if it does not meet the secondary peak criteria. If you prefer it to remain ambiguous, do not select this option.

If you wish to search just a range of bases, highlight them and check the **Search Selection Only** checkbox.

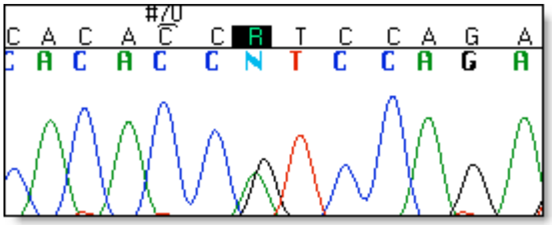
When you click **OK** to make the changes, a dialog notes how many bases will be changed; you can choose to **Cancel** or **Continue** at that point. If you continue, the changes take effect. Once they have taken effect, unlike normal edits to a base, they cannot be undone.

Once you have pressed the **Continue** button, you can find all the changes made during the session by selecting the first base in the sequence (you can click on it right on the chromatogram) and going to the **Select** menu and clicking on **Next Edited Base**. After you execute any **Select Next** command, **Sequencher** creates a shortcut for that command on the space bar so you can continue to search for your edits, using the same function over and over, by simply pressing the space bar.

SELECT NEXT COMMANDS

The **Select** menu provides you with several commands for navigating to bases which need further review or attention (see Table 15-1).

Figure 15-5 Base change generated by secondary peak



As you move your selection from each position that requires scrutiny to the next, you are likely to need to repeat these commands frequently. Therefore, **Sequencher** creates a shortcut for that command in the space bar after the first time you invoke a **Next** command. Press the space bar to move from one selected base to the next candidate.

To change any **Next** command to search in the reverse direction, combine the command with the **Shift** key.

If you use the **Next Met To Stop (>0bp)**, you can highlight the next pair of start and stop codons in one of the three forward reading frames. The number of bases shown in this command depends on whether you specified a preference for a minimum length in the **User Preferences**.

EDITING BASES FROM TRACES

Now that you can easily locate the bases that need attention, you can edit your data in either the **Contig Editor** window or the **Chromatogram** display. As you edit, both windows update automatically. (Refer to Chapter 12 “Editing Contigs” for more information on editing.)

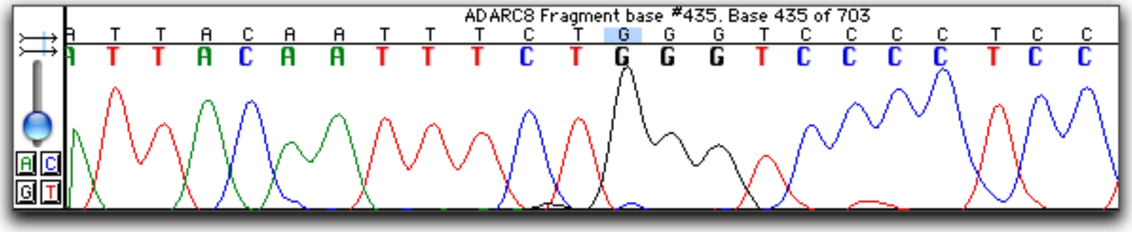
To edit a base while viewing a chromatogram, click on the current base call (upper line of text) you want to edit, then type the new base call (

Figure 15-6). You can also select multiple bases for deletion by dragging the cursor across them.

Table 15-1 Summary of Select commands

Select Command	Contig
Next Ambiguous Base	Ambiguous bases, Disagreements, First base in large gap run
Next Contig Disagree	Disagreements
Next Edited Base	Edited bases
Next Low Confidence Base	Any consensus base that is derived from a low confidence base
Next Met To Stop (>0bp)	Any pair of Start and Stop codons in one of forward ORFs
Next Inadequate Coverage	Finds consensus base that does not have a forward and reverse base that agree and are below low quality threshold

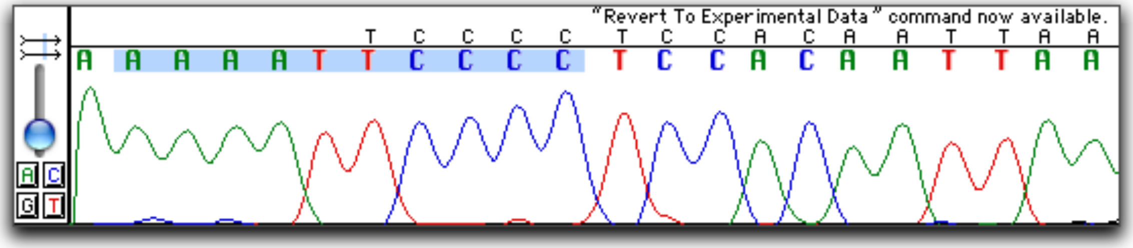
Figure 15-6 Clicking on a base displayed in a trace


ADARC8 Fragment base #435. Base 435 of 703
A T T A C A A T T T C T G G G T C C C C T C C
A T T A C A A T T T C T G G G T C C C C T C C
A C
G T

REVERTING TO EXPERIMENTAL DATA

If you have deleted too much of a sequence or have made edits you want to remove, you can copy the original base calls back into the current sequence. From within the chromatogram, select the original bases (lower line of text) you want to restore (Figure 15-7).

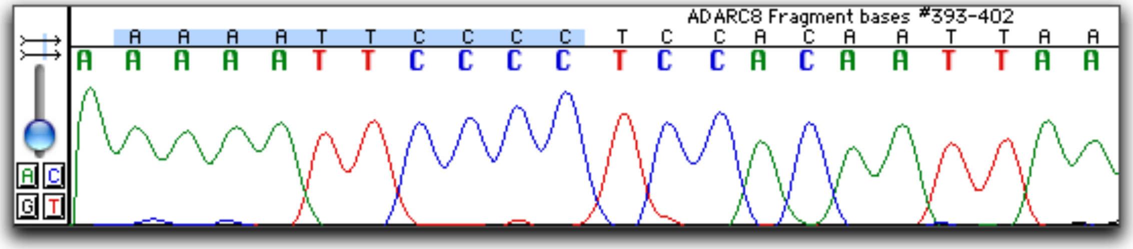
Figure 15-7 Original base calls selected


“Revert To Experimental Data” command now available.

T C C C C T C C A C A A T T A A
A A A A A T T C C C C T C C A C A A T T A A
A C
G T

Go to the **Sequence** menu and choose **Revert To Experimental Data**. Figure 15-8 shows the original bases restored.

Figure 15-8 Bases restored


ADARC8 Fragment bases #393-402

A A A A T T C C C C T C C A C A A T T A A
A A A A A T T C C C C T C C A C A A T T A A
A C
G T

Revert To Experimental Data is a powerful feature. It allows you to recover original data very precisely, as described above, or it can be used to globally revert an entire sequence or selection of sequences.

You can revert unassembled sequences to original data by selecting the sequences in the **Project Window** and executing the **Revert To Experimental Data** command.

REVERTING ASSEMBLED SEQUENCES BACK TO EXPERIMENTAL DATA

If you need to revert a sequence that has been assembled in a contig, select its icon from the left-hand side of the **Contig Editor**. Go to the **Contig** menu and use the **Remove Selected Sequence…** command. Now you can use the **Revert To Experimental Data** command as described above.

If your sequence has trace data, select all of the original bases. Go to the **Sequence** menu and choose **Revert To Experimental Data**.

Note: The **Extend Selection** command allows you to select all of the bases in a chromatogram without scrolling.

ADJUSTING TRACE POSITIONS

As the four dyes migrate in different lanes on certain automated sequencers, the synchronization among the four traces can deteriorate in output files. **Sequencher** allows for adjusting or "re-tracking" the position of one trace relative to the others. When you are looking at a chromatogram, specify which of the four traces you want to shift by selecting a base in the *original* data line (the *lower* of the two lines of sequence characters).

Go to the **Edit** menu and choose **Shift** to display a submenu that lets you shift the bases line left or right by a few pixels. The adjustments made are to aid you in reviewing the original data; they will not be saved to the file when you save the project.

16. NGS FOR DNA AND RNA-SEQ

In this chapter, we explain how to align or assemble Next-Generation sequences using **Maq**, **GSNAP**, **BWA**, or **Velvet**. We will show you how to use FastQ Quality Reports to check the quality of your data. If you are working with reference-guided alignment, then you only need to choose a Reference Sequence, choose your reads files, and begin the alignment. If you are working with de novo assembly, then you simply choose your reads files and begin the assembly. We also discuss viewing your results in an external viewer. You will learn how to hunt for SNPs, analyze methylation data, and perform RNA editing-tolerant alignments. You will also learn how to work with Multiplex ID data and use advanced settings.

In the second half of the chapter, you will learn how to use the programs in the **Cufflinks** suite for analysing RNA-Seq NGS data, normalizing your data, quantifying your data, and performing expression and differential expression analysis. You will also learn how to visualize your results using the RNA-Seq plots and charts and export the results in PNG (Portable Network Graphic) format.

For information on installing the External Alignment or RNA-Seq tools, please refer to the installation instructions for the tools that can be found on our website at http://genecodes.com/support.

SEQUENCE FILE FORMATS

Sequencher accepts many sequence formats. For Next-Generation sequencing, the most relevant formats are **FastA** and **FastQ**. Although most sequencers have their own native formats, as long as the data can be converted into **FastA** or **FastQ** format, **Sequencher** will be able to align or assemble those reads.

When performing reference-guided alignment, you will be using a Reference Sequence that will have been imported into **Sequencher**. It can be either a **FastA** file or a **GenBank** file. The advantage of using a GenBank sequence is that it will usually carry annotations, although you can mark up any sequence with annotations using **Sequencher's Mark Selection as Feature** command.

FASTA AND FASTQ FORMATS

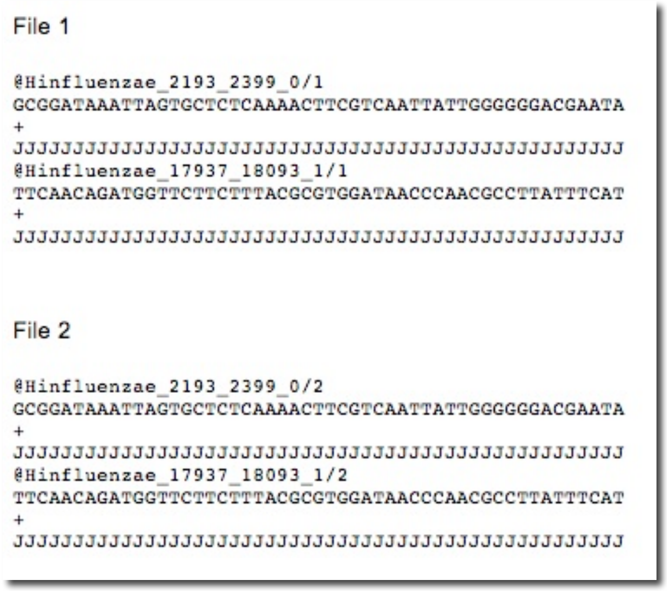
FastA is a simple format which contains a sequence name preceded by the > character and then followed by a sequence. The format can accommodate more than one sequence per file (concatenated FastA). The **FastQ** format contains the sequence name preceded by the @ character, the sequence, and the PHRED scores associated with the sequence. Again, the files can contain more than one sequence per file.

One thing to note, although it is possible to extract **FastQ** format files from Illumina and 454 sequencing runs, the formats may differ slightly. 454 conforms to what is popularly known as Sanger format. Illumina has several versions of the **FastQ** format. If your Illumina data has been produced by Casava 1.8 or later, then it will conform to Sanger **FastQ**.

PAIRED-END DATA

If you are dealing with paired-end data in **FastQ** format, then your data will need to be in two files. The read pairs must be in the same order. For example, in the image below you can see the data files. In each file, the pairs end in /1 in the first file and /2 in the second file.

Figure 16-1 Format of Paired-Ends reads files



File 1

```
@Hinfluenzae_2193_2399_0/1
GCGGATAAATTAGTGCTCTCAAAACTTCGTCAATTATTGGGGGGACGAATA
+
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@Hinfluenzae_17937_18093_1/1
TTCAACAGATGGTTCTTCTTTACGCGTGGATAACCCAACGCCTTATTTCAT
+
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

File 2

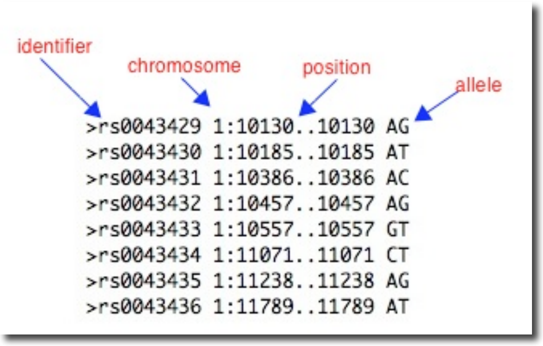
```
@Hinfluenzae_2193_2399_0/2
GCGGATAAATTAGTGCTCTCAAAACTTCGTCAATTATTGGGGGGACGAATA
+
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@Hinfluenzae_17937_18093_1/2
TTCAACAGATGGTTCTTCTTTACGCGTGGATAACCCAACGCCTTATTTCAT
+
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

In some instances, there may be other special formatting/layout of your files, ensure that this is done before submitting your sequences for alignment or the alignment will fail.

GSNAP LIST OF KNOWN SNPS FILE

If you plan to work with **GSNAP**, you will need to prepare a file containing a list of known SNPs. The file resembles a **FastA** file in that each line is prefixed with the > character.

Figure 16-2 Format of known SNPs file



```
>rs0043429 1:10130..10130 AG
>rs0043430 1:10185..10185 AT
>rs0043431 1:10386..10386 AC
>rs0043432 1:10457..10457 AG
>rs0043433 1:10557..10557 GT
>rs0043434 1:11071..11071 CT
>rs0043435 1:11238..11238 AG
>rs0043436 1:11789..11789 AT
```

GSNAP FASTA FORMAT FILE

If you plan to work with **FastA** files and wish to use the GSNAP algorithm, then you must format the paired-end reads files as follows. First you need to merge the two reads files into one file such that the paired-end reads are sequential. Then you need to remove the name of the second read.

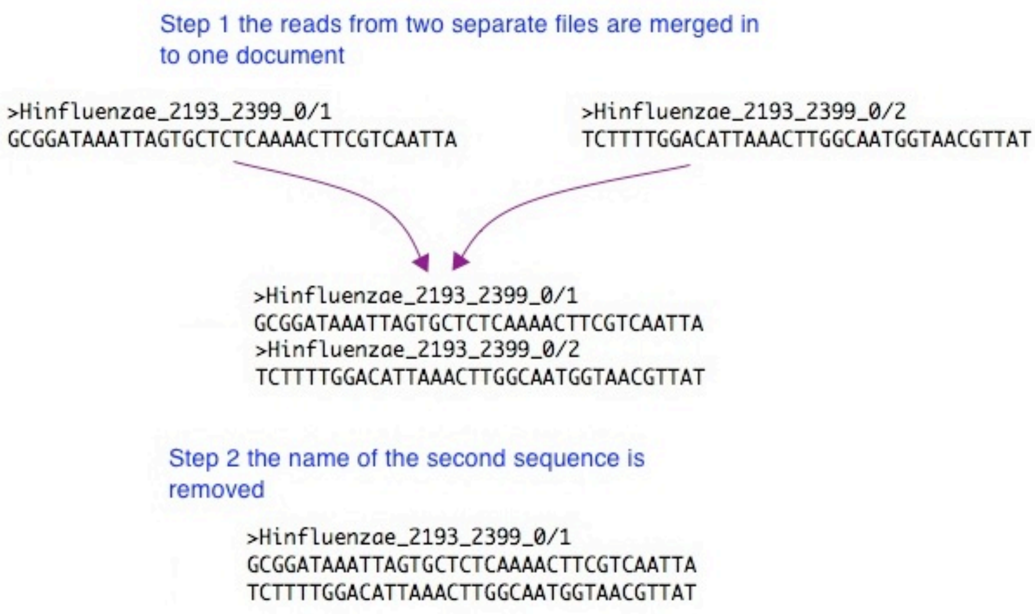
SAM FORMAT

SAM is the acronym for Sequence Alignment Map. The **SAM** file is a text format file but is capable of holding a rich set of information that can be used by a number of programs. It stores read alignments against Reference Sequences. Of interest to **Sequencher** users is that it is the resultant format for **GSNAP** and can be read by the **Tablet** program, which is an external Next-Generation assembly viewer.

AFG FORMAT

AFG is a text format file produced by **Velvet** that holds layout information relating to contigs, reads, and paired-ends. An **AFG** file can be viewed using the **Tablet** viewer.

Figure 16-3 Steps for creating paired-end FastA file



Step 1 the reads from two separate files are merged in to one document

```
>Hinfluenzae_2193_2399_0/1
GCGGATAAATTAGTGCTCTCAAAACTTCGTCAATTA
```

```
>Hinfluenzae_2193_2399_0/2
TCTTTTGGACATTAAACTTGGCAATGGTAACGTTAT
```

```
>Hinfluenzae_2193_2399_0/1
GCGGATAAATTAGTGCTCTCAAAACTTCGTCAATTA
>Hinfluenzae_2193_2399_0/2
TCTTTTGGACATTAAACTTGGCAATGGTAACGTTAT
```

Step 2 the name of the second sequence is removed

```
>Hinfluenzae_2193_2399_0/1
GCGGATAAATTAGTGCTCTCAAAACTTCGTCAATTA
TCTTTTGGACATTAAACTTGGCAATGGTAACGTTAT
```

LOCATION OF RESULTS FILES – THE EXTERNAL DATA HOME

The results may be saved as a contig in your project and/or a set of files in the Documents/Gene Codes/**Sequencher**/ folder. This location is referred to as the Home Directory. In the **Sequencher** directory, you will find a series of folders containing the results of any alignments and analyses performed using those external algorithms. In these folders, you will find a folder for each run you perform. The alignment and analyses run folders have a unique name assigned to them ensuring that previous runs are never overwritten. The **GSNAP database/BWA index** folders will be given the name you assign to them.

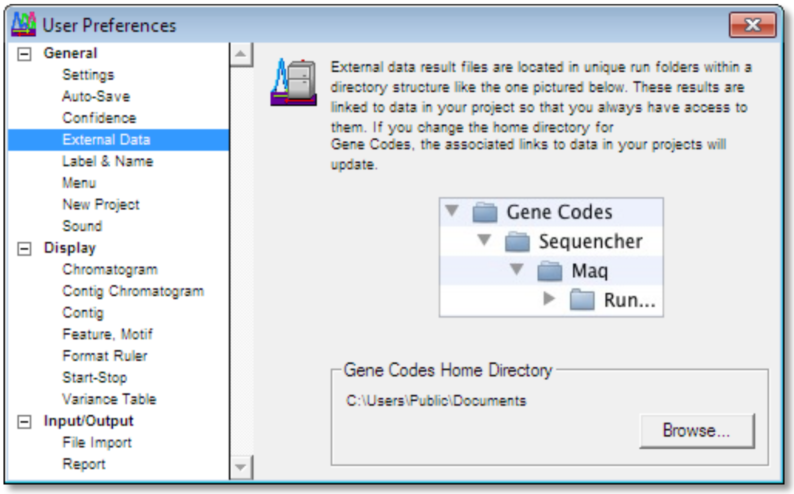
If you wish, you may change the location of this folder by going to the **External Data** pane in **User Preferences**.

Click on the **Browse…** button, browse to the desired location, and click the **OK** (Windows) or **Choose** (Mac) button. You may even create a new folder at your chosen location. Click the **Make New Folder** (Windows) or **New Folder** (Mac) button, then give it a name and click the **OK** (Windows) or **Choose** (Mac) button. The **User Preferences** pane will remember the new location of the Gene Codes Home Directory.

Table 16-1 Location of results in External Data Home

Algorithm	Type	Folder in External Data Home
Maq	Reference-guided aligner	Maq
BWA Index	Reference sequence indexes	BWA_Indexes
BWA-MEM	Reference-guided aligner	BWA
GSNAP Databases	Reference sequence database	GSNAP_Databases
GSNAP	Reference-guided aligner	GSNAP
Velvet	De novo assembler	Velvet
Cufflinks	RNA-Seq	Cufflinks
Cuffmerge	RNA-Seq merge of transcript.gtf	Cuffmerge
Cuffdiff	RNA-Seq differential expression	Cuffdiff
Cuffquant	RNA-Seq quantification	Cuffquant
Cuffnorm	RNA-Seq normalisation	Cuffnorm
MUSCLE	DNA multiple-sequence alignment	MUSCLE

Figure 16-4 Setting the Home directory



External data result files are located in unique run folders within a directory structure like the one pictured below. These results are linked to data in your project so that you always have access to them. If you change the home directory for Gene Codes, the associated links to data in your projects will update.

Gene Codes Home Directory: C:\Users\Public\Documents

Note: If you have already generated some results in the old location, you will need to move their corresponding Run folders from their current location to the new location so that referential integrity between the project, the contigs within it, and their Run folders is maintained.

STRATEGIES

Before you start, review how to use the **External Data Browser**. This is your access point to all the analyses you perform whether you are creating reference sequence databases/indexes or are using DNA-Seq or RNA-Seq tools. You can monitor and annotate your runs, view log files, delete old runs, or launch **Tablet** (where appropriate) to view alignments. For more information, see the next section entitled “External Data Browser”.

No matter how well prepared your samples were, it is always a good idea to check the quality of your reads data. Look at the section “FastQ Quality Control Reports” for more information on how to do this with **Sequencher**.

The alignment strategy itself can be divided into several parts. The first part consists of choosing whether you are going to perform reference-guided alignment or *de novo* assembly. If you are performing reference-guided alignment, then you will have a choice of algorithm. You also need to choose a reference sequence, and one or two reads files (in **FastA** or **FastQ** format), depending on whether you are working with single-end or paired-end data. If you want to, you can also perform SNP, methylation analysis, or RNA-tolerant editing analysis at the same time as the reference-guided alignment (**GSNAP** only). If you are working with *de novo* assembly, then you only need to know whether you are working with single-end or paired-end data (in **FastA** or **FastQ** format).

Sequencher can load your aligned results into the **Tablet** genome browser automatically. Whether you are performing an alignment or an assembly, you will therefore need to decide whether you want to view the alignment or assembly result immediately after the alignment has completed, using **Maqview** (**Maq** only) or **Tablet**, or to view the results later.

If you choose to perform any of the additional analyses, you can also view the results of your analysis later, using your web browser (the default viewer) or a text editor since the results are saved in the associated Run folder in both formats.

If the results are brought into the **Project Window**, you can also look at the results within your **Sequencher** project. You can use **Sequencher’s** features to examine the **Overview** or produce a **Summary Report**.

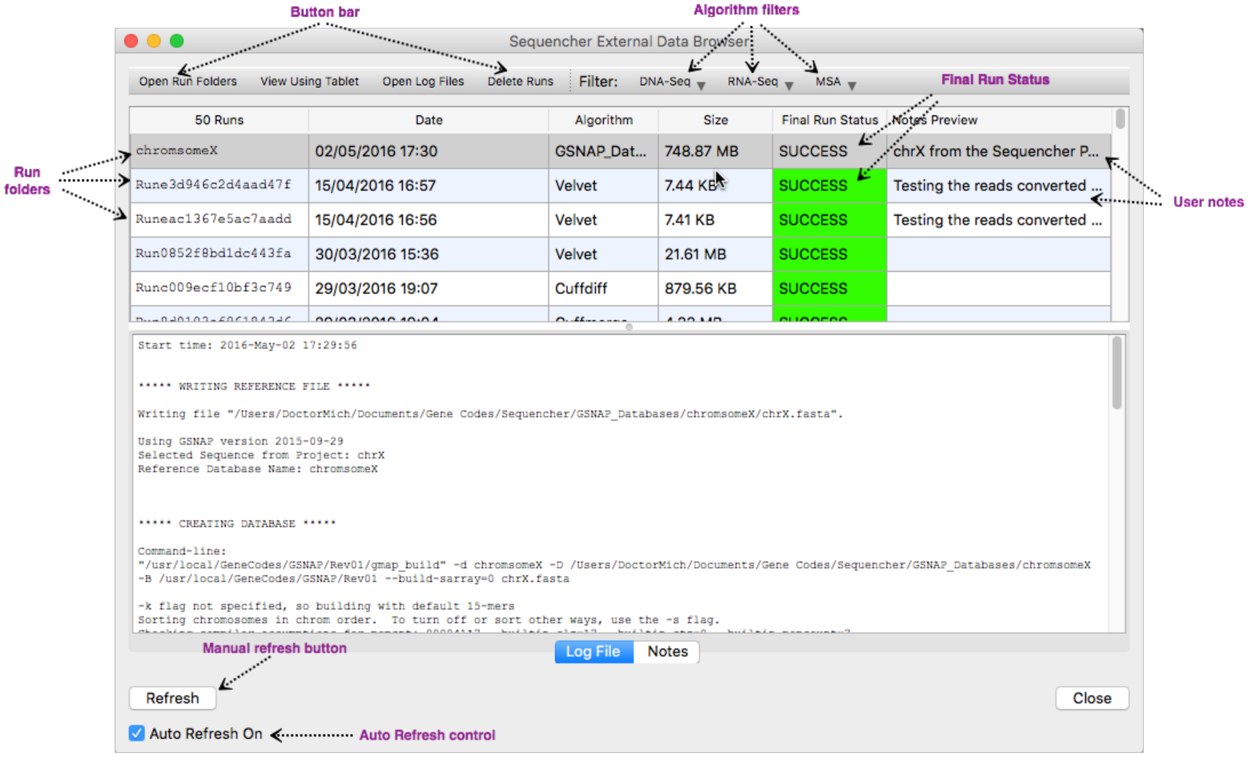
EXTERNAL DATA BROWSER

You can launch the **External Data Browser** using the **Window>Open External Data Browser** menu item. This opens the **Sequencher External Data Browser** dialog which consists of a button bar, a top pane divided into a number of columns, a lower pane, and finally some additional buttons. However, most of the time, you will not need to use this command as the **External Data Browser** opens automatically whenever you use any of the DNA-Seq or RNA-Seq tools such as **GSNAP** or **Cufflinks**.

USING THE BUTTON BAR

The button bar contains two types of controls; the first set is a set of buttons that perform specific functions such as **Open Run Folders** or **View Using Tablet**, while the second set is a set of drop-down menus that act to filter out specific algorithms from the Runs folder list. Clicking on an algorithm name in the drop-down menu toggles between removing and restoring the runs for that algorithm in the Runs list.

Figure 16-5 The External Data Browser



50 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
chromsomeX	02/05/2016 17:30	GSNAP_Dat...	748.87 MB	SUCCESS	chrX from the Sequencher P...
Rune3d946c2d4aad47f	15/04/2016 16:57	Velvet	7.44 KB	🟩 SUCCESS	Testing the reads converted ...
Runeac1367e5ac7aadd	15/04/2016 16:56	Velvet	7.41 KB	🟩 SUCCESS	Testing the reads converted ...
Run0852f8bd1dc443fa	30/03/2016 15:36	Velvet	21.61 MB	🟩 SUCCESS	
Runc009ecf10bf3c749	29/03/2016 19:07	Cuffdiff	879.56 KB	🟩 SUCCESS	

```
Start time: 2016-May-02 17:29:56


***** WRITING REFERENCE FILE *****

Writing file "/Users/DoctorMich/Documents/Gene Codes/Sequencher/GSNAP_Databases/chromsomeX/chrX.fasta".

Using GSNAP version 2015-09-29
Selected Sequence from Project: chrX
Reference Database Name: chromsomeX



***** CREATING DATABASE *****

Command-line:
"/usr/local/GeneCodes/GSNAP/Rev01/gmap_build" -d chromsomeX -D /Users/DoctorMich/Documents/Gene Codes/Sequencher/GSNAP_Databases/chromsomeX
-B /usr/local/GeneCodes/GSNAP/Rev01 --build-sarray=0 chrX.fasta

-k flag not specified, so building with default 15-mers
Sorting chromosomes in chrom order.  To turn off or sort other ways, use the -s flag.
```

RUNS PANE

This part of the dialog contains six columns. The first column lists all of the runs in your **Sequencher** External Data Folder (unless you have filtered one or more algorithms). The header of this column lists the number of runs available for viewing. This can be less than the total number of runs if you have already filtered out a specific algorithm. The second column contains the date and time the run was initiated. The third column lists the algorithm used for that particular run. The fourth column gives the size of the run folder and its contents. Column five is called Final Run Status and shows whether the run was a SUCCESS or if it FAILED. The final column contains any notes you may have added to the run information. Clicking the header name of any column will sort that column in ascending or descending order.

LOG FILE/NOTES PANE

The lower pane will list the contents of the log file for any selected run. To alternate between the log file view and the editable notes area, click the **Log File** button or the **Notes** button.

When you want to add notes, click on the row of interest and then click on the **Notes** button below the pane. A blank pane is revealed in place of the Log File view and a new **Save** button appears. You can type text notes into this pane. Clicking the **Save** button saves the notes and clicking the **Refresh** button refreshes the overall view. The **Auto Refresh On** control is checked by default, refreshes will happen automatically. The notes will be listed in the **Run** row for which they were entered.

Runs which generate a database/index will have a note automatically added. This Note contains the name of the reference sequence used to build the database/index.

You can always leave the **External Data Browser** by clicking on the **Close** button.

PERFORMING REFERENCE-GUIDED ALIGNMENTS

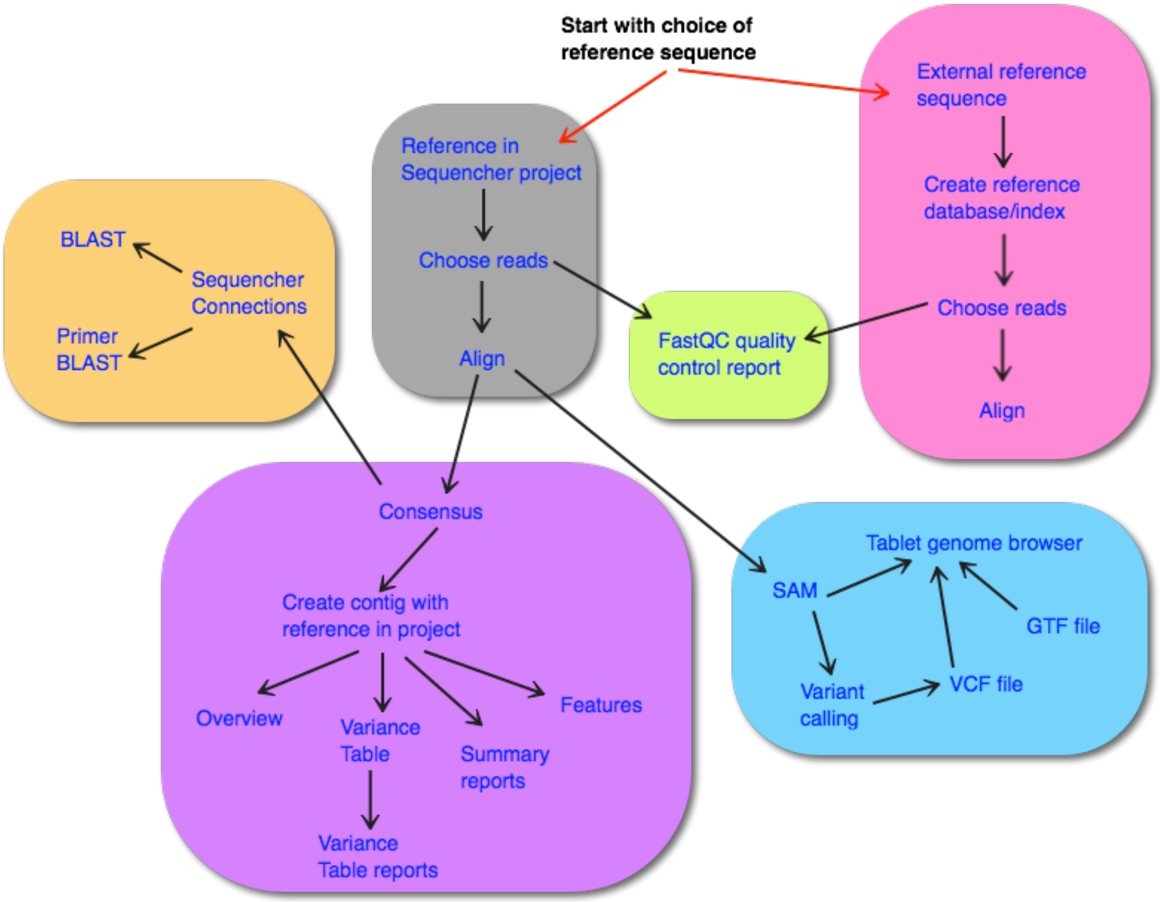
Reference-guided alignment differs from de novo assembly in that the reads are only compared to the reference sequence and not each other or nascent contigs. The actual alignment process forms part of a workflow which starts with the choice of reference sequence and algorithms. Figure 16-6 A generalized reference-guided alignment workflow gives an overview of the reference-guided workflow starting from the choice of a Reference Sequence.

FASTQ QUALITY CONTROL REPORTS

Before you start to work with your reads, it is always a good idea to check their quality. **Sequencher** makes use of the **FastQC** application to provide a series of reports on your **FastQ** data.

Choose **Sequence>Analyses>FastQ Quality Control Report**…. A new window appears. Click on the **Add File** button and browse to the file(s) you want to analyse.

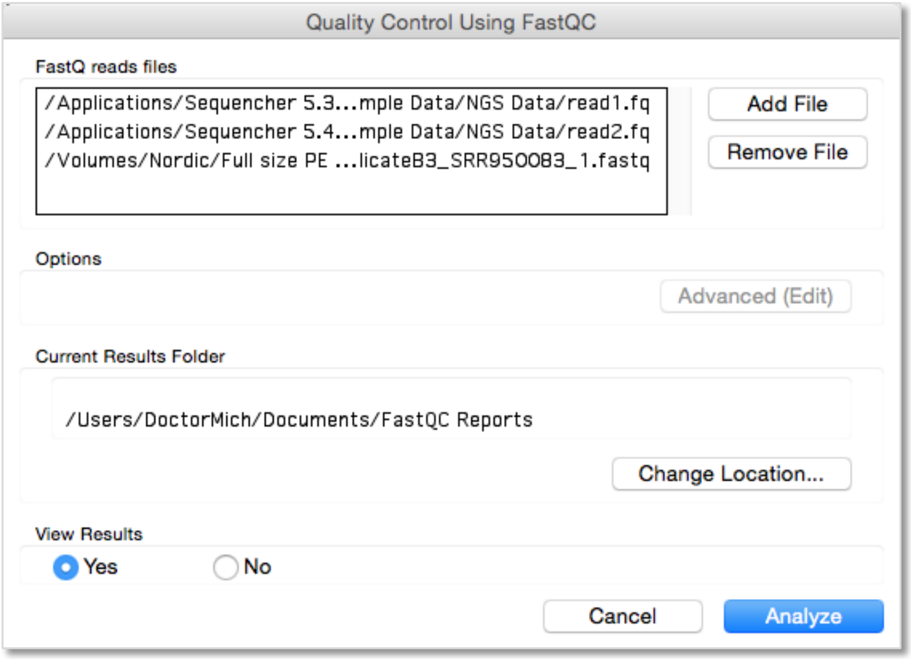
Figure 16-6 A generalized reference-guided alignment workflow



Select the file you want to add and click on the **Open** button. The file will now be listed in the **Quality Control Using FastQC** dialog. If you want to remove a file, highlight it and then click on the **Remove File** button. The results are saved automatically in **HTML** format in the **FastQC Reports** folder within your **Documents** folder. You can change the default location for your results by clicking on the **Change Location…** button and browsing to a new location. To view the results immediately after the analysis, click on the **Yes** radio button in the **View Results** groupbox. If you don't want to view the results immediately, then click on the **No** radio button, otherwise there are no other options to be set.

Note: If you have added 4 or fewer than 4 files to be analyzed, their reports will be opened automatically once the analysis is complete. If you have added greater than 4 files to analyze, and have specified to view them immediately after the analysis, you'll be presented with the **Report Selection** dialog asking you to select the reports you want to view at the end of the analysis.

Figure 16-7 Quality Control Using FastQC


Quality Control Using FastQC

FastQ reads files

/Applications/Sequencher 5.3...mple Data/NGS Data/read1.fq
/Applications/Sequencher 5.4...mple Data/NGS Data/read2.fq
/Volumes/Nordic/Full size PE ...licateB3_SRR950083_1.fastq

Options

Current Results Folder

/Users/DoctorMich/Documents/FastQC Reports

View Results

● Yes ○ No

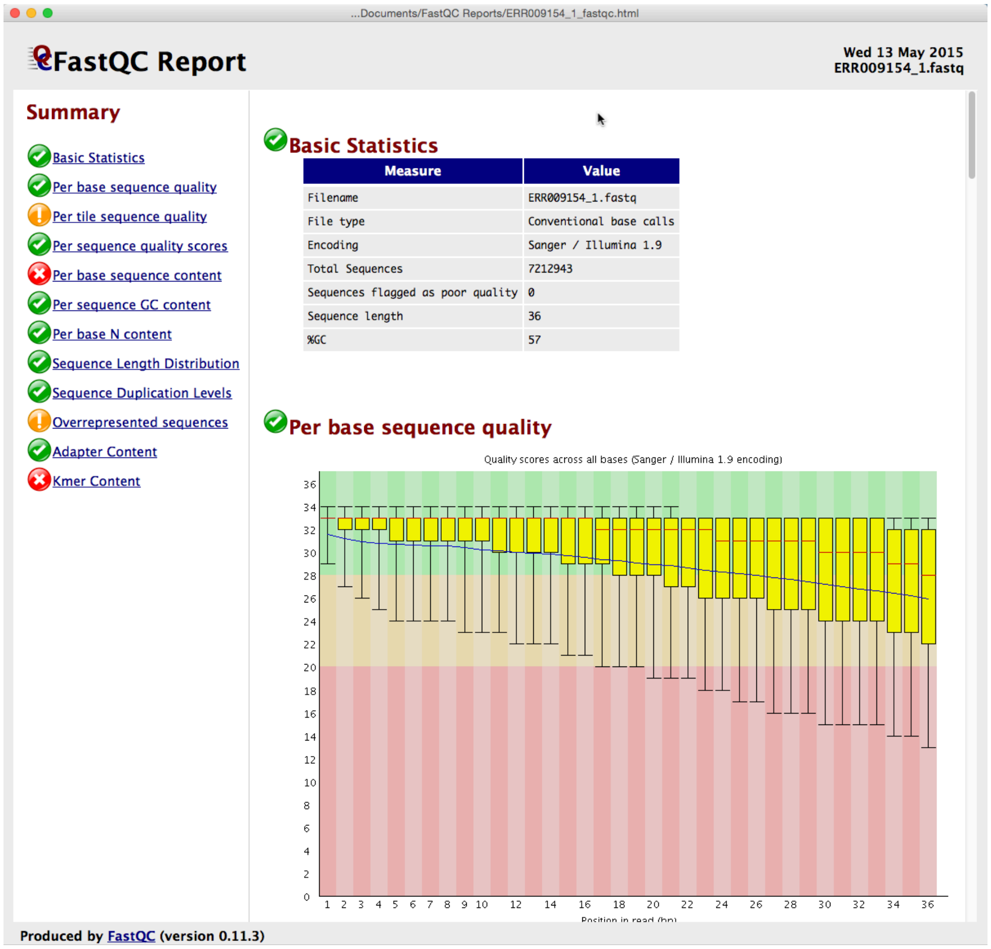
If you have previously created results you want to see, use the **View>Display FastQC Report...** menu item. A new window opens, choose the reports you want to open and click on the **Open** button. Whether you choose to see the analyses immediately or review them later, each report will open in its own window. On the left-hand side of the window is a clickable list of the analyses performed. The status of each test is indicated by a traffic lights system with a green check for pass, amber exclamation point for a warning, and a red cross for a failed test. You can find out more about each test by consulting the online documentation - FastQC Analysis Modules.

CREATING A REFERENCE SEQUENCE DATABASE OR INDEX

As the cost of sequencing has dropped and our ability to produce both longer reads and sequence longer genomes has increased, scientists have looked for ways to decrease the amount of time it takes to align each read on a reference sequence. With modern reference-guided aligners, the first step is to create a reference sequence database or index. The principle behind this is to enable the actual mapping of reads to the reference sequence to be speeded up. There are many techniques for doing this and each aligner will have its own.

From **Sequencher** 5.4.1 and later, you can now create databases/indexes that can be re-used. If you choose a sequence from your **Project Window**, you can choose to create a permanent database/index or just have **Sequencher** delete the database/index for you at the end of the alignment. For large genomes, having the database/index saved is especially useful since the time taken to build can be in the order of an hour (depending on the speed of the hardware used and the available RAM). Moreover, once a database/index has been built, you can share it with other **Sequencher** users.

Figure 16-8 FastQC Report



...Documents/FastQC Reports/ERR009154_1_fastqc.html

FastQC Report

Wed 13 May 2015
ERR009154_1.fastq

Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

✓ Basic Statistics

Measure	Value
Filename	ERR009154_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7212943
Sequences flagged as poor quality	0
Sequence length	36
%GC	57

✓ Per base sequence quality

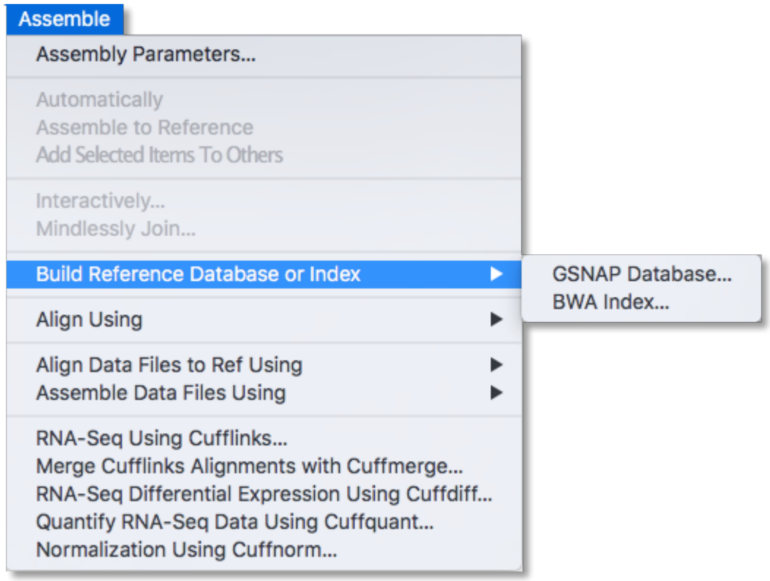
Produced by FastQC (version 0.11.3)

You can use a single sequence in **FastA** format or a series of sequences in **concatenated FastA** format (for example a series of chromosomes). If you plan to perform an *in silico* experiment such as comparing the **BWA-MEM** algorithm to the **GSNAP** algorithm, you will need to build a database/index for both aligners.

The files for the database/index are placed in the External Data Home folder (the default location for this is your Documents folder but you may change this through a **User Preference**). You will be able to give the database/index a name of your own choosing. In addition, information about the database/index run will be visible in the **External Data Browser** window.

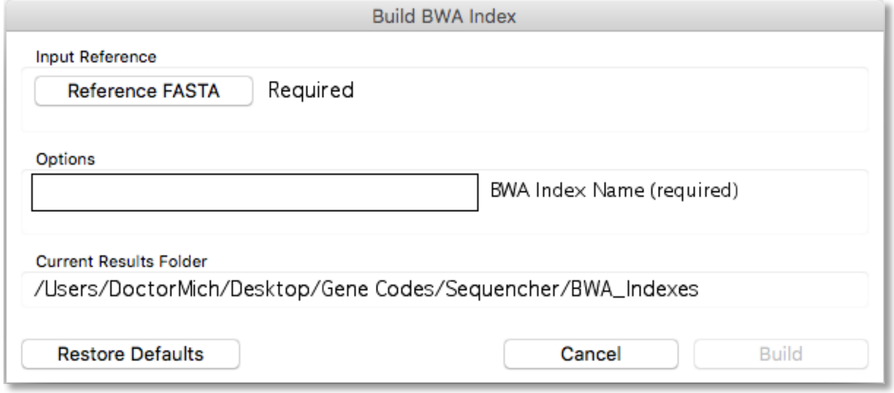
Choose **Assemble>Build Reference Database or Index>GSNAP Database…** to use a reference sequence with the **GSNAP** algorithm. If you want to use the reference sequence with **BWA-MEM**, choose **Assemble>Build Reference Database or Index>BWA Index…**. The **External Data Browser** window opens automatically with either command.

Figure 16-9 Build Reference Database or Index menu item


Assemble
Assembly Parameters...
Automatically
Assemble to Reference
Add Selected Items To Others
Interactively...
Mindlessly Join...
Build Reference Database or Index
Align Using
Align Data Files to Ref Using
Assemble Data Files Using
RNA-Seq Using Cufflinks...
Merge Cufflinks Alignments with Cuffmerge...
RNA-Seq Differential Expression Using Cuffdiff...
Quantify RNA-Seq Data Using Cuffquant...
Normalization Using Cuffnorm...
GSNAP Database...
BWA Index...

A new dialog appears. The look of the dialog is similar in both cases but the title of the dialog is dictated by the algorithm being used to generate the database/index. Click on the **Reference FASTA** button and browse to the **FastA** file containing your reference sequence(s). Click on the **Open** button. Give your database or index a name and click on the **Build** button.

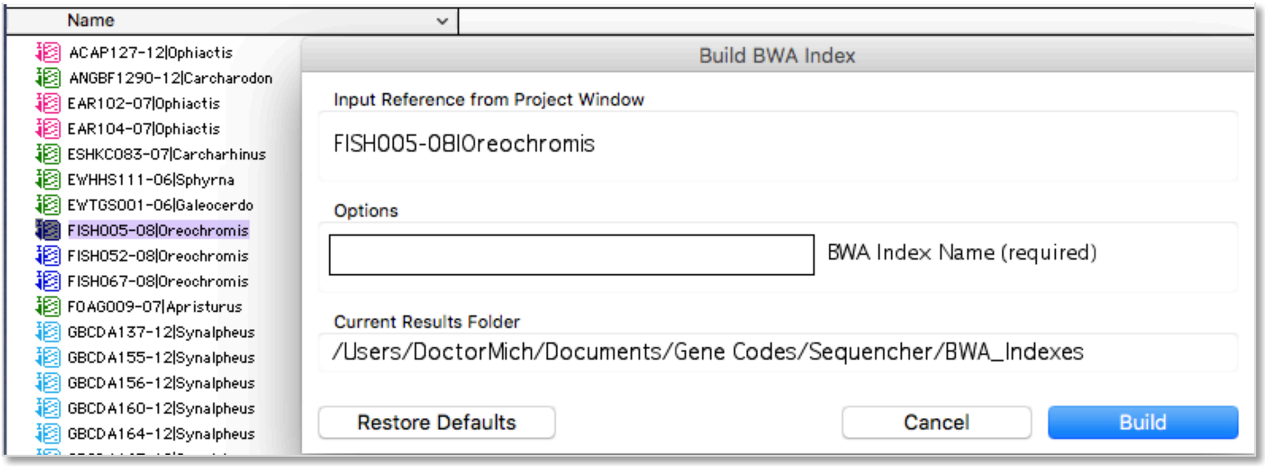
Figure 16-10 Build BWA Index dialog



/Users/DoctorMich/Desktop/Gene Codes/Sequencher/BWA_Indexes

If using a reference sequence from a selection in the **Project Window**, the following dialog will appear instead. In this case, as you can see, the reference sequence name replaces the **Reference FASTA** button.

Figure 16-11 Build BWA Index with selection from Project Window



Build BWA Index

Input Reference from Project Window

FISH005-08|Oreochromis

Options

BWA Index Name (required)

Current Results Folder

/Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA_Indexes

When you are using the **External Data Browser** to monitor the progress of the build (especially useful with large genomes), note that the log file pane refreshes automatically if the **Auto Refresh On** control is checked.

Once **Sequencher** has finished the creation of the database or index, you will see a green **SUCCESS** status in the Final Run Status column. If there was a problem with the creation of the database or index, you may see a red **FAILED** status. Consulting the log file will often give you a clue as to the reason for the failure. Very rarely, there will be no status message at all. This is also an indicator of problems with the creation/build run.

The **Notes** field allows you to add pertinent information about your reference sequence. This is a good place to record information such as the build version for complete or partial genomes. **Sequencher** will automatically add the name of the sequence file used to create the database or index to the **Notes** field.

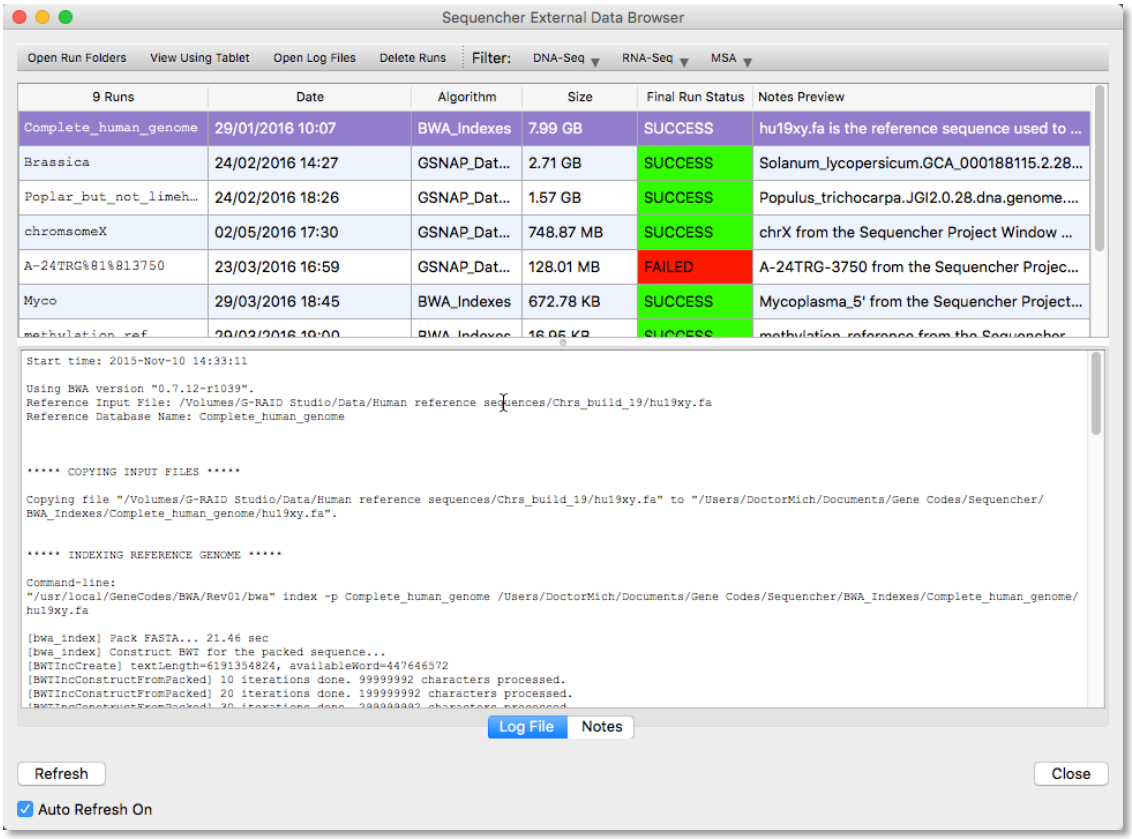
Once the database/index is built, you will see it listed in a drop-down menu the next time you use **GSNAP** or **BWA-MEM**.

REFERENCE-GUIDED ALIGNMENT USING THE MAQ ALGORITHM

The **Maq** reference-guided aligner is an older aligner but is kept for legacy purposes. Compared to other aligners, it has limitations. The major limitation is the number of reads (two million) it can deal with before its performance seriously degrades. This makes it impractical for large projects.

Launch **Sequencher** and import your Reference Sequence into a new project. You could use any sequence in your **Sequencher** project to act as the reference. It does not have to be marked as a Reference Sequence using the **Sequence>Reference Sequence** command first.

Figure 16-12 GSNAP Database and BWA Index listings in the External Data Browser

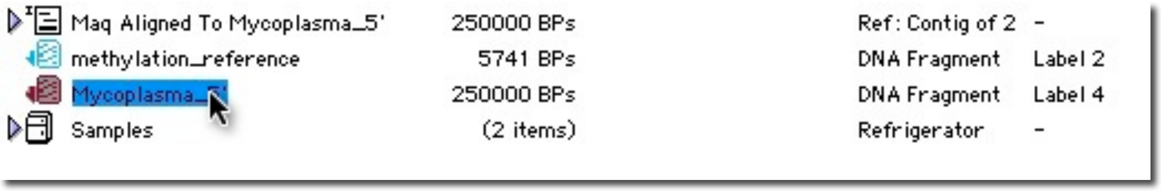


9 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Complete_human_genome	29/01/2016 10:07	BWA_Indexes	7.99 GB	SUCCESS	hu19xy.fa is the reference sequence used to ...
Brassica	24/02/2016 14:27	GSNAP_Dat...	2.71 GB	SUCCESS	Solanum_lycopersicum.GCA_000188115.2.28...
Poplar_but_not_limeh...	24/02/2016 18:26	GSNAP_Dat...	1.57 GB	SUCCESS	Populus_trichocarpa.JGI2.0.28.dna.genome....
chromsomeX	02/05/2016 17:30	GSNAP_Dat...	748.87 MB	SUCCESS	chrX from the Sequencher Project Window ...
A-24TRG%81%813750	23/03/2016 16:59	GSNAP_Dat...	128.01 MB	FAILED	A-24TRG-3750 from the Sequencher Projec...
Myco	29/03/2016 18:45	BWA_Indexes	672.78 KB	SUCCESS	Mycoplasma_5' from the Sequencher Project...
methylation_ref	29/03/2016 19:00	BWA_Indexes	16.95 KB	SUCCESS	methylation_reference from the Sequencher

```
Start time: 2015-Nov-10 14:33:11

Using BWA version "0.7.12-r1039".
Reference Input File: /Volumes/G-RAID Studio/Data/Human reference sequences/Chrs_build_19/hu19xy.fa
Reference Database Name: Complete_human_genome


***** COPYING INPUT FILES *****

Copying file "/Volumes/G-RAID Studio/Data/Human reference sequences/Chrs_build_19/hu19xy.fa" to "/Users/DoctorMich/Documents/Gene Codes/Sequencher/
BWA_Indexes/Complete_human_genome/hu19xy.fa".


***** INDEXING REFERENCE GENOME *****

Command-line:
"/usr/local/GeneCodes/BWA/Rev01/bwa" index -p Complete_human_genome /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA_Indexes/Complete_human_genome/
hu19xy.fa

[bwa_index] Pack FASTA... 21.46 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=6191354824, availableWord=447646572
[BWTIncConstructFromPacked] 10 iterations done. 99999992 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 199999992 characters processed.
[BWTIncConstructFromPacked] 30 iterations done. 299999992 characters processed.
```

Select the Reference Sequence and choose **Assemble>Align Data Files to Ref Using>Maq...**

Figure 16-13 Choosing a sequence to act as the reference from the Project Window

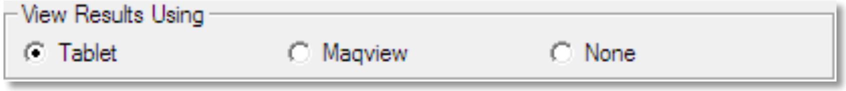


Maq Aligned To Mycoplasma_5'	250000 BPs	Ref: Contig of 2	-
methylation_reference	5741 BPs	DNA Fragment	Label 2
Mycoplasma_5'	250000 BPs	DNA Fragment	Label 4
Samples	(2 items)	Refrigerator	-

Choose your first reads file by clicking on the **Select File 1** button to browse to the file you want to use. Select the file and click on the **Open** button. Now choose the second reads file by clicking on the **Select File 2** button and click on the **Open** button in the same manner as for the first set of reads.

Select the viewer you want to use by clicking on its radio button. With **Maq**, you have the choice of the **Maqview** or **Tablet** viewers.

Figure 16-14 Choosing a viewer

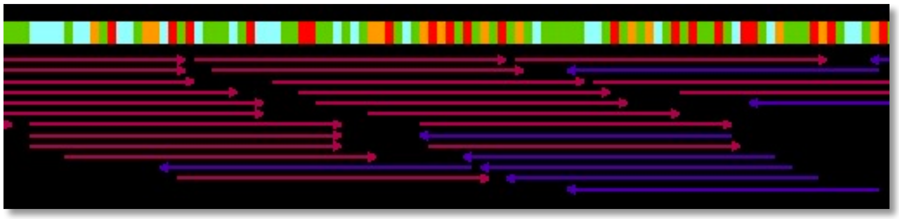

View Results Using

● Tablet ○ Maqview ○ None

Once you have clicked on the **Align** button, the alignment will take place. The consensus sequence that results is then formed into a contig with the reference sequence. If you chose to view your alignment immediately after the algorithm was finished, then the viewer you selected will launch automatically with the contig already loaded. If you prefer to view it later, then choose **Contig>Show NGS Data Using>Maqview.**

If you choose **Maqview**, then a Terminal window will open and display the reads. Maqview has different views, to move between these views, use the function keys on your keyboard**. F1** switches to the bases mode, **F2** switches to the colored blocks mode, and **F3** switches to the overview mode. In the **F2** and **F3** views, you can zoom in and out using **+** and **–** keys.

Figure 16-15 Maqview reads directionality

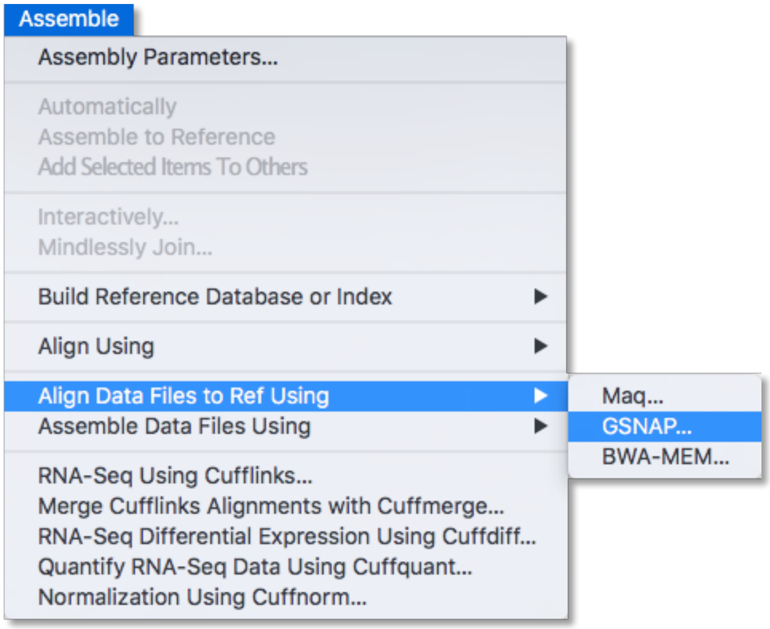


REFERENCE-GUIDED ALIGNMENT USING GSNAP OR BWA

When working with large NGS projects, you should use either **BWA-MEM** or **GSNAP**. There are two ways that you can provide a reference sequence, the first is to use a sequence from your current project (see Figure 16-13 Choosing a sequence to act as the reference from the Project Window) or you can use a prebuilt database/index. In most cases, these will be large sequence(s)/genome(s) which have been used to create the prebuilt database/index. Choose the database/index you want to use from the **BWA-MEM** or **GSNAP** dialog.

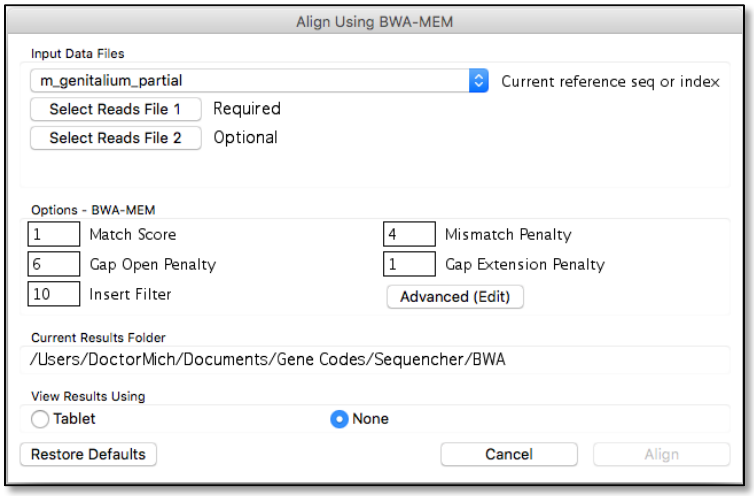
If you are going to use a sequence from your project, simply select it and then choose **Assemble>Align Data Files to Ref Using>GSNAP..**. or choose **Assemble>Align Data Files to Ref Using>BWA-MEM..**.. The name of the reference sequence will appear automatically in the **Current reference seq or index** (for **BWA**) or **Current reference seq or db** (for **GSNAP**) drop-down menu of the aligner's graphical user interface. If you are working with a prebuilt database, then choose it now from the **Current reference seq or db** or **Current reference seq or index** drop-down menu.

Figure 16-16 Choosing the GSNAP alignment algorithm


Assemble
Assembly Parameters...
Automatically
Assemble to Reference
Add Selected Items To Others
Interactively...
Mindlessly Join...
Build Reference Database or Index
Align Using
Align Data Files to Ref Using
Assemble Data Files Using
RNA-Seq Using Cufflinks...
Merge Cufflinks Alignments with Cuffmerge...
RNA-Seq Differential Expression Using Cuffdiff...
Quantify RNA-Seq Data Using Cuffquant...
Normalization Using Cuffnorm...
Maq...
GSNAP...
BWA-MEM...

Once the dialog has appeared, you will be able to choose the reads files and make changes to the various options and settings. If you want to use a prebuilt database/index, then choose **Assemble>Align Data Files to Ref Using>GSNAP...** or choose **Assemble>Align Data Files to Ref Using>BWA -MEM...** and then pick a reference sequence database/index from the drop-down menu at the top of the dialog. All databases or indexes for your chosen algorithm will be listed in this drop-down. The **External Data Browser** launches automatically with either of these commands.

Figure 16-17 Choosing reads files – BWA-MEM with chosen index


Align Using BWA-MEM

Input Data Files

m_genitalium_partial — Current reference seq or index

Select Reads File 1 — Required

Select Reads File 2 — Optional

Options - BWA-MEM

1 Match Score — 4 Mismatch Penalty

6 Gap Open Penalty — 1 Gap Extension Penalty

10 Insert Filter — Advanced (Edit)

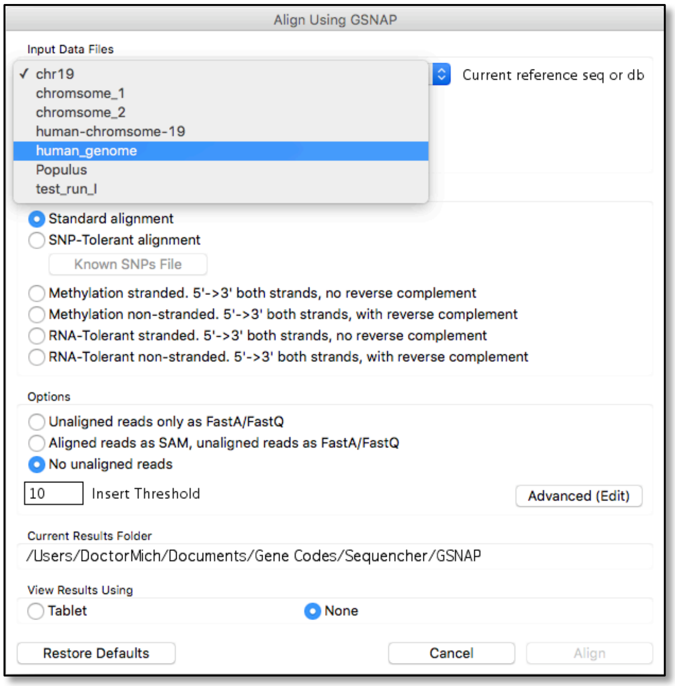
Current Results Folder

/Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA

View Results Using

○ Tablet ● None

Figure 16-18 Choosing reads files – GSNAP with menu showing available databases



● Standard alignment
○ SNP-Tolerant alignment
○ Methylation stranded. 5'->3' both strands, no reverse complement
○ Methylation non-stranded. 5'->3' both strands, with reverse complement
○ RNA-Tolerant stranded. 5'->3' both strands, no reverse complement
○ RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options
○ Unaligned reads only as FastA/FastQ
○ Aligned reads as SAM, unaligned reads as FastA/FastQ
● No unaligned reads
10 Insert Threshold

Current Results Folder
/Users/DoctorMich/Documents/Gene Codes/Sequencher/GSNAP

View Results Using
○ Tablet ● None

Note: If you have created databases/indexes in a different location than your current External Data Home setting, they will not be listed.

Once you have chosen your database/index, you can go on to choose your reads files and change any settings or options. When you are satisfied with your selections, click on the **Align** button to initiate the alignment.

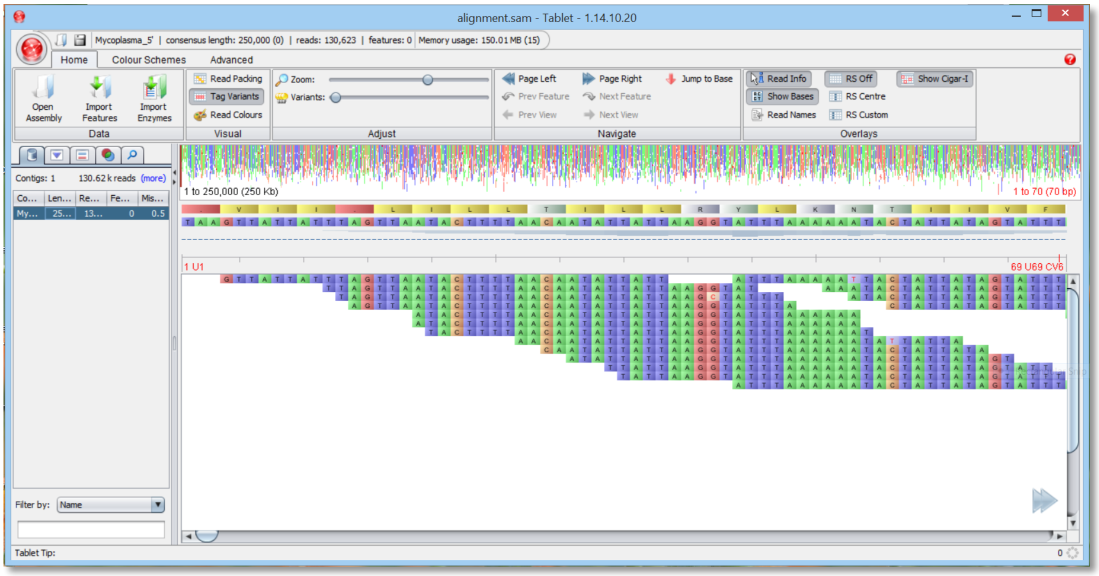
LOOKING AT REFERENCE-GUIDED ALIGNMENT RESULTS

If you chose to view your alignment at some other time by choosing the **None** option, you can view it at any time by choosing **Contig>Show NGS Data Using>Tablet** (which can be used with **BWA-MEM**, **GSNAP**, or **Velvet**).

If you chose **Tablet** as your view option, your results will be displayed in **Tablet**. You view a contig by clicking on its name in the contig list at the left-hand side of the window. You can move about and change the look of the view. For example, using the **Zoom** slider increases and decreases the size of the reads bases.

Using the **Variants** slider increases or decreases the background grey on which the bases are displayed so that moving the slider to the extreme right will effectively mask normal bases while potential SNPs will be highlighted. Clicking on the **Direction** button in the **Color Schemes** tab changes the color of the bases from their normal individual base color to dark blue or khaki depending on the direction of the read. The **Classic** view shows the bases without any background coloration. In order to return to the bases with colored backgrounds, click on the **Nucleotide** button in the **Color Schemes** tab.

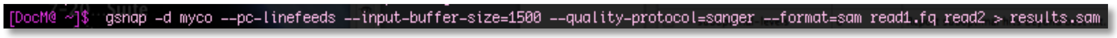
Figure 16-19 Tablet viewer



ADVANCED PARAMETERS FOR GSNAP

When you send your reads to **GSNAP**, **Sequencher** is protecting you from having to work from the command line.

Figure 16-20 Example of a command line

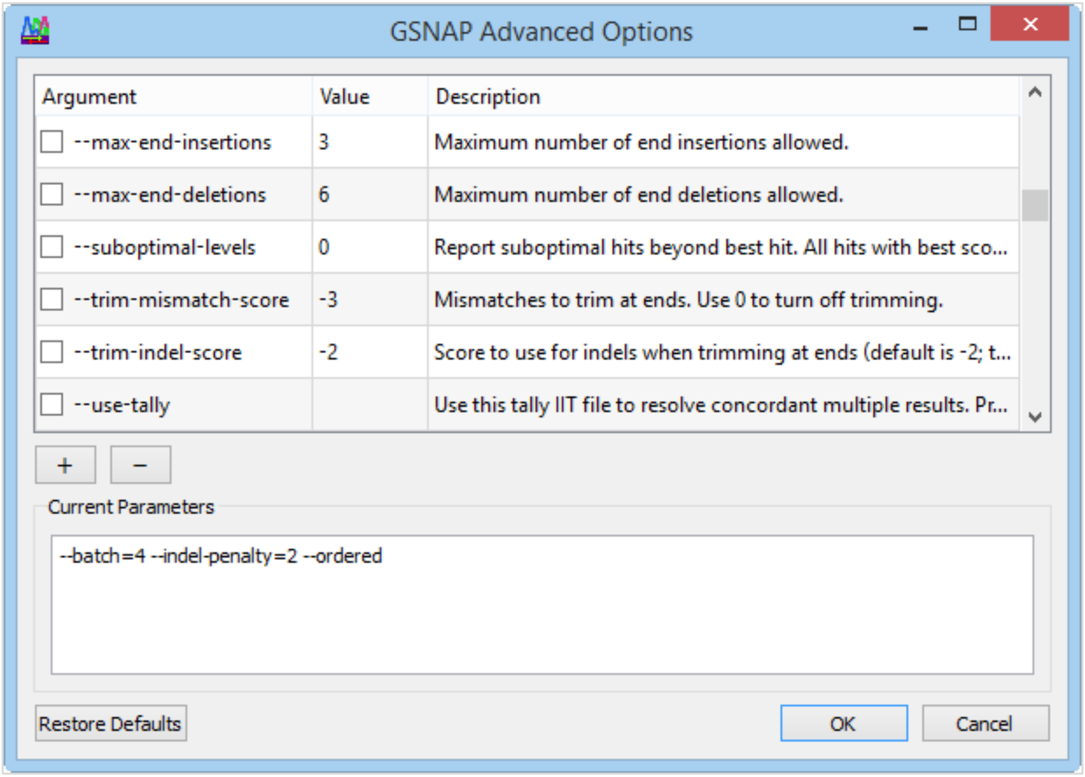


```
[DocM@ ~]$  gsnap -d myco --pc-linefeeds --input-buffer-size=1500 --quality-protocol=sanger --format=sam read1.fq read2 > results.sam
```

GSNAP has many parameters that can be set providing you use the command line. **Sequencher** has a novel GUI interface that allows you to choose advanced parameters and change their settings.

In order to see what the advanced parameters are, click on the **Advanced (Edit)** button. The **GSNAP Advanced Options** dialog appears. This is in the form of a table with 3 columns. The first column contains a checkbox and parameter, the second column contains the value of that parameter, and the third column contains its description.

Figure 16-21 GSNAP Advanced Options dialog



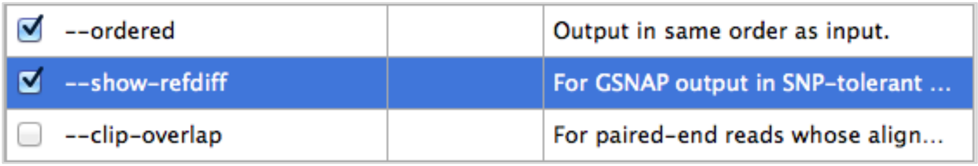
Argument	Value	Description
□ --max-end-insertions	3	Maximum number of end insertions allowed.
□ --max-end-deletions	6	Maximum number of end deletions allowed.
□ --suboptimal-levels	0	Report suboptimal hits beyond best hit. All hits with best sco...
□ --trim-mismatch-score	-3	Mismatches to trim at ends. Use 0 to turn off trimming.
□ --trim-indel-score	-2	Score to use for indels when trimming at ends (default is -2; t...
□ --use-tally		Use this tally IIT file to resolve concordant multiple results. Pr...

Current Parameters

```
--batch=4 --indel-penalty=2 --ordered
```

In order to use a parameter, simply click on the checkbox so that a check appears. If you want to change the value, then click in the Value cell for that parameter and edit the value. If you are unable to edit the value or the value cell is empty initially, it may be that this parameter is a flag. That is to say, it has no value but is used in an on or off state.

Figure 16-22 Enabled parameters



☑ --ordered		Output in same order as input.
☑ --show-refdiff		For GSNAP output in SNP-tolerant ...
□ --clip-overlap		For paired-end reads whose align...

Below this table are two buttons, a **+** (plus) and a **–** (minus). These are used to add or remove parameters, values, and their descriptions.

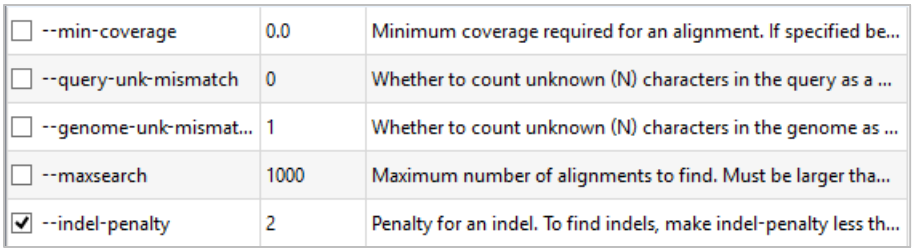
Beneath this is a text field titled **Current Parameters**. Any changes to the checkbox will cause a parameter to appear or disappear from this text field. Items in this text field will become part of the command line that **Sequencher** composes on your behalf. Together with the values that **Sequencher** controls, these values are sent to the **GSNAP** program.

If you find that you have removed a parameter by mistake or wish to restore all the settings to their original values, click on the **Restore Defaults** button.

GSNAP AND INDELS

The **GSNAP** algorithm allows you to control the appearance of insertions and deletions using several different parameters. These parameters can be set using the **Advanced Options**. The parameters you can set include --**indel-penalty** and **--min-coverage**.

Figure 16-23 Enabled parameters allowing indels in alignment

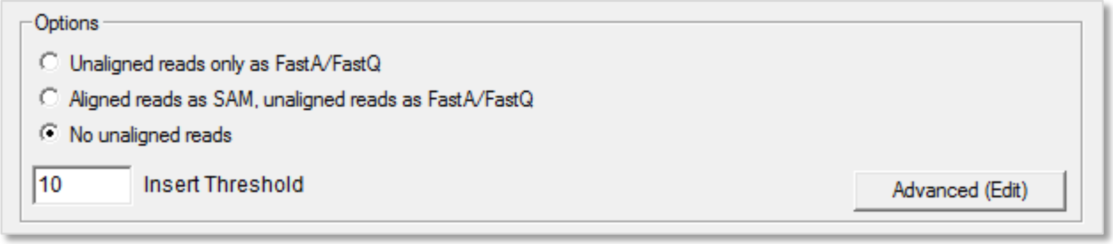


□ --min-coverage	0.0	Minimum coverage required for an alignment. If specified be...
□ --query-unk-mismatch	0	Whether to count unknown (N) characters in the query as a ...
□ --genome-unk-mismat...	1	Whether to count unknown (N) characters in the genome as ...
□ --maxsearch	1000	Maximum number of alignments to find. Must be larger tha...
☑ --indel-penalty	2	Penalty for an indel. To find indels, make indel-penalty less th...

In addition to setting these penalties, you also need to set a value in the **Insert Threshold** text field in the **Options** group box on the **Align Using GSNAP** dialog. The default value is **10** which is the same as the **Tablet** value. This means that at least **10** reads containing this insertion are required in order to see the insert in **Tablet**. If you change the value in **Sequencher**, there is a possibility that the views will diverge.

Note*:* In order to change the insertion threshold in **Tablet**, visit its home page at http://bioinf.scri.ac.uk/tablet/

Figure 16-24 Sequencher Insertion Threshold in GSNAP dialog



Options

○ Unaligned reads only as FastA/FastQ
○ Aligned reads as SAM, unaligned reads as FastA/FastQ
● No unaligned reads

10 Insert Threshold

ADVANCED PARAMETERS FOR BWA-MEM

BWA-MEM is able to align reads between 70bp and 1Mbp. While **BWA-MEM** does not have quite the same sophistication when it comes to controlling indels as **GSNAP**, it does have some advanced options which are worthy of note. To see what the advanced parameters are, click on the **Advanced (Edit)** button on the **Align Using BWA-MEM** dialog. The **BWA Advanced Options** dialog appears. This is in the form of a table with 3 columns with a pane below it. The first column contains a checkbox and parameter, the second column contains the value of that parameter, and the third column contains its description.

In order to use a parameter, simply click the checkbox so that the check mark appears. If you want to change the value, then click in the Value cell for that parameter and edit the value. If you are unable to edit the value or the value cell is empty initially, it may be that this parameter is a flag. That is to say, it has no value but is used in an on or off state. Below this table are two buttons, a **+** (plus) and a **–** (minus). These are used to add or remove parameters, values, and their associated descriptions.

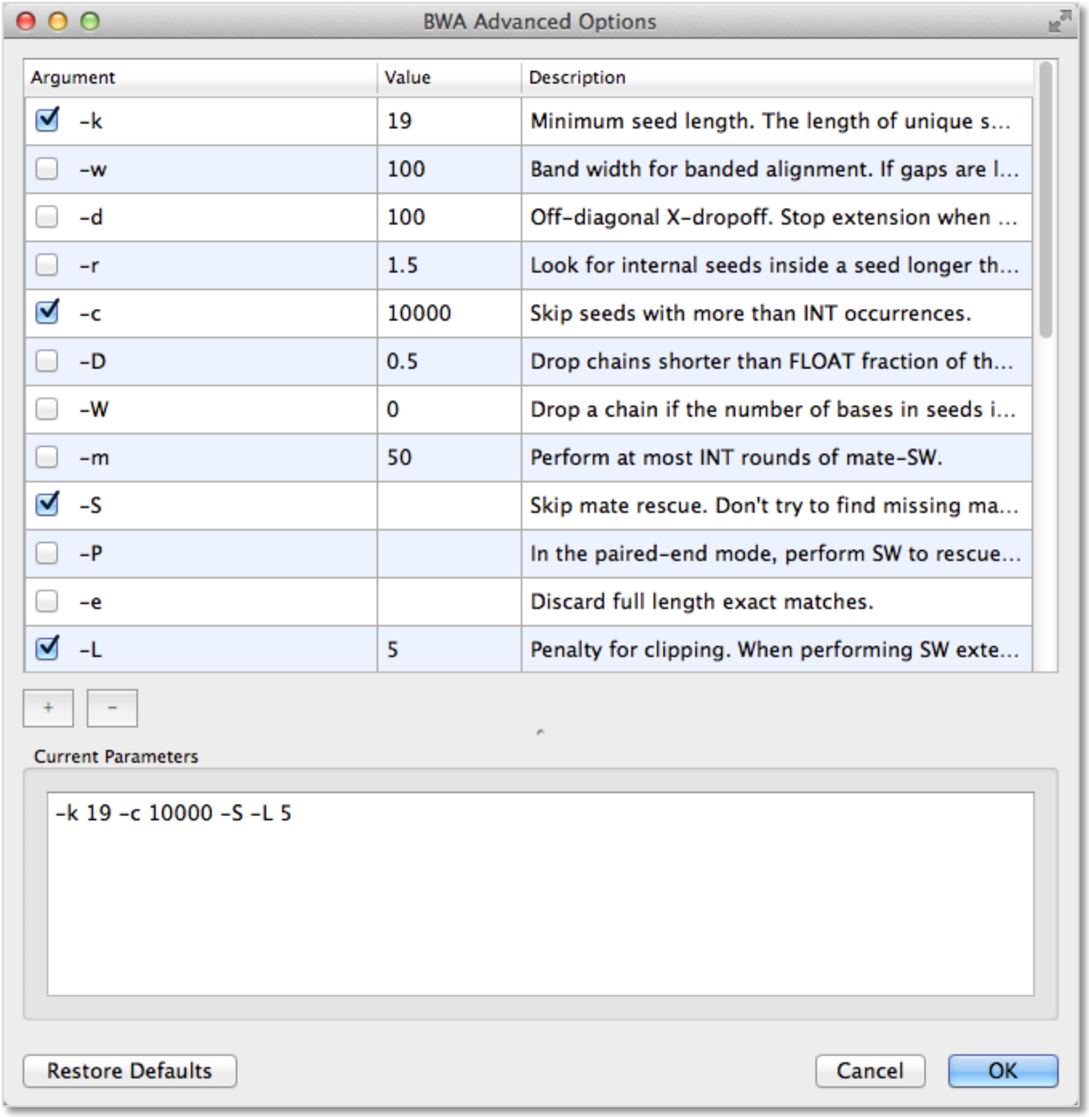
One example of controlling **BWA-MEM** using the **Advanced Options** is the **Band Width**, when this parameter is chosen and a value given, the **BWA-MEM** algorithm will ignore gaps longer than this value although other aspects of the algorithm also have some effect.

If you are working with paired-end (PE) data, you can use the **Paired End Mode** to use a more rigorous algorithm to rescue missing hits, that is, reads that have not already been aligned and properly paired.

If you prefer not to find too many unmatched reads pairs, you can choose an option to penalize unpaired read pairs. A default value is provided but you can change this if you prefer.

Beneath the table containing the **Advanced Options** is a text field entitled **Current Parameters**. A change to one of the checkboxes in the table will cause a parameter, and its value if it has one, to appear or disappear from this text field. Items in this text field will become part of the command line that **Sequencher** composes on your behalf. Together with the values that **Sequencher** controls, these values are sent to the **BWA-MEM** program. If you find that you have removed a parameter by mistake or wish to restore all the settings to their original values, click on the **Restore Defaults** button.

Figure 16-25 BWA Advanced options



Argument	Value	Description
☑ -k	19	Minimum seed length. The length of unique s…
□ -w	100	Band width for banded alignment. If gaps are l…
□ -d	100	Off-diagonal X-dropoff. Stop extension when …
□ -r	1.5	Look for internal seeds inside a seed longer th…
☑ -c	10000	Skip seeds with more than INT occurrences.
□ -D	0.5	Drop chains shorter than FLOAT fraction of th…
□ -W	0	Drop a chain if the number of bases in seeds i…
□ -m	50	Perform at most INT rounds of mate-SW.
☑ -S		Skip mate rescue. Don't try to find missing ma…
□ -P		In the paired-end mode, perform SW to rescue…
□ -e		Discard full length exact matches.
☑ -L	5	Penalty for clipping. When performing SW exte…

Current Parameters

```
-k 19 -c 10000 -S -L 5
```

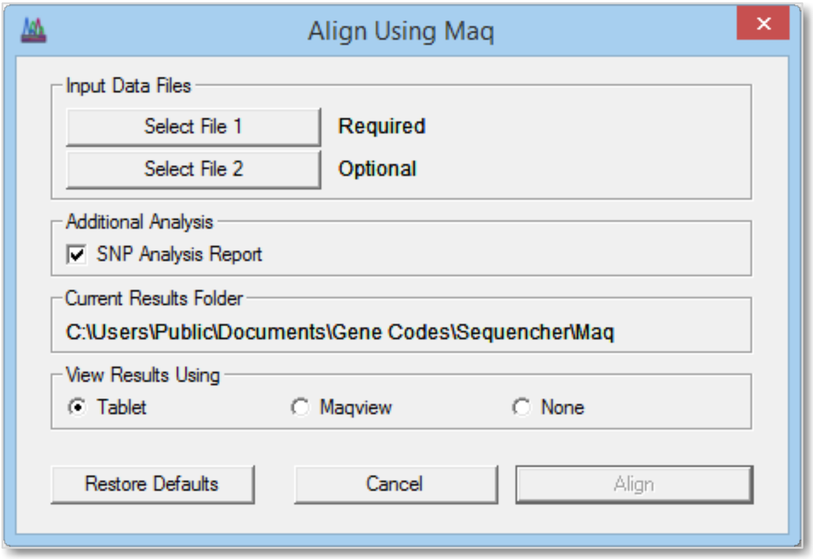
FURTHER ANALYSES WITH REFERENCE-GUIDED ALIGNMENTS

SNP HUNTING WITH MAQ

To perform this analysis, select your Reference Sequence from the **Project Window**. Now choose the command to perform SNP analysis with **Maq**, **Assemble>Align Data Files to Ref Using>Maq…** . The **Align Using Maq** dialog will appear.

From this dialog, click on the **Select File 1** button and browse to the reads file you want to use. If you are working with paired-end data, then click on the **Select File 2** button and browse to the second reads file you want to use.

Figure 16-26 Align Using Maq dialog



Align Using Maq

Input Data Files

Select File 1 Required

Select File 2 Optional

Additional Analysis

☑ SNP Analysis Report

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\Maq

View Results Using

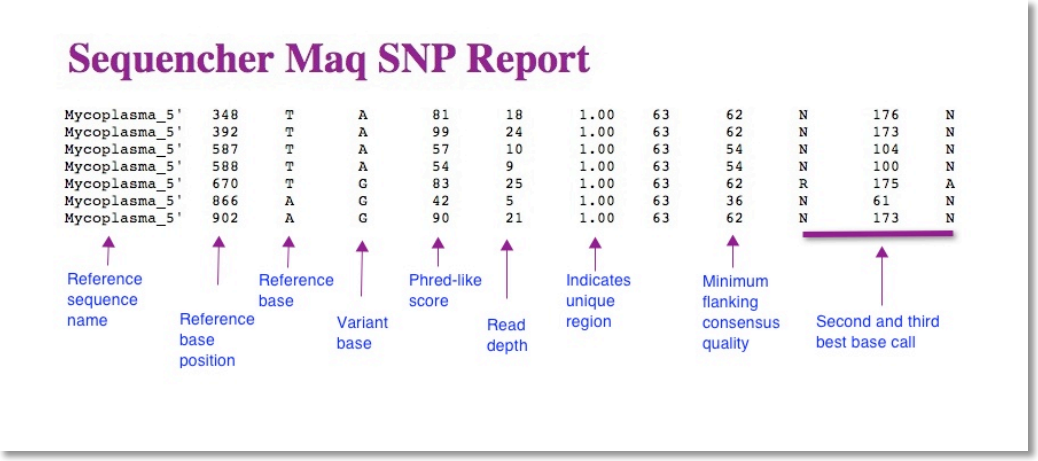
● Tablet ○ Maqview ○ None

Choose the **SNP Analysis Report** option by clicking on the **SNP Analysis Report** checkbox**.**

Finally, choose your viewing options. The results are saved as a contig in your project and a SNP report in the Gene Codes/Sequencher/Maq folder. If you chose **Tablet** as your view option, follow the instructions for using **Tablet** above. If you chose **None,** then you can view the contig assembly in **Tablet** (**Contig>Show NGS Data Using>Tablet**) or in **Maqview** (**Contig>Show NGS Data Using>Maqview**) at a later time.

To view the SNP report results, select the contig of interest and choose the **Sequence>Analyses>Maq SNP Report** command. The results will be displayed in your default web browser. The first column is the name of the Reference Sequence, the second column is the position of the reference base, next is the reference base itself followed by variant base. Column 5 contains a Phred-like quality value which will assist in determining the reliability of the SNP. Column 6 gives you the read depth. When the table reports 1.00 in column 7, this means that the region is near enough unique. The final columns relate to the second best and third best base calls at this position.

Figure 16-27 Explanation of MAQ SNP report fields



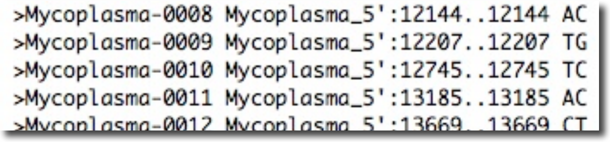
Sequencher Maq SNP Report

```
Mycoplasma_5'   348   T    A    81    18    1.00   63    62    N    176    N
Mycoplasma_5'   392   T    A    99    24    1.00   63    62    N    173    N
Mycoplasma_5'   587   T    A    57    10    1.00   63    54    N    104    N
Mycoplasma_5'   588   T    A    54    9     1.00   63    54    N    100    N
Mycoplasma_5'   670   T    G    83    25    1.00   63    62    R    175    A
Mycoplasma_5'   866   A    G    42    5     1.00   63    36    N    61     N
Mycoplasma_5'   902   A    G    90    21    1.00   63    62    N    173    N
```

SNP TOLERANT ALIGNMENT WITH GSNAP

GSNAP performs SNP hunting in a different way than **Maq**. You need to supply a file that contains a list of known SNPs, this has to be in a specific format which will be described below.

Figure 16-28 Format of known SNPs file



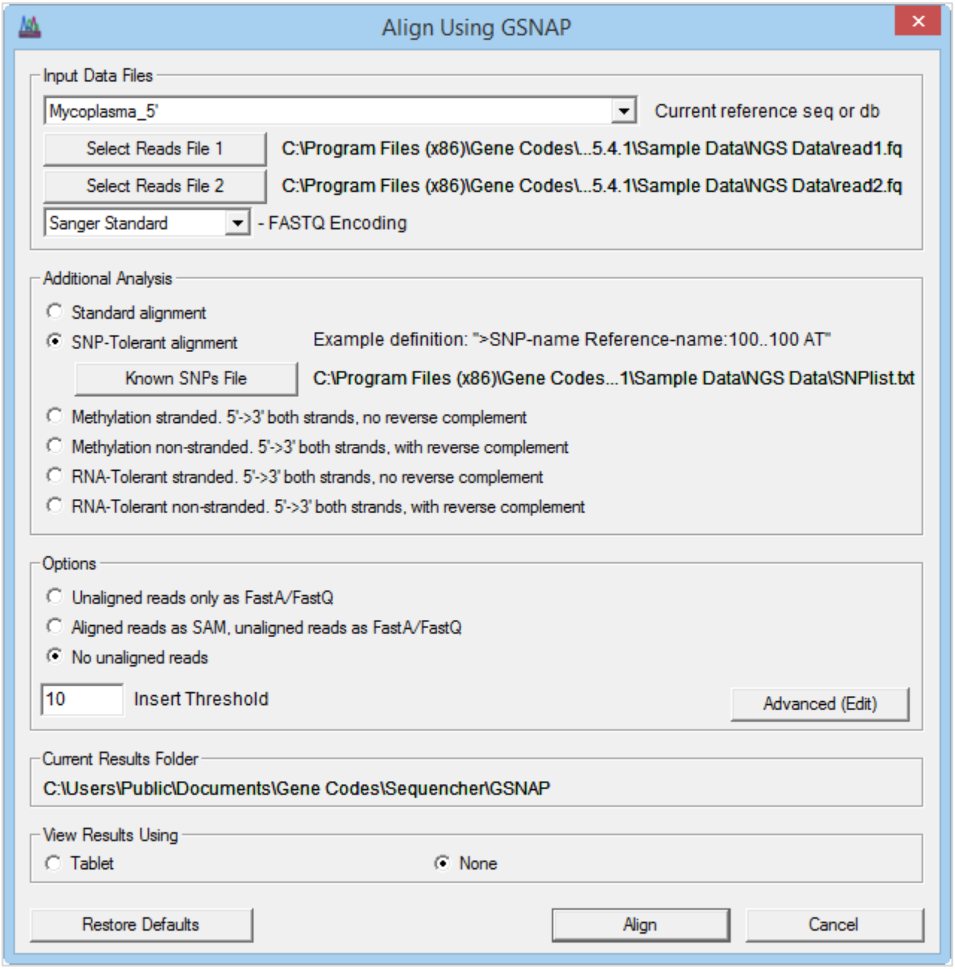
```
>Mycoplasma-0008 Mycoplasma_5':12144..12144 AC
>Mycoplasma-0009 Mycoplasma_5':12207..12207 TG
>Mycoplasma-0010 Mycoplasma_5':12745..12745 TC
>Mycoplasma-0011 Mycoplasma_5':13185..13185 AC
>Mycoplasma-0012 Mycoplasma_5':13669..13669 CT
```

The file is a text file and each SNP is placed on a separate line. Look at the image above.

Each line must begin with the > character. This is followed by an identifier, in this case it is a number preceded by the word Mycoplasma_5'. Next comes the Reference Sequence name and positional information in the format *name:xxx..xxx*. If there are spaces in the name, these should be replaced by underscores. The position information in this file always assumes that the first base of the Reference Sequence in **Sequencher** is 1, no matter its actual numbering relative to its chromosomal or contig position.

To perform the analysis, you need to select a Reference Sequence from the **Project Window** or a prebuilt database/index from the drop-down menu on the **GSNAP** dialog. Choose **Assemble>Align Data Files to Ref Using>GSNAP…**. The **Align Using GSNAP** and **External Data Browser** dialogs will appear. Click **Select Reads File 1** and browse to the first reads file you want to use. Click on **Select Reads File 2** if you are using paired-end data. Now choose **SNP-Tolerant alignment** from the **Additional Analysis** groupbox. The **Known SNPs File** button is now enabled. Click on this button and browse to the file containing your list of known SNPs.

Figure 16-29 Setting GSNAP SNP-tolerant alignment



Align Using GSNAP

Input Data Files

Mycoplasma_5' — Current reference seq or db

Select Reads File 1 — C:\Program Files (x86)\Gene Codes\..5.4.1\Sample Data\NGS Data\read1.fq

Select Reads File 2 — C:\Program Files (x86)\Gene Codes\..5.4.1\Sample Data\NGS Data\read2.fq

Sanger Standard - FASTQ Encoding

Additional Analysis

○ Standard alignment

● SNP-Tolerant alignment — Example definition: ">SNP-name Reference-name:100..100 AT"

Known SNPs File — C:\Program Files (x86)\Gene Codes...1\Sample Data\NGS Data\SNPlist.txt

○ Methylation stranded. 5'->3' both strands, no reverse complement

○ Methylation non-stranded. 5'->3' both strands, with reverse complement

○ RNA-Tolerant stranded. 5'->3' both strands, no reverse complement

○ RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options

○ Unaligned reads only as FastA/FastQ

○ Aligned reads as SAM, unaligned reads as FastA/FastQ

● No unaligned reads

10 Insert Threshold

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\GSNAP

View Results Using

○ Tablet ● None

Choose also whether or not you want to see the results file in **Tablet** immediately after the alignment has finished. If you choose to see the results immediately, the results file will be opened in **Tablet** if it is installed. The results file will also be saved in the Run folder. Now click on the **Align** button to initiate the analysis. You can review the results later by choosing the contig of interest in the **Project Window** and choosing the **Sequence>Analyses>GSNAP SNP Analysis** command.

METHYLATION STUDIES WITH GSNAP

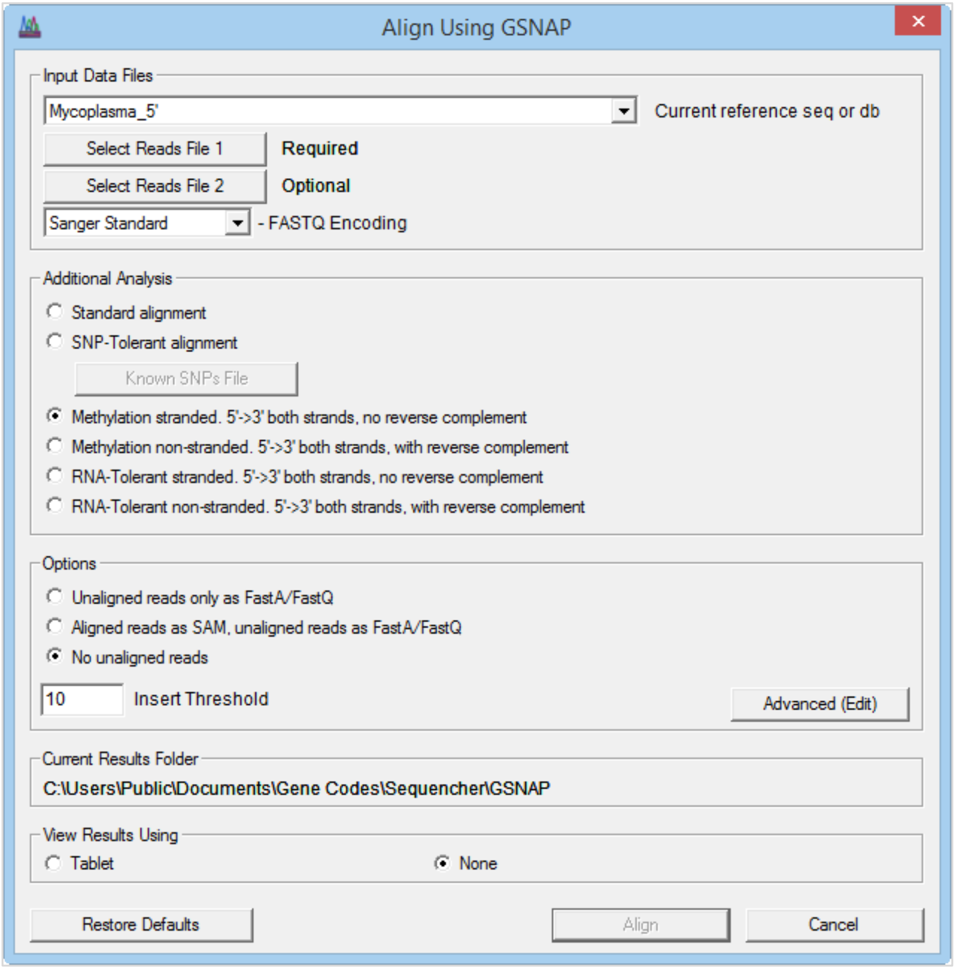
Studies of methylated parts of the genome involve using bisulfite treatment of the DNA followed by sequencing of the regions of interest. The bisulfite treatment chemically converts unmethylated Cs to Ts. **GSNAP** is capable of aligning the sequenced reads to the genomic (untreated) sequence.

To perform the analysis, you need to select a Reference Sequence from the **Project Window** or a prebuilt database/index from the **Align Using GSNAP** dialog and one or two (if paired-end) **FastA** or **FastQ** files containing reads from bisulfite treated DNA.

Load the Reference Sequence into **Sequencher** and ensure it is highlighted in the **Project Window**. Choose **Assemble**>**Align Data Files to Ref Using>GSNAP….** The **Align Using GSNAP** and **External Data Browser** dialogs will appear. If you are working with a prebuilt database, choose it now from the **Current reference seq or db** or **index** drop-down menu.

Next, click on the **Select Reads File 1** button and browse to the first reads file you want to use. Click on the **Select Reads File 2** button if you are using paired-end data. Two protocols are supported, referred to as stranded (for protocols which allow only 5’ -> 3’ RNA) and non-stranded (both sense and antisense reads are allowed). In **Sequencher**, you choose the stranded mode by clicking on the **Methylation stranded. 5’ -> 3’ both strands, no reverse complement** radio button. To choose the non-stranded protocol, click on the **Methylation non-stranded. 5’ -> 3’ both strands, with reverse complement** radio button.

Figure 16-30 GSNAP methylation settings



Align Using GSNAP

Input Data Files

Mycoplasma_5' — Current reference seq or db

Select Reads File 1 — Required

Select Reads File 2 — Optional

Sanger Standard - FASTQ Encoding

Additional Analysis

○ Standard alignment

○ SNP-Tolerant alignment

Known SNPs File

● Methylation stranded. 5'->3' both strands, no reverse complement

○ Methylation non-stranded. 5'->3' both strands, with reverse complement

○ RNA-Tolerant stranded. 5'->3' both strands, no reverse complement

○ RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options

○ Unaligned reads only as FastA/FastQ

○ Aligned reads as SAM, unaligned reads as FastA/FastQ

● No unaligned reads

10 Insert Threshold

Advanced (Edit)

Current Results Folder

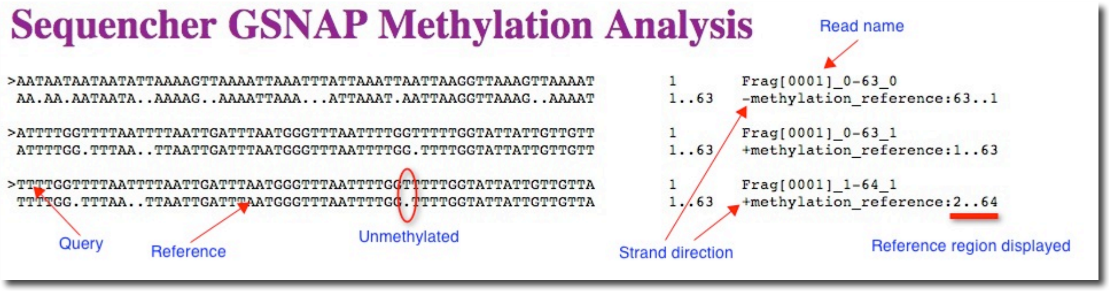
C:\Users\Public\Documents\Gene Codes\Sequencher\GSNAP

View Results Using

○ Tablet ● None

You can review the results later by choosing the contig of interest and choosing the relevant **Sequence>Analyses>GSNAP Methylation Analysis** drop-down menu item. The results will appear in your default browser.

Figure 16-31 Results File for Methylation Analysis



Sequencher GSNAP Methylation Analysis

```
>AATAATAATAATATTAAAAGTTAAAATTAAATTTATTAAATTAATTAAGGTTAAAGTTAAAAT      1       Frag[0001]_0-63_0
 AA.AA.AATAATA..AAAAG..AAAATTAAA...ATTAAAT.AATTAAGGTTAAAG..AAAAT      1..63   -methylation_reference:63..1

>ATTTTGGTTTTAATTTTAATTGATTTAATGGGTTTAATTTTGGTTTTTGGTATTATTGTTGTT      1       Frag[0001]_0-63_1
 ATTTTGG.TTTAA..TTAATTGATTTAATGGGTTTAATTTTGG.TTTTGGTATTATTGTTGTT      1..63   +methylation_reference:1..63

>TTTTGGTTTTAATTTTAATTGATTTAATGGGTTTAATTTTGGTTTTTGGTATTATTGTTGTTA     1       Frag[0001]_1-64_1
 TTTTGG.TTTAA..TTAATTGATTTAATGGGTTTAATTTTGG.TTTTGGTATTATTGTTGTTA     1..63   +methylation_reference:2..64
```

The image above explains the various parts of the report. The most informative part of the report is the character circled in red and called unmethylated. The . (period) character indicates a C in the Reference Sequence which occurs opposite a T in the query read. This indicates that this C was unmethylated and hence open to modification by the bisulfite reagent.

RNA A-TO-I TOLERANT ALIGNMENT WITH GSNAP

RNA editing is a process where some adenosines in mRNA are converted to inosine by adenosine deaminase acting on RNA (ADAR). The mutations caused by this process have been associated with diversification of the transcriptome. Aberrant activity is associated with a wide variety of human diseases, regulation of splicing and disease pathologies. The changes caused by ADAR activity may be detected using NGS and looking for adenosines that have been converted to guanosines.

To perform the analysis, you need to select a Reference Sequence from the **Project Window** or a prebuilt database/index from the **Align Using GSNAP** dialog and one or two (if paired-end) FastA or FastQ files. Load the Reference Sequence into **Sequencher** and ensure it is highlighted in the **Project Window**. Choose **Assemble>Align Data Files to Ref Using>GSNAP…**. The **Align Using GSNAP** and **External Data Browser** dialogs will appear. If you are working with a prebuilt database, then choose it now from the **Current reference seq db or index** drop-down menu. From the dialog, click on the **Select Reads File 1** button and browse to the first reads file you want to use. Click on the **Select Reads File 2** button if you are using paired-end data.

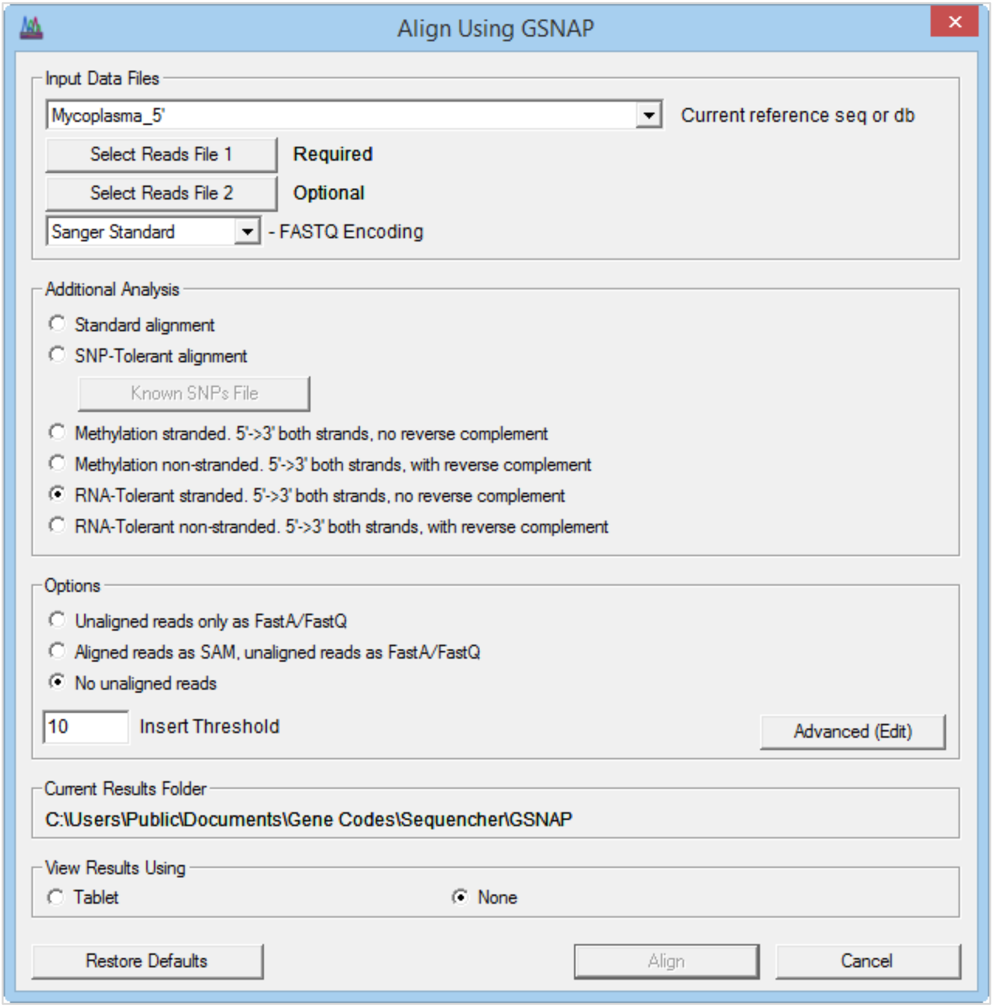
In order to use the alignment mode that is tolerant to the A-to-G changes correctly, you need to know which laboratory protocol was used in preparing your reads. Two protocols are supported, you may see these referred to as stranded and non-stranded. In **Sequencher**, you choose the stranded mode (for protocols which allow only 5’ -> 3’ RNA) by clicking on the **RNA-Tolerant stranded. 5’ -> 3’ both strands, no reverse complement** radio button. To choose the non-stranded protocol (both sense and antisense reads are allowed), click on the **RNA-Tolerant stranded. 5’ -> 3’ both strands, with reverse complement** radio button.

After you have set any other options or advanced options, you launch the analysis by clicking on the **Align** button.

You can review the results later by choosing the contig of interest in the **Project Window** and choosing the **Sequence>Analyses>GSNAP RNA-Tolerant Analysis** menu item. The results will appear in your default browser.

Note: If you are in Viewer mode and you request an HTML report for a newly run GSNAP SNP Tolerant alignment, Methylation Analysis, or RNA-Tolerant alignment, you will see an example report.

Figure 16-32 RNA-Tolerant stranded alignment



Align Using GSNAP

Input Data Files

Mycoplasma_5' — Current reference seq or db

Select Reads File 1 — Required

Select Reads File 2 — Optional

Sanger Standard - FASTQ Encoding

Additional Analysis

○ Standard alignment

○ SNP-Tolerant alignment

Known SNPs File

○ Methylation stranded. 5'->3' both strands, no reverse complement

○ Methylation non-stranded. 5'->3' both strands, with reverse complement

● RNA-Tolerant stranded. 5'->3' both strands, no reverse complement

○ RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options

○ Unaligned reads only as FastA/FastQ

○ Aligned reads as SAM, unaligned reads as FastA/FastQ

● No unaligned reads

10 Insert Threshold

Advanced (Edit)

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\GSNAP

View Results Using

○ Tablet ● None

Restore Defaults Align Cancel

VARIANT CALLING WITH SAMTOOLS

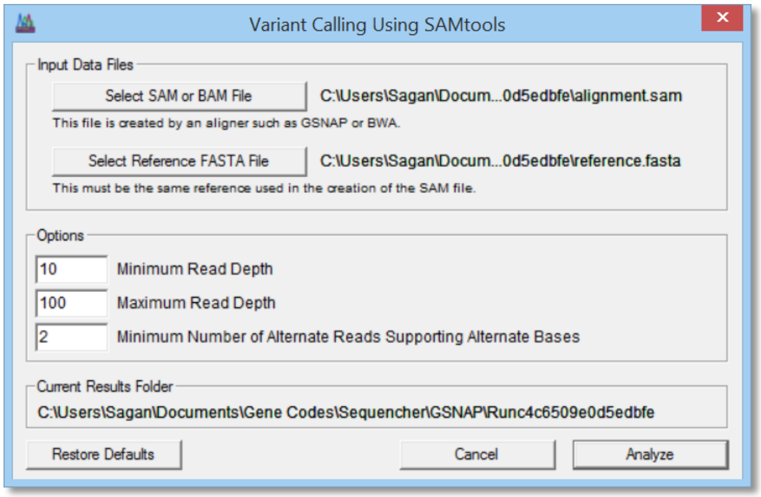
You can now use **SAMtools** to search for variants in any alignment where the results have been saved in **SAM** or **BAM** format. **SAMtools** does this by changing the mapped data to genome or chromosome numbering (based on your reference). It then calculates the Bayesian prior probability from the data. Next, **SAMtools** uses the prior probability distribution and the data to produce a genotype which is filtered and written out to a VCF file. You can view the VCF file and your aligned data in any genome browser that reads **SAM** and **VCF** formats.

There are two paths you can take for analysing your SAM/BAM files. In the first path, open a project containing contigs which have been created from aligned NGS data. Highlight a single contig, then choose **Sequence>Analyses>SAMtools Variant Calling…** A dialog opens. Notice that the path information for the SAM file and the reference sequence are already filled in. The second method is to choose **Sequence>Analyses>SAMtools Variant Calling…** and when the dialog opens, use the buttons to browse to your SAM/BAM file and then browse to your FastA reference sequence.

There are three options for you to set, the **Minimum Read Depth**, the **Maximum Read Depth**, and the **Minimum Number of Alternate Reads Supporting Alternate Bases**.

The **Minimum Read Depth** is self-explanatory. The **Maximum Read Depth** is used to provide a threshold since SNPs with very high coverage might represent variation from variable copy number repeats. The parameter should be altered to suit the coverage in your data. The **Minimum Number of Alternate Reads Supporting Alternate Bases** parameter is designed to set a lower threshold for the alternate reads supporting each alternative allele. Once you are happy with your selections, click on the **Analyze** button. A VCF (Variant Calling Format) file will be created in the same folder as the SAM/BAM file you have just analyzed.

Figure 16-33 Variant Calling with SAMtools dialog



Variant Calling Using SAMtools

Input Data Files

Select SAM or BAM File C:\Users\Sagan\Docum...0d5edbfe\alignment.sam

This file is created by an aligner such as GSNAP or BWA.

Select Reference FASTA File C:\Users\Sagan\Docum...0d5edbfe\reference.fasta

This must be the same reference used in the creation of the SAM file.

Options

10 Minimum Read Depth

100 Maximum Read Depth

2 Minimum Number of Alternate Reads Supporting Alternate Bases

Current Results Folder

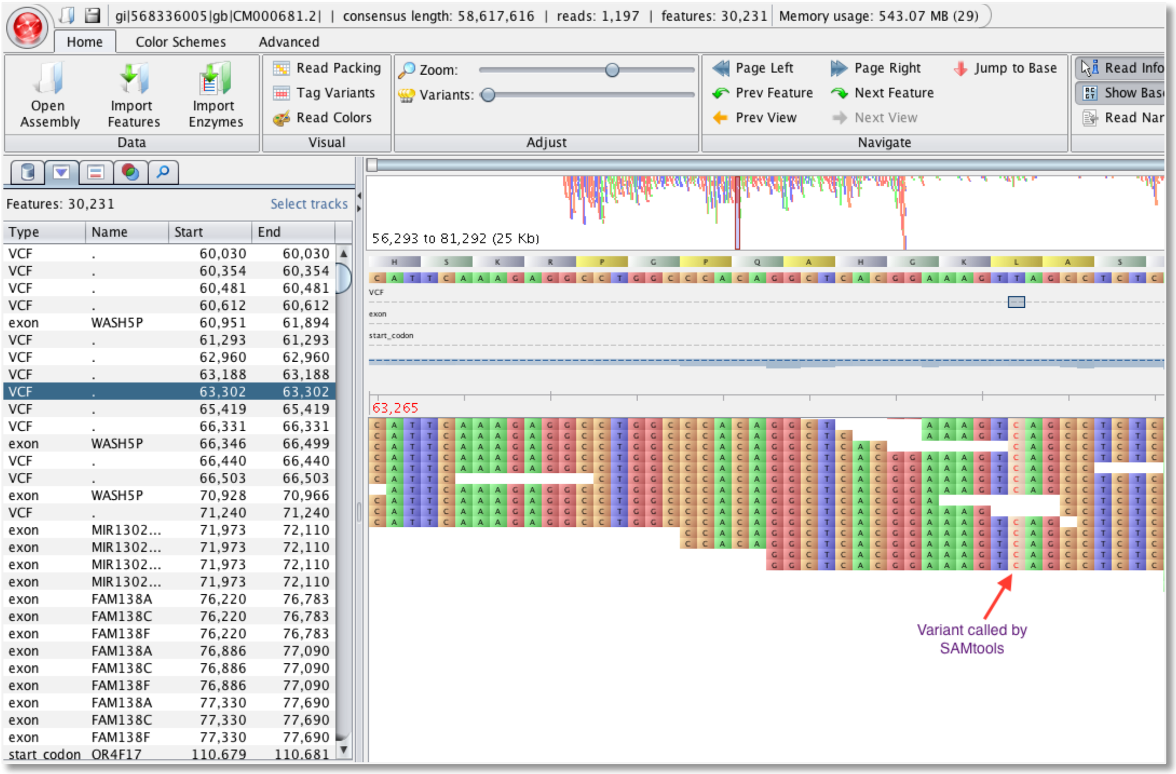
C:\Users\Sagan\Documents\Gene Codes\Sequencher\GSNAP\Runc4c6509e0d5edbfe

VIEWING VCF FILES IN TABLET

You can view the results using the **Tablet** genome browser. If you are working in a project and have aligned your reads and called the variants as described above, you can review the alignments as follows, highlight one contig in your project and choose **Contig>Show NGS Data Using>Tablet.** The **Tablet** genome browser will launch and the aligned reads will be loaded into the browser. You now load the VCF file into the browser by clicking on the **Import Features** button. Click on the second tab to see a list of the variants and their locations. Click on a variant to jump to its location in the contig. You can see an example of this in the image below.

Note: You can also load other VCF files and GTF files using this same method. This will give you a rich representation of known features in your assembled reads.

Figure 16-34 Example of VCF file import in Tablet with features from GTF file



gi|568336005|gb|CM000681.2| | consensus length: 58,617,616 | reads: 1,197 | features: 30,231 | Memory usage: 543.07 MB (29)

Features: 30,231

Type	Name	Start	End
VCF	.	60,030	60,030
VCF	.	60,354	60,354
VCF	.	60,481	60,481
VCF	.	60,612	60,612
exon	WASH5P	60,951	61,894
VCF	.	61,293	61,293
VCF	.	62,960	62,960
VCF	.	63,188	63,188
VCF	.	63,302	63,302
VCF	.	65,419	65,419
VCF	.	66,331	66,331
exon	WASH5P	66,346	66,499
VCF	.	66,440	66,440
VCF	.	66,503	66,503
exon	WASH5P	70,928	70,966
VCF	.	71,240	71,240
exon	MIR1302...	71,973	72,110
exon	MIR1302...	71,973	72,110
exon	MIR1302...	71,973	72,110
exon	MIR1302...	71,973	72,110
exon	FAM138A	76,220	76,783
exon	FAM138C	76,220	76,783
exon	FAM138F	76,220	76,783
exon	FAM138A	76,886	77,090
exon	FAM138C	76,886	77,090
exon	FAM138F	76,886	77,090
exon	FAM138A	77,330	77,690
exon	FAM138C	77,330	77,690
exon	FAM138F	77,330	77,690
start_codon	OR4F17	110,679	110,681

REFERENCE-GUIDED ALIGNERS AND MULTIPLEX IDS

Multiplexing reads is a way of maximizing reagent and sequencer use by mixing a number of separate samples together in one flowcell (for example) and sequencing them at the same time. The DNA from each sample is tagged with a unique DNA identifier. After the sequencing run, the reads are separated using the unique DNA identifier.

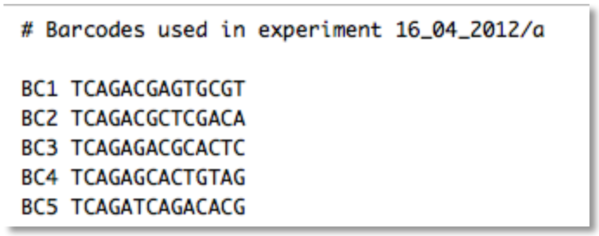
Sequencher can perform this separation or demultiplexing of the reads automatically for you. This is known as binning and the resultant files are called bins. **Sequencher** then takes each bin and aligns the reads it contains against the chosen Reference Sequence.

You need to supply a Reference Sequence, one FastA/FastQ file containing reads from a Multiplex ID experiment if you are working with single-end sequences, or two FastA/FastQ files containing reads if you are working with paired-end sequences.

FORMAT OF BARCODE FILE

You will also need to supply a text file containing the barcodes used in this Multiplex ID experiment. This file is called the barcodes file and it contains one barcode name and sequence per line. The file can accommodate a single comment line where the comment is preceded by a # mark.

Figure 16-35 Format of barcodes file



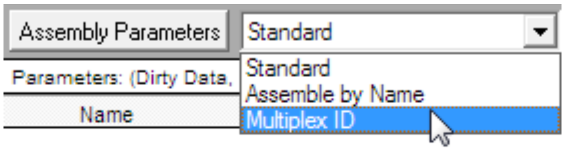
```
# Barcodes used in experiment 16_04_2012/a

BC1 TCAGACGAGTGCGT
BC2 TCAGACGCTCGACA
BC3 TCAGAGACGCACTC
BC4 TCAGAGCACTGTAG
BC5 TCAGATCAGACACG
```

***Note**:* The sequence should consist of the tag and barcode if the tag has not been removed by your sequencing pipeline.

ENABLING MULTIPLEX ID MODE

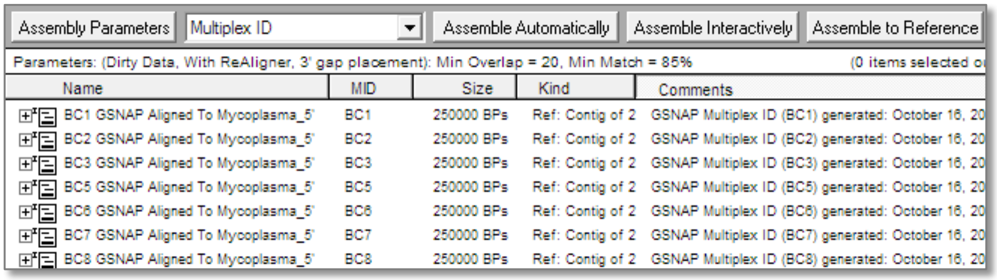
First you will need to enable the **Multiplex ID** mode by choosing **Multiplex ID** from the drop-down menu on the **Project Window** button bar.

Figure 16-36 Drop-down Mode menu



Once you have enabled this mode, a new column appears in the **Project Window**, this is called MI**D** and will contain the name of the barcode that originally formed part of the reads aligned to create this contig.

Figure 16-37 Project Window view showing MID column



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%   (0 items selected o

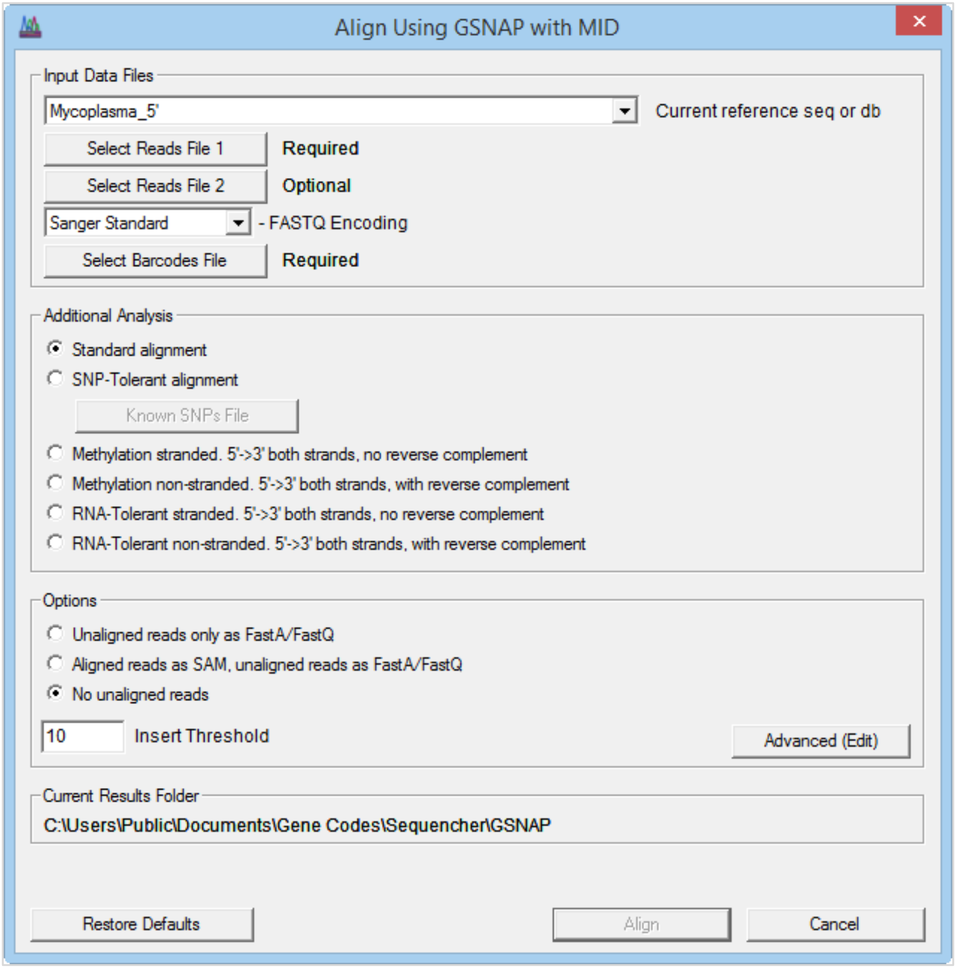
Name	MID	Size	Kind	Comments
BC1 GSNAP Aligned To Mycoplasma_5'	BC1	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC1) generated: October 16, 20
BC2 GSNAP Aligned To Mycoplasma_5'	BC2	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC2) generated: October 16, 20
BC3 GSNAP Aligned To Mycoplasma_5'	BC3	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC3) generated: October 16, 20
BC5 GSNAP Aligned To Mycoplasma_5'	BC5	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC5) generated: October 16, 20
BC6 GSNAP Aligned To Mycoplasma_5'	BC6	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC6) generated: October 16, 20
BC7 GSNAP Aligned To Mycoplasma_5'	BC7	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC7) generated: October 16, 20
BC8 GSNAP Aligned To Mycoplasma_5'	BC8	250000 BPs	Ref: Contig of 2	GSNAP Multiplex ID (BC8) generated: October 16, 20

If the **Multiplex ID** mode is still enabled when you save and close your project, **Sequencher** will remember this setting for you.

RUNNING GSNAP WITH MULTIPLEX DATA

Load the Reference Sequence into **Sequencher** and ensure it is highlighted in the **Project Window** or choose a database/index from the **Align Using GSNAP** dialog. Choose **Assemble>Align Data Files to Ref Using>GSNAP….** The **Align Using GSNAP with MID** and **External Data Browser** dialogs will appear. Now click on the **Select Reads File 1** button (if using single-end reads) and browse to the reads file you want to use. Click on the **Select Barcodes File** button and browse to your prepared barcodes file. Once you are ready, click on the **Align** button.

Figure 16-38 Align Using GSNAP with MID dialog



Align Using GSNAP with MID

Input Data Files

Mycoplasma_5' — Current reference seq or db

Select Reads File 1 — Required

Select Reads File 2 — Optional

Sanger Standard - FASTQ Encoding

Select Barcodes File — Required

Additional Analysis

● Standard alignment

○ SNP-Tolerant alignment

Known SNPs File

○ Methylation stranded. 5'->3' both strands, no reverse complement

○ Methylation non-stranded. 5'->3' both strands, with reverse complement

○ RNA-Tolerant stranded. 5'->3' both strands, no reverse complement

○ RNA-Tolerant non-stranded. 5'->3' both strands, with reverse complement

Options

○ Unaligned reads only as FastA/FastQ

○ Aligned reads as SAM, unaligned reads as FastA/FastQ

● No unaligned reads

10 Insert Threshold

Advanced (Edit)

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\GSNAP

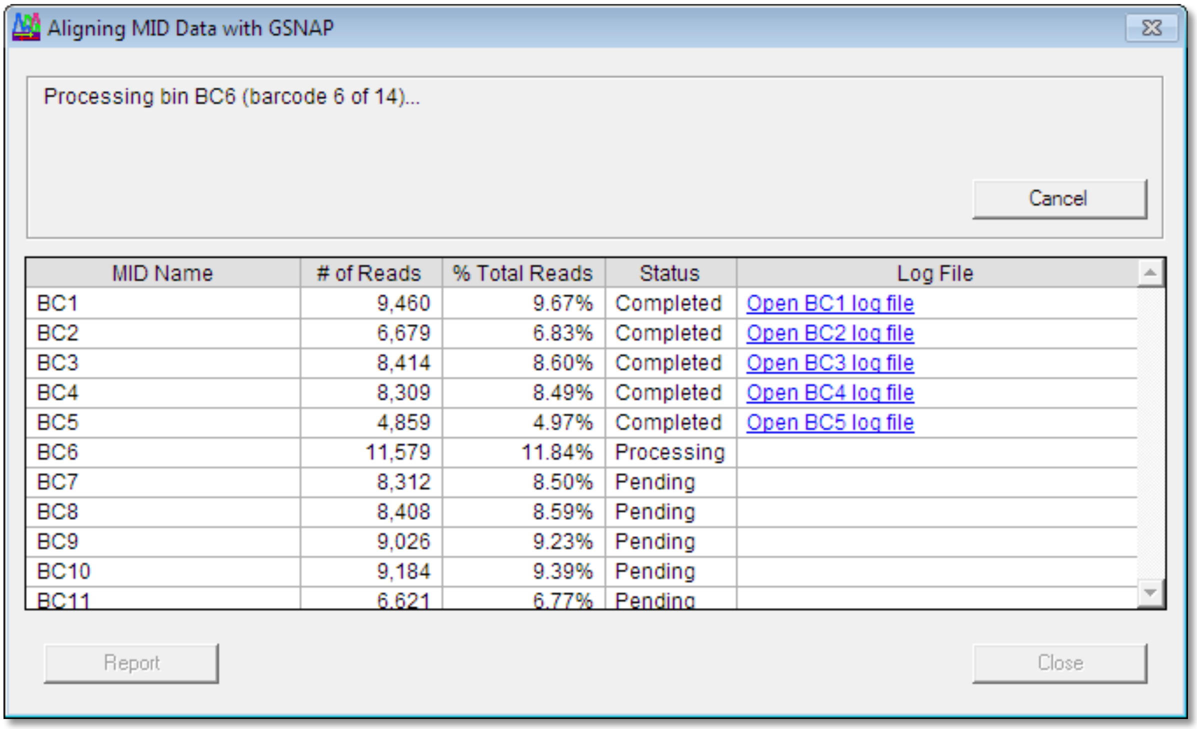
Restore Defaults

Align

Cancel

A new dialog appears (below). This dialog is divided into two panes. The top pane shows a constantly updated status line for the overall process. This pane also contains a **Cancel** button, if you click this button, you will cancel the current and all subsequent operations for the current MID run. The lower pane is a table that is divided into a number of columns (see figure below) showing the name of the Barcode, the number of reads in the bin, the % of the total reads sequenced that this represents, the status of the specific bin alignment, and finally a link to the log file. Click on this link to open the log file in the default viewer.

Figure 16-39 Multiplex ID with GSNAP status report



Processing bin BC6 (barcode 6 of 14)...

MID Name	# of Reads	% Total Reads	Status	Log File
BC1	9,460	9.67%	Completed	Open BC1 log file
BC2	6,679	6.83%	Completed	Open BC2 log file
BC3	8,414	8.60%	Completed	Open BC3 log file
BC4	8,309	8.49%	Completed	Open BC4 log file
BC5	4,859	4.97%	Completed	Open BC5 log file
BC6	11,579	11.84%	Processing	
BC7	8,312	8.50%	Pending	
BC8	8,408	8.59%	Pending	
BC9	9,026	9.23%	Pending	
BC10	9,184	9.39%	Pending	
BC11	6,621	6.77%	Pending	

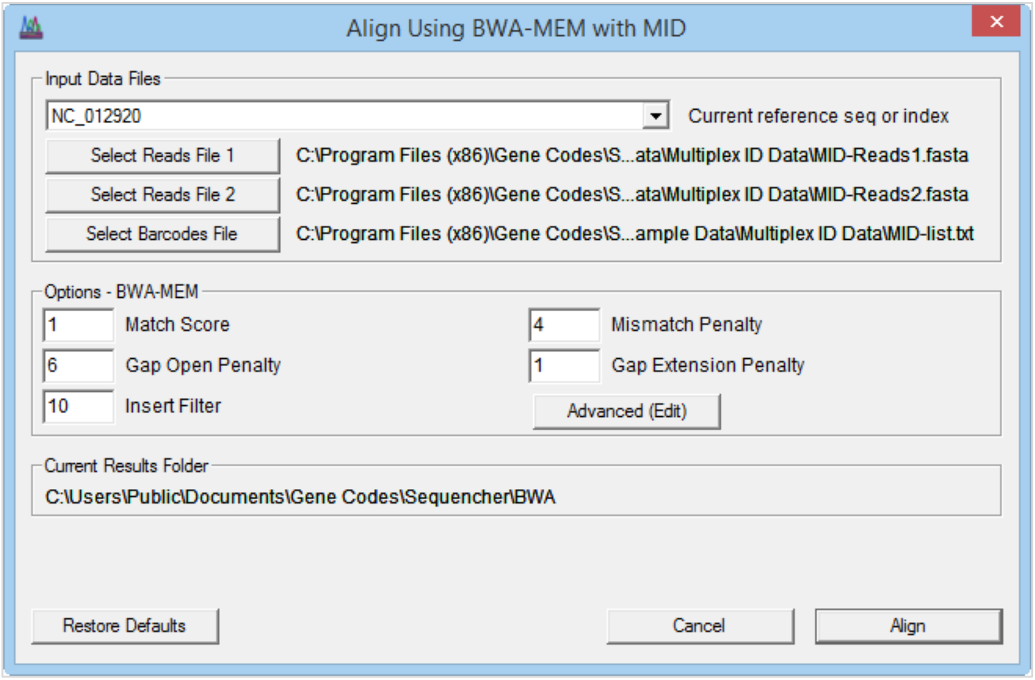
There is also a **Close** button that becomes enabled when the entire process has completed. Click on the **Report** button to obtain a formatted **Multiplex Status Report** that will open in a **Report Preview** window. You can save or print this report.

When you return to the **Project Window**, you will see the consensus sequence for each bin that has been returned to **Sequencher** and assembled to the Reference Sequence in order to create a contig. In the case of the above image, if 14 files containing reads were created during de-multiplexing, then 14 contigs will be created within **Sequencher**.

RUNNING BWA-MEM WITH MULTIPLEX DATA

Load the Reference Sequence into **Sequencher** and ensure it is highlighted in the **Project Window** or choose a database/index from the **Align Using BWA-MEM** dialog. Choose **Assemble>Align Data Files to Ref Using>BWA-MEM…**. The **Align Using BWA-MEM with MID** and **External Data Browser** dialogs will appear. Now click on the **Select Reads File 1** button (if using single-end reads) and browse to the reads file you want to use. Click on the **Select Barcodes File** button and browse to your prepared barcodes file. Once you are ready, click the **Align** button.

Figure 16-40 Align Using BWA-MEM with MID dialog



Input Data Files

NC_012920 — Current reference seq or index

Select Reads File 1: C:\Program Files (x86)\Gene Codes\S...ata\Multiplex ID Data\MID-Reads1.fasta

Select Reads File 2: C:\Program Files (x86)\Gene Codes\S...ata\Multiplex ID Data\MID-Reads2.fasta

Select Barcodes File: C:\Program Files (x86)\Gene Codes\S...ample Data\Multiplex ID Data\MID-list.txt

Options - BWA-MEM

1	Match Score	4	Mismatch Penalty
6	Gap Open Penalty	1	Gap Extension Penalty
10	Insert Filter		

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\BWA

A new dialog appears (see similar figure above, Figure 16-39 Multiplex ID with GSNAP status report).

This dialog is divided into two panes. The top pane shows a constantly updated status line for the overall process. This pane also contains a **Cancel** button, if you click this button you will cancel the current and all subsequent operations for the current MID run. The lower pane is a table that is divided into a number of columns showing the name of the Barcode, the number of reads in the bin, the % of the total reads sequenced that this represents, the status of the specific bin alignment, and finally a link to the log file. Click on this link to open the log file in the default viewer.

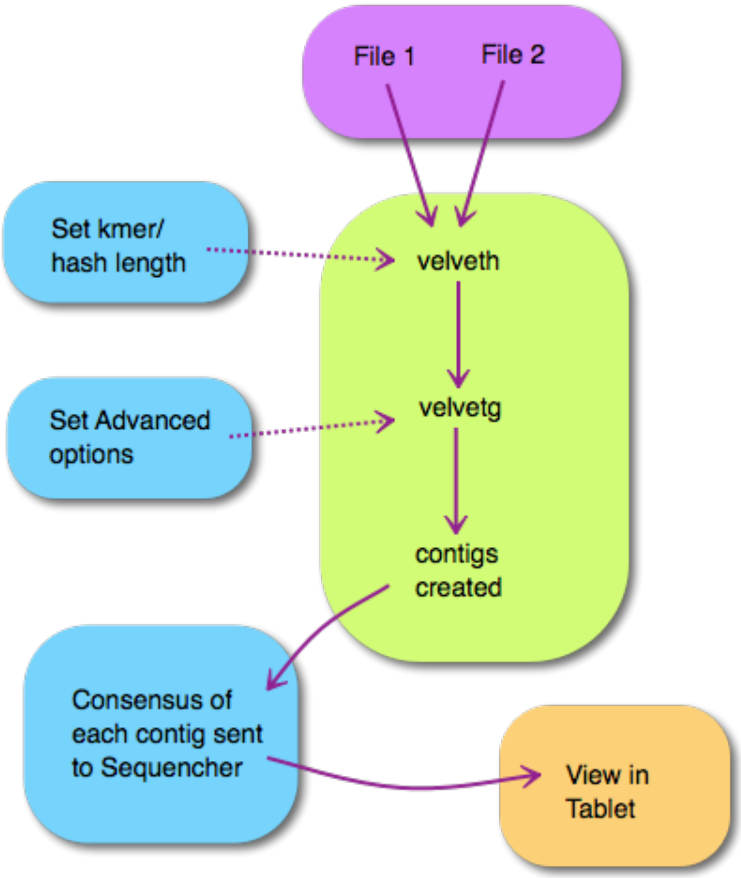
There is also a **Close** button that becomes enabled when the entire process has completed. Click on the **Report** button to obtain a formatted **Multiplex Status Report** that will open in a **Report Preview** window. You can save or print this report.

When you return to the **Project Window**, you will see the consensus sequence for each bin that has been returned to **Sequencher** and assembled to the Reference Sequence in order to create a contig. In the case of the above image, if 14 files containing reads were created during de-multiplexing, then 14 contigs will be created within **Sequencher**.

PERFORMING DE NOVO ASSEMBLIES

GSNAP and **BWA-MEM** are great programs for reference-guided alignment, that is to say you have a sequence that is identical or closely related to the region you are sequencing. If you are sequencing an entirely new organism for example, or sequencing a new region that has not been sequenced before, you may not have a reference sequence available. In this case, you will need to perform a De novo assembly. **Sequencher** allows you to perform a De novo assembly using the **Velvet** algorithm.

Figure 16-41 Velvet de novo assembly workflow



ASSEMBLING SINGLE-END DATA WITH VELVET

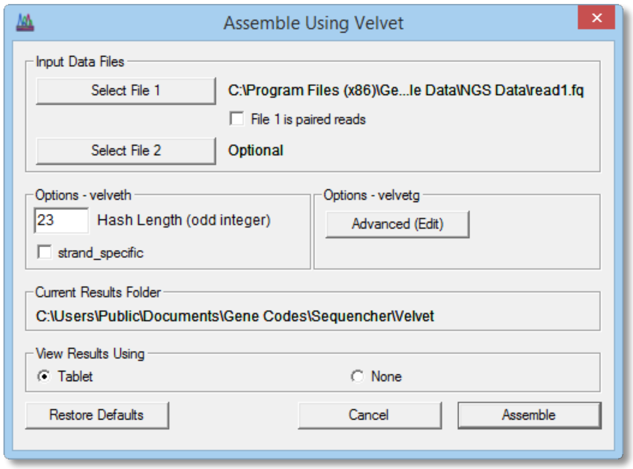
Since this is a *De novo* assembly, you do not need a Reference Sequence. Choose **Assemble>Assemble Data Files Using>Velvet…**. The **Assemble Using Velvet** and **External Data Browser** dialogs will appear. Click on the **Select File 1** button and browse to the reads file you want to use. **Hash Length** is also known as k-mer or word length. If a DNA string can be considered to consist of a series of characters (with an alphabet of four letters), then a word/k-mer/hash length will be a string of characters of a specified length.

The hash length is one of the key parameters for **Velvet** and is set in the **Assemble Using Velvet** dialog. The hash length must be an odd integer (23, 35, or 47 for example) and this value should be shorter than most of the reads in the reads file. Anything shorter than this value will be ignored. Anything equal to or longer than this value will be used in the assembly. The longer the hash length, the more specific it is. The shorter the hash length, the more sensitive it is. Finding the best hash length will require a balance between specificity and sensitivity. Enter the new value in the Hash Length input field and then run the assembly by clicking on the **Assemble** button.

Once the assembly is complete, the dialog is automatically dismissed and a number of sequences will appear in the **Project Window**. Each sequence will represent the consensus of a contig created by the **Velvet** application, of which there may be many. It may be worth using **Sequencher's** built-in assembly algorithms to form contigs from some or all of these consensus sequences.

If you want to review the contigs from which these consensus sequences were derived, then highlight one consensus sequence and choose **Contig>Show NGS Data Using>Tablet.**

Figure 16-42 Setting the Hash Length for Velvet in the Velvet dialog



Assemble Using Velvet

Input Data Files

Select File 1 C:\Program Files (x86)\Ge...le Data\NGS Data\read1.fq

□ File 1 is paired reads

Select File 2 Optional

Options - velveth

23 Hash Length (odd integer)

□ strand_specific

Options - velvetg

Advanced (Edit)

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\Velvet

View Results Using

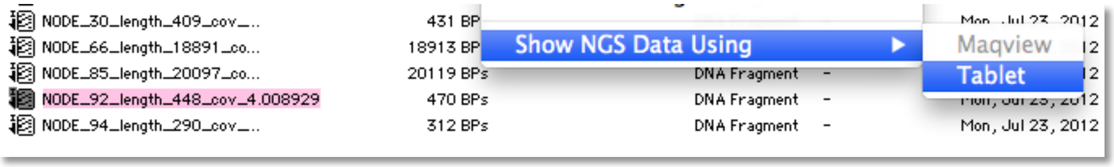
● Tablet ○ None

Restore Defaults Cancel Assemble

ASSEMBLING PAIRED-END DATA WITH VELVET

The **Velvet** assembler also works with paired-end data. Select menu command **Assemble>Assemble Data Files Using>Velvet…**, then choose the two paired-ends reads files by clicking on the **Select File 1** button and then the **Select File 2** button. Next, choose whether you want to view the results after assembly by selecting the **Tablet** or **None** radio button. Finally, click on the **Assemble** button.

Figure 16-43 Choosing to view the reads of a highlighted consensus in Tablet



If you know that your paired files have already been merged into a single file, then follow these steps instead. Click on the **Select File 1** button and locate the merged file. Next click in the **File 1 is paired reads** checkbox. The hash length is also known as k-mer or word length. If a DNA string can be considered to consist of a series of characters (with an alphabet of four letters), then a word/k-mer/hash length will be a string of characters of a specified length.

The hash length is one of the key parameters for **Velvet** and is set in the **Assemble Using Velvet** dialog. It must be an odd integer (23, 35, or 47 for example) and this value should be shorter than most of the reads in the reads file. Anything shorter than this value will be ignored. Anything equal to or longer than this value will be used in the assembly. The longer the hash length, the more specific it is. The shorter the hash length, the more sensitive it is. Finding the best hash length will require a balance between specificity and sensitivity.

Enter the new value in the Hash Length input field. Choose your viewing options and then click on the **Assemble** button.

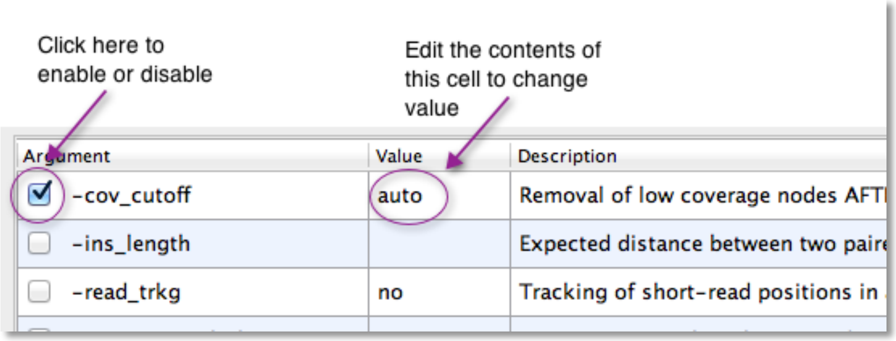
LOOKING AT VELVET RESULTS IN TABLET

Once you have submitted your data, the results come back into **Sequencher** as a series of consensus sequences representing each contig constructed by **Velvet**. If you want to explore the underlying data contributing to the consensus, then you need to use the **Tablet** browser. To do this, you need to select the consensus whose contig you want to explore, then choose menu command **Contig>Show NGS Data Using>Tablet**. The contig will open in the **Tablet** browser.

ADVANCED PARAMETERS FOR VELVET

In order to see what the advanced parameters are, click on the **Advanced (Edit)** button. The **Velvet Advanced Options** dialog appears. This is in the form of a table with 3 columns. The first column contains a checkbox and parameter, the second column contains the value of that parameter, and the third column its description. These parameters are specific to **Velvetg**. In order to use a parameter, simply click on the checkbox so that a check appears. If you want to change the value, then click in the **Value** cell for that parameter and edit the value. If you are unable to edit the value or the value cell is empty initially, it may be that this parameter is a flag. That is to say, it has no value but is used in an on or off state.

Figure 16-44 Enabling parameters and setting values in Velvet Advanced dialog



Click here to enable or disable

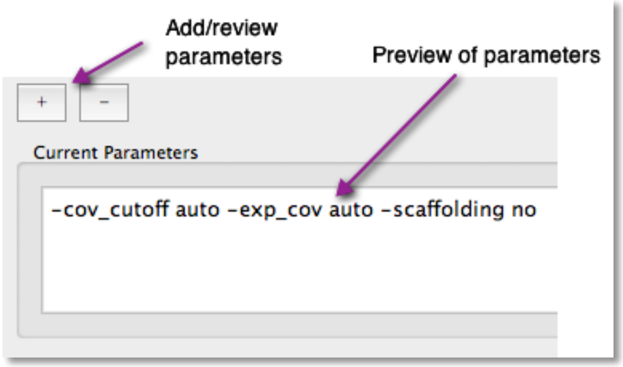
Edit the contents of this cell to change value

Argument	Value	Description
☑ -cov_cutoff	auto	Removal of low coverage nodes AFTI
□ -ins_length		Expected distance between two pair
□ -read_trkg	no	Tracking of short-read positions in

Below this table are two buttons, a **+** (plus) and a **–** (minus). These are used to add or remove parameters, values, and their descriptions.

Beneath this is a pane titled **Current Parameters**. Any changes to the states of checkboxes will cause a parameter to appear or disappear from this pane. Items in this pane will become part of the command line that **Sequencher** composes on your behalf. Together with the values that **Sequencher** controls, these values are sent to the **Velveth** program.

Figure 16-45 Add and Remove parameters with Preview window



Add/review parameters

Preview of parameters

Current Parameters

```
-cov_cutoff auto -exp_cov auto -scaffolding no
```

If you find that you have removed a parameter by mistake or wish to restore all the settings to their original values, then click on the **Restore Defaults** button. This will restore the parameters to the values that were set before you started editing.

*USING **SEQUENCHER** TO IMPROVE A VELVET ALIGNMENT*

To do this, you will need to select the consensus sequences you want to merge. Then choose an assembly algorithm and assembly parameters by clicking on the **Assembly Parameters** button that you can find on the **Project Window** button bar. Now click the **Assemble Automatically** button. **Sequencher** will use the parameters you set and try to assemble the reads. You will generally find some improvement in your results. However, setting the parameters to be too relaxed would not result in an improvement. The overlaps would probably contain too many gaps or mismatches.

VELVET AND MULTIPLEX IDS

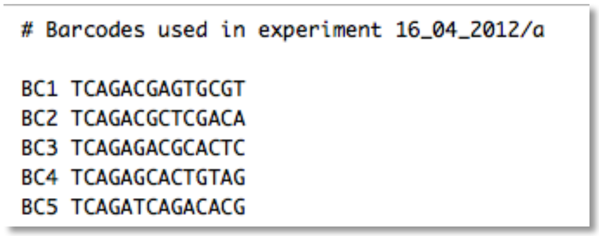
Multiplexing reads is a way of maximizing reagent and sequencer use by mixing a number of separate samples together in one flowcell, for example, and sequencing them at the same time. The DNA from each sample is tagged with a unique DNA identifier. After the sequencing run, the reads are separated using the unique DNA identifier.

Sequencher can perform this separation or demultiplexing of the reads automatically for you. This is known as binning and each reads file is called a bin.

You need to supply a FastA or FastQ file containing reads from a Multiplex ID experiment, and a text file containing the barcodes used in this Multiplex ID experiment. This file is called the barcodes file and it contains one barcode name and sequence per line. The file can accommodate a single comment line, where the comment is preceded by a # mark.

FORMAT OF BARCODE FILE

Figure 16-46 Format of barcode file



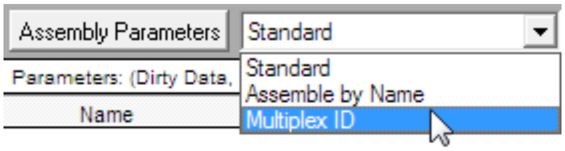
```
# Barcodes used in experiment 16_04_2012/a

BC1 TCAGACGAGTGCGT
BC2 TCAGACGCTCGACA
BC3 TCAGAGACGCACTC
BC4 TCAGAGCACTGTAG
BC5 TCAGATCAGACACG
```

***Note**:* The sequence should consist of the tag and barcode if the tag has not been removed by your sequencing pipeline.

ENABLING MULTIPLEX ID MODE

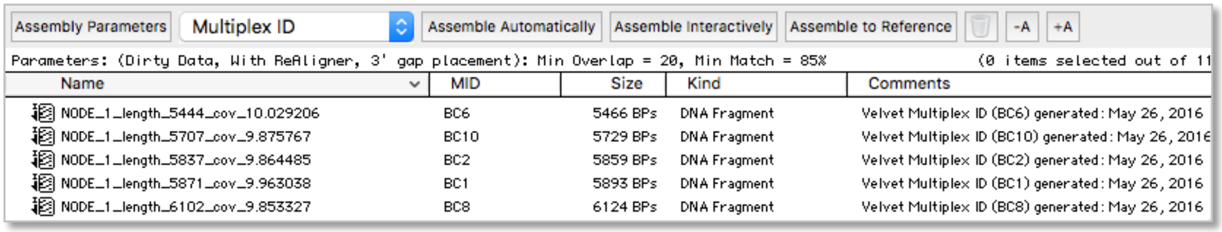
First you will need to enable the **Multiplex ID** mode by choosing **Multiplex ID** from the drop-down menu on the **Project Window** button bar.

Figure 16-47 Drop-down Mode menu



Once you have enabled this mode, a new MID column appears in the **Project Window** that will contain the name of the barcode that originally formed part of the reads aligned to create the consensus sequence. Depending on the settings you have chosen, there may be a number of consensus sequences that had the same barcode.

Figure 16-48 Project Window view showing MID column



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap = 20, Min Match = 85%   (0 items selected out of 11

Name	MID	Size	Kind	Comments
NODE_1_length_5444_cov_10.029206	BC6	5466 BPs	DNA Fragment	Velvet Multiplex ID (BC6) generated: May 26, 2016
NODE_1_length_5707_cov_9.875767	BC10	5729 BPs	DNA Fragment	Velvet Multiplex ID (BC10) generated: May 26, 2016
NODE_1_length_5837_cov_9.864485	BC2	5859 BPs	DNA Fragment	Velvet Multiplex ID (BC2) generated: May 26, 2016
NODE_1_length_5871_cov_9.963038	BC1	5893 BPs	DNA Fragment	Velvet Multiplex ID (BC1) generated: May 26, 2016
NODE_1_length_6102_cov_9.853327	BC8	6124 BPs	DNA Fragment	Velvet Multiplex ID (BC8) generated: May 26, 2016

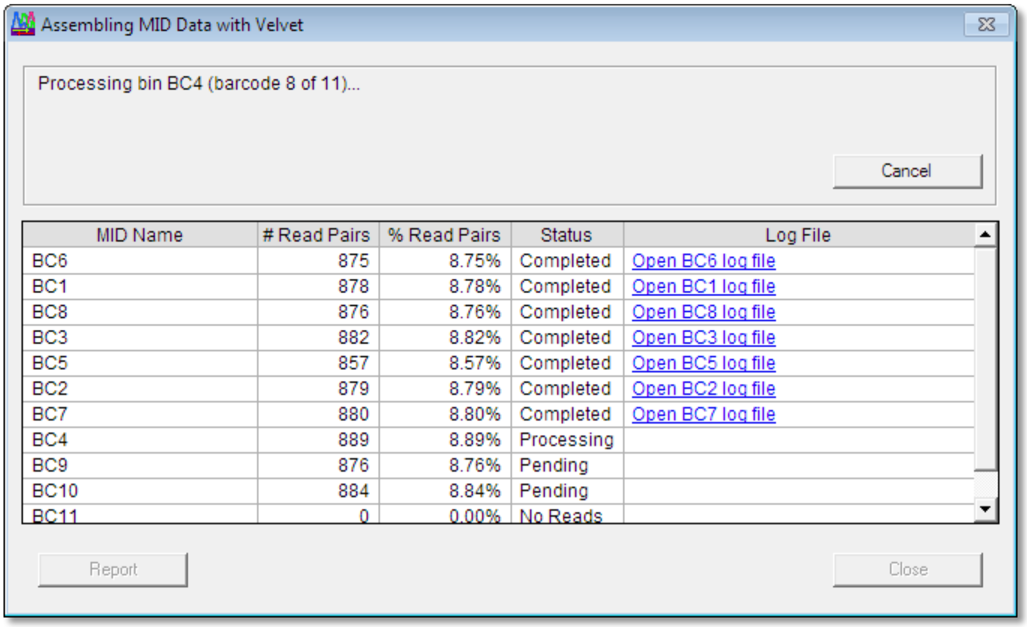
If this mode is still enabled when you save and close your project, **Sequencher** will remember this setting for you.

RUNNING VELVET WITH MULTIPLEX DATA

The first step is to click on the **Assembly Mode** drop-down menu which is found on the **Project Window** button bar. Next select the **Multiplex ID** menu item. Choose **Assemble>Align Data Files Using>Velvet…**. The **Assemble Using Velvet with MID** and **External Data Browser** dialogs will appear. Click on the **Select Reads File 1** button and browse to the reads file you want to use. Click on the **Select Barcodes File** button and browse to your prepared barcodes file. Once you are ready, click the **Assemble** button.

The **Assembling MID Data with Velvet** dialog appears. This dialog is divided into two panes. The top pane shows a constantly updated status line for the overall process. This pane also contains a **Cancel** button. If you click on this button, you will cancel the current and all subsequent operations for the current MID run.

Figure 16-49 Assembling MID data with Velvet Status Report



Processing bin BC4 (barcode 8 of 11)...

MID Name	# Read Pairs	% Read Pairs	Status	Log File
BC6	875	8.75%	Completed	Open BC6 log file
BC1	878	8.78%	Completed	Open BC1 log file
BC8	876	8.76%	Completed	Open BC8 log file
BC3	882	8.82%	Completed	Open BC3 log file
BC5	857	8.57%	Completed	Open BC5 log file
BC2	879	8.79%	Completed	Open BC2 log file
BC7	880	8.80%	Completed	Open BC7 log file
BC4	889	8.89%	Processing	
BC9	876	8.76%	Pending	
BC10	884	8.84%	Pending	
BC11	0	0.00%	No Reads	

The lower pane is a table that is divided into a number of columns (see figure above) showing the name of the Barcode, the number of reads in the bin, the % of the total reads sequenced that this represents, the status of the specific bin alignment, and finally a link to the log file. Click on this link to open the log file in the default viewer.

There is also a **Close** button that becomes enabled when the entire process has completed. Click on the **Report** button to obtain a formatted **Multiplex Status Report** that will open in a **Report Preview** window. You can save or print this report.

When you return to the **Project Window**, you will see that one or more consensus sequences for each bin have been brought into **Sequencher**. Check the MID column, this holds the name of the barcode for each consensus sequence. You can click on the MID column header to sort the column by barcode name so all consensus sequences belonging to a particular barcode name will be grouped together.

You can review your results in **Tablet**. To do this, you need to select the consensus in the **Project Window** whose contig you want to explore, then choose menu command **Contig>Show NGS Data Using>Tablet**. The contig will open in the **Tablet** viewer.

RNA-SEQ AND DIFFERENTIAL EXPRESSION

RNA-Seq is an alternative to microarray analysis for analyzing gene expression. RNA is extracted from the tissue or sample under examination and then sequenced using NGS technology. This produces a snapshot of the RNA in the sample. The millions of reads produced means a greater depth of coverage is achievable using NGS which in turn may enable a user to find rare transcripts.

We have added the **Cufflinks**[1] suite of programs enabling you to perform RNA-Seq analysis on your own computer. The **Cufflinks** suite is a series of command line programs. We have added a graphical user interface to these programs and their options, as well as adding plots and charts so that you can avoid using the command line altogether. You will need to align your RNA-Seq read data using **GSNAP** or **BWA-MEM** or another aligner that can produce results in **SAM** or **BAM** format and then perform the RNA-Seq analysis on these files.

In the simplest workflow for analyzing RNA-Seq data to test for differential expression, using the **Cufflinks** suite of programs is a three-stage process with the first step being repeated. The reason for this repetition is two-fold. First, assembly is more computationally expensive as we get increases in read numbers. So the first step, which requires a SAM/BAM file and a GTF annotation file, is performed on one set of data at a time. The second reason is that, with a pooled mixture of samples, **Cufflinks** faces the problem of trying to sort out and select from a complex mixture of spliced isoforms. It is easier to do this on one data set at a time.

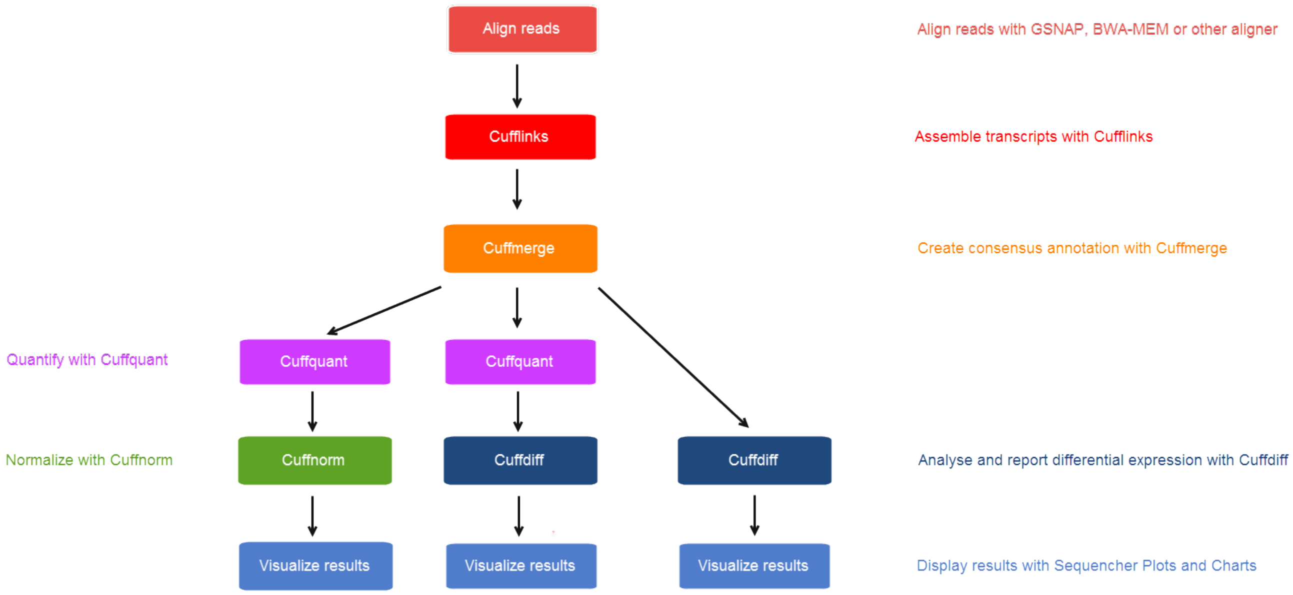
RNA-SEQ WORKFLOW

In the first step of the RNA-Seq workflow, the **Cufflinks** program quantifies the expression level of a single sample. This first stage is then repeated for each sample in your experiment. In the second stage, the transcript annotation results from analyzing each sample are merged. Finally, in the third stage, differential expression is analysed and reported. This final step consists of testing the statistical significance of each observed change in expression between the samples. The model evaluates changes assuming that the number of reads produced by each transcript is proportional to its abundance. The final output contains information including log2(fold_change), p_values, and q_values, and attributes such as gene_id and chromosomal location.

This is the simplest form of the workflow. However, it is possible to split this workflow so that the quantification of the number of reads aligned per transcript is undertaken by **Cuffquant**. This reduces the computational burden normally imposed by having the **Cuffdiff** program perform this step. This is also a necessary step if, instead of performing differential expression tests, you are simply looking at transcription levels. In this case, you would use **Cuffnorm** to normalize expression values after the quantification step.

[1] Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation** *Nature Biotechnology* doi:10.1038/nbt.1621

Figure 16-50 RNA-Seq workflow using Cufflinks suite


Align reads
Align reads with GSNAP, BWA-MEM or other aligner
Cufflinks
Assemble transcripts with Cufflinks
Cuffmerge
Create consensus annotation with Cuffmerge
Quantify with Cuffquant
Cuffquant
Cuffquant
Normalize with Cuffnorm
Cuffnorm
Cuffdiff
Cuffdiff
Analyse and report differential expression with Cuffdiff
Visualize results
Visualize results
Visualize results
Display results with Sequencher Plots and Charts

WHAT IS A GTF FILE

GTF stands for Gene Transfer Format. It is a type of annotation file with information in a tabbed text format. Unlike the more familiar **GenBank** format, each line contains one feature and its ancillary information.

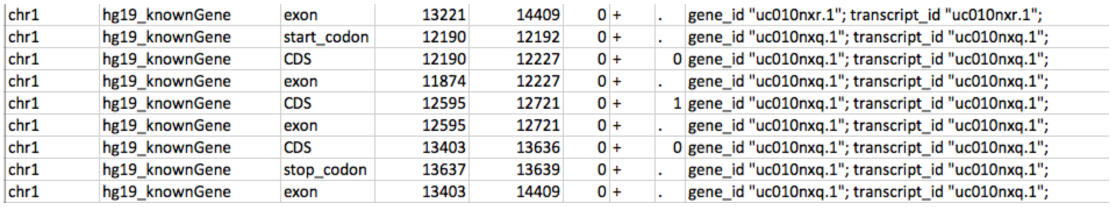
Table 16-2 GTF file fields and description

Column	Description
Seqname	Name of the chromosome or scaffold chromosome names can be given with or without the 'chr' prefix.
Source	Name of the program or data source that generated this feature.
Feature	Feature type name, e.g. Gene
Start	Start position of the feature
End	End position of the feature

Score	A floating point value
Strand	Defined as + (forward) or - (reverse)
Frame	One of '0', '1', or '2'. '0' indicates that the first base of the feature is the first base of a codon and so on.
Attribute	A semicolon-separated list of tag-value pairs of additional information

The image below shows a few lines from a GTF file. This was obtained from UCSC and is part of the annotation for human chromosome 1. It shows four feature types, exon, start codon, stop codon, and CDS. Columns four and five contain the start and end positions of each feature. Column seven indicates that these features are all on the forward (+) strand. Column eight indicates that, for only two of the three CDS features, the first base of the position is also the first base of the codon.

Figure 16-51 Example of GTF file for human genome



chr1	hg19_knownGene	exon	13221	14409	0	+	.	gene_id "uc010nxr.1"; transcript_id "uc010nxr.1";
chr1	hg19_knownGene	start_codon	12190	12192	0	+	.	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	CDS	12190	12227	0	+	0	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	exon	11874	12227	0	+	.	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	CDS	12595	12721	0	+	1	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	exon	12595	12721	0	+	.	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	CDS	13403	13636	0	+	0	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	stop_codon	13637	13639	0	+	.	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";
chr1	hg19_knownGene	exon	13403	14409	0	+	.	gene_id "uc010nxq.1"; transcript_id "uc010nxq.1";

OBTAINING GTF FILES FROM UCSC

We will only talk about the procedure for obtaining a GTF file from the UCSC browser but there are other resources from which you can obtain this type of annotation file, such as ENSEMBL.

Go to the UCSC browser website at http://genome.ucsc.edu/. You will see a menu bar just below the words USCS Genome Bioinformatics. Click on the item called Tables. This will take you to a new web page with the options you need to set for obtaining your GTF file.

Choose the correct Clade, Genome, and Assembly you want to use from the drop-down menus provided. Set the Group and Track types if required. Now set the Region. If you are choosing a position on a chromosome, you will need to specify it in the following manner: chrX:1-150000000, for example. Ensure the Output format is GTF - gene transfer format. Do not click in any of the checkboxes to the right of this menu as you will be downloading the file, not sending it to another website.

If you are not setting any other values on the Table Browser, then give your file a name by typing in the **Output file** input field. If the file is likely to be large, then also click on the **gzip**

compressed radio button. Most computers are able to open this type of file. Finally, click on the **Get Output** button. The file will be downloaded to your default downloads location.

Figure 16-52 UCSC Table Browser

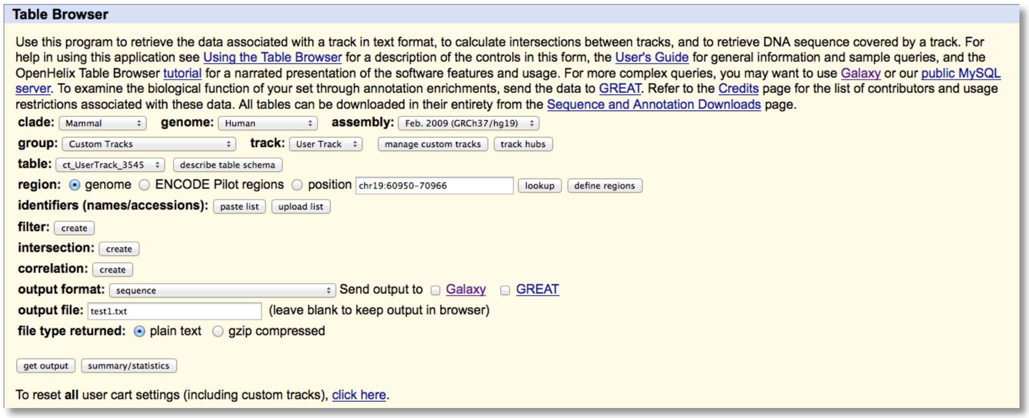


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Custom Tracks **track:** User Track [manage custom tracks] [track hubs]

table: ct_UserTrack_3545 [describe table schema]

region: ● genome ○ ENCODE Pilot regions ○ position chr19:60950-70966 [lookup] [define regions]

identifiers (names/accessions): [paste list] [upload list]

filter: [create]

intersection: [create]

correlation: [create]

output format: sequence Send output to □ Galaxy □ GREAT

output file: test1.txt (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

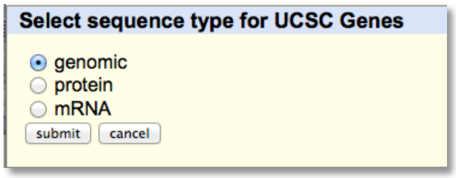
[get output] [summary/statistics]

To reset **all** user cart settings (including custom tracks), click here.

OBTAINING A FASTA FILE OF TRANSCRIPTS

The process is similar to that for obtaining the GTF file. Follow all the steps until you get to the Output format, this time choose sequence as your format. When you click on the **Get Output** button, you will be directed to a new webpage. Here you will be asked to choose between genomic, protein, or mRNA. Finally, click on the **Submit** button. The file will be downloaded to your default downloads location.

Figure 16-53 Choosing an export format for the UCSC browser



Select sequence type for UCSC Genes

● genomic
○ protein
○ mRNA

[submit] [cancel]

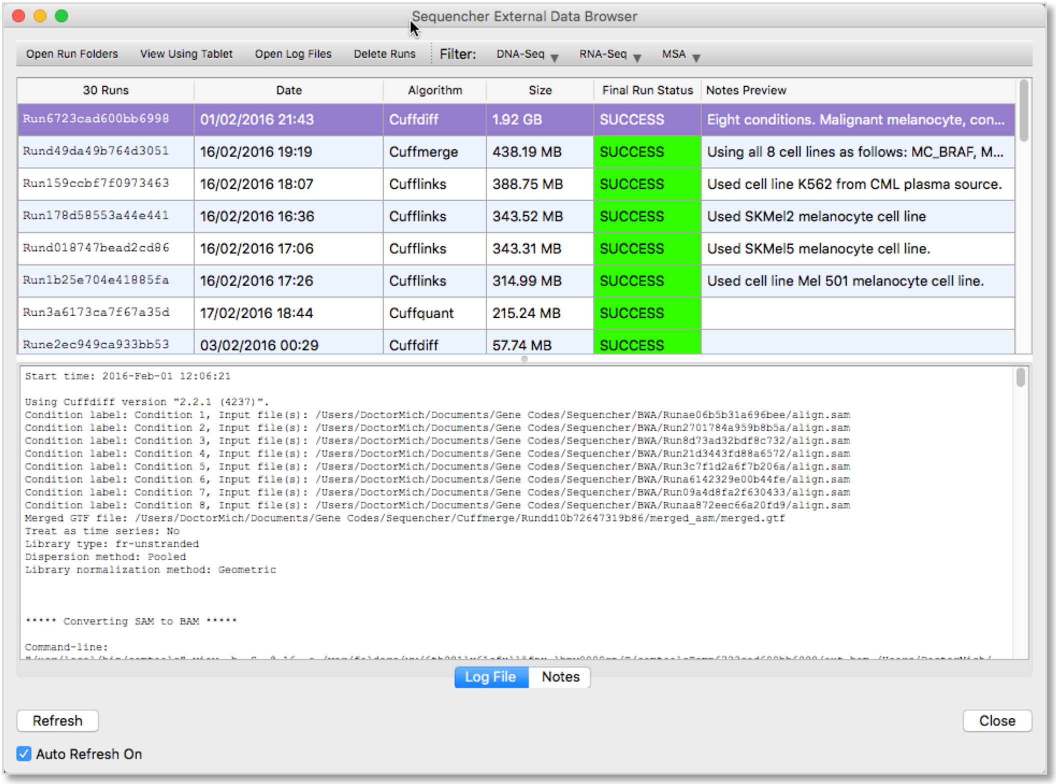
MANAGING YOUR RNA-SEQ ANALYSES WITH THE EXTERNAL DATA BROWSER

The results for each step of the RNA-Seq workflow can be found in individual run folders in your **External Data Home** folder. The default location is in your **Documents** folder. Contained within the **Documents** folder is another folder called **Gene Codes**. Inside that folder you will find the **Sequencher** folder which holds the **Cufflinks**, **Cuffmerge**, **Cuffdiff**, **Cuffquant**, and **Cuffnorm** folders.

The **External Data Browser** is an important tool when managing any of your NGS or RNA-Seq analyses and results. You can track the progress of a current run by viewing its log file in the **External Data Browser** log file pane. You can add notes to remind you of important details relating to the data files you used or even experimental details such as the source of the data. All Run folders are assigned names containing random alphanumeric characters to prevent you from accidentally over-writing results which may have taken hours to gather. The notes you add will enable you to distinguish which results should be merged together during the RNA-Seq workflow as well as allowing you to review your results at a later date.

The **Extenal Data Browser** is launched automatically whenever you use any of the RNA-Seq tools and also by going to the **Window** menu and choosing **Open External Data Browser**. The dialog that opens contains a button bar, a table of your current NGS and RNA-Seq runs, and a lower pane which can be used to display the log file for the chosen run or any notes you have made. The drop-down menus labeled **DNA-Seq**, **RNA-Seq**, and **MSA** are used to filter out Run folders associated with the different analysis types. If you click on the DNA-Seq drop-down menu, you can click on any item to remove or include a specific set of runs in the browser. They are not removed from your computer. If you want to remove a Run from your computer, click on the Run name and then click **Delete Run.**

Figure 16-54 The External Data Browser



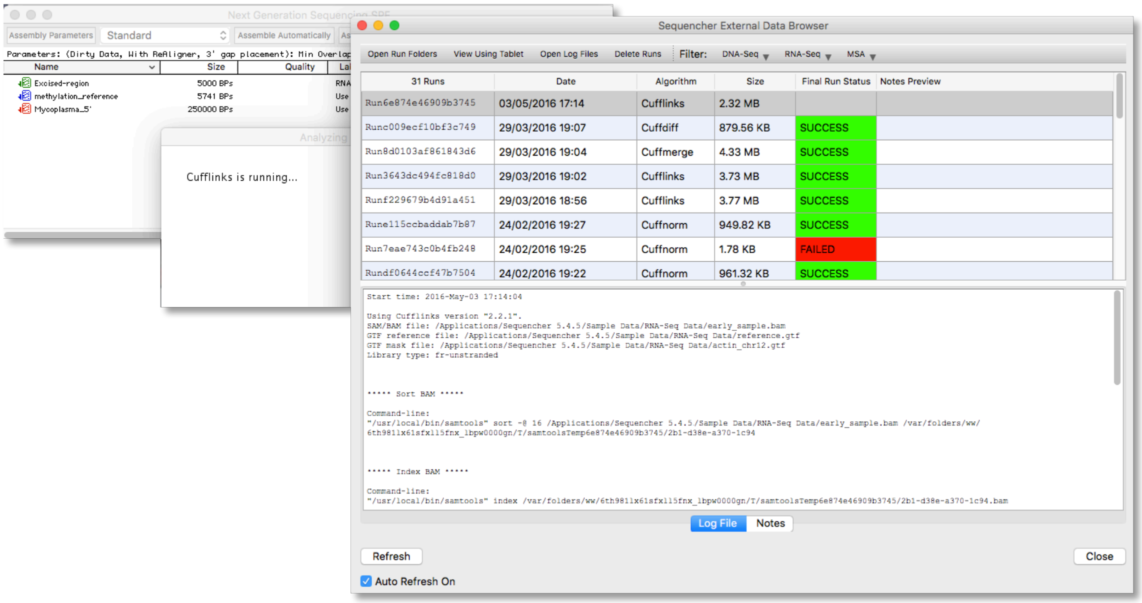
30 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Run6723cad600bb6998	01/02/2016 21:43	Cuffdiff	1.92 GB	SUCCESS	Eight conditions. Malignant melanocyte, con...
Rund49da49b764d3051	16/02/2016 19:19	Cuffmerge	438.19 MB	SUCCESS	Using all 8 cell lines as follows: MC_BRAF, M...
Run159ccbf7f0973463	16/02/2016 18:07	Cufflinks	388.75 MB	SUCCESS	Used cell line K562 from CML plasma source.
Run178d58553a44e441	16/02/2016 16:36	Cufflinks	343.52 MB	SUCCESS	Used SKMel2 melanocyte cell line
Rund018747bead2cd86	16/02/2016 17:06	Cufflinks	343.31 MB	SUCCESS	Used SKMel5 melanocyte cell line.
Run1b25e704e41885fa	16/02/2016 17:26	Cufflinks	314.99 MB	SUCCESS	Used cell line Mel 501 melanocyte cell line.
Run3a6173ca7f67a35d	17/02/2016 18:44	Cuffquant	215.24 MB	SUCCESS	
Rune2ec949ca933bb53	03/02/2016 00:29	Cuffdiff	57.74 MB	SUCCESS	

```
Start time: 2016-Feb-01 12:06:21

Using Cuffdiff version "2.2.1 (4237)".
Condition label: Condition 1, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Runae06b5b31a696bee/align.sam
Condition label: Condition 2, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Run2701784a959b8b5a/align.sam
Condition label: Condition 3, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Run8d73ad32bdf8c732/align.sam
Condition label: Condition 4, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Run21d3443fd88a6572/align.sam
Condition label: Condition 5, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Run3c7f1d2a6f7b206a/align.sam
Condition label: Condition 6, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Runa6142329e00b44fe/align.sam
Condition label: Condition 7, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Run09a4d8fa2f630433/align.sam
Condition label: Condition 8, Input file(s): /Users/DoctorMich/Documents/Gene Codes/Sequencher/BWA/Runaa872eec66a20fd9/align.sam
Merged GTF file: /Users/DoctorMich/Documents/Gene Codes/Sequencher/Cuffmerge/Rundd10b72647319b86/merged_asm/merged.gtf
Treat as time series: No
Library type: fr-unstranded
Dispersion method: Pooled
Library normalization method: Geometric



***** Converting SAM to BAM *****

Command-line:
```

☑ Auto Refresh On

Once you have started a new analysis, you can see its progress in the **External Data Browser**. The log file pane updates automatically when the **Auto Refresh On** control is checked or by periodically clicking on the **Refresh** button. This will update the log file view so you can see the exact progress of your analysis.

Figure 16-55 The External Data Browser displaying log from running Cufflinks algorithm



Parameters: (Dirty Data, With ReAligner, 3' gap placement): Min Overlap

Name	Size	Quality	La
Excised-region	5000 BPs		RNA
methylation_reference	5741 BPs		Use
Mycoplasma_5'	250000 BPs		Use

Cufflinks is running...

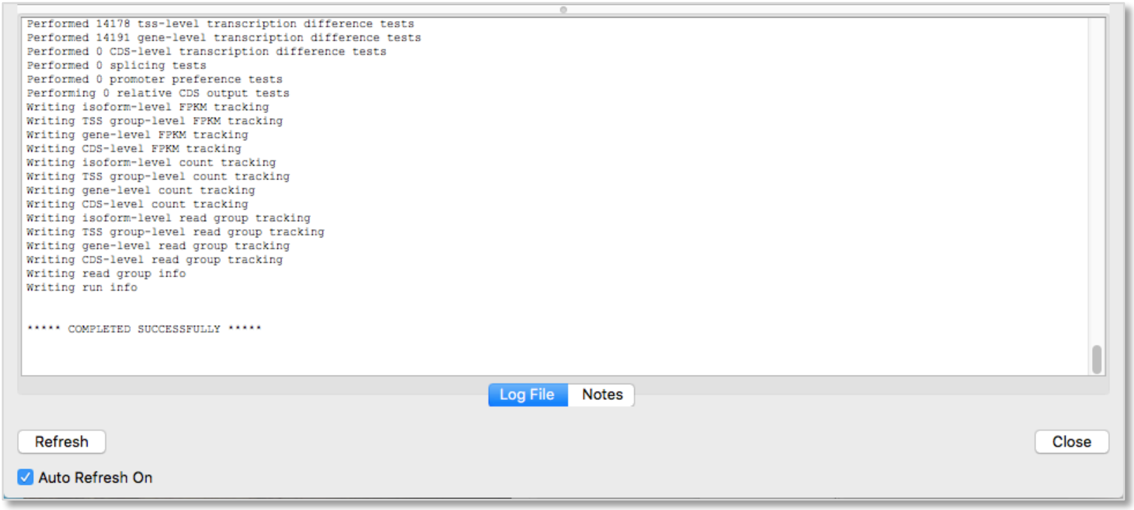
31 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Run6e874e46909b3745	03/05/2016 17:14	Cufflinks	2.32 MB		
Runc009ecf10bf3c749	29/03/2016 19:07	Cuffdiff	879.56 KB	SUCCESS	
Run8d0103af861843d6	29/03/2016 19:04	Cuffmerge	4.33 MB	SUCCESS	
Run3643dc494fc818d0	29/03/2016 19:02	Cufflinks	3.73 MB	SUCCESS	
Runf229679b4d91a451	29/03/2016 18:56	Cufflinks	3.77 MB	SUCCESS	
Rune115ccbaddab7b87	24/02/2016 19:27	Cuffnorm	949.82 KB	SUCCESS	
Run7eae743c0b4fb248	24/02/2016 19:25	Cuffnorm	1.78 KB	FAILED	
Rundf0644ccf47b7504	24/02/2016 19:22	Cuffnorm	961.32 KB	SUCCESS	

```
Start time: 2016-May-03 17:14:04

Using Cufflinks version "2.2.1".
SAM/BAM file: /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/early_sample.bam
GTF reference file: /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/reference.gtf
GTF mask file: /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/actin_chr12.gtf
Library type: fr-unstranded



***** Sort BAM *****

Command-line:
"/usr/local/bin/samtools" sort -@ 16 /Applications/Sequencher 5.4.5/Sample Data/RNA-Seq Data/early_sample.bam /var/folders/ww/
6th981lx61sfx115fnx_lbpw0000gn/T/samtoolsTemp6e874e46909b3745/2b1-d38e-a370-1c94



***** Index BAM *****

Command-line:
"/usr/local/bin/samtools" index /var/folders/ww/6th981lx61sfx115fnx_lbpw0000gn/T/samtoolsTemp6e874e46909b3745/2b1-d38e-a370-1c94.bam
```

☑ Auto Refresh On

The log file may accumulate many rows of information so be aware that you might need to scroll down to the bottom of the file in order to see the most recent information.

Figure 16-56 External Data Browser viewing a log file

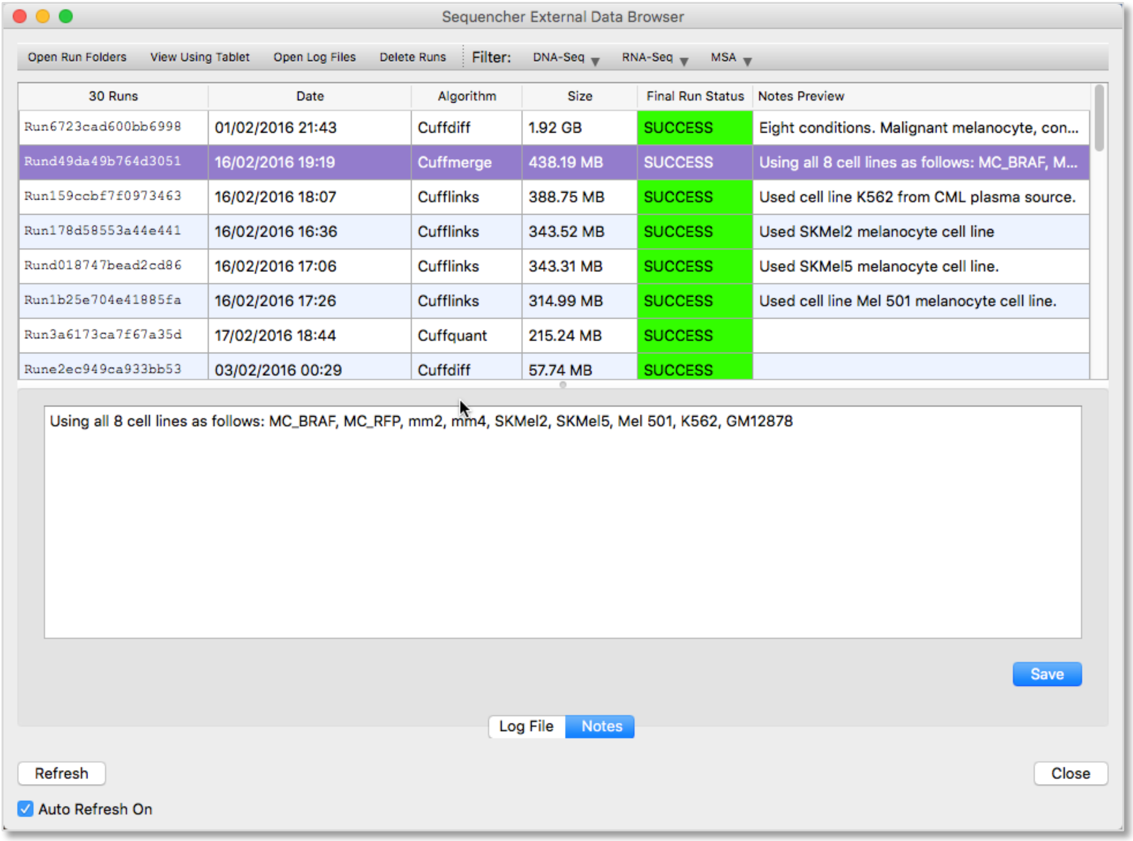


```
Performed 14178 tss-level transcription difference tests
Performed 14191 gene-level transcription difference tests
Performed 0 CDS-level transcription difference tests
Performed 0 splicing tests
Performed 0 promoter preference tests
Performing 0 relative CDS output tests
Writing isoform-level FPKM tracking
Writing TSS group-level FPKM tracking
Writing gene-level FPKM tracking
Writing CDS-level FPKM tracking
Writing isoform-level count tracking
Writing TSS group-level count tracking
Writing gene-level count tracking
Writing CDS-level count tracking
Writing isoform-level read group tracking
Writing TSS group-level read group tracking
Writing gene-level read group tracking
Writing CDS-level read group tracking
Writing read group info
Writing run info


***** COMPLETED SUCCESSFULLY *****
```

☑ Auto Refresh On

To add a note to your run, click on the **Notes** tab and enter the information. Click on the **Save** button and then click on the **Refresh** button. The new **Notes** information will now appear next to the Run information in the **Notes Preview** column. If you have entered several lines of information, it may not all appear in the default table view but it will be saved in the body of the Note and will be seen in a tooltip for the note as you hover over the note field. You can also drag the window out to change its size or proportions.

Figure 16-57 The External Data Browser - adding a note



30 Runs	Date	Algorithm	Size	Final Run Status	Notes Preview
Run6723cad600bb6998	01/02/2016 21:43	Cuffdiff	1.92 GB	SUCCESS	Eight conditions. Malignant melanocyte, con...
Rund49da49b764d3051	16/02/2016 19:19	Cuffmerge	438.19 MB	SUCCESS	Using all 8 cell lines as follows: MC_BRAF, M...
Run159ccbf7f0973463	16/02/2016 18:07	Cufflinks	388.75 MB	SUCCESS	Used cell line K562 from CML plasma source.
Run178d58553a44e441	16/02/2016 16:36	Cufflinks	343.52 MB	SUCCESS	Used SKMel2 melanocyte cell line
Rund018747bead2cd86	16/02/2016 17:06	Cufflinks	343.31 MB	SUCCESS	Used SKMel5 melanocyte cell line.
Run1b25e704e41885fa	16/02/2016 17:26	Cufflinks	314.99 MB	SUCCESS	Used cell line Mel 501 melanocyte cell line.
Run3a6173ca7f67a35d	17/02/2016 18:44	Cuffquant	215.24 MB	SUCCESS	
Rune2ec949ca933bb53	03/02/2016 00:29	Cuffdiff	57.74 MB	SUCCESS	

Using all 8 cell lines as follows: MC_BRAF, MC_RFP, mm2, mm4, SKMel2, SKMel5, Mel 501, K562, GM12878

☑ Auto Refresh On

CUFFLINKS

A requirement of quantifying expression levels is to identify accurately which reads are produced by a given isoform. **Cufflinks** uses RNA-Seq reads previously aligned to a genome and assembles individual transcripts from that data. The program attempts to infer the splicing structure for each gene. However, many alternate splicing events are possible, so **Cufflinks** takes a frugal approach when assembling the transcriptome assembly. It reports as few transcript fragments as possible when trying to satisfy the explanation of the possible splicing events in the original data.

In order to run **Cufflinks**, a reference file is required. This reference will be a **GTF** format file, not the more familiar GenBank file. This is used to guide the initial re-alignment of your data.

This type of alignment is called Reference-Annotation Based Transcript Assembly[2]. The GTF file provides information on all reference transcripts as well as any novel genes and isoforms that are assembled.

Alternatively, you can use a GTF file simply to estimate isoform expression only. Transcripts that cannot be explained by the reference transcripts file are ignored.

It is also possible to tell the program to ignore certain transcripts, such as mitochondrial transcripts, or other transcripts you wish to ignore in your analysis. This is called the mask file and is also in **GTF** format.

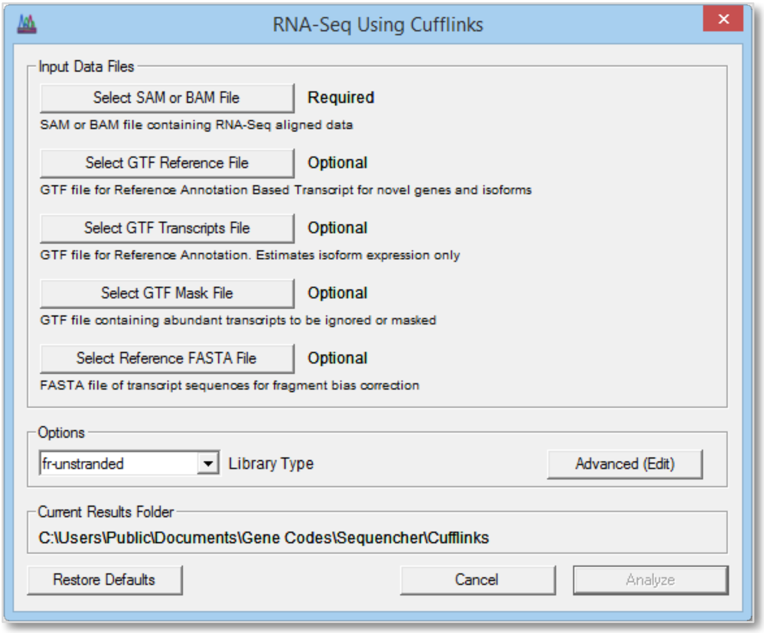
Cufflinks includes an option to detect and correct any bias and this can improve overall estimates of transcript abundance. This file is a **FastA** format file and generally contains multiple sequences.

This first step, aligning a SAM and GTF file, needs to be repeated one or more times for each of your sample data sets. The output from this step is taken forward to the next stage. This strategy is adopted over a pooled read method because it reduces the computational cost, and additionally it lowers the complexity of splice events that need to be accounted for, reducing the probability of incorrect transcript assembly.

To run **Cufflinks**, go to the **Assemble** menu and choose **RNA-Seq Using Cufflinks….** The dialog in the following figure appears along with the **External Data Browser**. Click on the **Select SAM or BAM File** button and navigate to the first of the files that you are going to analyse. You will also need to provide a GTF file and optionally a FastA file if you are performing fragment bias correction. You will need to decide whether you are performing a reference-guided alignment looking for all transcripts including novel ones, estimating isoform expression, or masking out certain transcripts such as those emanating from mitochondria or highly abundant genes such as actin. Then choose the appropriate button. You will also need to choose a library type from that drop-down menu.

[2] Roberts A, Pimentel H, Trapnell C, Pachter L. **Identification of novel transcripts in annotated genomes using RNA-Seq** *Bioinformatics* doi:10.1093/bioinformatics/btr355

Figure 16-58 Cufflinks dialog



Input Data Files

Required
SAM or BAM file containing RNA-Seq aligned data

Optional
GTF file for Reference Annotation Based Transcript for novel genes and isoforms

Optional
GTF file for Reference Annotation. Estimates isoform expression only

Optional
GTF file containing abundant transcripts to be ignored or masked

Optional
FASTA file of transcript sequences for fragment bias correction

Options

Library Type: fr-unstranded

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\Cufflinks

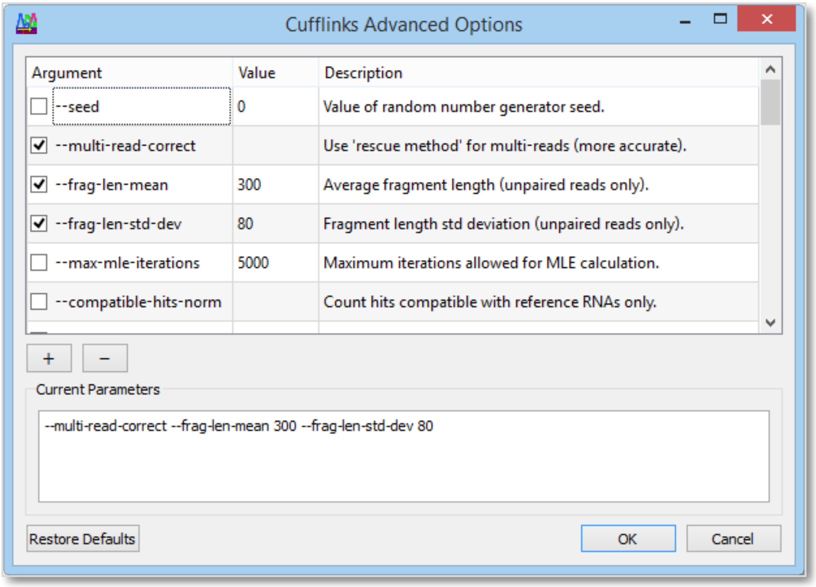
Each run will produce a transcripts.gtf file amongst its output files, it is these files that are required for the next step with **Cuffmerge**. Each **Cufflinks** run automatically creates a separate run folder with a unique name. You will find the transcripts.gtf files in the Run folders situated within the **Cufflinks** folder of your **External Data Home** folder.

ADVANCED OPTIONS FOR CUFFLINKS

Cufflinks is provided with a rich set of options. These can be found by clicking on the **Advanced (Edit)** button. You will see a new dialog (see below). To select an option, click on the checkbox adjacent to the option you wish to use. The parameter will be added to the **Current Parameters** window at the bottom of the dialog.

Some parameters require you to change a value. For example, the default value for –frag-len-mean is 200 but, in this example, the user has changed it to 300. Others are on/off switches. In this example, the user has specified the use of multi-read-correct which is used when reads map to more than one location in the genome and a more accurate method of assigning reads to locations is required. Once you have chosen all the Advanced Options you wish to apply, click on the **OK** button. The **Advanced Options** dialog closes. Click on the **Analyze** button on the **RNA-Seq Using Cufflinks** dialog to start the analysis of your data

Figure 16-59 Cufflinks Advanced Options



Argument	Value	Description
□ --seed	0	Value of random number generator seed.
☑ --multi-read-correct		Use 'rescue method' for multi-reads (more accurate).
☑ --frag-len-mean	300	Average fragment length (unpaired reads only).
☑ --frag-len-std-dev	80	Fragment length std deviation (unpaired reads only).
□ --max-mle-iterations	5000	Maximum iterations allowed for MLE calculation.
□ --compatible-hits-norm		Count hits compatible with reference RNAs only.

--multi-read-correct --frag-len-mean 300 --frag-len-std-dev 80

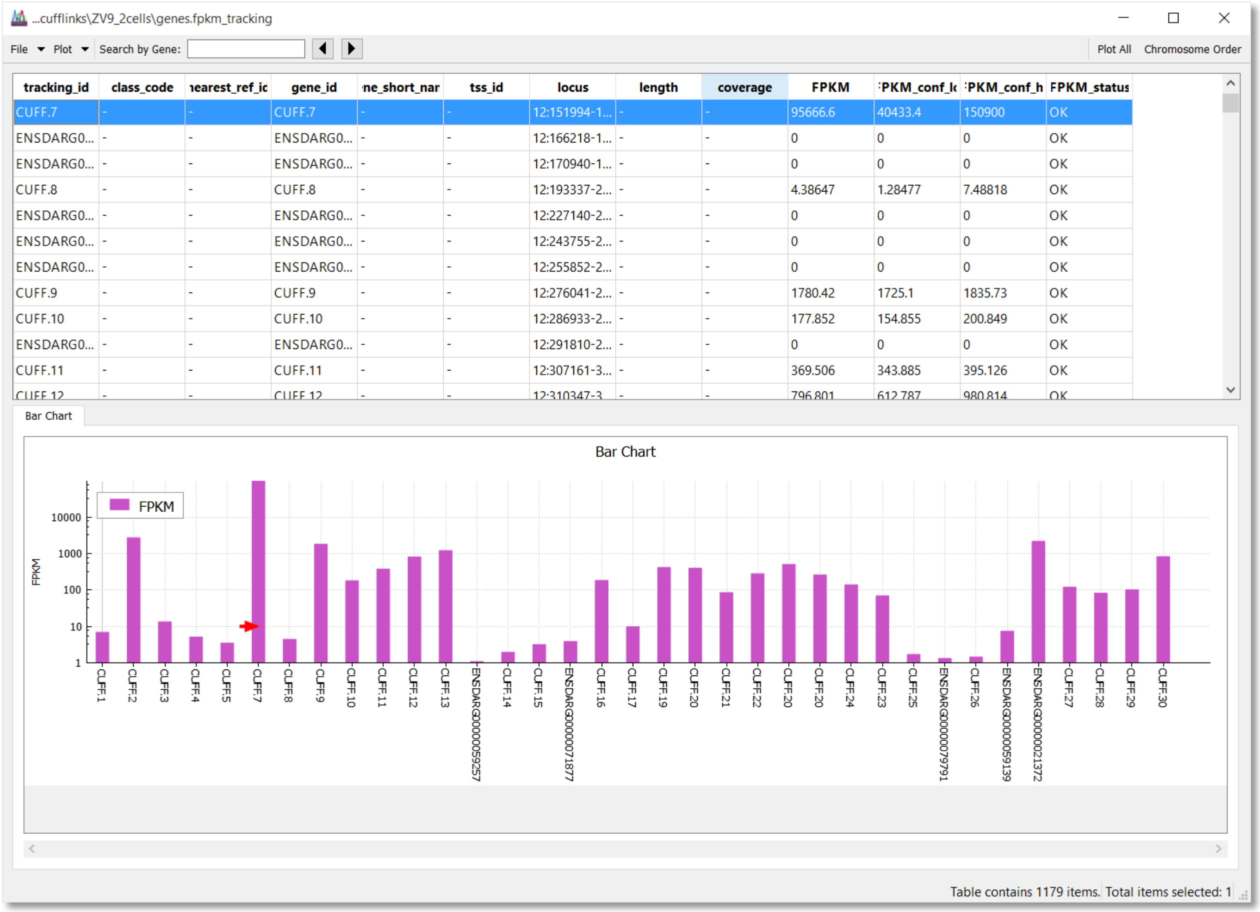
VIEWING CUFFLINKS RESULTS FILES

Cufflinks produces four main results files. These are called *skipped.gtf, transcripts.gtf, isoforms.fpkm_tracking,* and *genes.fpkm_tracking.* The *skipped.gtf* file contains any skipped loci, this depends on the maximum number of fragments a locus may have, which is set in the Advanced (Edit) options. The *transcripts.gtf* file contains isoforms detected by **Cufflinks** and is used in the **Cuffmerge** step. The *isoforms.fpkm_tracking file* contains the estimated isoform expression values. The *genes.fpkm_tracking file* contains the estimated gene expression values.

If you are not studying differential expression or you want to view the intermediate steps in the differential expression pipeline, you can view the contents of the *.fpkm_tracking* files and explore the results as a table and bar chart. To view the contents of the two fpkm_tracking files, use **Sequencher's View>Display RNA-Seq Data and Plots…** menu command.

When you hold the cursor over a bar in the bar chart, the FPKM value will be displayed. Clicking on the bar will display and highlight the data row linked to that bar. We have also added a colored background to the column from which the data is taken. As well as a gene ID, the row also contains the gene short name, its locus, the length of the transcript, and FPKM value.

Figure 16-60 Cufflinks Bar chart



...cufflinks\ZV9_2cells\genes.fpkm_tracking

tracking_id	class_code	ıearest_ref_ic	gene_id	ıe_short_nar	tss_id	locus	length	coverage	FPKM	ᵎPKM_conf_lc	ᵎPKM_conf_h	FPKM_status
CUFF.7	-	-	CUFF.7	-	-	12:151994-1...	-	-	95666.6	40433.4	150900	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:166218-1...	-	-	0	0	0	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:170940-1...	-	-	0	0	0	OK
CUFF.8	-	-	CUFF.8	-	-	12:193337-2...	-	-	4.38647	1.28477	7.48818	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:227140-2...	-	-	0	0	0	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:243755-2...	-	-	0	0	0	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:255852-2...	-	-	0	0	0	OK
CUFF.9	-	-	CUFF.9	-	-	12:276041-2...	-	-	1780.42	1725.1	1835.73	OK
CUFF.10	-	-	CUFF.10	-	-	12:286933-2...	-	-	177.852	154.855	200.849	OK
ENSDARG0...	-	-	ENSDARG0...	-	-	12:291810-2...	-	-	0	0	0	OK
CUFF.11	-	-	CUFF.11	-	-	12:307161-3...	-	-	369.506	343.885	395.126	OK
CUFF.12	-	-	CUFF.12	-	-	12:310347-3	-	-	796.801	612.787	980.814	OK

Table contains 1179 items. Total items selected: 1

CUFFMERGE

In this stage, the program takes the output files from the repeated **Cufflinks** runs and merges them. You provide **Cuffmerge** with the transcripts.gtf files from each of your **Cufflinks** runs and a GTF reference file. You can optionally provide **Cuffmerge** with a FASTA file for excluding artefacts and classifying transcript fragments. **Cuffmerge** then produces a 'consensus' GTF file that can be used by **Cuffdiff** to estimate differential expression.

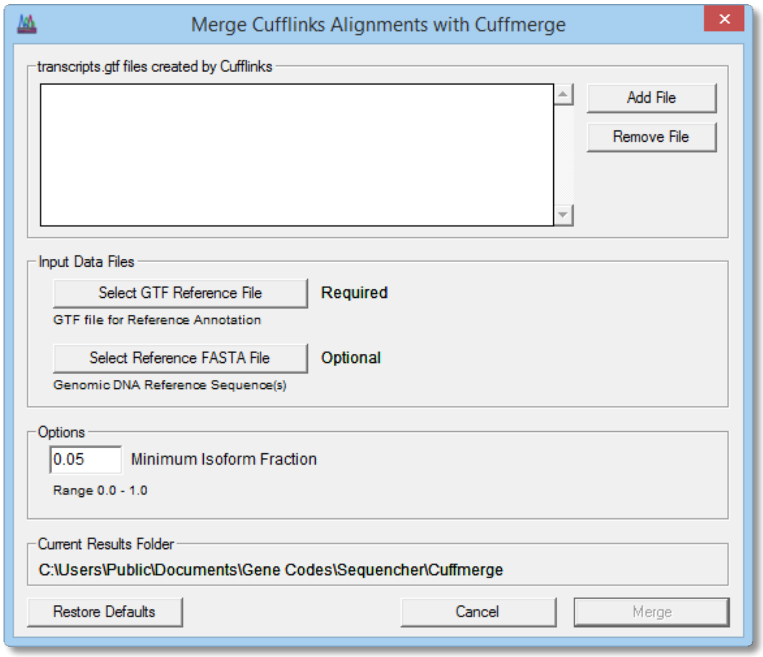
Cuffmerge performs a Reference-Annotation Based Transcript assembly, producing a single merged.gtf file from the transcripts.gtf files and the reference GTF file you provided.

Choose the **Merge Cufflinks Alignments with Cuffmerge...** menu item from the **Assemble** menu. The following dialog appears along with the **External Data Browser**. Click on the **Add File** button and navigate to the **Cufflinks** folder and choose a Run folder, double-click on the Run folder to open it and select the transcripts.gtf file. Click on the **Open** button. Repeat these steps for each transcripts.gtf file you want to merge.

Choose the reference GTF file you intend to use by clicking on the **Select GTF Reference File** button. Browse to your reference GTF file and, once you have selected it, click on the **Open** button. Set any other options and then click on the **Merge** button.

This step produces the merged.gtf file that you will use in the final step with **Cuffdiff**. You will find these files in the **Cuffmerge** folder in your **External Data Home** folder. Each **Cuffmerge** run automatically creates a separate Run folder with a unique name.

Figure 16-61 Cuffmerge dialog


Merge Cufflinks Alignments with Cuffmerge

transcripts.gtf files created by Cufflinks

Input Data Files

Select GTF Reference File — Required

GTF file for Reference Annotation

Select Reference FASTA File — Optional

Genomic DNA Reference Sequence(s)

Options

Minimum Isoform Fraction: 0.05

Range 0.0 - 1.0

Current Results Folder

C:\Users\Public\Documents\Gene Codes\Sequencher\Cuffmerge

CUFFDIFF

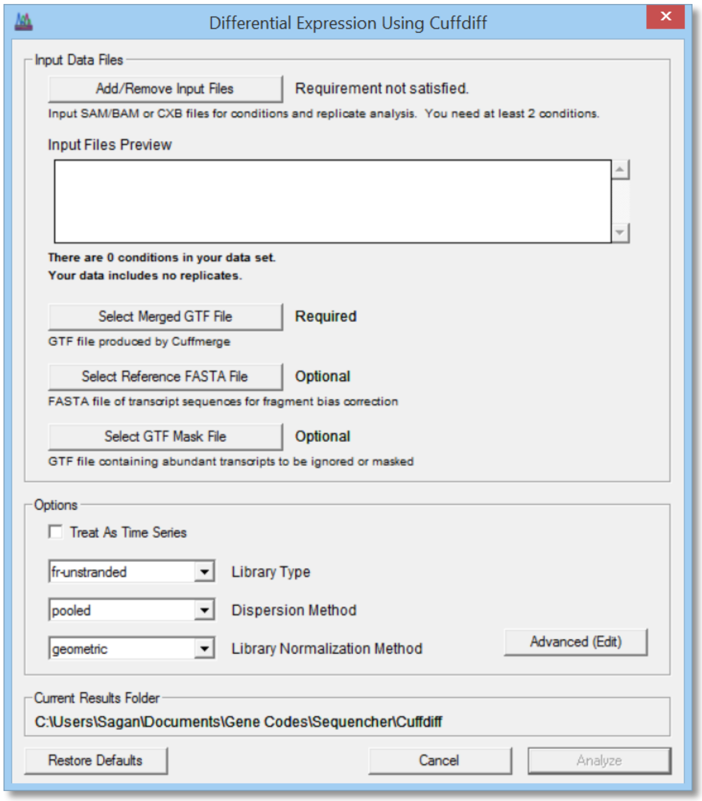
In the final stage, the merged.gtf file and the original SAM/BAM or CXB files are analysed to find genes that are differentially expressed. The results files contain statistical information such as p-values and log2(fold change) in expression between the samples. The outputs (.diff files and .fpkm_tracking files) from this stage may be viewed as tables or charts.

Cufflinks has the ability to filter out transcripts that are very low in abundance. The authors explain that this is due to the fact that analyzing expression, at such low levels, is not reliable. You can change the threshold for filtering out these low-abundance transcripts by altering the value of the minimum isoform fraction.

Cuffdiff supports different library types and you need to know which protocol was used to generate the data you are analyzing. The standard Illumina protocol is specified by fr-unstranded and the standard Solid protocol is specified by fr-secondstrand. You can find more details in the **Cufflinks** online manual[3].

Go to the **Assemble** menu and choose the **RNA-Seq Differential Expression Using Cuffdiff…** menu item.

Figure 16-62 Cuffdiff dialog



Requirement not satisfied.

Input SAM/BAM or CXB files for conditions and replicate analysis. You need at least 2 conditions.

There are 0 conditions in your data set.
Your data includes no replicates.

Select Merged GTF File: Required
GTF file produced by Cuffmerge

Select Reference FASTA File: Optional
FASTA file of transcript sequences for fragment bias correction

Select GTF Mask File: Optional
GTF file containing abundant transcripts to be ignored or masked

☐ Treat As Time Series

fr-unstranded — Library Type
pooled — Dispersion Method
geometric — Library Normalization Method

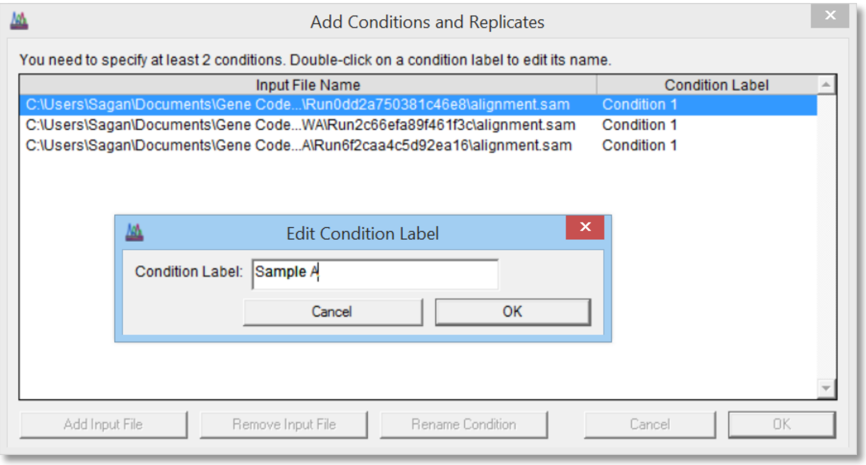
Current Results Folder: C:\Users\Sagan\Documents\Gene Codes\Sequencher\Cuffdiff

You will need to tell **Sequencher** which SAM/BAM or CXB files you used in the **Cufflinks** or **Cuffquant** steps. Click on the **Add/Remove Input Files** button. The **Add Conditions and Replicates** dialog opens. Click on the **Add Input File** button and browse to the location of the file you want to add. The files you select with the **Add Input File** button will appear in the Input File Name list. If you chose the wrong files, simply highlight them and click on the **Remove Input File** button to delete them from the list. Initially, each file will have a default condition label of Condition 1. You can edit this as you add files or change the label after you have added all your files. Change a Condition Label by double-clicking on it and entering a new name in the

[3] http://cufflinks.cbcb.umd.edu/manual.html

Edit Condition Label dialog. Click on the **OK** button to dismiss the dialog. To dismiss the **Add Conditions and Replicates** dialog, click on its **OK** button.

Figure 16-63 Add Conditions and Replicates editing a conditions label



You need to specify at least 2 conditions. Double-click on a condition label to edit its name.

Input File Name	Condition Label
C:\Users\Sagan\Documents\Gene Code...\Run0dd2a750381c46e8\alignment.sam	Condition 1
C:\Users\Sagan\Documents\Gene Code...WA\Run2c66efa89f461f3c\alignment.sam	Condition 1
C:\Users\Sagan\Documents\Gene Code...A\Run6f2caa4c5d92ea16\alignment.sam	Condition 1

Condition Label: Sample A

You will notice that the status of a number of items in **Differential Expression Using Cuffdiff** dialog have changed. These status changes display the list of files you have added for analysis, whether you have fulfilled the requirements for differential expression (at least two conditions), and whether your data includes replicates. Click on the **Select Merged GTF File** button and navigate to the location of the merged.gtf file for your data. You will find it in the **Cuffmerge** Run folder that was created from the corresponding **Cuffmerge** run. These Run folders are in your **External Data Home** folder. If you are providing a FASTA file for fragment bias correction, click on the **Select Reference FASTA File** button, browse to the location of the FASTA file, and click on the **Open** button. **Cuffdiff** tries to correct[4] for positional fragment bias (a local effect) where reads are located preferentially at the beginning and end of a transcript, as well as sequence-specific bias (a global effect).

Ensure that you have the correct values set for **Library Type**, **Dispersion Method**, and **Library Normalization Method** by making selections from the drop-down menus. If your data is from a time-series, then click on the **Treat As Time Series** button. Now click on the **Analyze** button.

[4] Adam Roberts, Cole Trapnell, Julie Donaghey, John L. Rinn and Lior Pachter, Improving RNA-Seq expression estimates by correcting for fragment bias Genome Biology, Volume 12, R22 (2011)

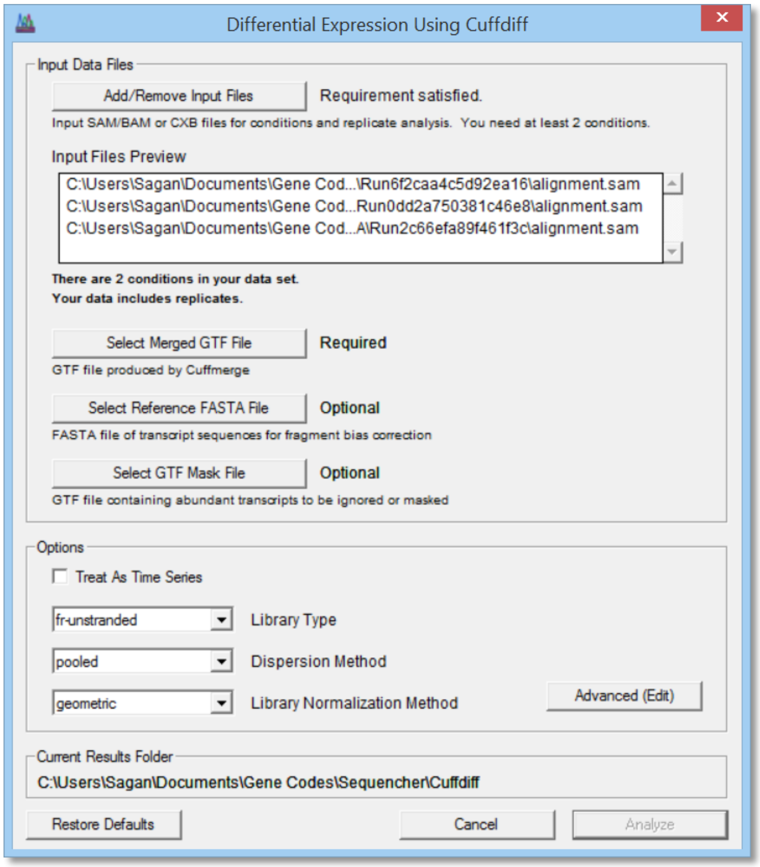
ADVANCED OPTIONS FOR CUFFDIFF

Cuffdiff is provided with a rich set of advanced options. These can be found by clicking on the **Advanced (Edit)** button on the **Differential Expression Using Cuffdiff** dialog.

You will see a new dialog (see below). The table in the dialog contains a list of the options, descriptions, and checkboxes. To select an option, click on the checkbox adjacent to the option you wish to use. The parameter will be added to the **Current Parameters** window at the bottom of the dialog.

Some parameters require you to change a value. For example, the default value for --frag-len-mean (fragment mean length for unpaired reads) is 200 but in this example the user has changed it to 300.

Figure 16-64 Status changes on Cuffdiff dialog



Requirement satisfied.

Input SAM/BAM or CXB files for conditions and replicate analysis. You need at least 2 conditions.

Input Files Preview

```
C:\Users\Sagan\Documents\Gene Cod...\Run6f2caa4c5d92ea16\alignment.sam
C:\Users\Sagan\Documents\Gene Cod...Run0dd2a750381c46e8\alignment.sam
C:\Users\Sagan\Documents\Gene Cod...A\Run2c66efa89f461f3c\alignment.sam
```

There are 2 conditions in your data set.
Your data includes replicates.

Required
GTF file produced by Cuffmerge

Optional
FASTA file of transcript sequences for fragment bias correction

Optional
GTF file containing abundant transcripts to be ignored or masked

☐ Treat As Time Series

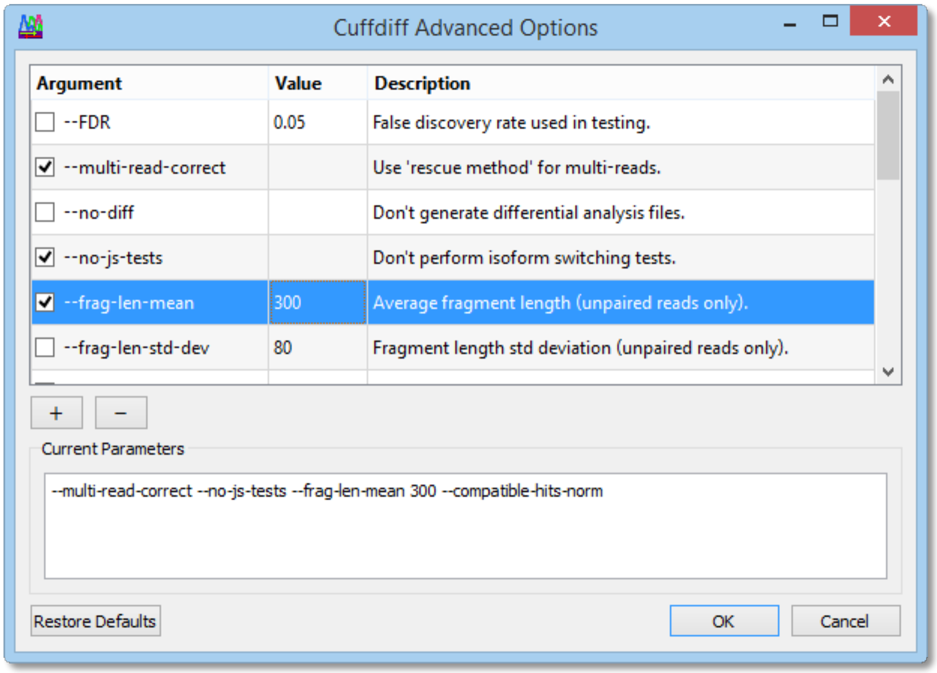
fr-unstranded — Library Type
pooled — Dispersion Method
geometric — Library Normalization Method

C:\Users\Sagan\Documents\Gene Codes\Sequencher\Cuffdiff

Other options are on/off switches. In the example below, the user has specified the use of --multi-read-correct which is used when reads map to more than one location in the genome and a more accurate method of assigning reads to locations is required. Once you have chosen all the Advanced Options you wish to apply, click on the **OK** button.

The **Cuffdiff Advanced Options** dialog closes. Click on the **Analyze** button on the **Differential Expression Using Cuffdiff** dialog to start the analysis of your data.

Figure 16-65 Cuffdiff Advanced Options



Argument	Value	Description
☐ --FDR	0.05	False discovery rate used in testing.
☑ --multi-read-correct		Use 'rescue method' for multi-reads.
☐ --no-diff		Don't generate differential analysis files.
☑ --no-js-tests		Don't perform isoform switching tests.
☑ --frag-len-mean	300	Average fragment length (unpaired reads only).
☐ --frag-len-std-dev	80	Fragment length std deviation (unpaired reads only).

Current Parameters

```
--multi-read-correct --no-js-tests --frag-len-mean 300 --compatible-hits-norm
```

VIEWING RNA-SEQ DIFFERENTIAL EXPRESSION RESULTS

Once **Cuffdiff** has completed its run, you will find a new Run folder has been created within the **Cuffdiff** folder located in your **External Data Home** folder. This run folder contains 23 files. These files contain information on FPKM tracking, Count tracking, Read group tracking, Differential expression, Differential splicing, Differential coding sequence output, Differential promoter use, Read group info, and Run info. The files that contain information on differential expression testing all have the file extension .diff.

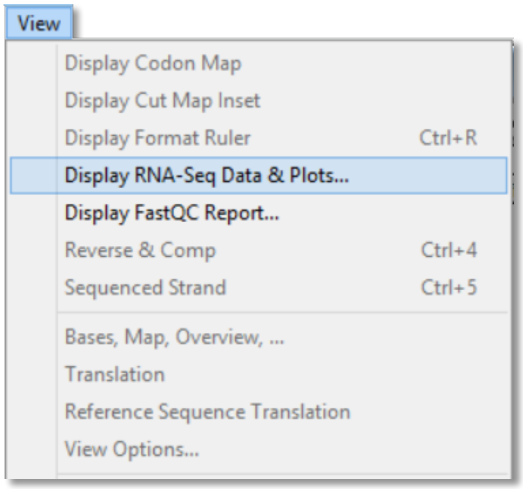
Table 16-3 Differential Expression results files

	Diff	Read group	FPKM tracking	Count tracking	Info
CDS	✓	✓	✓	✓	
Isoform	✓	✓	✓	✓	
Gene	✓	✓	✓	✓	
Tss	✓	✓	✓	✓	
Promotors	✓				
Splicing	✓				
Bias params					✓
Run					✓
Read groups					✓
Var model					✓

Note: Some files may not contain any results. This does not mean the analysis has failed and is dependant on the initial analysis choices or data.

You can view data from files with a .diff file extension by going to the **View** menu and selecting the **Display RNA-Seq Data & Plots…** menu item.

Figure 16-66 Display RNA-Seq plots & data in view menu



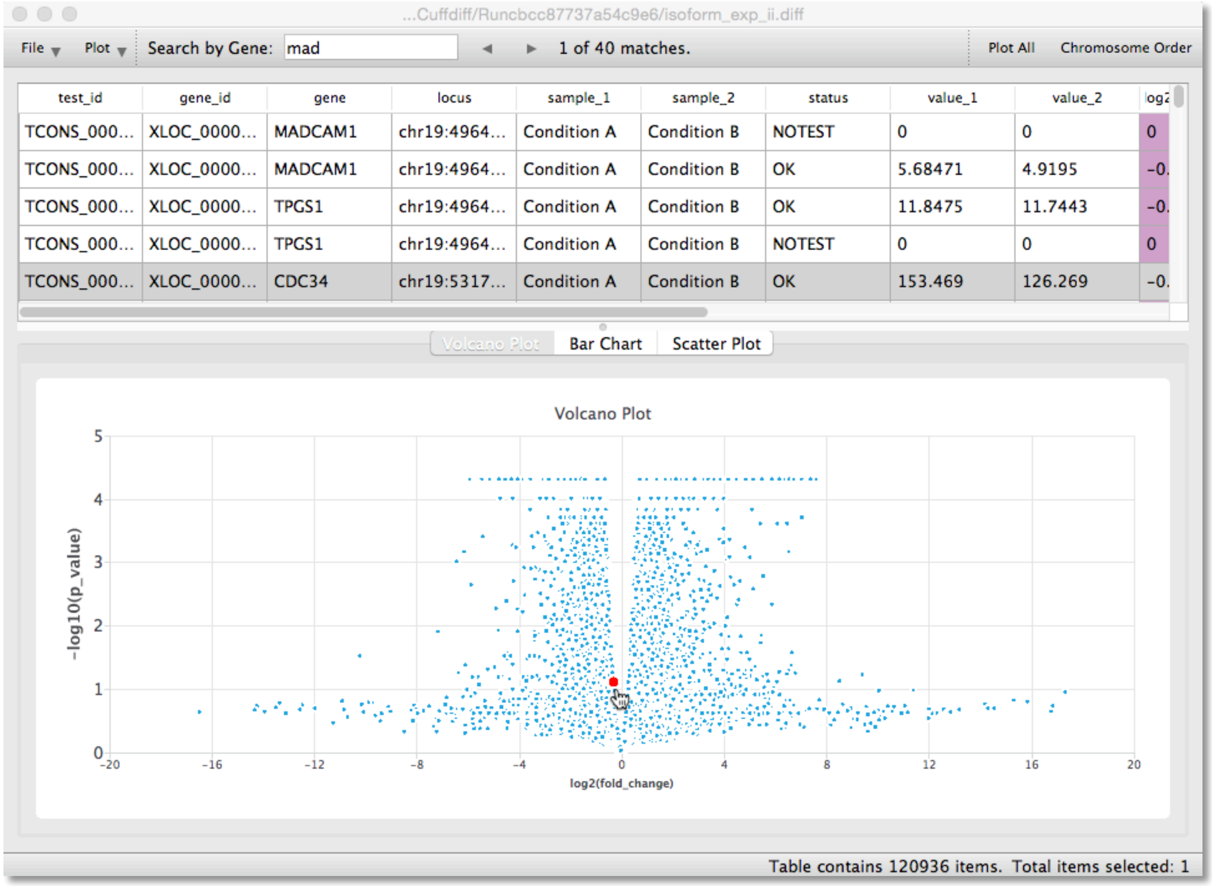
A file picker dialog appears, browse to a .diff file such as gene_exp.diff or isoform_exp.diff and click on the **Open** button. A new window opens with a table of your data in the top pane and a plot in the lower pane. Note that if there is no differential expression reported, then the plot will be empty. Check by scrolling down through the table of data.

THE VOLCANO AND SCATTER PLOT

There are several plots you can view. The **Volcano Plot** is a type of scatter plot where the log2(fold-change) is plotted on the x axis and the –log(p-value) is plotted on the y axis. Spots that are further left or right of the origin have a greater fold-change. Spots that have a higher value on the y-axis are likely to have a more significant p-value.

To assist in locating the data used to create the plot, we have added colored backgrounds to those columns. When your cursor moves over a spot, it changes from an arrow to a hand, if you click on it the spot turns red. The hand-shaped cursor also indicates that, when you click on the spot, its data row will be highlighted in the table. Notice also that when you click on a data row, if the data has been plotted it will be highlighted with a red spot.

Figure 16-67 Differential Expression Volcano Plot



...Cuffdiff/Runcbcc87737a54c9e6/isoform_exp_ii.diff

Search by Gene: mad

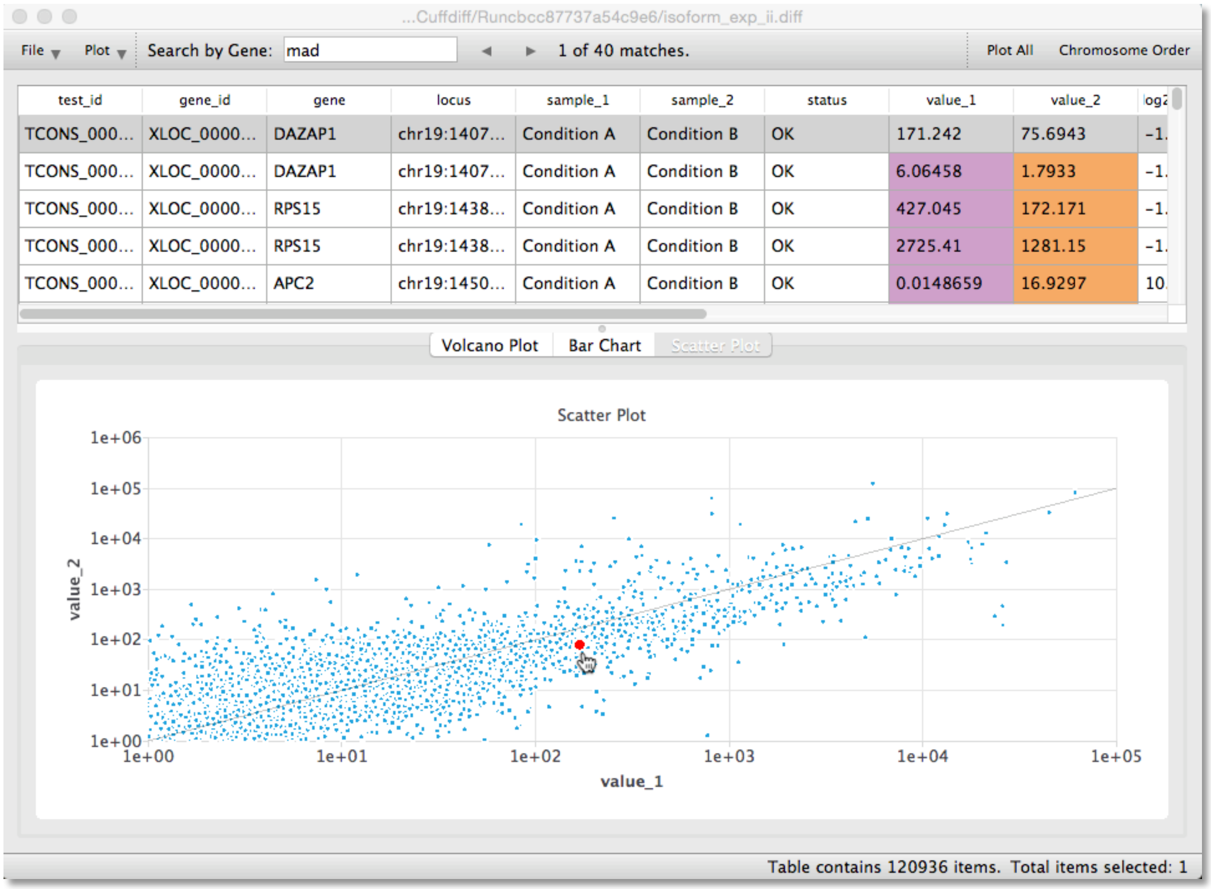
1 of 40 matches.

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2
TCONS_000...	XLOC_0000...	MADCAM1	chr19:4964...	Condition A	Condition B	NOTEST	0	0	0
TCONS_000...	XLOC_0000...	MADCAM1	chr19:4964...	Condition A	Condition B	OK	5.68471	4.9195	-0.
TCONS_000...	XLOC_0000...	TPGS1	chr19:4964...	Condition A	Condition B	OK	11.8475	11.7443	-0.
TCONS_000...	XLOC_0000...	TPGS1	chr19:4964...	Condition A	Condition B	NOTEST	0	0	0
TCONS_000...	XLOC_0000...	CDC34	chr19:5317...	Condition A	Condition B	OK	153.469	126.269	-0.

Table contains 120936 items. Total items selected: 1

Another plot you can use is the standard **Scatter Plot**, this takes the log10 FPKM values from the two samples in a .diff file and plots them against each other. When your cursor moves over a spot, it changes from an arrow to a hand, if you click on it the spot turns red. The hand-shaped cursor also indicates that, when you click on the spot, its data row will be highlighted in the table. Notice also that when you click on a data row, if the data has been plotted, it will be highlighted with a red spot.

Figure 16-68 Differential Expression Scatter Plot



test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2
TCONS_000...	XLOC_0000...	DAZAP1	chr19:1407...	Condition A	Condition B	OK	171.242	75.6943	-1.
TCONS_000...	XLOC_0000...	DAZAP1	chr19:1407...	Condition A	Condition B	OK	6.06458	1.7933	-1.
TCONS_000...	XLOC_0000...	RPS15	chr19:1438...	Condition A	Condition B	OK	427.045	172.171	-1.
TCONS_000...	XLOC_0000...	RPS15	chr19:1438...	Condition A	Condition B	OK	2725.41	1281.15	-1.
TCONS_000...	XLOC_0000...	APC2	chr19:1450...	Condition A	Condition B	OK	0.0148659	16.9297	10.

Table contains 120936 items. Total items selected: 1

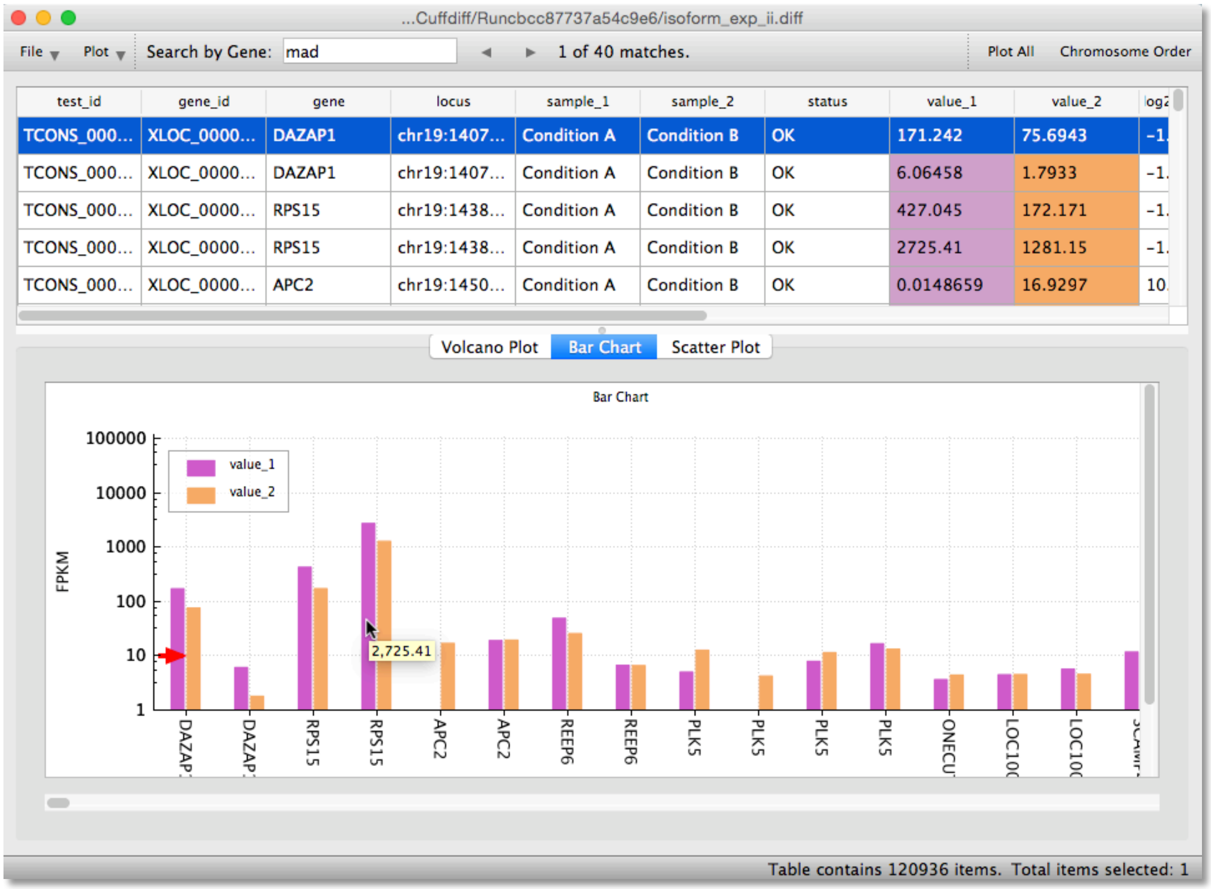
Note: The title bar of each plot is the name of the file used to create it.

THE BAR CHART

There is also a **Bar Chart**. This plot displays changes in expression based on the FPKM metric. This is defined as Fragments Per Kilobase of exon per Million fragments mapped. You can switch to this by clicking on the **Bar Chart** tab. The axes displayed in this chart are log10(FPKM) vs gene_id or gene.

As you move your cursor over a bar, its FPKM value is displayed in a tool tip. If there are two bars plotted for a particular gene or locus, by moving the cursor over each bar, you can compare the FPKM values. When you click on a bar, its data row is highlighted in the table and it is marked with a red arrow. Likewise, if you click on a data row and its data has been plotted, it will be highlighted with a red arrow. Note that we have added colored backgrounds to those columns whose data is plotted in the chart.

Figure 16-69 Differential Expression Bar Chart



...Cuffdiff/Runcbcc87737a54c9e6/isoform_exp_ii.diff

File ▾ Plot ▾ Search by Gene: mad ◂ ▸ 1 of 40 matches. Plot All Chromosome Order

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2
TCONS_000...	XLOC_0000...	DAZAP1	chr19:1407...	Condition A	Condition B	OK	171.242	75.6943	–1.
TCONS_000...	XLOC_0000...	DAZAP1	chr19:1407...	Condition A	Condition B	OK	6.06458	1.7933	–1.
TCONS_000...	XLOC_0000...	RPS15	chr19:1438...	Condition A	Condition B	OK	427.045	172.171	–1.
TCONS_000...	XLOC_0000...	RPS15	chr19:1438...	Condition A	Condition B	OK	2725.41	1281.15	–1.
TCONS_000...	XLOC_0000...	APC2	chr19:1450...	Condition A	Condition B	OK	0.0148659	16.9297	10.

Table contains 120936 items. Total items selected: 1

SEARCHING FOR SPECIFIC GENES IN A RESULTS TABLE

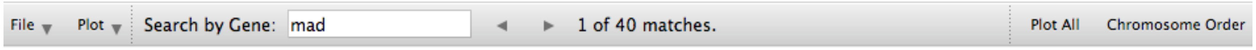
Depending on your experiment, the results table could contain tens of thousands of rows. You can search for specific genes in your results table by typing a few characters into the **Search by Gene** input field. As you type a few characters, you will see that a status message reports the number of items matching those characters. You can navigate the list of items using the two buttons to the right of the **Search by Gene** search field. The status message indicates which hit is currently selected from the hit list.

SORTING AND PLOTTING SELECTED RESULTS

While you may want to view all your results in the **Volcano** or **Scatter Plot**, you may want to plot a subset of results in the **Bar Chart**. As a prelude to plotting only some of your results, you can sort the data table simply by clicking in the header row of your chosen column. You can now select the rows that you want to include or exclude from the plot by using **Shift+click** or **Ctrl/Cmd+click**. Once you are satisfied with the selection, choose **Selected** or **Unselected** from the **Plot** menu. You can restore all the plottable items to the plots by clicking on the **Plot All** button.

When mixed letters and numbers are sorted on a computer, the ASCII sort order is used. Suppose you had three items called A1, A2, and A11. These would be sorted A1, A11, A2 by a computer although you might consider the correct order to be A1, A2, A11. Clicking on the **locus** column in the results table will cause it to be sorted in this fashion. To avoid this problem, we have created the **Chromosome Order** button. Click this button when you want to restore the original chromosomal locus sort order.

Figure 16-70 Plot All and Chromosome Order buttons



THE ALTERNATE ROUTE TO ANALYSING YOUR RNA-SEQ DATA

Performing RNA-Seq analysis, especially expression or differential expression with large data sets and replicates, can be very compute intensive. It is not unusual for runs to take tens of hours.

The **Cufflinks** suite contains two programs, **Cuffquant** and **Cuffnorm**, which can help to reduce the analysis time-scale. **Cuffquant** performs some of the elements of **Cuffdiff** by quantifying gene and transcript expression for a SAM/BAM file. The results are saved in the **CXB** format which is a binary format like BAM files and not human readable like SAM files. The results can then be sent to **Cuffdiff** and its run is both sped up and requires less memory.

Cuffnorm can also use the output from **Cuffquant**. Its function is to calculate normalized expression values for genes and transcripts. It is not used for differential expression but is designed to be used in the preparation of expression values for genes, transcripts, CDS groups, or TSS groups.

CUFFQUANT

Go to the **Assemble** menu and choose the **Quantify RNA-Seq Data Using Cuffquant…** menu item. For each **Cuffquant** run, you will need to tell **Sequencher** which SAM/BAM files you analyzed in the **Cufflinks** step. Click on the **Select SAM or BAM File** button and browse to the location of the SAM/BAM file. Click on the **Open** button. You will also need to provide a GTF file for the reference annotation, and optionally a GTF Mask file for abundant transcripts or FastA file if you are performing fragment bias correction, choosing the appropriate button(s). You will also need to choose a library type from its drop-down menu.

Each run will produce a CXB file. The result of a **Cuffquant** analysis cannot be directly viewed, instead it must be further analysed by **Cuffdiff** or **Cuffnorm**. Each **Cuffquant** run automatically creates a uniquely named run folder within the **Cuffquant** folder of your **External Data Home** folder.

ADVANCED OPTIONS FOR CUFFQUANT

Cufflquant is provided with a set of Advanced Options. These can be found by clicking on the **Advanced (Edit)** button. You will see a new dialog. To select an option, click on the checkbox adjacent to the option you wish to use. The parameter will be added to the **Current Parameters** window at the bottom of the dialog.

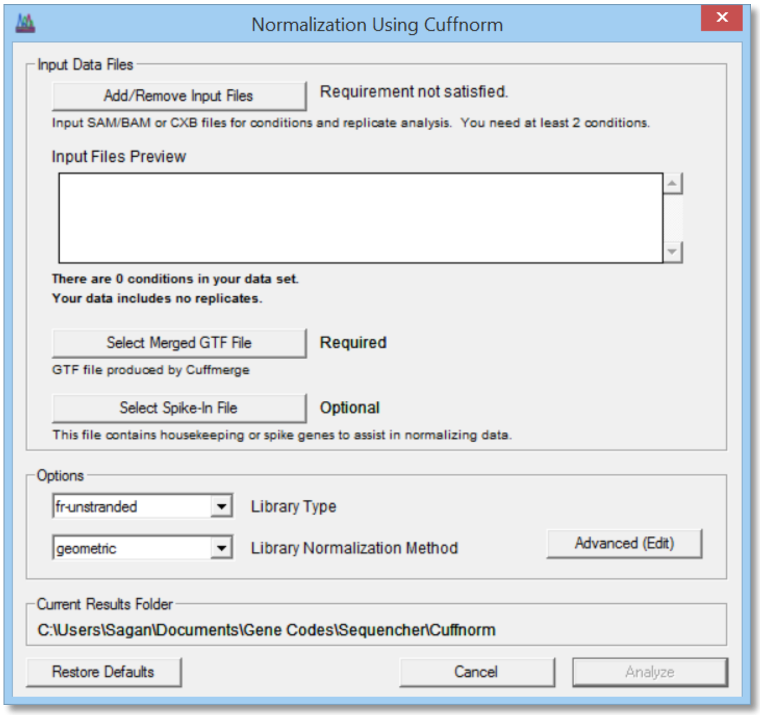
Some parameters require you to change a value. For example, the default value for –frag-len-mean is 200. Others are on/off switches. Once you have chosen all the Advanced Options you wish to apply, click on the **OK** button. The **Advanced Options** dialog closes. Click on the **Analyze** button on the **Quantify RNA-Seq Data Using Cuffquant** dialog to start the analysis of your data.

CUFFNORM

Go to the **Assemble** menu and choose the **Normalization Using Cuffnorm…** menu item.

You will need to tell **Sequencher** which SAM/BAM files you used in the **Cufflinks** step. Click on the **Add/Remove Input Files** button. The **Add Conditions and Replicates** dialog opens. Click on the **Add Input File** button and browse to the location of the SAM, BAM, or CXB file you want to add. The files you select with the **Add Input File** button will appear in the **Input File Name** list. If you chose the wrong files, simply highlight them and click on the **Remove Input File** button to delete them from the list. Initially, each file will have a default condition label of Condition 1. You can edit this as you add files or change the label after you have added all your files. Change a Condition Label by double-clicking on it and entering a new name in the **Edit Condition Label** dialog. Click on the **OK** button to dismiss the dialog. To dismiss the **Add Conditions and Replicates** dialog, click on its **OK** button.

Figure 16-71 Cuffnorm dialog



Requirement not satisfied.

Input SAM/BAM or CXB files for conditions and replicate analysis. You need at least 2 conditions.

There are 0 conditions in your data set.
Your data includes no replicates.

Required

GTF file produced by Cuffmerge

Optional

This file contains housekeeping or spike genes to assist in normalizing data.

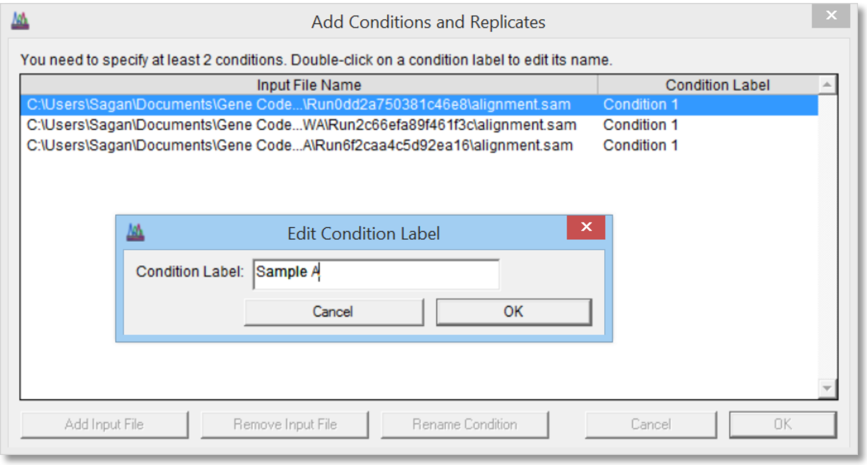
Library Type: fr-unstranded

Library Normalization Method: geometric

C:\Users\Sagan\Documents\Gene Codes\Sequencher\Cuffnorm

You will notice that the status of a number of items in the **Normalization Using Cuffnorm** dialog have changed. These status changes display the list of files you have added for analysis, whether you have fulfilled the requirements for normalization (at least two conditions), and whether your data includes replicates. Click on the **Select Merged GTF File** button and navigate to the location of the merged.gtf file for your data. You will find it in the **Cuffmerge** Run folder that was created from the corresponding **Cuffmerge** run. These Run folders are in your **Cuffmerge External Data Home** folder. If you are providing a spike-in file to assist in normalizing your data, click on the **Select Spike-In File** button, browse to the location of the spike-in file, and click on the **Open** button.

Figure 16-72 Add Conditions and Replicates editing a conditions label



Add Conditions and Replicates

You need to specify at least 2 conditions. Double-click on a condition label to edit its name.

Input File Name	Condition Label
C:\Users\Sagan\Documents\Gene Code...\Run0dd2a750381c46e8\alignment.sam	Condition 1
C:\Users\Sagan\Documents\Gene Code...WA\Run2c66efa89f461f3c\alignment.sam	Condition 1
C:\Users\Sagan\Documents\Gene Code...A\Run6f2caa4c5d92ea16\alignment.sam	Condition 1

Edit Condition Label

Condition Label: Sample A

Ensure that you have the correct values set for **Library Type** and **Library Normalization Method** by making selections from the drop-down menus. Now click on the **Analyze** button.

ADVANCED OPTIONS FOR CUFFNORM

There are only three advanced options for **Cuffnorm**, unlike the other members of the **Cufflinks** suite. Probably the most important options are **–compatible-hits-norm**, which only counts the fragments compatible with a reference transcript toward the number of mapped fragments in the FPKM denominator, and **–total-hits-norm**, which counts all the fragments.

17. RESTRICTION MAPS

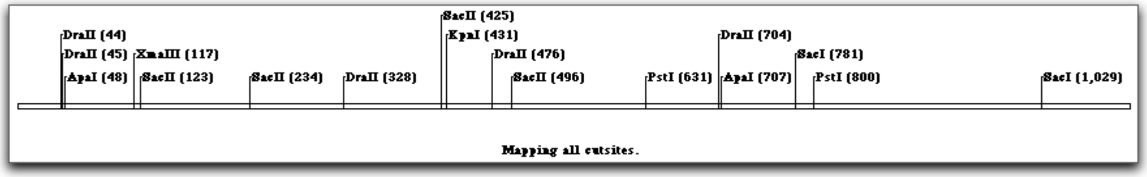
In this chapter, we discuss restriction maps and how to work with them. We discuss choosing Restriction Map display options, selecting and changing selected enzymes, saving new or modified enzymes, and copying the map.

DISPLAYING A RESTRICTION MAP

You can view a Restriction Map for a sequence or a contig by selecting its icon in the **Project Window** and opening an editor. Go to the button bar of the editor and click on the **Cut Map** button. Alternately, from the editor, go the **View** menu and choose **Bases, Map, Overview, …**, then choose **Restriction Map** from the submenu.

Figure 17-1 shows a restriction map display. The number in parentheses shows the cut location.

Figure 17-1 A single restriction map



Mapping all cutsites.

In this type of restriction map, clicking the name of an enzyme highlights all the locations where that enzyme cuts.

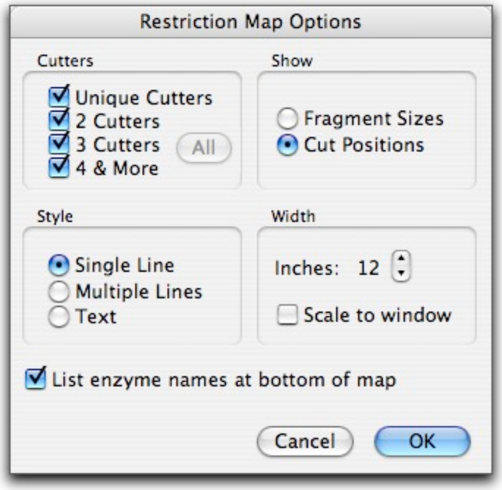
RESTRICTION MAP DISPLAY OPTIONS

HOW TO SELECT ENZYMES

To change the restriction map display, click on the **Options** button in the button bar or go the **View** menu and choose **View Options…**

A dialog (Figure 17-2) lets you change several attributes of the restriction map.

Figure 17-2 Restriction Map Options



Restriction Map Options

Cutters

☑ Unique Cutters
☑ 2 Cutters
☑ 3 Cutters
☑ 4 & More

Show

○ Fragment Sizes
● Cut Positions

Style

● Single Line
○ Multiple Lines
○ Text

Width

Inches: 12

☐ Scale to window

☑ List enzyme names at bottom of map

Use the **Cutters:** checkboxes to display enzymes according to the number of times they would cut. For example, you could select just **2 Cutters**, that is, only those enzymes that cut in two locations. If you select the **All** button, all of the checkboxes are turned on.

HOW TO SHOW CUT POSITIONS OR FRAGMENT SIZES

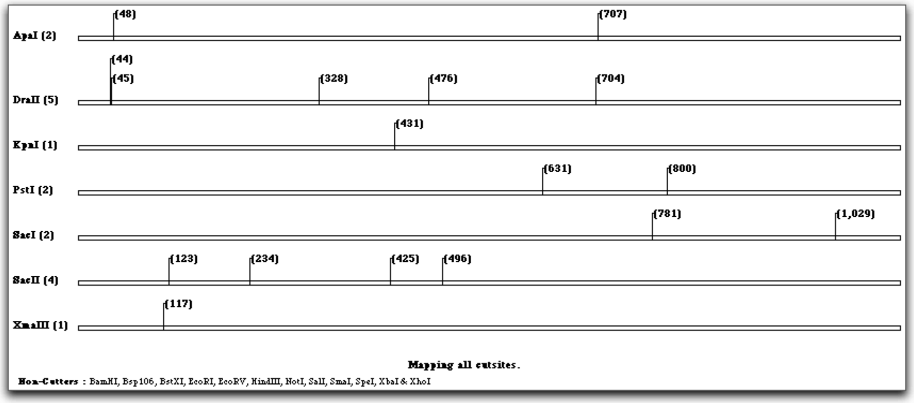
Use the Show options to choose the Fragment Sizes or the Cut Positions display by clicking on one of the radio buttons.

HOW TO SET THE MAP STYLE

Use a radio button from the **Style:** pane to choose one of the three basic map styles. The **Single Line** map is the default display.

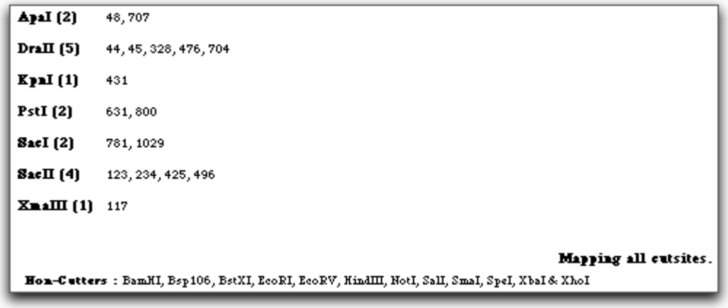
Choose **Multiple Lines** map (Figure 17-3) if you want to show each enzyme on its own line.

Figure 17-3 Multiple line Restriction Map



Choose the **Text** representation (Figure 17-4) to list the cut sites in tabular form. In this view, the number in parentheses following each enzyme shows the number of times that particular enzyme cuts.

Figure 17-4 Text Restriction Map



ApaI (2)	48, 707
DraII (5)	44, 45, 328, 476, 704
KpnI (1)	431
PstI (2)	631, 800
SacI (2)	781, 1029
SacII (4)	123, 234, 425, 496
XmaIII (1)	117

Mapping all cutsites.

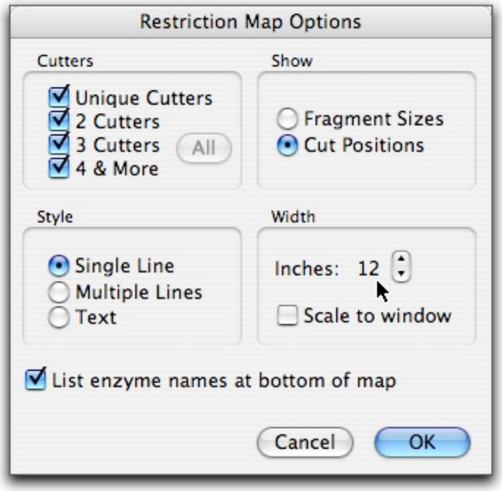
Non-Cutters : BamHI, Bsp106, BstXI, EcoRI, EcoRV, HindIII, NotI, SalI, SmaI, SpeI, XbaI & XhoI

SETTING THE MAP WIDTH

The restriction map normally scales to the width of the window. When you print a restriction map, it scales to the size of the printed page.

To set the width of the diagram in inches, turn off the **Scale to window** checkbox and use the elevator button to set the width in inches (see Figure 17-5). This option is particularly useful when copying a map to paste into a drawing or presentation program.

Figure 17-5 Map options showing width in inches



Restriction Map Options

Cutters

☑ Unique Cutters
☑ 2 Cutters
☑ 3 Cutters
☑ 4 & More

Show

○ Fragment Sizes
● Cut Positions

Style

● Single Line
○ Multiple Lines
○ Text

Width

Inches: 12

□ Scale to window

☑ List enzyme names at bottom of map

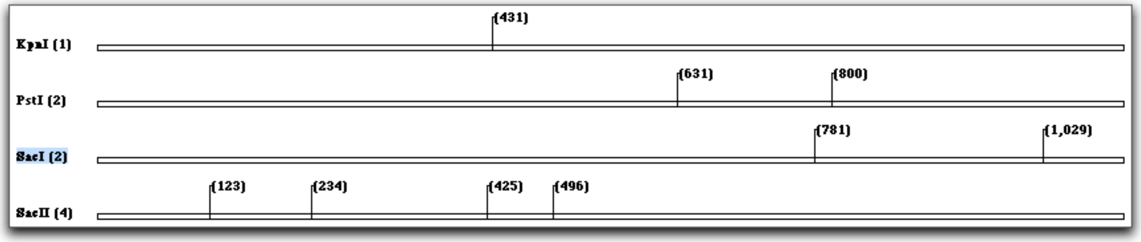
SETTING THE MAP CAPTION

Use the checkbox **List enzymes names at bottom of map** to toggle the caption at the bottom of the restriction map. It lists all of the enzymes used to make the map (see Figure 17-5 above). The caption also lists as “Non-Cutters” enzymes that were not mapped. For example, if you were displaying only unique cutters, then enzymes that cut more than once would be excluded. They would be listed as “Cutters that are not mapped.” Enzymes that did not cut at all would be listed as “Non-Cutters.”

GETTING MORE INFORMATION ABOUT AN ENZYME

You can get more information on the enzymes in your restriction map or edit the enzymes included in the map. In Figure 17-6, where the **Multiple Lines** style is being used, the SacI enzyme is selected.

Figure 17-6 SacI enzyme selected

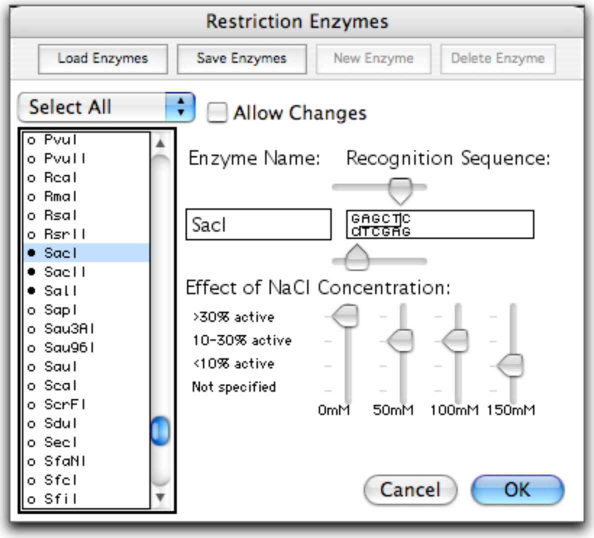


If you click on an enzyme label in the map and go the **File** menu and then choose **Get Info…**, you can bring up the **Restriction Enzymes** dialog. You can also use the keyboard shortcut **Ctrl+I** (Windows) or **Cmd+I** (Mac). Available enzymes appear in the scrolling list on the left side of the window. **Sequencher** highlights the enzyme you have selected and puts the appropriate information in the **Enzyme Name:** and **Recognition Sequence:** fields.

CHANGING THE SELECTED ENZYME

You can also bring up the dialog by clicking on the **Select Enzymes…** button in the button bar of any cut map. Alternately, you can go to the **Window** menu and choose **Specify Restriction Enzymes.** The enzyme editor dialog is shown in Figure 17-7.

Figure 17-7 Restriction Enzymes Editor



○ PvuI
○ PvuII
○ RcaI
○ RmaI
○ RsaI
○ RsrII
● SacI
● SacII
● SalI
○ SapI
○ Sau3AI
○ Sau96I
○ SauI
○ ScaI
○ ScrFI
○ SduI
○ SecI
○ SfaNI
○ SfcI
○ SfiI

□ Allow Changes

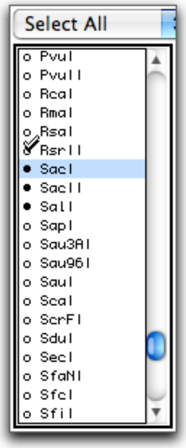
Enzyme Name: SacI

Recognition Sequence: GAGCTC / CTCGAG

Available enzymes appear in the scrolling list on the left side of the window. To view information about a particular enzyme, click on its name. **Sequencher** highlights the enzyme and puts the appropriate information in the **Enzyme Name** and **Recognition Sequence** fields.

If an enzyme is currently selected for display in the map, the bullet to the left of the name on Mac is filled in or the checkbox to the left of the name on Windows will be filled in with an x. An enzyme not selected for the map has an open bullet ‘o’ on Mac or an empty checkbox on Windows. When you move the cursor to the bullet/checkbox column, it changes to a check mark. To change your map selections, click on the bullets (Mac) or checkboxes (Windows) to toggle them on or off. Figure 17-8 shows the bulleted list in detail.

Figure 17-8 Bulleted Enzyme list



○ PvuI
○ PvuII
○ RcaI
○ RmaI
○ RsaI
○ RsrII
● SacI
● SacII
● SalI
○ SapI
○ Sau3AI
○ Sau96I
○ SauI
○ ScaI
○ ScrFI
○ SduI
○ SecI
○ SfaNI
○ SfcI
○ SfiI

If you are interested in enzymes with special characteristics, such as blunt ends or 3’ overhangs, explore the commands in the **Select** menu within the dialog.

Start by choosing **Select None** to clear all selections. Each menu item you choose after that will add all of the enzymes with the specified characteristic. Selections are *cumulative*—no command (except **Select None**) removes any enzymes from the list.

If, for example, you want all the enzymes that recognize sequences either 4 or 5 bases in length, choose **Select None** followed by **4-base** and then **5-base**.

When you are finished selecting enzymes, click **OK** in the lower right corner of the dialog or press the **Return** key. **Sequencher** redraws your restriction map with the newly selected enzymes.

CHANGING THE DEFAULT ENZYMES

To change the default enzyme selections for items in a project, select the sequences or contigs you want to change. Go to the **Window** menu and choose **Specify Restriction Enzymes** while viewing the **Project Window**. **Sequencher** will warn you if no sequences are chosen. The enzymes you choose at this time are now the default set for any selected sequences or contigs in your project.

ADDING AND EDITING ENZYMES

Sequencher lets you add new enzymes to the list. Go to the top of the enzyme editor and select **Allow Changes**. Click on **New Enzyme** to open a blank entry in the enzyme editor. Note that the **OK** button changes to **Done**.

You can now enter information in the **Enzyme Name** and **Recognition Sequence** fields. When you click **Done**, **Sequencher** adds the new enzyme to the list in alphabetical order.

CHANGING THE RECOGNITION SEQUENCE

The recognition sequence field shows a double-stranded sequence. If you type in the sense strand, the anti-sense strand will be calculated automatically.

Sliders above and below the recognition sequence window are used to move the lines which indicate the cleavage positions. You can use the mouse to move the sliders left and right.

Note: The maximum length for a recognition sequence is fifteen bases. If you exceed this length, the right end will be truncated.

SALT CONCENTRATION EFFECTS

You can record buffer sensitivity with the Effect of Salt Concentration: slider controls.

The default is an unspecified value for all concentrations of NaCI. Use the slider controls to indicate the percentage activity for a given NaCl concentration.

SAVING NEW OR MODIFIED ENZYMES

If you enter an enzyme that you will want to use later, click **Save Enzymes** to save the enzyme set as a file. Click **Load Enzymes** to retrieve a saved file.

SETTING THE SEQUENCE SELECTION IN A SEQUENCE EDITOR

When you click on the **Cut Map** button, a full-page cut-map is displayed. If you click between two cut sites, **Sequencher** highlights the resulting restriction fragment. Hold down the **Shift** key and click in the rectangle between other cut sites. Your selection is extended from the original site you selected. When you toggle back to the **Bases View**, you will see that the sequence representing the chosen region also has been highlighted.

If you wish to display the sequence and the cut map at the same time, click on the button showing the staggered enzyme cut icon. A window with the restriction map appears. You can adjust the size of the panes by dragging the splitter bar that separates the scroll bars for each pane.

Like the full-page cut map, this window is linked to the **Sequence Editor**. If you click on the name of an enzyme, **Sequencher** boxes the recognition sites. If you click between two cut sites, **Sequencher** highlights the resulting restriction fragment.

Hold down the **Shift** key and click in the rectangle between other cut sites. Your selection is extended from the original site you selected. You can copy the bases between the two cut sites by going to the **Edit** menu and choosing **Copy Selection**.

You can use this command to excise a particular restriction fragment by clicking on it in the cut map and switching back to the **Bases View**. Once you are in the **Bases View**, you can remove the highlighted fragment by going to the **Edit** menu and using the **Cut Selection** command.

SETTING THE SEQUENCE SELECTION IN A CONTIG EDITOR

When you are working in the **Bases View** of a **Contig Editor**, you can display a restriction enzyme map by clicking on the **Cut Map** button in the editor's button bar. You can also display this view by going to the **View** menu, selecting **Bases, Map, Overview, ...,** and then choosing **Restriction Map** from the submenu.

If you click between two cut sites, **Sequencher** highlights the resulting restriction fragment. Hold down the **Shift** key and click in the rectangle between other cut sites.

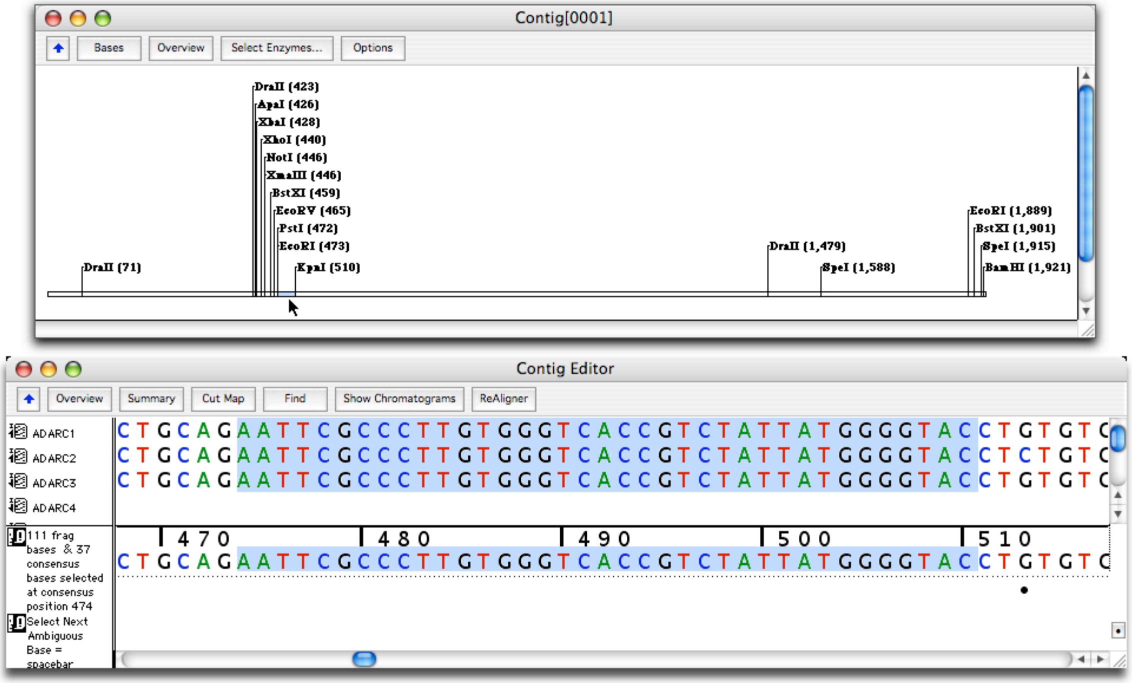
Your selection is extended from the original site you selected. When you return to the **Bases View**, you will see that the sequence representing the chosen region has been highlighted also. (See Figure 17-9.)

COPYING THE MAP

After you have made a selection in the restriction map, as in Figure 17-9, you can copy the bases between the two cut sites by going to the **Edit** menu and choosing **Copy Selection**.

You can also copy the entire diagram for use in an illustration program by going to the **Edit** menu and choosing **Copy As** and **Picture** from the submenu.

Figure 17-9 Fragment selected in map view and its equivalent bases highlighted



```
ADARC1  CTGCAGAATTCGCCCTTGTGGGTCACCGTCTATTATGGGGTACCTGTGTC
ADARC2  CTGCAGAATTCGCCCTTGTGGGTCACCGTCTATTATGGGGTACCTCTGTC
ADARC3  CTGCAGAATTCGCCCTTGTGGGTCACCGTCTATTATGGGGTACCTGTGTC
ADARC4
          470       480       490       500       510
        CTGCAGAATTCGCCCTTGTGGGTCACCGTCTATTATGGGGTACCTGTGTC
```

111 frag bases & 37 consensus bases selected at consensus position 474

18. SIMPLE BLAST

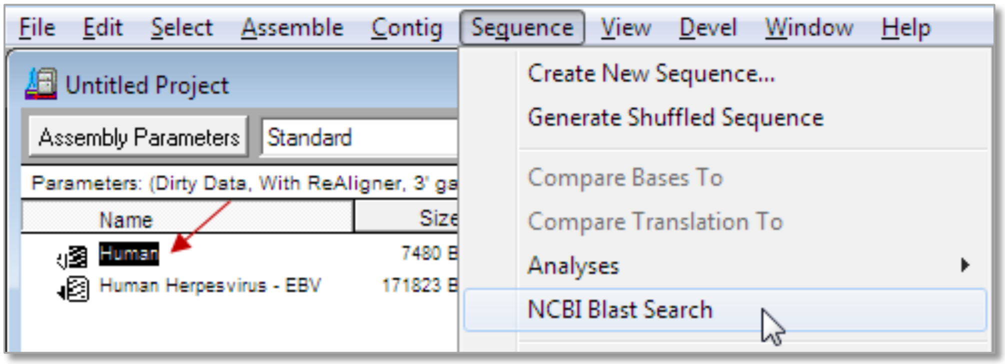
In this chapter, we describe how to use the built-in Blast feature to verify your sequence.

Depending on the speed of your internet connection, you will be able to query and verify short sequences rapidly, locate features, check primers, and so on, whilst still working on your project.

HOW TO USE BLAST WITH A SEQUENCE

In the image below, a sequence icon has been highlighted (indicated by red arrow). The entire sequence will be BLASTed. To initiate the search, choose the **Sequence** menu and then select the **NCBI Blast Search** menu item.

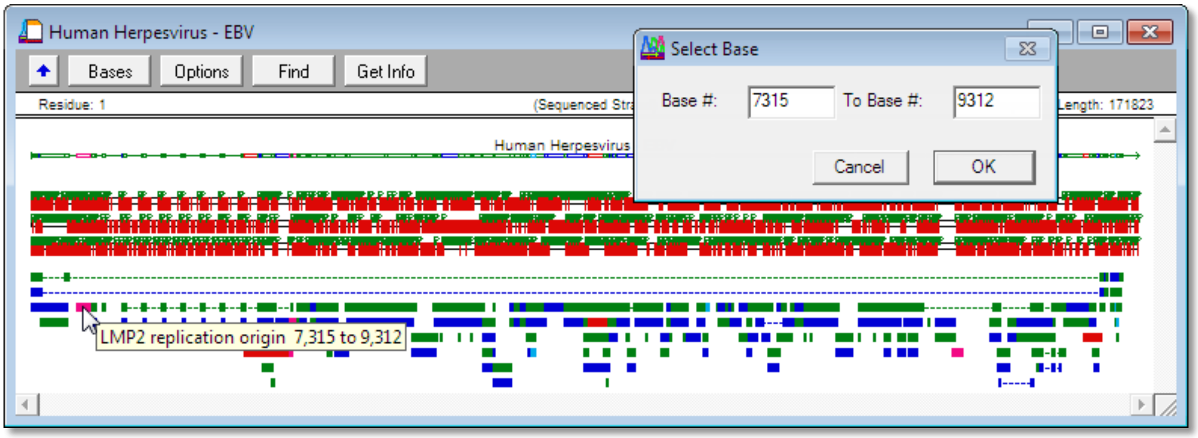
Figure 18-1 NCBI Blast menu command



HOW TO USE BLAST WITH A RANGE OF BASES

Instead of selecting a whole sequence, you may also choose to send part of a sequence. In this image, hovering the mouse over a feature in the **Overview** gives the location of the feature (yellow pop-up). You can then specify these coordinates by choosing the **Select** menu and then clicking on the **Bases By Number…** menu item. Next choose the **Sequence** menu and then select the **NCBI Blast Search** menu item.

Figure 18-2 Selecting bases for a BLAST search



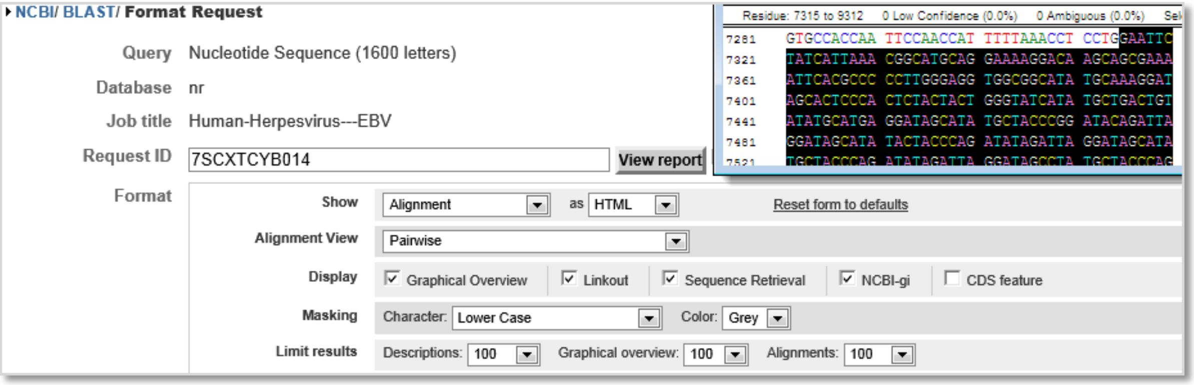
Base #: 7315 To Base #: 9312

HOW TO USE BLAST WITH A SPECIFIC SEGMENT OF SEQUENCE

You may also choose a subset of a sequence by highlighting bases in the **Sequence Editor** or **Contig Editor**. Next choose the **Sequence** menu and then select the **NCBI Blast Search** menu item to initiate the search.

This image shows this action (highlighted in blue) together with the BLAST Format Request page which will be shown in your default web browser. Click on the **View Report** button in the web page to see the actual results.

Figure 18-3 Selecting bases in a Sequence Editor for a BLAST search



Query Nucleotide Sequence (1600 letters)

Database nr

Job title Human-Herpesvirus---EBV

Request ID 7SCXTCYB014

Format

Show: Alignment as HTML

Alignment View: Pairwise

Display: ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi ☐ CDS feature

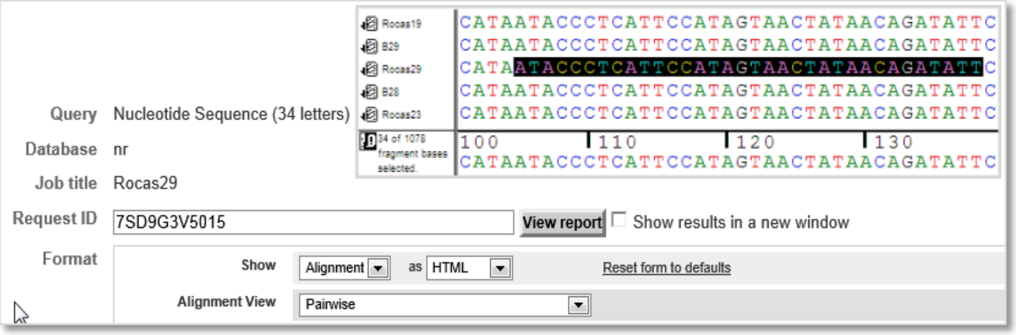
Masking: Character: Lower Case Color: Grey

Limit results: Descriptions: 100 Graphical overview: 100 Alignments: 100

```
Residue: 7315 to 9312    0 Low Confidence (0.0%)    0 Ambiguous (0.0%)    Sel
7281    GTGCCACCAA TTCCAACCAT TTTTAAACCT CCTGGAATTC
7321    TATCATTAAA CGGCATGCAG GAAAAGGACA AGCAGCGAAA
7361    ATTCACGCCC CCTTGGGAGG TGGCGGCATA TGCAAAGGAT
7401    AGCACTCCCA CTCTACTACT GGGTATCATA TGCTGACTGT
7441    ATATGCATGA GGATAGCATA TGCTACCCGG ATACAGATTA
7481    GGATAGCATA TACTACCCAG ATATAGATTA GGATAGCATA
7521    TGCTACCCAG ATATAGATTA GGATAGCCTA TGCTACCCAG
```

In this image, you can see that a segment from a sequence in the **Contig Editor** has been highlighted and sent to NCBI BLAST.

Figure 18-4 Selecting bases in a Contig Editor for a BLAST search



Query Nucleotide Sequence (34 letters)

Database nr

Job title Rocas29

Request ID 7SD9G3V5015

HOW TO USE BLAST WITH A CONTIG CONSENSUS

If you choose a contig icon from the **Project Window** and then choose the **Sequence** menu and select the **NCBI Blast Search** menu item, then only the consensus of the contig will be used as the search query.

19. MOTIFS AND FEATURES

In this chapter, we describe **Sequencher**'s ability to locate and highlight subsequences that you specify as motifs. You will learn how to create, edit, and display features and list a sequence's features.

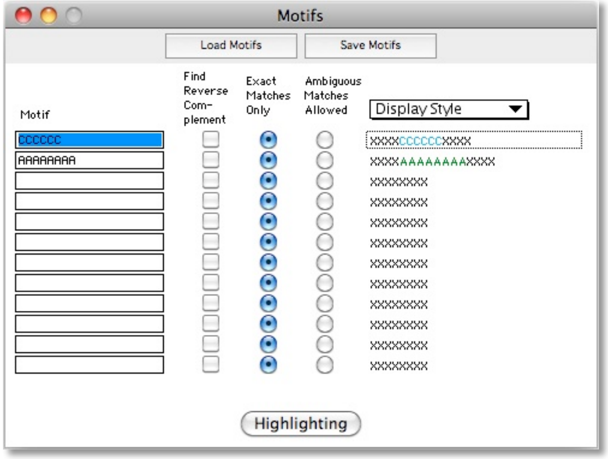
MOTIFS

Sequencher lets you define up to 12 patterns of 50 or fewer bases as motifs.

ENTERING MOTIFS

You define new motifs by going to the **Window** menu and clicking on the **Motif Definitions…** command.

Figure 19-1 Motif entry window



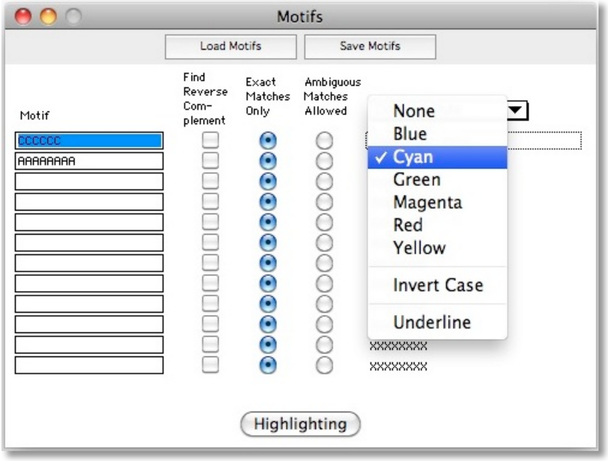
Motif	Find Reverse Com-plement	Exact Matches Only	Ambiguous Matches Allowed	Display Style
CCCCCC	□	●	○	XXXXCCCCCCXXXX
AAAAAAAA	□	●	○	XXXXAAAAAAAAXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX
	□	●	○	XXXXXXXX

The default setting shows exact matches only. To show the reverse complement for a motif, click on the **Find Reverse Complement** checkbox. Use the radio buttons to specify whether you want to highlight the motif using the exact bases you entered or using ambiguous matches.

To specify how you would like the motif you are entering to be highlighted, click on the **Display Style** pull-down menu (Figure 19-2) to choose color, case, and underlining while the cursor is still blinking in the entry field.

In the example (Figure 19-2), every instance of 'AAAAA' will be underlined and every start codon (ATG) will be shown in lower-case letters. You can save the motifs in a file on your hard disk by clicking on the **Save Motifs** button. Enter a suitable name for your motif file in the **Save** dialog.

Figure 19-2 Display style drop-down menu



QUICK MOTIF ENTRY

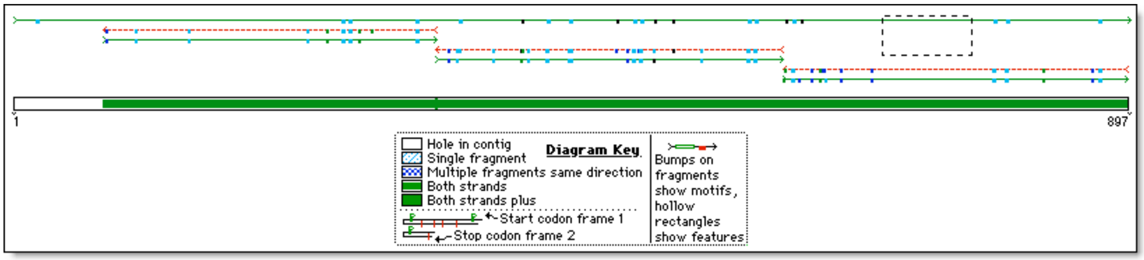
Select the bases in a **Sequence Editor**, then go to the **Edit** menu and click on **Enter Selection As Motif**. If you choose this command, the default style will be used (see Chapter 23 “Customizing Sequencher and User Preferences”).

DISPLAYING MOTIFS

To use the motifs you have entered, go the **View** menu and choose **Display Motifs**.

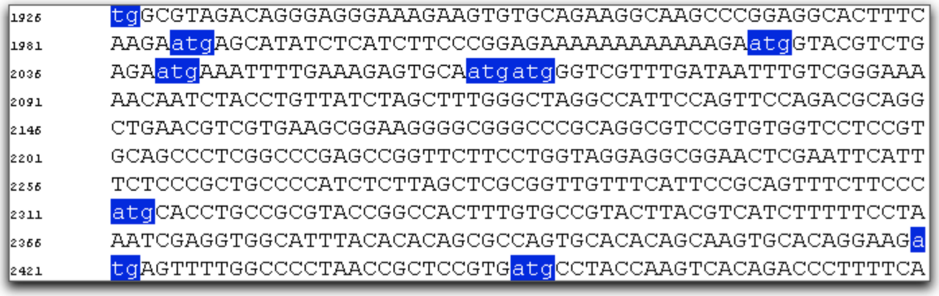
Figure 19-3 shows how motifs are displayed in the **Contig Overview**: they are the small downward facing colored blocks on the arrows representing the sequences. The colors are those you set in the **Motif Default Style** menu (see Chapter 23 “Customizing **Sequencher** and User Preferences”).

Figure 19-3 Overview showing motifs



In the **Bases View** of a sequence, motifs will appear highlighted in the manner you specified. You may need to deselect **Display Color Bases**, **Colors As Backgrounds**, or **Display Base Confidences** to see the Motifs more clearly. In Figure 19-4, for example, all start codons are in lower case type and blue with only **Colors as Backgrounds** on.

Figure 19-4 Sequence Editor showing motifs



```
1926    tgGCGTAGACAGGGAGGGAAAGAAGTGTGCAGAAGGCAAGCCCGGAGGCACTTTC
1981    AAGAatgAGCATATCTCATCTTCCCGGAGAAAAAAAAAAAAGAatgGTACGTCTG
2036    AGAatgAAATTTTGAAAGAGTGCAatgatgGGTCGTTTGATAATTTGTCGGGAAA
2091    AACAATCTACCTGTTATCTAGCTTTGGGCTAGGCCATTCCAGTTCCAGACGCAGG
2146    CTGAACGTCGTGAAGCGGAAGGGGCGGGCCCGCAGGCGTCCGTGTGGTCCTCCGT
2201    GCAGCCCTCGGCCCGAGCCGGTTCTTCCTGGTAGGAGGCGGAACTCGAATTCATT
2256    TCTCCCGCTGCCCCATCTCTTAGCTCGCGGTTGTTTCATTCCGCAGTTTCTTCCC
2311    atgCACCTGCCGCGTACCGGCCACTTTGTGCCGTACTTACGTCATCTTTTTCCTA
2366    AATCGAGGTGGCATTTACACACAGCGCCAGTGCACACAGCAAGTGCACAGGAAGa
2421    tgAGTTTTGGCCCCTAACCGCTCCGTGatgCCTACCAAGTCACAGACCCTTTTCA
```

If you have a certain set of motifs you want to use for several projects (for example, if you use motif definitions to highlight certain restriction sites), you can save your definitions to a file. To do so, click **Save Motifs** at the top of the **Motifs** dialog and follow the instructions. Then click on the **Load Motifs** button to use those definitions in other projects.

HIGHLIGHTING

The **Highlighting** aspect of **Motifs** gives you extended functionality where you will obtain a colored base only if it meets certain criteria. To access **Highlighting**, you must first click on the **Highlighting** button at the bottom of the **Motifs** dialog. Unlike **Motifs** which are restricted to a palette of colors, with highlighting, you can use the system colors which greatly extends the range of colors available to you.

To enable **Highlighting,** you must check the **Enable Highlighting** checkbox. Then choose whether you wish to have any of the criteria or all of the criteria you choose met. Choose the value from the **Highlight bases that meet** drop-down menu. Click on the color picker box to assign your color. Now choose the criteria you want. For example, you may want to see how many bases have a Phred score of 50 or above. Check the **Confidence between** checkbox and then assign the confidence range values. Finally, decide whether you want the coloring to override any other type of decoration, such as a feature, by clicking on the appropriate radio button.

FEATURES

WHAT IS A FEATURE

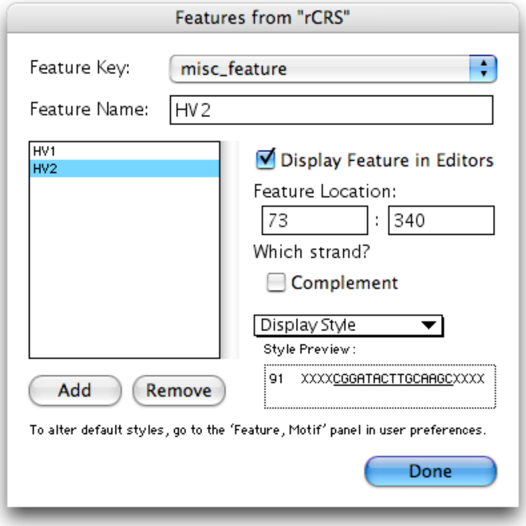
A feature represents a subsequence or region of a sequence which has some biological significance. Assigning a feature is a method of describing and naming such a region. **Sequencher** supports both personal features and GenBank standardized features. The **Feature Editor** enables you to name and highlight sequence regions of special interest.

Note: Features will not appear unless you go the **View** menu and select **Display Features**.

CREATING AND EDITING FEATURES

To create or edit features in a sequence, double-click on the sequence icon. This will open a **Sequence Editor** for your selected sequence. Go to the **Sequence** menu and choose **Edit Features…** to display the **Feature Editor** (Figure 19-5).

Figure 19-5 Feature Editor dialog



Features from "rCRS"

Feature Key: misc_feature

Feature Name: HV2

HV1
HV2

☑ Display Feature in Editors

Feature Location: 73 : 340

Which strand?

☐ Complement

Style Preview:

91 XXXXCGGATACTTGCAAGCXXXX

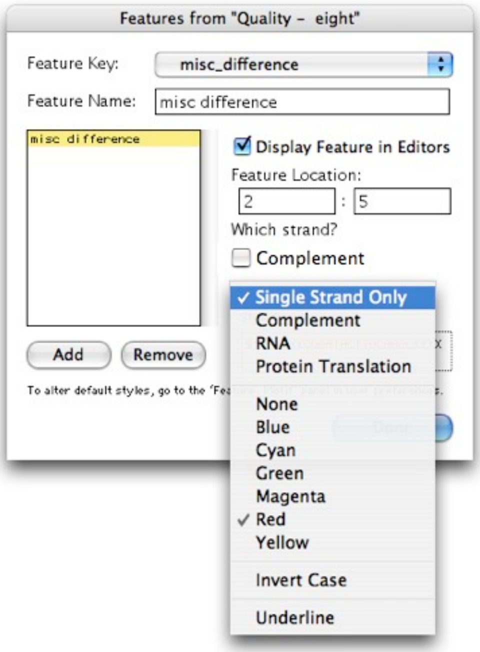
To alter default styles, go to the 'Feature, Motif' panel in user preferences.

To create a new feature, click on the **Add** button in the dialog. From the **Feature Key:** menu, choose **Sequencher** if you are creating personal annotations, or choose one of the GenBank keys.

If you have chosen a GenBank **Feature Key:**, **Sequencher** will insert a default name in the **Feature Name:** field using the Feature Key and Feature Qualifier. You may edit this if you wish. Enter the starting and ending base numbers for the feature in the **Feature Location** input fields. The default strand is **5'-> 3'**but you can change this using the **Complement** checkbox. Use the **Display Style** pull down menu to specify strand type, color, case and underlining. You may select one attribute from each of the four style categories. When you have finished entering or editing your features, click **Done.**

Check the **Display Feature in Editors** checkbox and then go to the **View** menu and click on **Display Features** to have your feature appear in the **Sequence** or **Contig Editor** and the associated overviews. When you set a feature to include Protein or DNA-RNA, this will only be displayed in appropriate views.

Figure 19-6 The Display Style menu



Features from "Quality – eight"

Feature Key: misc_difference

Feature Name: misc difference

misc difference

☑ Display Feature in Editors

Feature Location: 2 : 5

Which strand?

□ Complement

To alter default styles, go to the 'Fe

EDITING AN EXISTING FEATURE

To edit an existing feature, go to the **Sequence** menu and choose **Edit Features…** To display the **Feature Editor**, click on the name of the feature in the feature list. Use the **Display Style** pull-down menu to specify new display attributes. You can also change the Feature Key, edit the Feature Name, and the Feature Location. For example, you could change a **Sequencher** annotation into a GenBank feature.

If you want to remove a feature, select its name from the list and click on the **Remove** button.

When you have finished entering or editing your features, click **Done.** The features will appear in the **Sequence Editor** if you have checked the **Display Features in Editors** checkbox and **Display Features** in the **View** menu.

QUICK FEATURE CREATION

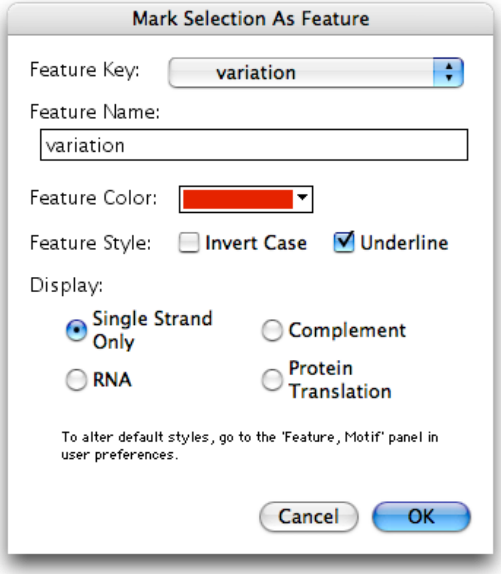
To create a feature quickly, select a region in the **Sequence** or **Contig Editor** and then go to the **Sequence** menu and choose **Mark Selection As Feature**. **Sequencher** displays a dialog called **Mark Selection As Feature** that lets you assign a Feature Key, name the feature, and specify display attributes. When you have finished assigning the feature attributes click the **OK** button.

If you create a feature while selecting bases in the consensus of a contig, you will actually create a feature for each of the sequences that contribute to the consensus at the selected bases.

Note: If a sequence spans only a portion of the bases selected for a feature, that sequence will not carry the new feature.

If you do not define display properties, the display attributes will default to the ones you specified in the **Feature, Motif** section of the **User Preference** settings. (See Chapter 23 “Customizing Sequencher and User Preferences” for more information.)

Figure 19-7 The Mark Selection As Feature dialog



Mark Selection As Feature

Feature Key: variation

Feature Name: variation

Feature Color: 🟥

Feature Style: ☐ Invert Case ☑ Underline

Display:

● Single Strand Only ○ Complement

○ RNA ○ Protein Translation

To alter default styles, go to the 'Feature, Motif' panel in user preferences.

SHOWING AND HIDING FEATURES

Go to the **View** menu and use **Display Features** to toggle feature display on and off. For each individual feature, you must also enable or disable its display in **User Preferences**. (See Chapter 23 “Customizing Sequencher and User Preferences” for more information.)

LISTING A SEQUENCE’S FEATURES

If you want to get a complete list of all your sequence’s features, go to the **Sequence** menu and choose **Feature Listing**. This will display a text list of each feature giving its name, Feature Key, range, style, and any Feature Qualifiers. You can print out this listing for your records.

LISTING SEVERAL SEQUENCE’S FEATURES

You can also view the listings for several sequences at once by selecting the sequences using **Ctrl+click** (Windows) or **Cmd+click** (Mac). Then go to the **Sequence** menu and choose **Feature**

Listing. A new dialog will appear asking if you want to open a feature listing window for each of the selected sequences. Click on the **OK** button to see the listings.

Note: If you have imported a GenBank sequence and **Sequencher** does not recognize any Feature Keys, this list may differ.

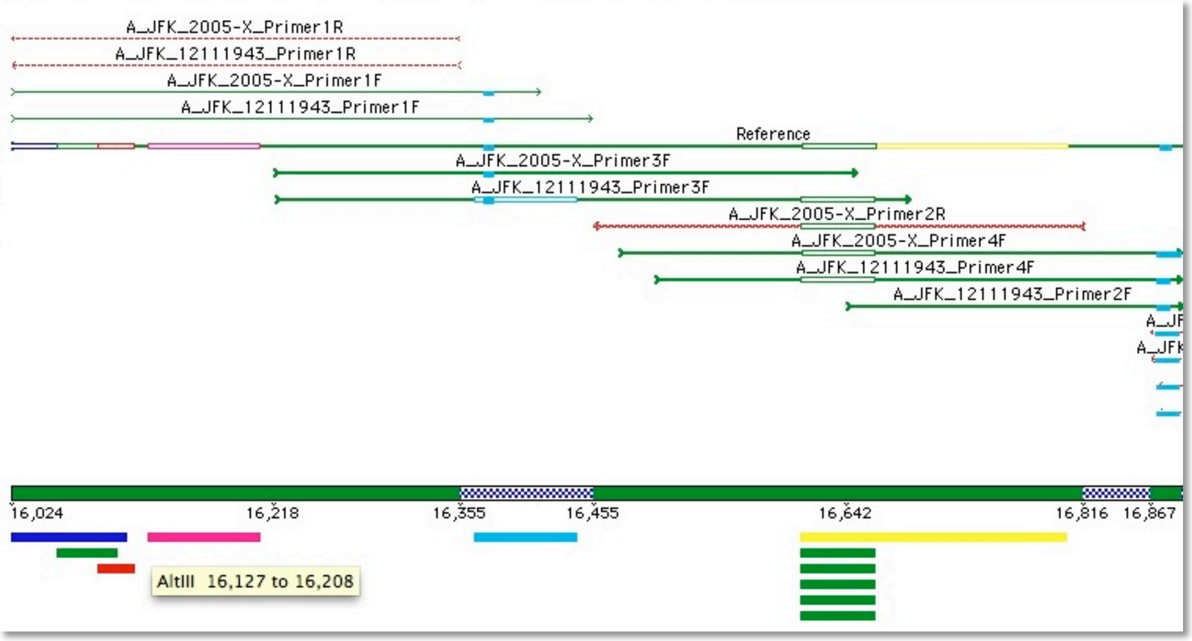
HOW FEATURES ARE DISPLAYED

When you annotate your sequence with features or if you have imported a sequence containing a GenBank Feature Table, **Sequencher** will present a graphical representation of the features under the display of overviews and restriction maps.

The graphical feature map has two parts. Above the sequences (red/green lines), the features are presented on one line; in this example, it is the Reference Sequence which has been marked up. Below the coverage map, the feature may be presented over several lines. You can see this at the five prime end of the sequence in the image below, where there are three features in blue, red, and green which overlap. In the bottom map, the features are displayed on separate lines.

As you hold your cursor over a feature, a tooltip appears with information about that feature. In the image above the cursor was left over the pink colored feature. The tool lets you know that it is called AltIII and its range is from base position 16, 127 to 16, 208. Note also the green features around base position 16, 642. Here the same feature exists on all the reads which overlap at this position.

Figure 19-8 Overlapping Features on the Overview



20. FINDING ITEMS

In this chapter, you will learn how to find items according to specific criteria. It also discusses how to refine your subsequence search and search for and find particular items in the **Project Window**. This chapter also discusses how to find open windows and editors.

FINDING THE PROJECT WINDOW

You can always bring the **Project Window** to the front by going to the **Window** menu and choosing **Project Window**.

FINDING OPEN WINDOWS WITH MENU COMMANDS

Sequencher has a number of commands in the **Window** menu for finding open windows hidden behind other windows—**Project Window**, **Variance Table**, **Translated Variance Table**, **Chromatograms**, **Contig Editors**, and **Sequence Editors**. Choose the appropriate command to bring the window you want to the front of your screen. If one or another of the commands is dimmed, you do not currently have those editors open. If you have several windows of a type open, these will be listed as submenus of that type.

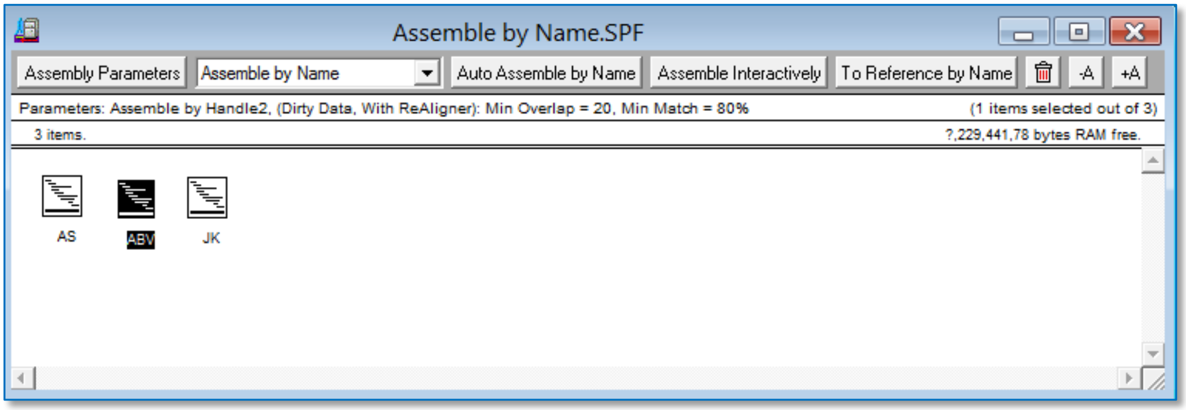
FINDING OPEN WINDOWS IN THE PROJECT VIEW

If you are looking at a **Project Window**, any icons that already have open editors are dimmed. To move an open editor to the front, just double-click the dimmed icon, for example, the icon for **ABV** shown in Figure 20-1 A dimmed icon.

FINDING THE CURRENT SELECTION

In a **Sequence** or C**ontig Editor**, you can always scroll to the currently selected bases by using the **Enter** key on the number keypad.

Figure 20-1 A dimmed icon



Assemble by Name.SPF

Parameters: Assemble by Handle2, (Dirty Data, With ReAligner): Min Overlap = 20, Min Match = 80%

(1 items selected out of 3)

3 items.

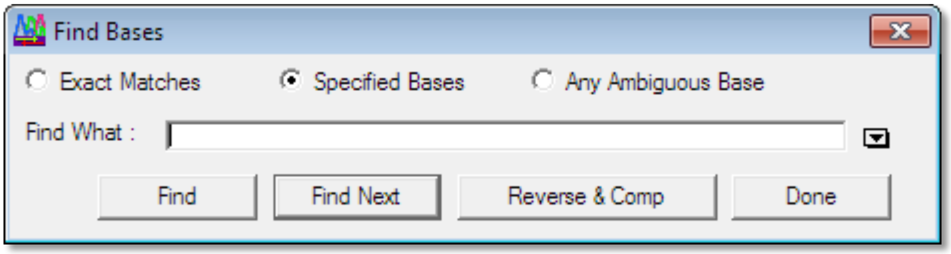
?,229,441,78 bytes RAM free.

SEARCHING IN OPEN WINDOWS

FINDING A SUBSEQUENCE

You can search for a specific subsequence in an open **Sequence** or **Contig Editor**. Go to the **Select** menu and choose **Find Bases…**; **Sequencher** displays the **Find** dialog shown in Figure 20-2. In the **Find What:** input field, type the string of bases you want to find. Use the drop-down menu to specify the recognition sequence of one of a subset of restriction enzymes available. Use the **Other** option in the drop-down menu to type in the name of your enzyme of interest. If it is in the database, the recognition sequence will be placed in the **Find What:** input field.

Figure 20-2 Find Bases dialog



Find Bases

○ Exact Matches ● Specified Bases ○ Any Ambiguous Base

Find What :

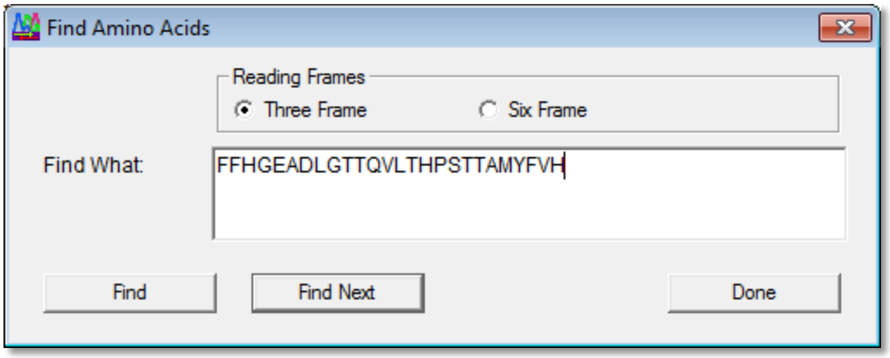
Find locates the first occurrence of the search string anywhere and then selects the string in the editor. **Find Next** finds the next occurrence after the current selection, whether it is a match found by **Sequencher** or a selection that you marked.

The **Find Bases** dialog stays in front until you click **Done**. If you want to repeat the same subsequence search but without invoking the **Find Bases** dialog, you can go to the **Select** menu and use the **Find Bases Again** command.

FIND AMINO ACIDS

Find Amino Acids is a protein identity search feature. You can use it to search in three or six frames in a **Sequence Editor** or **Contig Editor**. You need to use the single letter amino acid codes. As well as the standard set of codes, you can also use ? (question mark) or . (dot character). The ? (question mark) is used when an amino acid cannot be determined from the codon. The . (dot) is used to represent a stop codon. To search from the beginning of your sequence(s), use the **Find** button. To search from the position of the cursor, use the **Find Next** button. Click on the **Done** button when you have finished searching. If you later find that you wish to repeat the search, you can use the **Select>Find Again** command. The results of the search are shown in the **Sequence** or **Contig Editor** as a highlighted region.

Figure 20-3 Find Amino Acids dialog



Reading Frames: ● Three Frame ○ Six Frame

Find What: FFHGEADLGTTQVLTHPSTTAMYFVH

REFINING A SUBSEQUENCE SEARCH

You can refine any of your searches by using one of the radio buttons. The default setting is **Exact Matches.**

EXACT MATCHING

If you click **Exact Matches**, the search is based on exact character matches and disregards matches that occur due to ambiguities. For example, AGGTW would only match AGGTW.

MATCHING AMBIGUOUS BASES

If you click **Specified Bases,** the search finds only matches that are identical or more specific than the sequence you entered in the **Find What:** input field. For example, the string AGGTW would match AGGTT or AGGTA, since W is the ambiguity code for T *or* A.

ANY AMBIGUOUS MATCH

If you click **Any Ambiguous Match**, the search will find any degenerate DNA that may match the sequence you have entered, including matches consisting of all unknown bases. For example, AGGTW would match NNGTN or even NNNNN.

INCLUDE REFERENCE SEQUENCE IN FIND

If you have a Reference Sequence assembled into your contig, then you have the option of including it in your **Find Bases** search. Open the contig of interest to the **Bases View** and choose **Select>Find Bases**. The **Find Bases** dialog opens, but with an additional checkbox called **Include Reference Sequence in Find**. Click on this checkbox to include the Reference Sequence.

FINDING BASES BY NUMBER

To select a range of bases quickly, go to the **Select** menu and choose **Bases by Number**. A dialog lets you enter the range of bases you want to highlight.

EXTEND SELECTION

You can extend the current selection to the beginning or end of your sequence by going to the **Select** menu and choosing **Extend Selection.** Then choose the suboption **To Left End** or **To Right End.**

Note: In a **Contig Chromatogram**, the **Extend Selection** command can extend your selection across the entire sequence in one click.

OTHER SEARCH OPTIONS

You can also search for ambiguous bases by going to the **Select** menu and using the **Next Ambiguous Base** command. You can search for edited bases using the **Next Edited Base** option. (These are described in more detail in Chapter 8 "The Sequence Editor".) You can also search for potential open reading frames using the **Next Met to Stop (> 0b)** option. (This is described in more detail in Chapter 23 "Customizing Sequencher and User Preferences.")

PROJECT WINDOW SELECTION

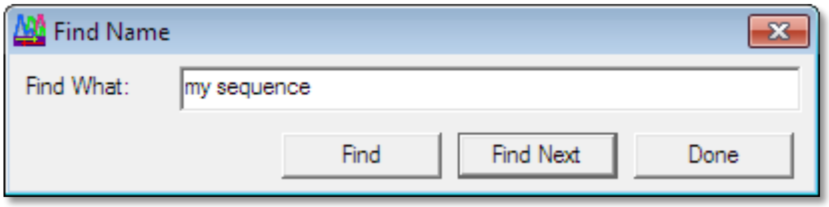
FINDING AN ITEM BY NAME

To locate a specific sequence or contig, click in the **Project Window** and type the name of the item you are trying to find, for example, 'Sequence C.' As you type, **Sequencher** starts looking in the **Project Window** for an icon with the name you are typing and highlights a match. If it finds more than one, it highlights the first one.

Note: If you pause between characters for more than a few seconds, the program assumes you've started typing a different name and looks for the new name.

Another way to locate a sequence is to go to the **Select** menu and choose **Item Named....** This brings up the dialog shown below.

Figure 20-4 Item Named… dialog



Find What: my sequence

Type the name of the sequence you want to find and click **Find**. **Sequencher** searches through all the sequences displayed in the **Project Window** for a name that matches the name you entered.

- When it finds a sequence, the sequence is displayed and highlighted in the **Project Window**.
- If **Sequencher** does not find a sequence with the exact name you typed, it looks for a sequence with a name starting with the first few letters you typed.
- If **Sequencher** still does not find a match, it looks at all the sequences contained in contigs. If it finds a matching sequence within a contig, it highlights the contig that contains the matching sequence.
- If **Sequencher** can find no matches anywhere, it stops searching and informs you that the item was not found.
- If there is more than one sequence with a matching name, **Sequencher** selects and highlights the first one. If that is not the one you want, click **Find Next**. This continues the search in alphabetical order.

Note: The **Find** window is movable.

SELECTING ALL ITEMS THAT…

The **Select** menu's **All Items That** command provides a series of suboptions based on different criteria. For example, you can choose sequences that have chromatograms or that include a particular annotation in the comments field. The menus are context sensitive so not all options may be available for a particular view.

Most options will have a dialog containing a number of important buttons that control the results of your search. The **Select All Matches** button will select matches based on your search criteria. If you change your criteria and search again, these results will be added to your initial search. If you don't want this information added, you must choose **Select None**.

A status line will indicate how many items matched out of the total number of items available. You can dismiss the window by clicking on **Done** or by simply clicking on the **Project Window** or performing another action.

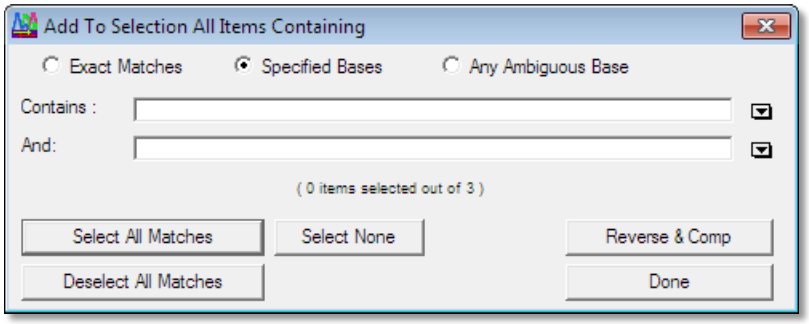
If you want to select all the items that *do not* contain the specified string, go to **Select All Items That**, click on the **All Items That** command, and click on the **Done** button. Then go to the **Select** menu and choose **Invert Selection**.

CONTAIN A SUBSEQUENCE

To select all the sequences and contigs in the **Project Window** that contain a specific string of bases, you can use the **Contain Subsequence** command. Go to the **Select** menu, click on the **All Items That** command, and then choose the **Contain Subsequence** menu item to display a dialog (Figure 20-5).

Note: You can do a Boolean 'and' search using two specifiers for the subsequence. Enter the first specifier in the **Contains:** field and the second in the **And:** field. There are also two drop-down menus to assist you in the search for specific restriction enzyme sites.

Figure 20-5 Contain subsequence… dialog



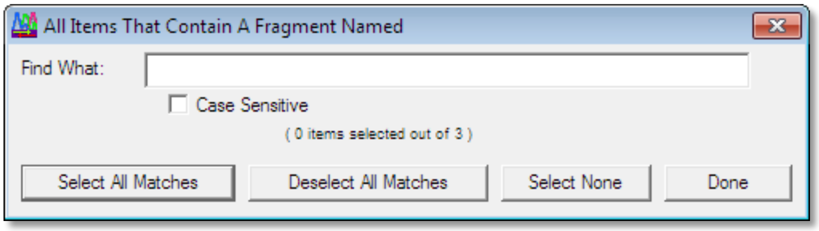
○ Exact Matches ● Specified Bases ○ Any Ambiguous Base

(0 items selected out of 3)

CONTAIN ITEMS NAMED

If you want to find which contig contains a specific sequence, go to the **Select** menu and choose **All Items That** and then **Contain Items Named…** to display a dialog (Figure 20-6). This lets you select items in the **Project Window** that contain a sequence bearing the specified name.

Figure 20-6 Contain Items Named… dialog



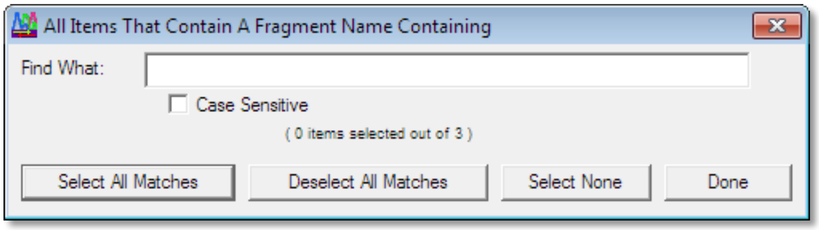
□ Case Sensitive

(0 items selected out of 3)

CONTAIN ITEMS WITH NAMES CONTAINING

If you are looking for several sequences with related or similar names and need to check which contig or refrigerator they are in, go to the **Select** menu and choose **All Items That.** Click on the submenu option **Contain Items With Names Containing…** to display a dialog (Figure 20-7). This lets you select items in the **Project Window** where the search string is based on only part of the complete name.

Figure 20-7 Contain Items With Names Containing… dialog



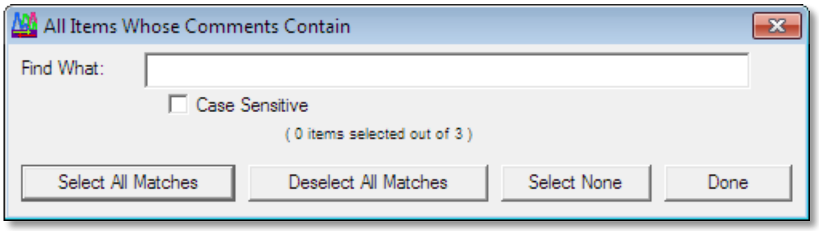
HAVE CHROMATOGRAMS

If you want to find out which sequences have a chromatogram or which contigs contain sequences with chromatograms, go to the **Select** menu and choose **All Items That** and then **Have Chromatograms**.

HAVE COMMENTS CONTAINING…

If you want to select all the sequences in the **Project Window** whose comments contain a specific text string, go to the **Select** menu and choose **All Items That** and then **Have Comments Containing…** to display a dialog (Figure 20-8).

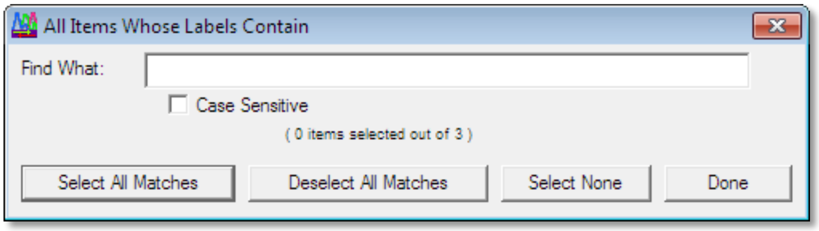
Figure 20-8 Have Comments Containing… dialog



HAVE LABELS CONTAINING…

If you want to select all the sequences in the **Project Window** whose label contains a specific text string, go to the **Select** menu and choose **All Items That** and then **Have Labels Containing…** to display a dialog (Figure 20-9).

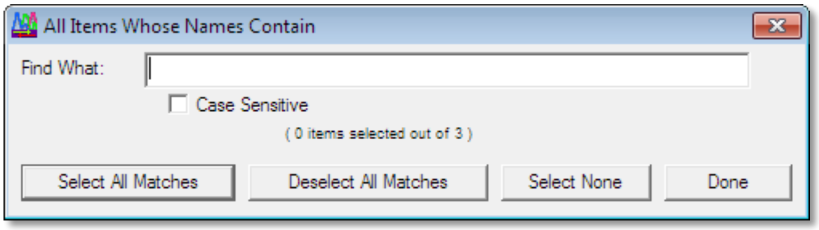
Figure 20-9 Have Labels Containing… dialog



HAVE NAMES CONTAINING…

If you want to select all the sequences in the **Project Window** whose names contain a specific text string, go to the **Select** menu and choose **All Items That** and then **Have Names Containing…** to display a dialog (Figure 20-10).

Figure 20-10 Have Names Containing… dialog


All Items Whose Names Contain

Find What:

□ Case Sensitive

(0 items selected out of 3)

WERE EDITED ON OR SINCE…

If you want to select all the sequences and contigs in the **Project Window** that were edited on a specified day, go to the **Select** menu and choose **All Items That** and then **Were Edited On Or Since…** to display a dialog. You can use the drop-down menu to specify a preset day such as today, yesterday, 1 week ago, or even 2 months ago. You can also type a date in the **Find What:** box.

Note: You can also include items edited on all the days since a specific date by checking the **And All Days Since** checkbox.

21. EXPORTING DATA

In this chapter, you will learn how to export your work. This chapter discusses how to export sequences, consensuses, and contigs, and also how to export defined data such as selected bases and subprojects.

EXPORTING DATA FROM **SEQUENCHER**

Sequencher has a rich set of export formats. You can export sequence(s), contig(s), or consensus sequence(s), subsets of your data, the Variance Table, the Summary Report and even protein translations.

Sequencher offers you a number of options depending on the type of data you choose to export. For each data type, you can select a choice of format and the location for the export. For some data types, there are additional options. **Sequencher** may offer a default for the export. Unless you deselect the option to use default names, **Sequencher** saves the file(s) in the default name. The **Export** dialog will default to the last location and choice of format used.

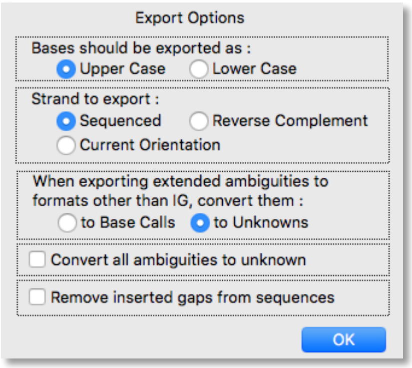
All of the export commands are accessible from the **File** menu.

EXPORTING SEQUENCES

To export a sequence or a number of sequences, click on the item(s) you want to export. Go to the **File** menu and click on **Export** and choose **Sequence(s)…** from the submenu. **Sequencher** provides a context-sensitive dialog for changing the export options.

The **Browse…** button opens a dialog called **Choose a Folder** (Mac) or **Browse for Folder** (Windows). From this dialog, you can choose where to store the new text file(s). You may create a new folder in which to store your sequences by clicking on the **Make New Folder** (Windows) or **New Folder** (Mac) button at the bottom left of the window. Click on the **OK** (Windows) or **Choose** (Mac) button to record your chosen location.

Figure 21-1 Export Options dialog



Export Options

Bases should be exported as :
● Upper Case ○ Lower Case

Strand to export :
● Sequenced ○ Reverse Complement
○ Current Orientation

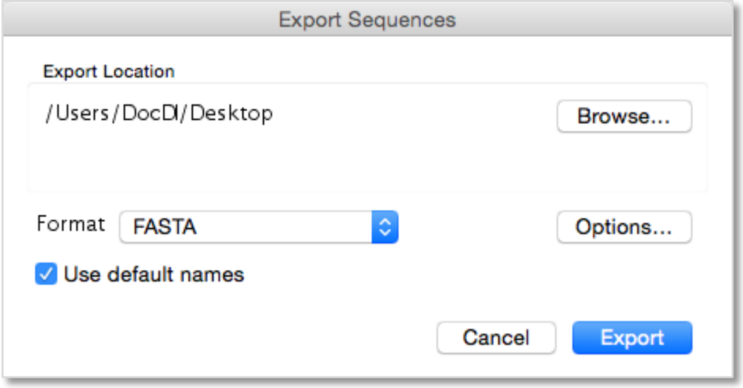
When exporting extended ambiguities to formats other than IG, convert them :
○ to Base Calls ● to Unknowns

□ Convert all ambiguities to unknown

□ Remove inserted gaps from sequences

The **Format:** drop-down menu allows you to choose a format type for your exported sequence. You may set the case of the exported sequence, the sequence strand to export, and the treatment of ambiguities and gaps. To choose these settings, click on the **Options…** button, select the options you want, and click **OK** (see Figure 21-2). Finally, click on the **Export** button to execute the command.

Figure 21-2 Export Sequences drop-down menu



Export Location
/Users/DocDI/Desktop

Format FASTA

☑ Use default names

Note: If you have added features to a sequence and chosen GenBank as your export format, these will be exported as a GenBank Feature Table.

EXPORTING A CONSENSUS

To export a consensus sequence, click on the contig icon(s) in the **Project Window**. Then go to the **File** menu and choose **Export** and **Consensus…** from the submenu. The **Format:** drop-down menu allows you to choose the format type for your exported sequence. Clicking on the **Options…** button displays the window where you turn options on or off. In most cases, you will want to just click on **Export** and use the original consensus name.

The **Browse…** button opens a window called **Choose a Folder** (Mac) or **Browse for Folder** (Windows)**.** From this window, you can choose where to store the new text file(s). You may create a new folder in which to store your sequences by clicking on the **Make New Folder** (Windows) or **New Folder** (Mac) button at the bottom left of the window. Click on the **OK** (Windows) or **Choose** (Mac) button to record your chosen location.

Uncheck the **Use Default Names** checkbox if you wish to set a specific export name.

EXPORTING CONTIGS

To export a contig, click on the item you want to export. Go to the **File** menu and click on **Export** and **Contig…** from the submenu. **Sequencher** displays the export dialog. The browse button allows you to choose a disk and folder to store the new text file(s). The format button allows you to choose the format type for your exported contig. You can choose from MSF, Nexus Interleaved, Nexus Sequential, Aligned FastA, and CAF. You may also set a number of options. Clicking on the **Options…** button displays the window where you turn options on or off. Finally, clicking on the **Export** button executes the command.

The **Use Default Names** option gives each export the name of its contig. Uncheck the **Use Default Names** checkbox if you wish to set a specific export name.

The **Browse…** button opens a window called **Choose a Folder** (Mac) or **Browse for Folder** (Windows)**.** From this window you can choose where to store the new text file(s). You may create a new folder in which to store your sequences by clicking on the **Make New Folder** (Windows) or **New Folder** (Mac) button at the bottom left of the window. Click on the **OK** (Windows) or **Choose** (Mac) button to record your chosen location.

If you execute the **Export** command with the **Sequence…** suboption while a contig is selected, **Sequencher** will create a folder with the name of the contig. The folder contains all of the sequences in the contig.

EXPORTING SELECTION AS SUBPROJECT

If you wish to share your data with colleagues who may have older versions of **Sequencher** or another program, choose **Export** from the File menu and then **Selection As Subproject…,** then select the appropriate version or type from the **Format** drop-down menu.

In certain cases, you may want to move some but not all items from one **Sequencher** project into another. Working in the **Project Window**, select the sequences and contigs you want to export. Go to the **File** menu and choose **Export** and then choose **Selection As Subproject**. **Sequencher** creates a new project containing only the items you selected. Now you can import the newly created sub-project into the chosen **Sequencher** project.

You should remember to provide a relevant name to define the new project because you have derived it from a selection of multiple items.

Note: Unlike other **Export** commands which may create a number of files on your hard disk, this command will create only one new project.

EXPORT SELECTED BASES

You may want to export selected bases rather than an entire sequence. To do so, drag the mouse to highlight your selected bases or go to the **Select** menu and use the **Bases by Number** command. Then go to the **File** menu and choose **Export** and click on **Selected Bases…** from the submenu. You will then see the **Save As** window which allows you to choose your format options and preferred location.

EXPORT OPTIONS

If you click on the **Options** button in an export dialog, **Sequencher** provides a context-sensitive dialog for changing the export options. Select the options you want and click **Export**.

EXPORT FORMATS

When you export a file, you may want to specify a file format so other software programs can access it. Click on the **Format:** drop-down menu in the export dialog. The menu will display a list of the available formats.

The **Sequencher Export** menu is context sensitive, which means that it varies based on whether your selection contains a sequence, a contig, or both. For each kind of export, you will be prompted to apply appropriate formats. See the Table below for more information.

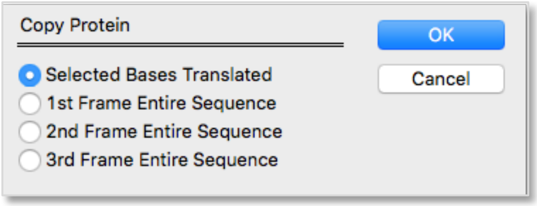
Note: SCF, Standard Chromatogram File format, is a file format for exchanging data between different sequencing systems that use four color chromatograms to show sequencing data. Most current sequencing hardware supports this file standard.

EXPORTING PROTEIN TRANSLATIONS

Sequencher can automatically translate bases into proteins as it copies items to the clipboard. These protein sequences can then be pasted in to another file or program.

To perform this, select a portion or all of the sequence and then go to the **Edit** menu and choose **Copy As and Protein Translation…** from the submenu. **Sequencher** asks you whether you want to translate only the selected bases or whether you want the entire sequence but in a specific frame. The default setting is to translate only the bases you selected. The **Selected Bases Translated** radio button is only available if you have selected some bases.

Figure 21-3 Copy As Protein Translation dialog

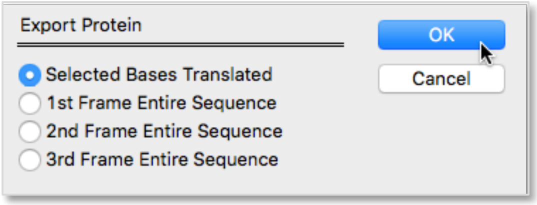


Copy Protein

● Selected Bases Translated
○ 1st Frame Entire Sequence
○ 2nd Frame Entire Sequence
○ 3rd Frame Entire Sequence

You can also export the translation. Go to the **File** menu. Click on the **Export** command and choose **Selected as Protein…** from the submenu. Pick from the following options:

Figure 21-4 Selected As Protein options



Export Protein

● Selected Bases Translated
○ 1st Frame Entire Sequence
○ 2nd Frame Entire Sequence
○ 3rd Frame Entire Sequence

The **Selected Bases Translated** radio button is only available if you have selected some bases. You will then be prompted with a **Save As:** dialog.

Table 21-1

	Export format type	**Description**
Sequence & Consensus formats	ASCII plain, Unformatted,	Plain text
	AFDIL, Fitch, Genentech, IG, Strider	Specialist and legacy formats
	GenBank, NBRF, EMBL	Database formats
	FASTA, concatenated FASTA, FASTQ	Commonly used formats
	Nexus/PAUP interleaved, Nexus/PAUP sequential, Phylip, Phylip 3, Phylip4	Phylogenetic program format
	SCF 2.0, SCF 3.0	Standard Chromatogram Format
Contig formats	CAF	Common Assembly Format
	MSF	Multiple Sequence Format
	FASTA aligned	Phylogenetic program format
	Nexus/Paup	Phylogenetic program format
Subproject formats	CAF	Common Assembly Format
	FASTA	Commonly used formats
	Sequencher	Older project file formats
Special import formats	ACE, CAF, CEP, FASTA aligned, GCG	Project formats from other programs such as Contig Express

EXPORT OVERVIEWS

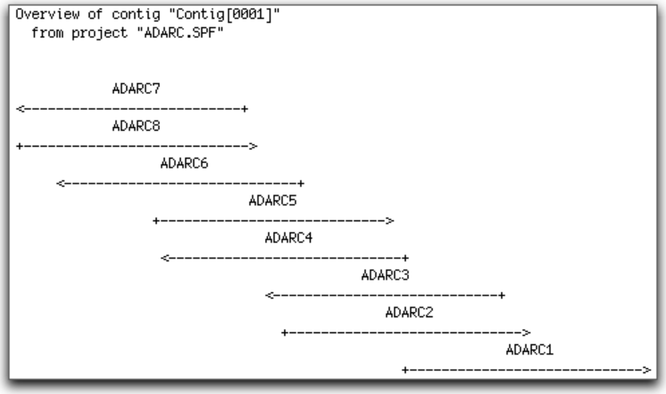
If you are looking at the schematic view of a contig, you can generate a text version map of the sequences, as shown in Figure 21-5, for viewing in other programs or word processors. There are two ways to do this.

1. Go to the File menu and choose **Export**, then **Overview as Text** from the submenu .**Sequencher** will treat the map as any other exported file.

2. Alternatively, you can copy the overview as a picture and paste it into other programs. Select the overview window by clicking on it, then go to the **Edit** menu and choose **Copy As** and **Picture** from the submenu.

Note: When viewing the output, you must specify a fixed-width font to preserve the scale of the arrows in relation to each other and the sequence names.

Figure 21-5 ASCII map of sequence



```
Overview of contig "Contig[0001]"
  from project "ADARC.SPF"


            ADARC7
<----------------------------+
            ADARC8
+----------------------------->
                  ADARC6
     <-----------------------------+
                             ADARC5
                 +----------------------------->
                               ADARC4
                  <-----------------------------+
                                           ADARC3
                               <-----------------------------+
                                              ADARC2
                                 +------------------------------>
                                                             ADARC1
                                                +----------------------------->
```

EXPORTING CONTIG SUMMARIES

When you are displaying the **Summary** view of a contig (see “Summary Report View” in Chapter 12, “Editing Contigs”), you can export the contents of the report to a text file. Go to the **File** menu and choose **Export** then the **Summary…** submenu. This is particularly useful if you wish to add annotations to a report of your assembly in standard page format.

22. OUTPUT

In this chapter, we discuss printing options. We will explain how to set up pages, copy pictures, print, and use different report formats.

COPY PICTURE

Sequencher can copy the images from many of its windows, including text windows, to a picture on the clipboard. Select the window you want copied, go to the **Edit** menu and choose **Copy As,** then click on **Picture.**

PAGE SETUP

Before printing, go to the **File** menu and choose **Print Setup…** (Windows) or **Page Setup…** (Mac) to change the options that are specific to your printer. Some options that may be available on your printer include portrait or landscape printing (vertical or horizontal orientation), paper size, reduction, enlargement, or special handling for color printing. (See Chapter 23, "Customizing Sequencher and User Preferences" for more information.)

PRINT TRACE IN ONE PAGE

This command appears within the **File** menu and allows you to print an individual trace so that the entire sequence fits on one page.

PRINTING IN DETAIL

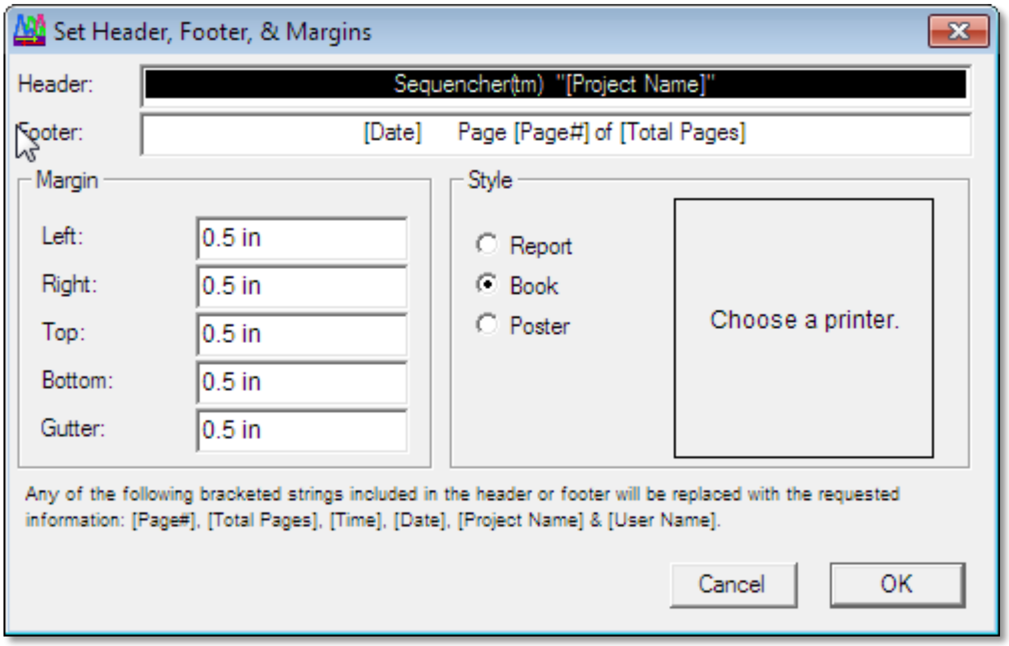
Most windows that contain data can be printed. To print, select a window by bringing it to the front. Then go to the **File** menu and choose **Print** and enter your print specifications.

SETTING HEADER, FOOTER AND MARGIN OPTIONS

Specify header, footer, and margins for any printout by going to the **File** menu and choosing **Set Header & Footer….** You enter the text of the header and footer and a code style for elements into a dialog (Figure 22-1) that **Sequencher** will automatically increase incrementally. The code elements are [Date], [Time], [Page], [Total Pages], [Project Name], and [User Name]. You can also add your own free text in either the header or the footer. Headers and footers are supported in all reports.

Set the margins by typing appropriate numbers into the **Left, Right, Top, Bottom**, and **Gutter** boxes in the **Margin** pane. Remember not to set margins smaller than those your printer can handle.

Figure 22-1 Setting Margins for printouts



Header: Sequencher(tm) "[Project Name]"

Footer: [Date] Page [Page#] of [Total Pages]

Margin

Left: 0.5 in

Right: 0.5 in

Top: 0.5 in

Bottom: 0.5 in

Gutter: 0.5 in

Style

○ Report

● Book

○ Poster

Choose a printer.

Any of the following bracketed strings included in the header or footer will be replaced with the requested information: [Page#], [Total Pages], [Time], [Date], [Project Name] & [User Name].

REPORT FORMATS

Special report formats are used for printing the contents of the **Contig Editor**. Specify the report format you want by clicking on the **Report, Book,** or **Poster** radio buttons from the **Style** pane of the **Set Header, Footer, & Margins** dialog. **Sequencher** then adjusts the report accordingly.

The **Report** option is for reports printed on single-sided pieces of paper that need to be displayed one page at a time. Each page is fully labeled.

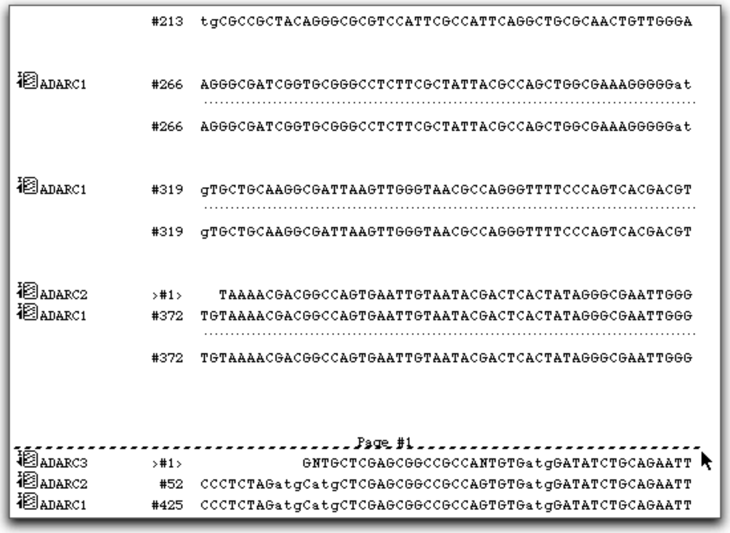
The **Book** option is for two-sided printouts when the user can look at a left and a right page simultaneously. Only the left-hand page of each pair will have all the labels for each sequence printed in a report; the right page uses the additional space for more data.

The **Poster** option is for displaying all your data together. It minimizes the amount of labeling and assumes that the entire report is to be laid out as one big rectangle for paste-up and display.

PAGE BREAKS

Page breaks, which are shown on screen as in Figure 22-2, appear in the **Sequence Editor** and in the **Summary Report** view of a contig.

Figure 22-2 A page break in the Summary Report



```
                 #213  tgCGCCGCTACAGGGCGCGTCCATTCGCCATTCAGGCTGCGCAACTGTTGGGA


ADARC1           #266  AGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGat
                       .....................................................
                 #266  AGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGat


ADARC1           #319  gTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGT
                       .....................................................
                 #319  gTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGT


ADARC2           >#1>    TAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGG
ADARC1           #372  TGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGG
                       .....................................................
                 #372  TGTAAAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTGGG



-------------------------------------Page #1--------------------------------------
ADARC3           >#1>             GNTGCTCGAGCGGCCGCCANTGTGatgGATATCTGCAGAATT
ADARC2            #52  CCCTCTAGatgCatgCTCGAGCGGCCGCCAGTGTGatgGATATCTGCAGAATT
ADARC1           #425  CCCTCTAGatgCatgCTCGAGCGGCCGCCAGTGTGatgGATATCTGCAGAATT
```

23. CUSTOMIZING SEQUENCHER & USER PREFERENCES

In this chapter, we discuss how to set your view preferences such as colors, custom codes, features and motifs, and the genetic codes you want to use. We also discuss how to set your user preferences such as confidence scores, labels and names, assembly parameters, and how you want to display chromatograms. You will learn that all of these preferences can be saved as individual or group preferences.

VIEW PREFERENCES

Sequencher offers you a number of options for specifying how you want to view your data in the **Sequence** and **Contig Editors**. These options can provide visual cues which help to differentiate your data.

AMBIGUOUS BASES

Sequencher can underline all ambiguous bases in the Sequence or Contig Editors. Go to the View menu and choose Base Ambiguities As. Go to the submenu and choose Underlined to mark ambiguities or Not Highlighted for unmarked ambiguities.

EDITED BASES

Sequencher can highlight bases that have been edited, which helps you audit the integrity of your data. Go to the **View** menu and choose **Base Edits As**. Select **Not Highlighted** for unmarked edits, **Bold Magenta** for colored edits, and **bOLD & cASE cHANGE** for bold and a changed case.

DISPLAY COLOR BASES

Sequencher can display bases in color. To turn on this feature for either a Sequence or Contig Editor display, choose Display Color Bases from the View menu. (For information on how to change the color assignments, see the section “Ambiguity Codes” in this chapter.)

COLORS AS BACKGROUNDS

Sometimes, displaying bases in color does not differentiate them enough. You can further enhance the colored bases by going to the **View** menu and choosing **Colors As Backgrounds**. **Sequencher** then displays the base as a colored rectangle with the base character printed over it.

DISPLAY BASE CONFIDENCES

Sequencher can represent imported confidence scores as a colored background. In order to view base confidence values as colored backgrounds, go to the **View** menu and click on **Display Base Confidences**. **Sequencher** will only display these backgrounds if the original data included base quality scores. (For information on how to change the threshold values, see the section “User Preferences” in this chapter.) You may need to turn off **Colors As Backgrounds**.

DISPLAY FEATURES

Sequencher can display features you have entered. Go to the **View** menu and click on **Display Features** to toggle this view on and off. For each individual feature, you must also enable or disable its display in User Preferences. (See also Chapter 19, “Motifs and Features”, for more information.)

DISPLAY MOTIFS

Sequencher can display motifs you have entered. Go to the **View** menu and click on **Display Motifs** to toggle this view on and off. (See also Chapter 19, “Motifs and Features”, for more information.)

SETTING UP CUSTOM CODES

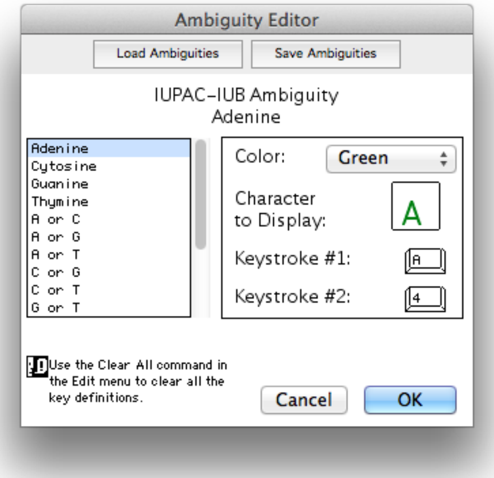
STANDARD CODING SYSTEM

The IUPAC-IUB ambiguity set is the most common coding system for representing DNA. **Sequencher** uses that coding system. See Appendix 28 for a table of the IUPAC-IUB code system.

CONFIGURING AMBIGUITY CODES

If you wish to configure **Sequencher**’s ambiguity coding system yourself, go to the **Window** menu and choose **Ambiguity/Key Codes...**

Figure 23-1 Ambiguity Editor



IUPAC–IUB Ambiguity
Adenine

Color: Green

Character to Display: A

Keystroke #1: A

Keystroke #2: 4

Use the Clear All command in the Edit menu to clear all the key definitions.

Each base and each ambiguity has a display character which is drawn on the screen and sent to the printer to represent that base or ambiguity. The display character interprets ASCII text files imported into **Sequencher**. Each code is also assigned a color which can be displayed in sequences or contigs by going to the **View** menu and clicking on **Display Color Bases**.

Select a base or ambiguity from the scrolling list at the left by clicking on its name. When the item is highlighted, information about it appears on the right side of the dialog.

To change the display character, click on the box to highlight it.

REPLACING AN EXISTING CHARACTER KEYSTROKE CODE

Each base code can be assigned up to two alternate keys from the keyboard. Using either of these keys generates the desired base code when using the **Sequence** or **Contig Editors**. These keystroke settings do not affect importing of data.

To change a keystroke setting for a base code, select the base code from the list on the left of the window, click on one of the keyboard keys to highlight it, and then press the key you want to assign to that base code.

To change a color setting for a base code, select the base code from the list on the left of the window and then select a different color from the **Color** drop-down menu.

SAVING AND LOADING CUSTOM CODES

You can load and save your customized ambiguity settings by going to the button bar at the top of the **Ambiguity Editor** and clicking on **Load Ambiguities** and then **Save Ambiguities**.

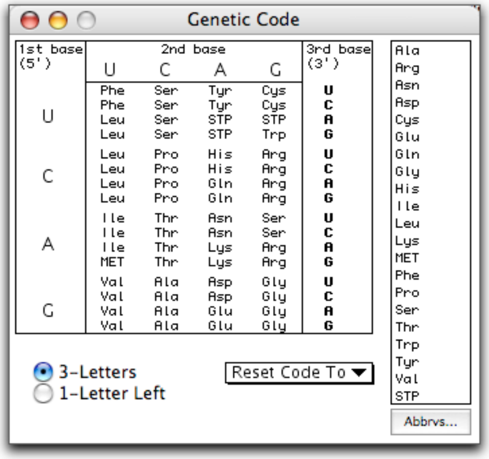
SHOW AND HIDE AMBIGUITY HELPER

Sequencher can show a small quick-reference window that lists the current ambiguity symbols. To bring up this window, go to the **Window** menu and click on **Show Ambiguity Helper.** Close the window with the close box in the upper left corner.

CHOOSING A GENETIC CODE

You will find that **Sequencher** already has the major genetic codes built in. You can choose from among these predefined coding systems by going to the **Window** menu and choosing **Genetic Code….** Select the code you want from **Reset Code To** pull down menu. The **Genetic Code** dialog is shown in Figure 23-2. Use the buttons at the bottom of the box to choose three letter or one-letter abbreviations. If you chose one-letter abbreviations, the letter will be aligned with the first (left) base of the codon.

Figure 23-2 Genetic Code Editor



1st base (5')	2nd base				3rd base (3')
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STP	STP	A
	Leu	Ser	STP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	MET	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Ala
Arg
Asn
Asp
Cys
Glu
Gln
Gly
His
Ile
Leu
Lys
MET
Phe
Pro
Ser
Thr
Trp
Tyr
Val
STP

● 3-Letters
○ 1-Letter Left

EDITING A GENETIC CODE

You can take any one of the pre-existing genetic codes and edit it in the **Genetic Code Editor** to change the codon translations. The **Genetic Code Editor** has two main parts: **Abbreviations** and **3-Letters** or **1-Letter Left** codes.

Go to the **Window** menu and choose **Genetic Code**. Choose the code you want to edit and then click on the table entry you want to change. Finally, click on the abbreviation you want in that position.

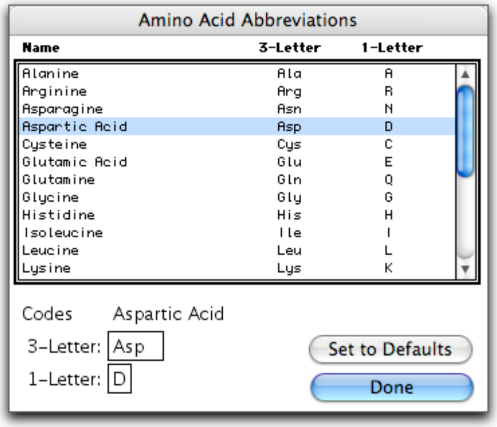
To undo any changes, select the appropriate genetic code from the **Reset Code** pull-down menu.

ABBREVIATIONS

To edit the abbreviations, click on the **Abbrvs…** button in the bottom-right corner of the dialog.

Another dialog (Figure 23-3) then lets you change both the three-letter abbreviations for the amino acid names and the corresponding one-letter codes. The standard abbreviations and one-letter codes are available as the default cases.

Figure 23-3 Abbreviations editor



Name	3-Letter	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K

Codes Aspartic Acid

3-Letter: Asp

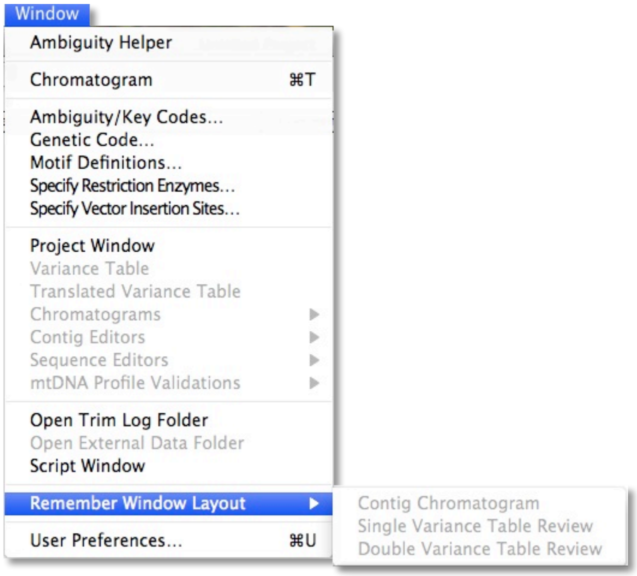
1-Letter: D

REMEMBER WINDOW LAYOUT

The **Remember Window Layout** command lets you define default positions for the **Contig Editor** and **Contig Chromatogram** when you are editing in the **Bases View**. First you must organize the windows. Then go to the **Window** menu and choose the command **Remember Window Layout.**

Select the submenu option called **Contig Chromatogram**. Every time you select the **Show Chromatograms** button, **Sequencher** will open the windows in your layout positions.

Figure 23-4 Remember Window Layout



You can also use the **Remember Window Layout** command to define default positions for the **Contig Editor**, **Contig Chromatogram**, **Variance Table**, or **Translated Variance Table** when you are in **Review Mode**. First you must organize the windows. Then go to the **Window** menu and choose the command **Remember Window Layout**. Select the submenu option **Single Variance Table Review** if you have one **Variance Table** open and **Double Variance Table Review** if both are open. Every time you select the **Review** button from the **Variance Table**, **Sequencher** will open the windows in your layout positions.

USER PREFERENCES

You can specify your preferences for many program settings in **Sequencher**. To set these, go to the **Window** menu and choose **User Preferences…**.

The **User Preferences** window contains a hierarchical list of preference settings. The main categories of settings are **General**, **Display** and **Input/Output**. Click on the triangle icon next to a category name to expand the list and show the various preference topics within that category. When a category list is expanded, the triangle changes to point downwards.

When you click a preference topic name, a Preference Pane appears on the right side of the window and enables you to specify your preferences for that topic.

Note: Each preference setting you select will be stored when you quit the program.

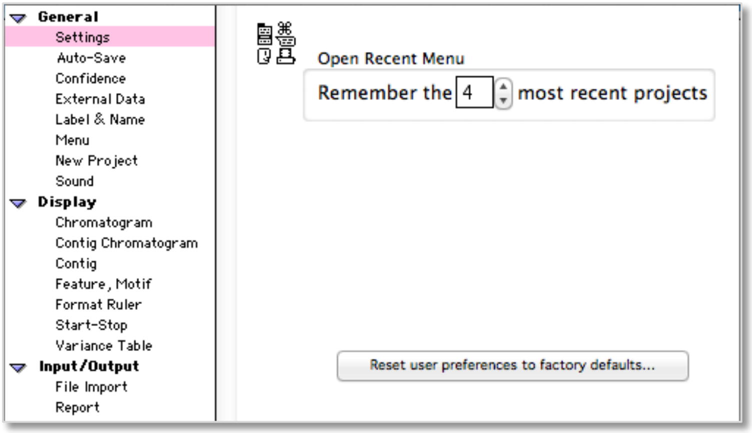
GENERAL

SETTINGS

On the **File** menu, you will find a command called **Open Recent**. The submenu for this command is a list of projects. You can specify the number of recent projects **Sequencher** will remember in this submenu. The maximum number is 99.

Sequencher is installed with a set of default values. As you work with the program, it replaces the default values with your preferences. However, if you want to restore the original settings, you will find a command button at the bottom of the **General > Settings** user preference pane called **Reset user preferences to factory defaults**. This option resets *all* of the user preferences to their original status.

Figure 23-5 Settings preferences



Open Recent Menu

Remember the 4 most recent projects

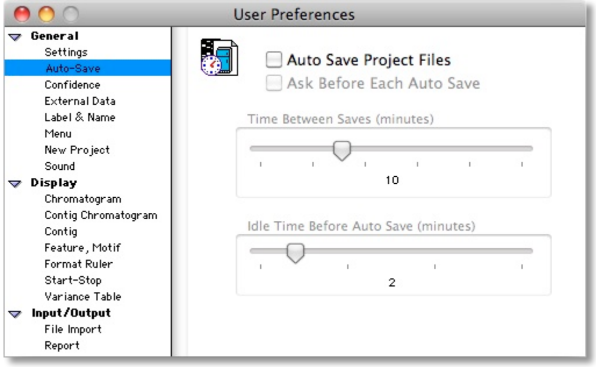
AUTO-SAVE

Auto-Save preferences control the timing of automatic saving of project files. To use this preference, you must have saved your data and given your project a name at least once.

To activate **Auto-Save,** click on the **Auto-Save Project Files** checkbox. Click on the **Ask Before Each Auto-Save** checkbox if you want to confirm the save each time.

Use the **Time Between Saves** slider to set how long **Sequencher** should wait between each Auto-Save. Use the **Idle Time Before Auto Save** slider to set how long your computer should be idle before **Sequencher** executes an Auto-Save. The latter setting will prevent **Sequencher** from performing an Auto-Save while you are typing.

Figure 23-6 AutoSave preference settings



□ Auto Save Project Files
□ Ask Before Each Auto Save

Time Between Saves (minutes): 10

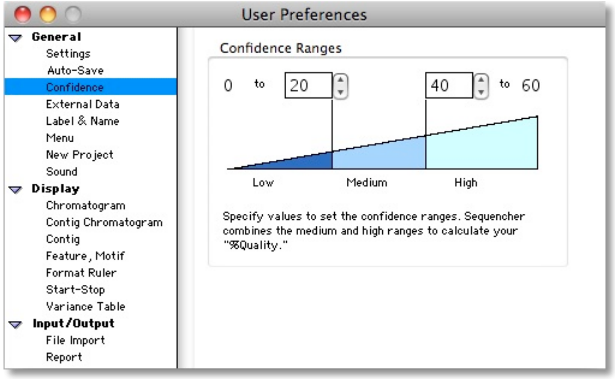
Idle Time Before Auto Save (minutes): 2

CONFIDENCE

Many of the new base callers generate confidence values associated with each base. **Sequencher** supports confidence values from PHRED, ACE, Trace Tuner, SCF, ABI, ESD and compatible files. The **Confidence** preferences control the display of base calls and confidence values from these files.

Sequencher's default settings show confidences below 20 as dark blue, the midrange values as a medium blue, and high confidences, that is above 40, as a light blue. Change the threshold values by altering the numbers in either of the **Confidence Ranges** boxes. If you have a **Sequence** or **Contig Editor** visible behind the **User Preference** pane, you will notice that the colors change as you set the new confidence ranges.

Figure 23-7 Confidence Ranges settings



Confidence Ranges

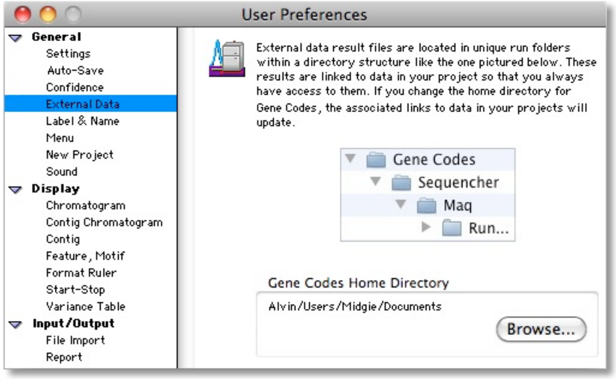
0 to 20   40 to 60

Specify values to set the confidence ranges. Sequencher combines the medium and high ranges to calculate your "%Quality."

EXTERNAL DATA

Results from external algorithms such as **Clustal**, **Maq**, and **GSNAP** are located in individual run folders each with a unique Run number. The default location for the **Home Directory** which contains these folders is in your **Documents** folder. Using the **External Data User Preference**, you can browse to a new location and use the **Choose** (Mac) or **OK** (Windows) buttons to confirm your choice.

Figure 23-8 External Data preference settings



External data result files are located in unique run folders within a directory structure like the one pictured below. These results are linked to data in your project so that you always have access to them. If you change the home directory for Gene Codes, the associated links to data in your projects will update.

Gene Codes Home Directory

Alvin/Users/Midgie/Documents

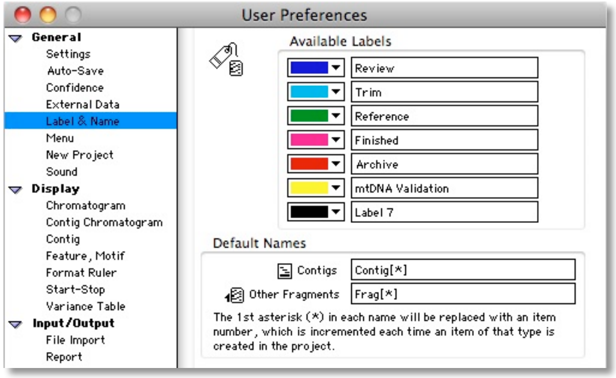
LABEL & NAME

The **Label & Name** preferences control the labels and descriptions used to mark sequences and contigs. Use the boxes grouped under **Available Labels** to choose label colors (use the pull-down color menus at the left of the label name) and descriptions for marking individual sequences.

Once you have set your preferences here, you can apply them to sequences, contigs, and refrigerators while you are working in a project by selecting the icon(s) in the **Project Window**. Go to the **Edit** menu and choose the **Label** submenu.

Under the **Default Names** section of the **Label & Name** pane, you can enter default names for the new contigs and sequenes. A bracketed asterisk tells **Sequencher** to add a number to the name and increase it for each new contig or sequence you add to the project.

Figure 23-9 Label and Names preference settings



Available Labels

🟦	Review
🟦	Trim
🟩	Reference
🟪	Finished
🟥	Archive
🟨	mtDNA Validation
⬛	Label 7

Default Names

Contigs	Contig[*]
Other Fragments	Frag[*]

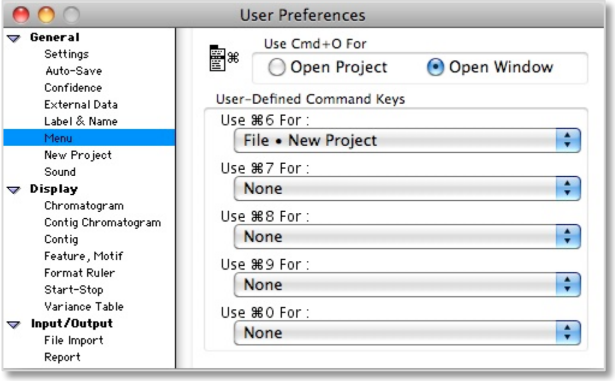
The 1st asterisk (*) in each name will be replaced with an item number, which is incremented each time an item of that type is created in the project.

MENU

The **Menu** preferences control command options that let you optimize the way you work in your own lab. Use the buttons under **Use Ctrl+O For** (Windows) or **Use Cmd+O For** (Mac) to program the **Ctrl+O** (Windows) or **Cmd+O** (Mac) key combination to open a project or open a window.

Use the pull-down menus grouped under **User-Defined Command Keys** to create keyboard shortcuts for any of the menu commands. After you have assigned a keyboard shortcut to a menu item, the shortcut will be listed to the right of the menu command on the drop-down menus.

Figure 23-10 Menu preference settings



Use Cmd+O For
○ Open Project ● Open Window

User-Defined Command Keys

Use ⌘6 For : File • New Project

Use ⌘7 For : None

Use ⌘8 For : None

Use ⌘9 For : None

Use ⌘0 For : None

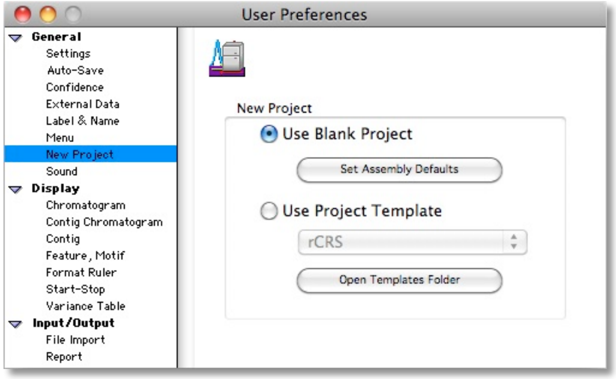
NEW PROJECT

This User Preference determines what type of project will be opened by the **New Project** command.

Click on the **Use Blank Project** radio button if you want your new projecy to be blank. Click on the **Use Project Template** radio button if you want the project settings to be controlled by a template. Specify the template you want by clicking on a name in the drop-down menu.

Templates are stored in a folder on your system. If you click onthe **Open Templates Folder** button, **Sequencher** will open a window displaying the templates. You can store up to 1,000 templates on your system. (For more information about templates, see Chapter 3, “The Project Window.”)

Figure 23-11 New Project User preference settings



New Project

● Use Blank Project

○ Use Project Template

rCRS

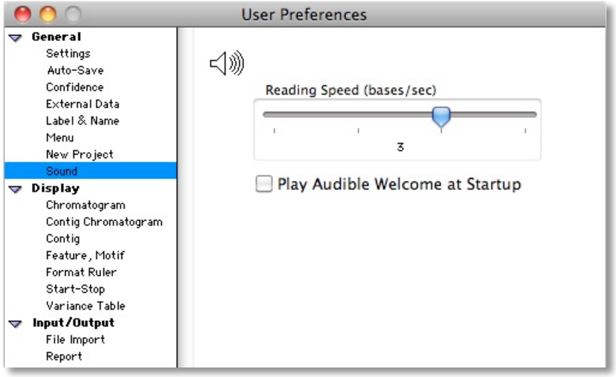
SOUND

Sequencher provides audio feedback to help you maintain the quality of data input.

When you select the **Sound** preference, you will see the preference pane shown in Figure 23-12. The slider controls how quickly **Sequencher** reads bases to you.

Click on the **Play Audible Welcome at Startup** checkbox if you want **Sequencher** to say “hello” at startup.

Figure 23-12 Sound preference settings



Reading Speed (bases/sec)

3

□ Play Audible Welcome at Startup

Note: You have to select a voice file to get audio output. (For more details, see Chapter 8 “The Sequence Editor” and review the section on “Voice Verification.”)

DISPLAY

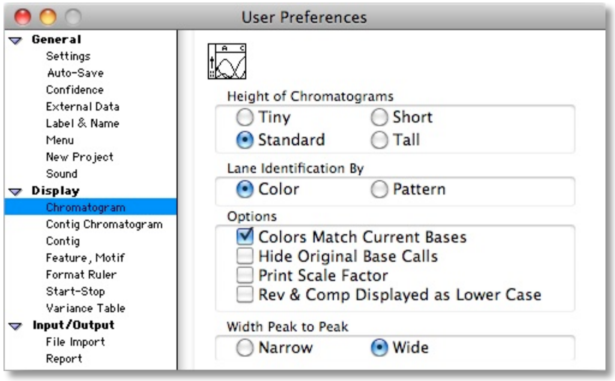
CHROMATOGRAM

Chromatogram preferences control the display of chromatograms. You can specify the **Height Of Chromatograms** by clicking one of the radio buttons from **Tiny** to **Tall**. If you use a black and white monitor, click on the **Pattern** radio button for **Lane Identification By**.

You can set lanes to be identified by color or pattern by selecting the appropriate radio button.

When viewing multiple sequences, you will notice that a reversed and complemented sequence displays the lane colors according to the original data. Click on the **Colors Match Current Bases** checkbox to display all current base calls in the same color. Click on the **Hide Original Base Calls** checkbox if you do not want them to appear in your chromatogram windows.

Figure 23-13 Chromatogram preference settings



User Preferences

Height of Chromatograms

○ Tiny ○ Short

● Standard ○ Tall

Lane Identification By

● Color ○ Pattern

Options

☑ Colors Match Current Bases

☐ Hide Original Base Calls

☐ Print Scale Factor

☐ Rev & Comp Displayed as Lower Case

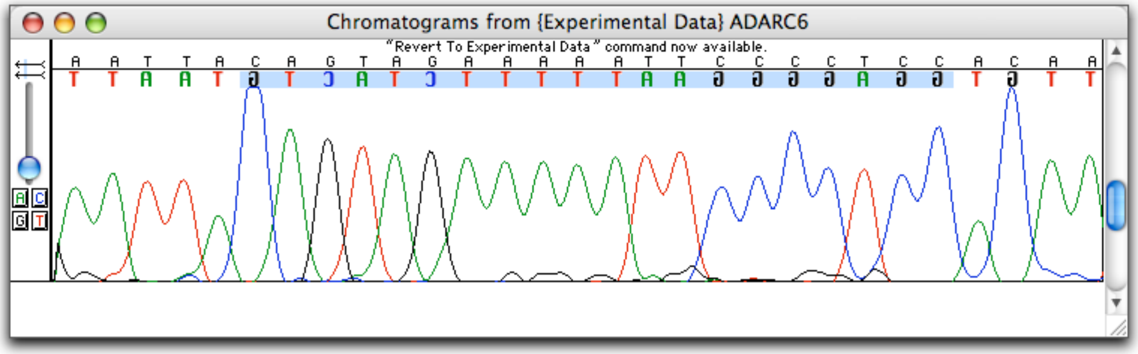
Width Peak to Peak

○ Narrow ● Wide

If you click on the **Print Scale Factor** checkbox and you used the slider shown at the left in Figure 23-12, the scale factor will appear on any printout of that chromatogram.

If you click on the **Rev & Comp Displayed as Lower Case** checkbox, the original base calls for a reversed and complemented sequence are shown in lower case (as opposed to backwards) as the second of two lines of bases above the traces (as in Figure 23-14).

Figure 23-14 Reversed and Complemented sequence



Chromatograms from {Experimental Data} ADARC6

“Revert To Experimental Data” command now available.

Under **Width Peak to Peak**, you can stretch your traces horizontally by selecting the **Wide** button.

CONTIG CHROMATOGRAM

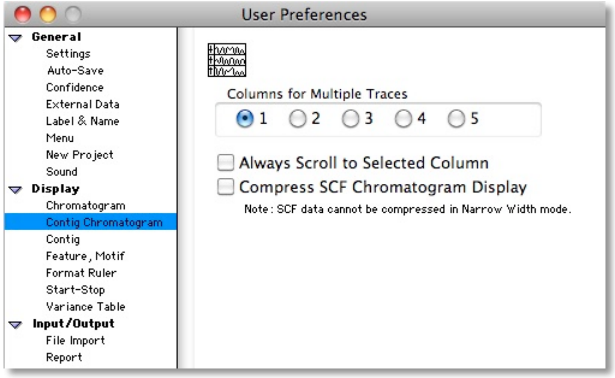
The **Contig Chromatogram** preferences control the display of chromatograms of sequences from a contig. When you have aligned several sequences from a contig and displayed the chromatograms, several traces appear in a single window, one above the other across the width of the window.

Use the **Columns for Multiple Traces** radio buttons to tile the data into columns so you can look at more traces at once. For example, if you select **3**, you’ll see three columns of traces. If your screen and window size allow you to see three rows of traces at once, you will be able to view a total of up to nine traces without having to scroll the window.

Click on the **Always Scroll to Selected Column** checkbox to have **Sequencher** automatically update the **Chromatogram** window whenever you select a new column of data in the **Contig Editor** window.

You can have **Sequencher** remember your window position. **Sequencher** can also compress SCF chromatogram displays.

Figure 23-15 Contig Chromatogram preference settings



User Preferences

- General
  - Settings
  - Auto-Save
  - Confidence
  - External Data
  - Label & Name
  - Menu
  - New Project
  - Sound
- Display
  - Chromatogram
  - Contig Chromatogram
  - Contig
  - Feature, Motif
  - Format Ruler
  - Start-Stop
  - Variance Table
- Input/Output
  - File Import
  - Report

Columns for Multiple Traces

● 1 ○ 2 ○ 3 ○ 4 ○ 5

□ Always Scroll to Selected Column

□ Compress SCF Chromatogram Display

Note: SCF data cannot be compressed in Narrow Width mode.

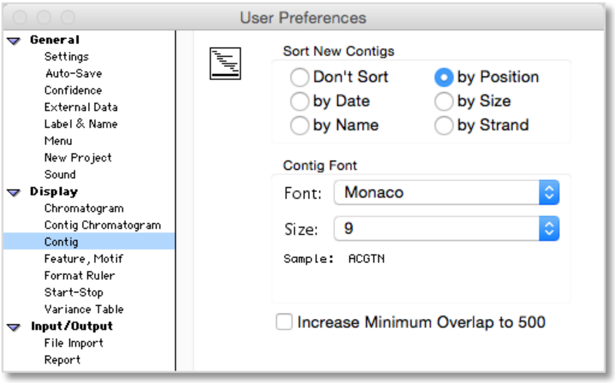
CONTIG

The **Contig** preferences control the vertical sorting of the sequences when a new contig is created or when sequences are added to an existing contig.

You can use the **Contig Font** option to change both the font and the font size. This allows you to use fonts other than the default Monaco 9 point font in the **Contig Editor**. You can see a sample of your current font selection below the font **Size:** drop-down menu in the **Sample** portion of the preference pane.

You can increase the **Minimum Overlap** range up to 500 by clicking on the **Increase Minimum Overlap to 500** checkbox.

Figure 23-16 Contig preference settings



Sort New Contigs

○ Don't Sort ● by Position
○ by Date ○ by Size
○ by Name ○ by Strand

Contig Font

Font: Monaco

Size: 9

Sample: ACGTN

□ Increase Minimum Overlap to 500

FEATURE, MOTIF

The **Feature, Motif** preferences control the default settings for new features and motifs. When you import a sequence from GenBank into **Sequencher**, these settings determine the name and appearance of the feature within both the **Overvie**w **and Sequencher's** editor windows. (For more information on GenBank Feature tables, refer to the Appendix 30 "Feature Keys and Qualifiers".)

When you click on the **Define Feature Key Default Styles…** button, a new window, also called **Define Feature Key Default Styles**, appears. On the left-hand side of the window is a pane listing Feature Keys.

To set new attributes, select the appropriate **Feature Key**. On the right-hand side, you will see a pane where you can set attributes such as the **Default Name**, **Display Style** and **Color**. You can restore the default settings at any time by clicking on the **Reset Style Defaults** button. Click on the **Done** button when you have finished.

When you click on an item in the left hand **Feature Key:** pane, you will see the same text appear in the **Default Name:** field. In most instances, there will be extra text enclosed in square brackets. The enclosed text is the default feature qualifier. GenBank have defined several Feature Qualifiers for each Feature Key. You can replace the text within the square brackets with your preferred Feature Qualifier. (For more on Feature Qualifiers, see Appendix

30 “Feature Keys and Qualifiers” and, for a listing of **Sequencher**’s default Qualifiers, see Appendix 29 “Default Feature Qualifiers”.)

Note: If a Label (indicated by /label= “some text”) is used in a Feature Table, it takes precedence over any default keys or qualifiers since it unambiguously identifies a feature or item.

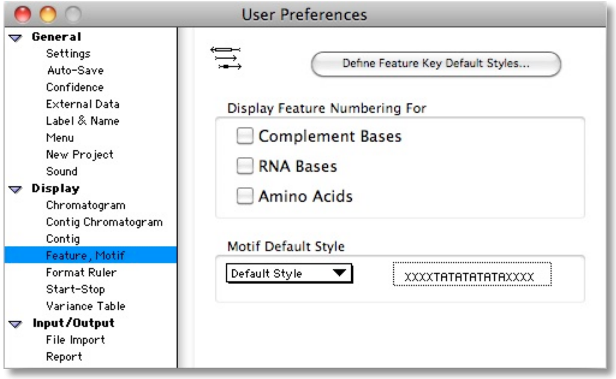
Click on the **Display Feature in Editors** checkbox if you want a specific feature to be displayed in the **Sequence** or **Contig Editors**.

If you are making changes to the way **Feature Keys** are currently displayed, you can click on the **Update Project…** button to update your project**.** You can decide whether to apply the current styles, name, or both to **All** features in your project. If you do not want to do so, just limit your choices to **Selected** features in your project by clicking on the appropriate button. When you are finished, click on the **Done** button to return to **User Preferences**.

Use one or more of the three **Display Feature Numbering For** checkboxes in the **Feature, Motif** pane to indicate which of the second-line features **Complement Bases, RNA Bases**, or **Amino Acids** should have its own numbering.

Use the **Motif Default Style** pull-down menu to set the color, case, and underlining style for new motifs. You can see a preview of the motif style in a box on the right-hand side of the **Motif Default Style** groupbox.

Figure 23-17 Features and Motif preference settings



Display Feature Numbering For

□ Complement Bases

□ RNA Bases

□ Amino Acids

Motif Default Style

Default Style ▼ XXXXTATATATATATAXXXX

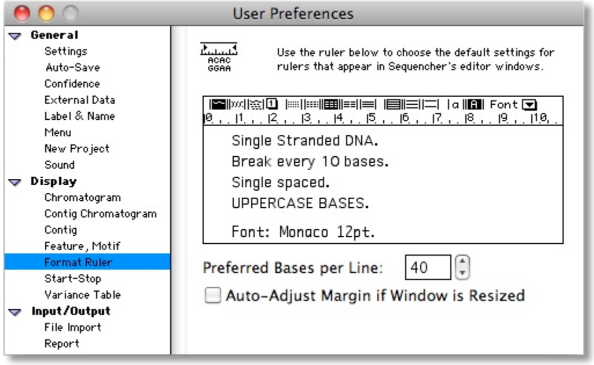
FORMAT RULER

The **Format Ruler** preferences control the display defaults for any view which has a format ruler. These are just defaults; you can change these settings later from within a specific editor.

When you click an icon in the sets of icons above the ruler, your selection is explained and confirmed in the panel below the ruler (Figure 23-18). For example, use the **Preferred Bases per Line** counter to increase or decrease the preferred number of bases in each line. Note that, if you request a line length that will not fit on the printed page, **Sequencher** uses the maximum line length that will fit.

If you want the margin to be automatically adjusted when you increase the size of a window, check the **Auto-Adjust Margin if Window is Resized** checkbox.

Figure 23-18 Format Ruler preference settings



Use the ruler below to choose the default settings for rulers that appear in Sequencher's editor windows.

Single Stranded DNA.
Break every 10 bases.
Single spaced.
UPPERCASE BASES.

Font: Monaco 12pt.

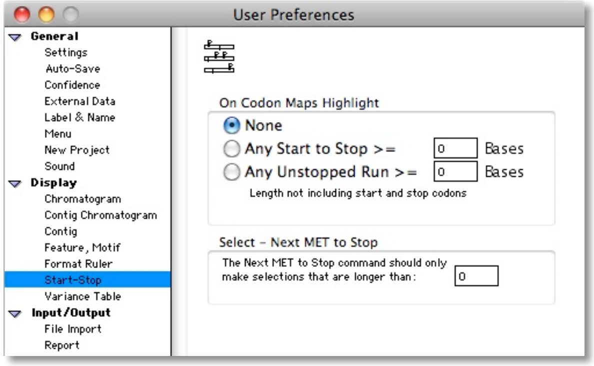
Preferred Bases per Line: 40

□ Auto-Adjust Margin if Window is Resized

START-STOP

The **Start-Stop** preferences control the locating of open reading frames from an editor. Use the buttons grouped under **On Codon Maps Highlight** to indicate whether you want **None**, which gives you the plain map of start and stop codons, **Any Start to Stop >=**, or **Any Unstopped Run >=** to give you a shaded box showing open frames of a length you specify. Figure 23-19 shows selections in the **Start-Stop** preference pane that will make **Sequencher** highlight ORFs beginning with a MET and longer than 60 bases.

Figure 23-19 Start-stop codon preference settings



On Codon Maps Highlight

● None
○ Any Start to Stop >= 0 Bases
○ Any Unstopped Run >= 0 Bases
Length not including start and stop codons

Select – Next MET to Stop

The Next MET to Stop command should only make selections that are longer than: 0

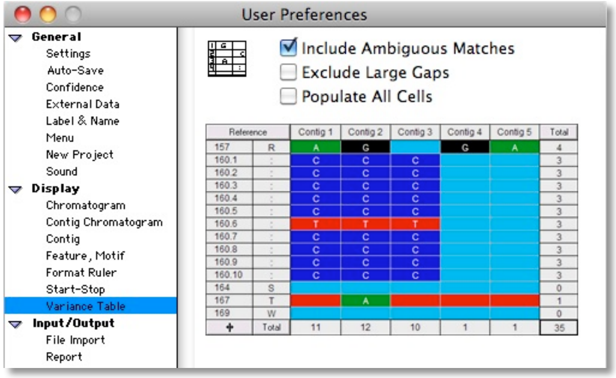
Under **Select - Next MET to Stop**, you can define the minimum length of ORF selections in a **Sequence Editor**. If you specify a minimum length here, for example 30 bases, the **Next MET to Stop** command under the **Select** menu in the **Sequence Editor** will change to read **Next MET To Stop (>30b)**.

VARIANCE TABLE

The **Variance Table** preferences control how ambiguous matches and large gaps will be treated. If you wish to include ambiguous matches in your table, click on the **Include Ambiguous Matches** checkbox. If you wish to exclude large gaps from your table, click on the **Exclude Large Gaps** checkbox.

When you generate a **Variance Table**, you would expect to have blank cells where the base at that position matches the exemplar. If you would like to see all bases across the comparison range, whether they match or are different to the exemplar, then click on the **Populate All Cells** checkbox (see Figure 23-20).

Figure 23-20 Variance Table preference settings



☑ Include Ambiguous Matches
□ Exclude Large Gaps
□ Populate All Cells

Reference		Contig 1	Contig 2	Contig 3	Contig 4	Contig 5	Total
157	R	A	G		G	A	4
160.1	:	C	C	C			3
160.2	:	C	C	C			3
160.3	:	C	C	C			3
160.4	:	C	C	C			3
160.5	:	C	C	C			3
160.6	:	T	T	T			3
160.7	:	C	C	C			3
160.8	:	C	C	C			3
160.9	:	C	C	C			3
160.10	:	C	C	C			3
164	S						0
167	T		A				1
169	W						0
+	Total	11	12	10	1	1	35

INPUT/OUTPUT

FILE IMPORT

File Import preferences control how **Sequencher** applies the automatic trim function to imported files.

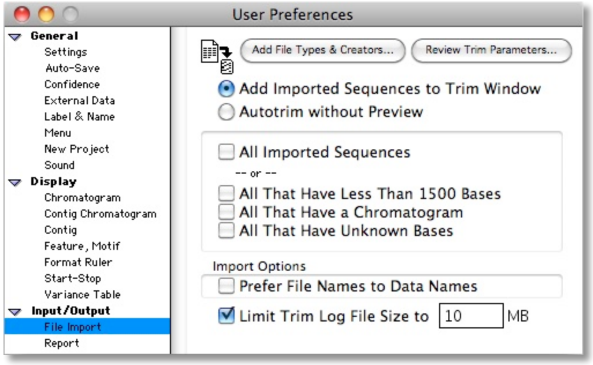
Use the checkboxes grouped under **Add Imported Sequences To Trim Window** to specify which sequences **Sequencher** should examine for poor quality ends and then trim.

You can also specify how you want to resolve conflicts between the names of imported files and names stored with particular sequences. If you click on the **Prefer File Names to Data Names** checkbox, **Sequencher** uses the file name.

Clicking on the **Review Trim Parameters…** button displays the **Ends Trimming** dialog in which you can set trim criteria. (See Chapter 6, “Preparing your Data for Assembly,” for details.)

Information about sequences which have been trimmed is recorded in a special file. This happens automatically. If you want to prevent the file from growing too large, then click on the **Limit Trim Log File Size to** checkbox. Then enter a value in the input field.

Figure 23-21 File Import preference settings



● Add Imported Sequences to Trim Window
○ Autotrim without Preview

□ All Imported Sequences
-- or --
□ All That Have Less Than 1500 Bases
□ All That Have a Chromatogram
□ All That Have Unknown Bases

Import Options
□ Prefer File Names to Data Names
☑ Limit Trim Log File Size to 10 MB

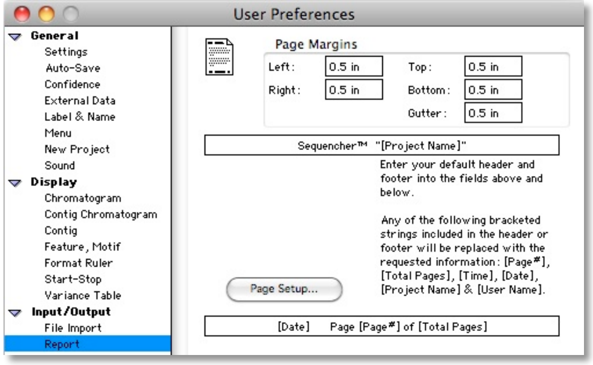
REPORT

The **Report** preferences control the default page setup for report printouts. Use the fields in the **Page Margins** groupbox to set the margins and gutter. Click on **Page Setup…** to open the printer **Page Setup** dialog. The page setup settings will apply to all windows you open thereafter.

You can specify headers and footers on reports generated from any window in which you are working. Any text you type into the field above the page layout will appear in the header and text you type in the field below the layout will appear in the footer.

For variable elements (date, time, page, and total number of pages), enclose the word— typed with an initial capital—in square brackets, as shown in Figure 23-22.

Figure 23-22 Report Margin preference settings



Page Margins
Left: 0.5 in Top: 0.5 in
Right: 0.5 in Bottom: 0.5 in
Gutter: 0.5 in

Sequencher™ "[Project Name]"

Enter your default header and footer into the fields above and below.

Any of the following bracketed strings included in the header or footer will be replaced with the requested information: [Page#], [Total Pages], [Time], [Date], [Project Name] & [User Name].

[Date] Page [Page#] of [Total Pages]

24. FORENSIC FEATURES

If you have the forensic version of **Sequencher**, you have access to special functions for DNA-based identification. In this chapter, we describe how you can validate mtDNA profiles, create reports, and export your results.

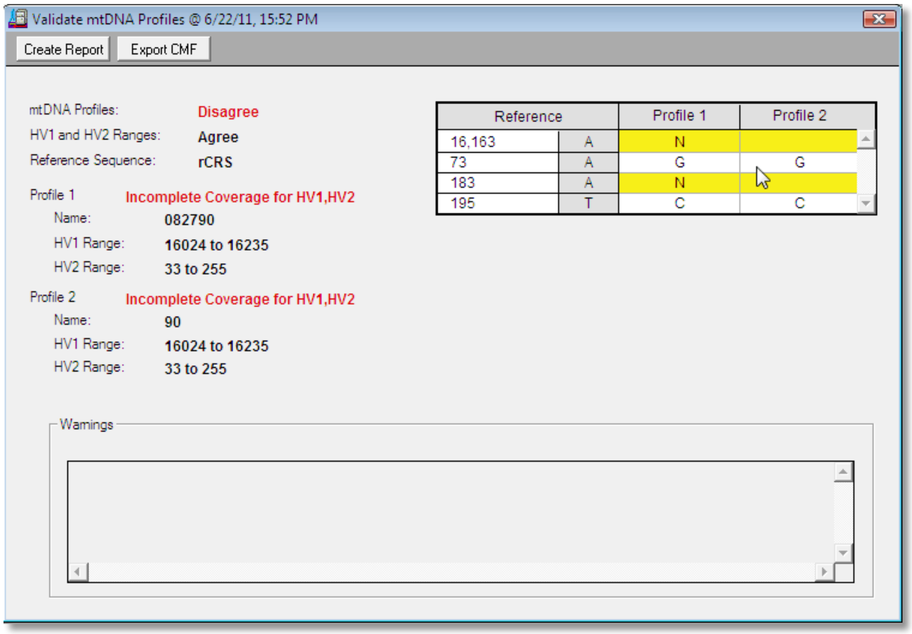
WORKING WITH MTDNA PROFILES

VALIDATING MITOCHONDRIAL DNA PROFILES

The **Validate mtDNA Profiles** command allows you to compare the results of separate analysts. Select two contigs that have been assembled to the same Reference Sequence. Go to the **Contig** menu and click on the **Validate mtDNA Profiles** item. This command creates a report in a new window. The title bar includes a date and time stamp as shown in Figure 24-1 below. The report includes the name of the Reference Sequence used, the name of the contigs being examined, and the HV1 and HV2 ranges. Underneath the main body of the report is a window pane that displays pertinent warnings.

In the report shown in Figure 24-1, the two mtDNA profiles disagree. The bases in conflict are highlighted in yellow. Also, there are positions where there is data for one mtDNA profile but not for the other. The base position is distinguished by a yellow highlight, and the cell with the missing data has a pink X though it. The cells are unshaded where the two profiles both agree and confirm each other.

Figure 24-1 The Validate mtDNA profiles window



Validate mtDNA Profiles @ 6/22/11, 15:52 PM

mtDNA Profiles: Disagree
HV1 and HV2 Ranges: Agree
Reference Sequence: rCRS

Profile 1 Incomplete Coverage for HV1,HV2
Name: 082790
HV1 Range: 16024 to 16235
HV2 Range: 33 to 255

Profile 2 Incomplete Coverage for HV1,HV2
Name: 90
HV1 Range: 16024 to 16235
HV2 Range: 33 to 255

Reference		Profile 1	Profile 2
16,163	A	N	
73	A	G	G
183	A	N	
195	T	C	C

CREATE REPORTS

At the top of the **Validate mtDNA Profiles** window is a button bar. Click on the **Create Report** button. A new window called **Analyst Report** opens. This window contains a version of the **Validate mtDNA** report that you can send to a printer by clicking on the **Print** button. You may also save this report by clicking on the **Save** button.

Note: The report includes spaces for the analysts' and reviewers' signatures.

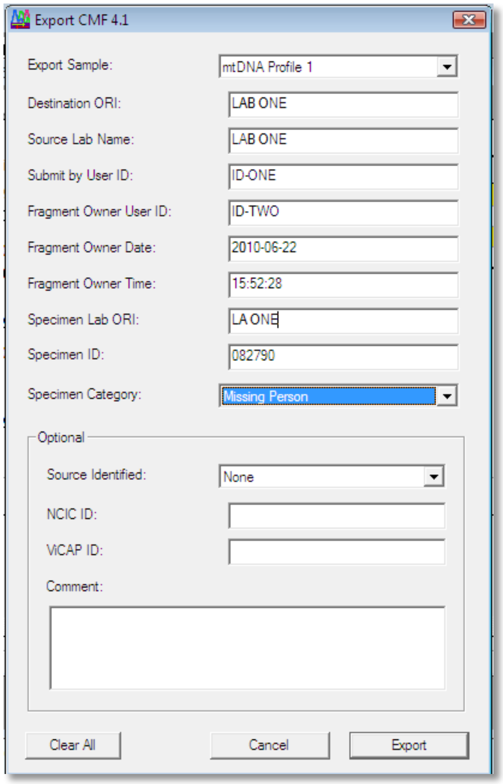
EXPORT CMF

You can create a CMF Report by clicking on the **Export CMF** button. A new window called **Export CMF** will open. Choose the **Export Sample:** and **Specimen Category:** from the relevant drop-down menus. The **Fragment Owner Date**, **Fragment Owner Time**, and **Specimen ID** fields are filled in automatically. You will need to enter information into the remaining text fields.

There are a number of optional fields. The **Source Identified** field is a drop-down menu. There are two text fields for **NCIC ID** and **ViCAP ID**. There is also a free text **Comment** field. When you are finished filling in the form, click on the **Export** button. If you have made an error, you can click on the **Clear All** button to reset the form, or you can click on the **Cancel** button to return to the **Validate mtDNA Profiles** window.

Note: If you move your cursor to pause over a field, you will see a brief description summarizing the information required for that field.

Figure 24-2 Export CMF report



Export CMF 4.1

Export Sample:	mtDNA Profile 1
Destination ORI:	LAB ONE
Source Lab Name:	LAB ONE
Submit by User ID:	ID-ONE
Fragment Owner User ID:	ID-TWO
Fragment Owner Date:	2010-06-22
Fragment Owner Time:	15:52:28
Specimen Lab ORI:	LA ONE
Specimen ID:	082790
Specimen Category:	Missing Person

Optional

Source Identified:	None
NCIC ID:	
ViCAP ID:	
Comment:	

FURTHER COMMANDS OF INTEREST

NEW PROJECT FROM TEMPLATE

The **New Project From Template** command allows you to open a new project containing all the sequences, settings, and preferences associated with your chosen template. Go to the **File** menu and click on **New Project From Template**. Select a template from the submenu. A new blank **Project Window** will open.

Note: A template called **rCRS** is provided. This template contains the Revised Cambridge Reference Sequence and is the default template for this command.

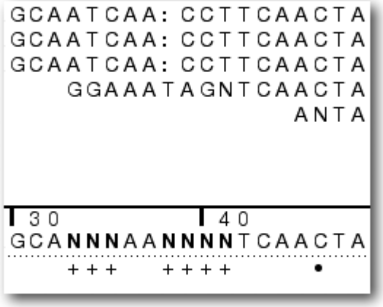
For more information on Templates, see Chapter 3 “The Project Window.” There is a tutorial on analyzing mtDNA with **Sequencher** in the **Tutorials** folder.

CONSENSUS TO FORENSIC STANDARDS

This calculation is based on the coverage of sequences at a given position and is taken from work at the U.S. Armed Forces DNA Identification Laboratory (AFDIL) in Rockville, Maryland.

- All positions with coverage of only one sequence are marked as N (ambiguous).
- All IUPAC codes apart from ACG and T are treated as N.
- If any positions disagree where the coverage is between two and four, this position is marked as N. If the coverage is between five and seven, any position with a disagreement will be marked as ambiguous.
- If there are two or more disagreeing bases at a position, this will be called an N regardless of the coverage.

Figure 24-3 Consensus to Forensic Standards example



```
GCAATCAA:CCTTCAACTA
GCAATCAA:CCTTCAACTA
GCAATCAA:CCTTCAACTA
   GGAAATAGNTCAACTA
               ANTA

30        40
GCANNNAANNNNTCAACTA
   +++   ++++    •
```

For more information on other types of consensus calculations, see Chapter 11, “The Contig Editor”.

DISPLAY SECONDARY PEAK

You can identify heterozygotes by identifying the second-highest peak beneath the primary peak. If you have a particular candidate for a heterozygote, first open the **Chromatogram** for the individual sequence and then click on the base call(s).

Go to the **Sequence** menu and choose **Call Secondary Peaks…**. A dialog lets you specify how the secondary peak should be called.

The slider lets you specify how significant the second-highest peak must be to generate a change.

If you want Ns to be replaced, then select the **Allow Ns to be replaced** checkbox.

If you’ve already edited bases by hand and don't want them changed, you must *deselect* the **Allow edited bases to be replaced** checkbox.

If you select **Only make changes that result in an ambiguity**, **Sequencher** may change an ambiguous base call to an A, C, G, or T if it does not meet the secondary peak criteria. If you prefer it to remain ambiguous, do not select this option.

If you wish to search just a range of bases, highlight them and check the **Search Selection Only** checkbox.

For more information, see Chapter 15 “Chromatograms”.

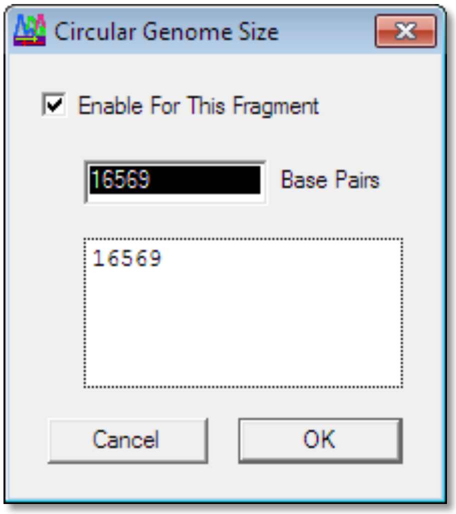
SET CIRCULAR GENOME SIZE

The **Set Circular Genome Size…** command allows you to set the number of bases in your DNA circle.

Select a sequence that has already been defined as a Reference Sequence. Then go to the **Sequence** menu and choose the **Set Circular Genome Size…** command. The dialog below appears. Enable the circular number by clicking on the **Enable For This Fragment** checkbox. If you have enabled the Cambridge Reference Sequence as your Reference Sequence, the number of base pairs in your sequence will automatically be set to 16,569. Dismiss the dialog by clicking on the **OK** button.

Note: Circular numbering can only be used in conjunction with a Reference Sequence. You must designate your sequence as a Reference Sequence before you can enable circular numbering.

Figure 24-4 The Circular Genome Size dialog



Circular Genome Size

☑ Enable For This Fragment

16569 Base Pairs

16569

HOW TO MARK A SEQUENCE AS A REFERENCE SEQUENCE

To designate a sequence as a Reference Sequence, select the sequence icon in the **Project Window** and, under the **Sequence** menu, select the **Reference Sequence** command**.** From now on, that sequence icon will include a small letter "R" to remind you that it has been marked as a Reference.

For more information on Reference Sequence properties and how to work with a Reference Sequence, see Chapter 7 “The Reference Sequence”.

25. SEQUENCHER CONNECTIONS

Sequencher Connections heralds an entirely new way of working with your data. Previously, you were able to work within **Sequencher** using its internal algorithms, then we introduced Plugins. These allowed you to choose and work with NGS algorithms such as **Maq**, **GSNAP**, **BWA**, and **Velvet** from **Sequencher**. These algorithms were still located on your local computer. Now you are able to use remote services such as **BLAST** and **Primer-BLAST** or local algorithms such as **MUSCLE** or **Local-BLAST** using the **Connections** extensible system.

In this chapter, you will learn about **Sequencher** Connections and how to apply the algorithms and services. You will also learn how to apply these to single or multiple sequences using Blast and Primer-BLAST. In addition, you will learn how to save primers to your project. You will also learn how to work with a group of Sequences using the **MUSCLE** algorithm. Finally you will learn how to view previous sessions.

SESSIONS

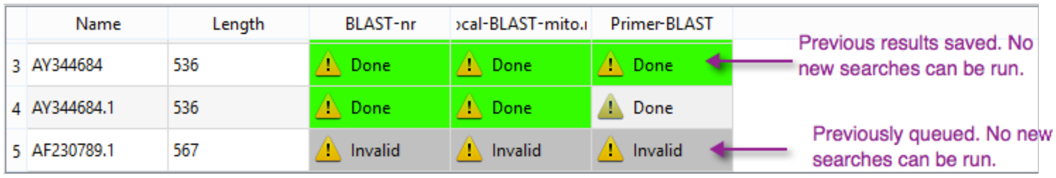
A session is a coordinated set of data and the analyses that operate on that data. Grouped sessions operate on sequences as a whole, for example when they are multiply aligned. In an Individual session, each analysis operates on a single sequence so that many single sequences can be analyzed in parallel.

Sequencher remembers sessions you have created and you can open them again and review results you obtained previously. Or run new analyses on the data in the session.

The data is a snapshot of its state as it was when added to the session. If there is a problem with data in an existing session, for example it has been corrupted, any existing results for that data will not be affected. However, you will not be able perform any new analyses and **Sequencher** will warn you that the data is invalid. This warning is only displayed for the duration of the session.

Note: Since the data in a session is a snapshot, you may edit the sequence copy in your **Sequencher** project without affecting the copy in Connections. Additionally, if data in the session becomes invalid, the data in your project is unharmed and you can use this in a new session.

Figure 25-1 Invalid Data warning



	Name	Length	BLAST-nr	›cal-BLAST-mito.ı	Primer-BLAST
3	AY344684	536	⚠ Done	⚠ Done	⚠ Done
4	AY344684.1	536	⚠ Done	⚠ Done	⚠ Done
5	AF230789.1	567	⚠ Invalid	⚠ Invalid	⚠ Invalid

Previous results saved. No new searches can be run.

Previously queued. No new searches can be run.

CHANNELS

Each analysis takes place in what is termed a **Channel** and may act on a single sequence or multiple sequences. Since you can modify analyses with different options, you can set up a series of Channels with each representing a new analysis or collection of settings. In the case of BLAST for example, this means that you can set up several searches against different databases or use different search settings against a single database. This will let you compare searches against say, the human EST database, the mouse EST database, and dbsts, while comparing results between blastn and megablast.

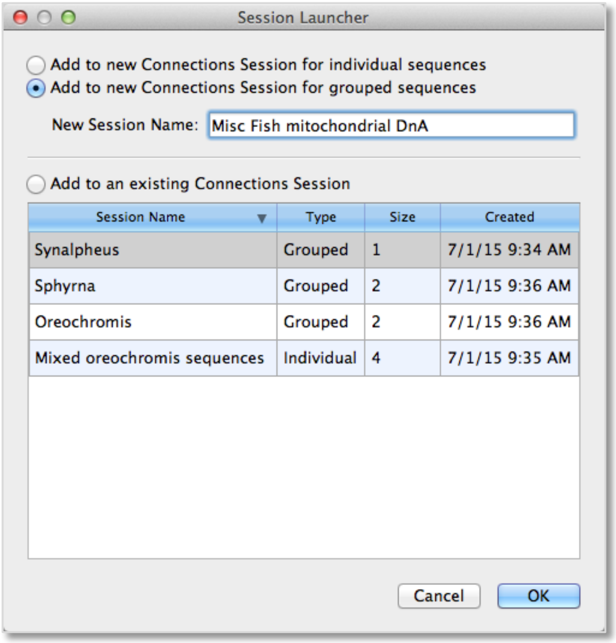
INVOKING A SEQUENCHER CONNECTIONS SESSION FOR THE FIRST TIME

In order to launch **Sequencher** Connections, you will need to select one or more items from the **Project Window**. You may send any number of sequences or even contigs to **Sequencher** Connections. If you want to send large numbers of sequences, be aware that the Blast server imposes restrictions to ensure that everyone may use the service without problems. **Sequencher** Connections has been written to take these regulations into account.

Go to the **Window** menu and choose **Add to Connections Session...**

You will see the **Sequencher** Connections **Session Launcher** dialog. Decide whether you want to add your sequences to a session for individual sequences (BLAST, Primer-BLAST, or Local-BLAST if you have installed it on your computer) or whether you want to add your sequences to a session for grouped sequences (MUSCLE). Click on the radio button next to the session type you want to launch. **Sequencher** gives your session a name automatically. You will probably want to replace this with something more meaningful. Do this by typing the desired name into the **New Session Name:** input field. Finally, click on the **OK** button.

Figure 25-2 The Session Launcher



○ Add to new Connections Session for individual sequences
● Add to new Connections Session for grouped sequences

New Session Name: Misc Fish mitochondrial DnA

○ Add to an existing Connections Session

Session Name	Type	Size	Created
Synalpheus	Grouped	1	7/1/15 9:34 AM
Sphyrna	Grouped	2	7/1/15 9:36 AM
Oreochromis	Grouped	2	7/1/15 9:36 AM
Mixed oreochromis sequences	Individual	4	7/1/15 9:35 AM

The session now opens in a new window. The upper half of the window contains a grid with the data you have selected and two default analysis Channels. Each row represents a sequence you are submitting for analysis and each column represents a "Channel" (described above) that handles a different kind of analysis. The default Channels are BLAST and Primer-BLAST.

The lower half of the window is where you see the results of those analyses in a series of tabs. (Hint: If you hover your mouse just above the tabs in the data section, your cursor will turn to a splitter allowing you to adjust the size of the two parts of the window.)

Initially your sequences will show in the grid as Queued. Click on one of the cells containing the word Queued. The default tab is the **Web View**, which opens to the Gene Codes website, but if you click on the **Sequence**, tab you will see the sequence you are about to submit. The **Text** and the **XML** views are blank at this time since there are no results to display.

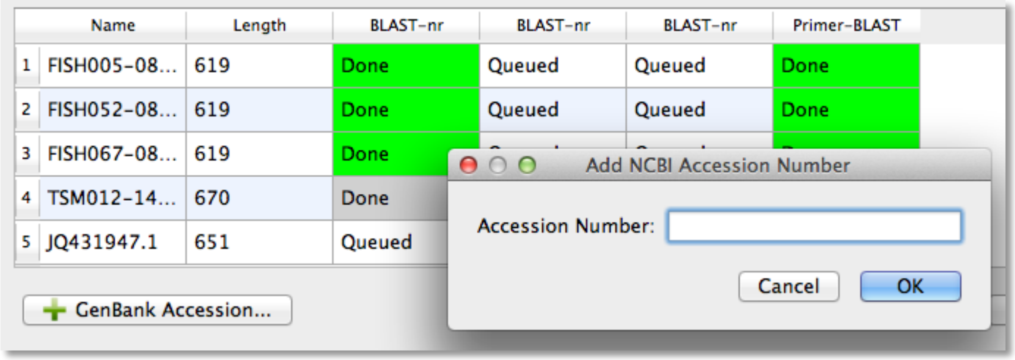
USING GENBANK ACCESSION NUMBERS

Once you have some sequences in the **Sequencher** Connections dialog, you may add further sequences if you know their GenBank accession numbers. The accession number is a unique identifier for a sequence record and is usually formed by two text characters followed by a series of numbers. Accession numbers don't change, but if the record changes, you may see a version number added as in the example below. (The GI number will change.) Connections will only allow you to add a sequence once to a specific session. If you need to have another copy of a sequence, you will need to add it to a different session.

Click on the GenBank Accession... button and type the accession number into the **Add NCBI Accession Number** dialog. To dismiss the dialog, click on the **OK** button.

In the image below, the sequence with accession number JQ431947.1 has already been added (row 5 in the table) and the user is about to add another accession number. If you click in any cell of an Accession Number sequence row, a new tab labeled GenBank appears, this tab contains the feature table for the sequence.

Figure 25-3 Using the Accession Number dialog to add sequences to a session



	Name	Length	BLAST-nr	BLAST-nr	BLAST-nr	Primer-BLAST
1	FISH005-08...	619	Done	Queued	Queued	Done
2	FISH052-08...	619	Done	Queued	Queued	Done
3	FISH067-08...	619	Done			
4	TSM012-14...	670	Done			
5	JQ431947.1	651	Queued			

Add NCBI Accession Number

Accession Number:

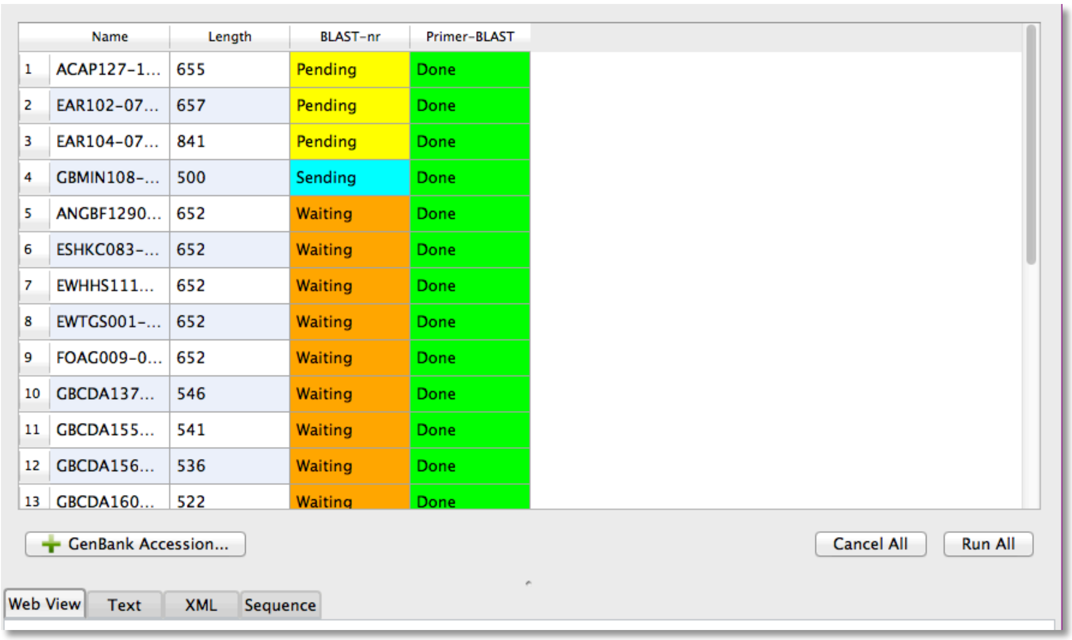
Sequencher Connections will fetch the record and add the sequence to the session window. You can also remove a sequence from **Sequencher** Connections by right-clicking on the number of the row.

Note: The outline colors for individual and group input Connections windows are different to help you distinguish sessions when you are working with several of them.

RUNNING SEQUENCHER CONNECTIONS

There are a number of different ways to send your data for analysis. Click on the **Run All** button. It is between the two panes, to the right of the dialog. This will send all your data to all of the Channels you have set up. Alternatively, you can send one sequence by right-clicking on the number to the left of the sequence name in the grid and choosing **Run on Each Channel** from the contextual menu. You can also send all sequences for a specific Channel by right-clicking on the column header of the Channel and choosing **Run on Each Sequence**. And finally, you can also send a single sequence for a single Channel by right-clicking in a single Channel cell and choosing **Run**.

Figure 25-4 Session status messages



	Name	Length	BLAST-nr	Primer-BLAST
1	ACAP127-1...	655	Pending	Done
2	EAR102-07...	657	Pending	Done
3	EAR104-07...	841	Pending	Done
4	GBMIN108-...	500	Sending	Done
5	ANGBF1290...	652	Waiting	Done
6	ESHKC083-...	652	Waiting	Done
7	EWHHS111...	652	Waiting	Done
8	EWTGS001-...	652	Waiting	Done
9	FOAG009-0...	652	Waiting	Done
10	GBCDA137...	546	Waiting	Done
11	GBCDA155...	541	Waiting	Done
12	GBCDA156...	536	Waiting	Done
13	GBCDA160...	522	Waiting	Done

Once you have chosen **Run All** or any of the other **Run** commands, your data is submitted to your chosen resource(s) and you can monitor the progress. The cells will change status and color as the analysis progresses, though some states change so quickly you may not see the change. The cells change status through Queued (white), Waiting (orange; data is in the queue before being sent to the connected resource), Sending (turquoise; data being sent to the connected resource), Pending (yellow; waiting for the resource to send results back), and Done (green). There is a further Done status for previously saved results. Finally, there is a Failed (red) status that indicates that no results were received.

You may also re-run your data by pressing the **Run All** button a second time. This will delete your existing results and run all of the analyses again. If you cancel before the new results appear, your previous results will be restored.

VIEWING RESULTS WITH SEQUENCHER CONNECTIONS

As soon as the cells change to Done or if you are working with older results and the status is Done, your results are ready for viewing. You can view the results as all the Channels for an individual sequence reach this state or you can wait till all the results have been received.

There are multiple different views depending on the Channel. The contents of the **Web View** changes from Connections-related text to display the BLAST or Primer-BLAST results view. This view contains the graphical overview of BLAST or Primer-BLAST results, the one-line descriptions of the hits, and the text alignments of query versus hit. The **Text** view displays the one-line descriptions and text alignments only. The **XML** view shows the alignments but with XML formatting. The **Sequence** view displays the query sequence.

When you add a sequence to the Connections Session by clicking on the **+ GenBank Accession…** button, a **GenBank** tab will appear. This view contains the Feature table for that sequence.

If you have created a Primer-BLAST analysis, then you will see a **Primer Picker** tab when you click on any cell containing the status of Done or Done. Finally, there is a **Schematic** view that allows you to compare analyses with different parameters including Primer-BLAST so you can view results from BLAST and Primer-BLAST at the same time within the Schematic. In order to view any particular tab, simply click on it.

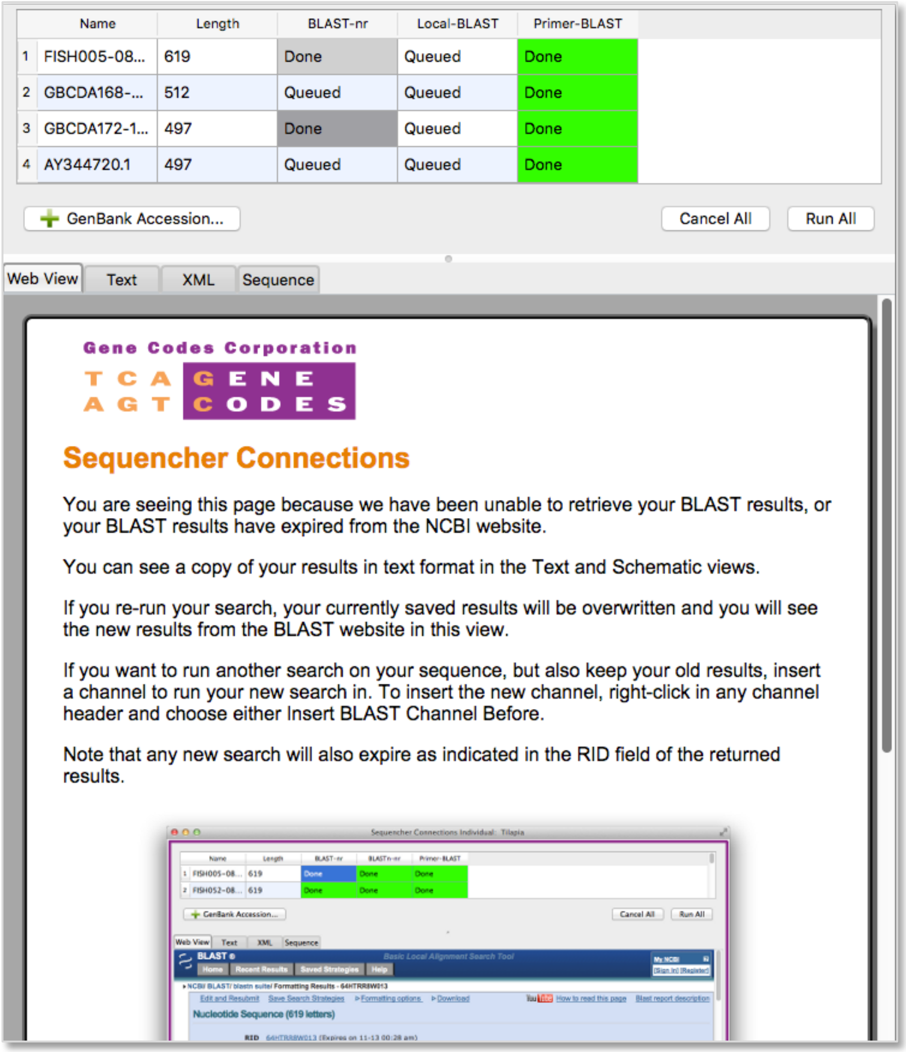
Once you have opened a tab, you may also save the results to a file by using the right-click contextual menu.

BLAST and Primer-BLAST results are available on the NCBI website for a maximum of 36 hours. You can check exactly when your BLAST results will expire by looking next to the RID in the **Web View**. Unfortunately, NCBI does not provide similar information for Primer-BLAST. However, once your results have expired from the NCBI website, Connections saves your results and displays a message to that effect in the **Web View**, which normally displays the NCBI website view of your results. You can still view the alignments or the primers. To view your results, click on the **Text** tab. You can also view the alignments by right-clicking on the name of the sequence and choosing **Show Schematic** from the context-sensitive menu.

ADDING OR REMOVING CHANNELS

In a **Sequencher** Connections session, a column represents a Channel (analysis) and the data to be analyzed is represented by a row. The simplest type of session would contain a single column and a single row. More complex sessions might contain many rows (sequences) and columns (Channels).

Figure 25-5 BLAST results expired from NCBI



	Name	Length	BLAST-nr	Local-BLAST	Primer-BLAST
1	FISH005-08...	619	Done	Queued	Done
2	GBCDA168-...	512	Queued	Queued	Done
3	GBCDA172-1...	497	Done	Queued	Done
4	AY344720.1	497	Queued	Queued	Done

Sequencher Connections

You are seeing this page because we have been unable to retrieve your BLAST results, or your BLAST results have expired from the NCBI website.

You can see a copy of your results in text format in the Text and Schematic views.

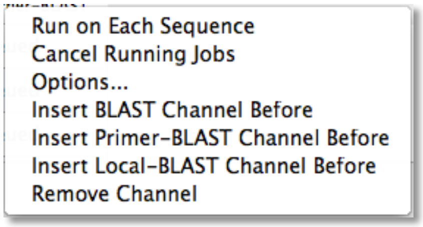
If you re-run your search, your currently saved results will be overwritten and you will see the new results from the BLAST website in this view.

If you want to run another search on your sequence, but also keep your old results, insert a channel to run your new search in. To insert the new channel, right-click in any channel header and choose either Insert BLAST Channel Before.

Note that any new search will also expire as indicated in the RID field of the returned results.

You can add more Channels by right-clicking in an existing Channel header row. A contextual menu appears that has menu items for adding new BLAST Channels, Primer-BLAST Channels, and Local-BLAST Channels. Each menu item adds the Channel before the Channel you have selected. If you want to remove a Channel, right-click in its column header and choose the **Remove Channel** menu item.

Figure 25-6 Right-click Channel menu



Run on Each Sequence
Cancel Running Jobs
Options...
Insert BLAST Channel Before
Insert Primer-BLAST Channel Before
Insert Local-BLAST Channel Before
Remove Channel

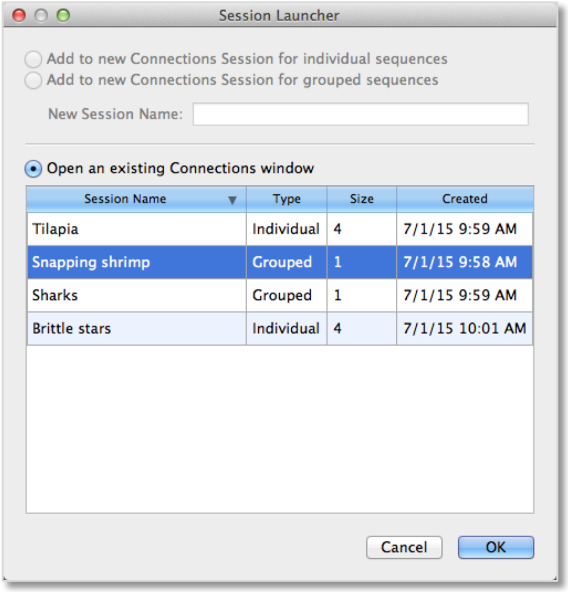
Once you have added a new Channel, you can create a new series of settings using **Options…** in the contextual menu (right-click in a Channel column header). You can see an example using BLAST in the next section.

You can also cancel running Channels by right-clicking on their headers and choosing the **Cancel Running Jobs** menu item.

VIEWING SAVED SESSIONS

Your sessions are saved with your project. This means that you can view your results again and again. In order to see a saved session, go to the **Window** menu and choose **Open Existing Connections Session…**. The **Session Launcher** dialog appears where you can choose a session from the list.

Figure 25-7 Viewing saved sessions list in the Session Launcher



○ Add to new Connections Session for individual sequences
○ Add to new Connections Session for grouped sequences

New Session Name:

● Open an existing Connections window

Session Name ▼	Type	Size	Created
Tilapia	Individual	4	7/1/15 9:59 AM
Snapping shrimp	Grouped	1	7/1/15 9:58 AM
Sharks	Grouped	1	7/1/15 9:59 AM
Brittle stars	Individual	4	7/1/15 10:01 AM

Clicking on a header in the **Session Launcher** dialog sorts the sessions by that column.

Once the session has opened, you will be able to review existing results in the **Web View**. Although results do expire from the NCBI website after 36 hours, for BLAST and Primer-BLAST, you can still see your results in the **Text** tab. You can also view the alignments by right-clicking on the name of the sequence and choosing **Show Schematic** from the context-sensitive menu.

DELETING SAVED SESSIONS

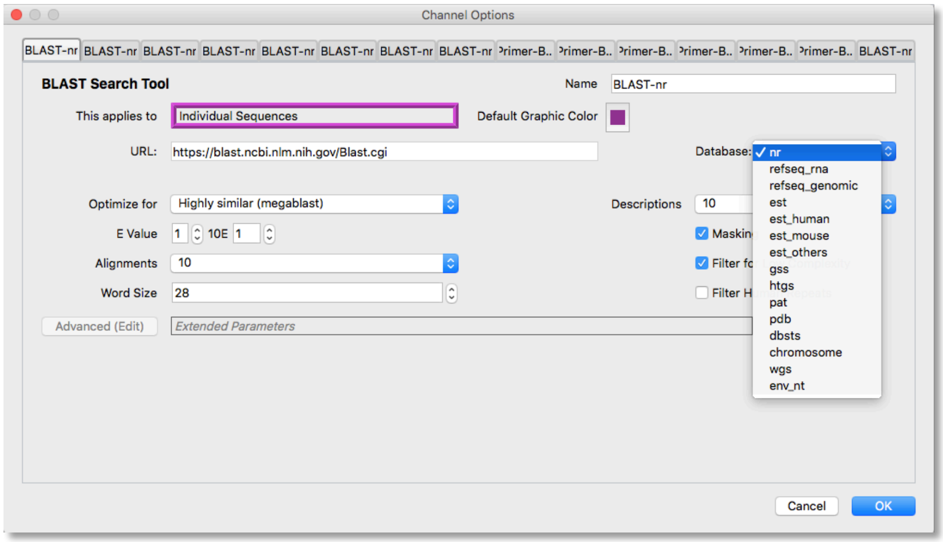
You can delete an existing Connections session from your project. To do this, go to **Window > Delete Existing Connections Session…**. A dialog appears listing all your current sessions. Select or multi-select the sessions you want to remove and then click on the **Delete Selected Session(s)** button. You cannot undo this action once the sessions have been deleted.

BLAST

For each new Channel you add, you will see a new tab in the **Channel Options** dialog. There are a number of options that can be set for BLAST. Right-click on a column header in the **Sequencher Connections** dialog and choose **Options...** from the contextual menu. A **Channel Options** dialog appears in which you can set the following: algorithm (blastn or megablast), number of Alignments (drop-down menu), Databases (drop-down menu), number of Descriptions (drop-down menu), the E (Expectation) Value, Filters (checkboxes), and Word size.

If you choose to change Alignments and Descriptions, you set an integer (10, 100, 250 for example). For the E or Expectation number, you set a number above 0. You cannot set a negative value but you can set a very small number such as 0.0001. The E value describes the number of hits you would expect to see by chance. The lower the E value, the more significant the match will be. If the E value is increased (larger than the default value of 10), you should get a larger list of hits but the hits have a less significant score.

Figure 25-8 BLAST Channel options



BLAST Search Tool

Name: BLAST-nr

This applies to: Individual Sequences

Default Graphic Color: 🟪

URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Database: nr

Optimize for: Highly similar (megablast)

Descriptions: 10

E Value: 1 10E 1

☑ Maskin

Alignments: 10

☑ Filter fo

Word Size: 28

☐ Filter H

Extended Parameters

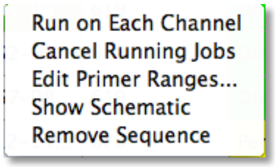
If you want to change the database you wish to search, then you can choose a different one from a series in the Database drop-down menu. The database you choose is also used to name the Channel. You can apply filters to your query sequence, for example low complexity or repeat elements, by clicking one or more of the checkboxes. If you include Masking as well as a Filter, the mask only applies to the look-up phase of the BLAST search. It does not apply to the extension phase of the BLAST algorithm.

PRIMER-BLAST

The Primer-BLAST resource at NCBI uses the well-known Primer3 program to design primers. It has an added advantage in that the BLAST algorithm is then used to screen the predicted primers against a user-selected database website to see whether there could be any potential for unintended amplification.

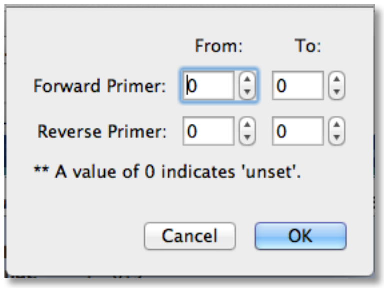
You have the option of sending the entire sequence for analysis or choosing a subset. Right-click on the number to the left of the sequence whose range you want to specify. **Choose Edit Primer Ranges…** from the contextual menu.

Figure 25-9 Right-click Sequence menu



You are presented with a dialog where you can specify From and To ranges for the forward and reverse primers. This is so you can specify when you want primers to be located on specific sites. For example, if you specified a Forward Primer range From: 200, To: 500, then your forward primers would be located within these coordinates. You can specify just one primer if you prefer (e.g. just the forward primer).

Figure 25-10 Edit Primer Ranges



	From:	To:
Forward Primer:	0	0
Reverse Primer:	0	0

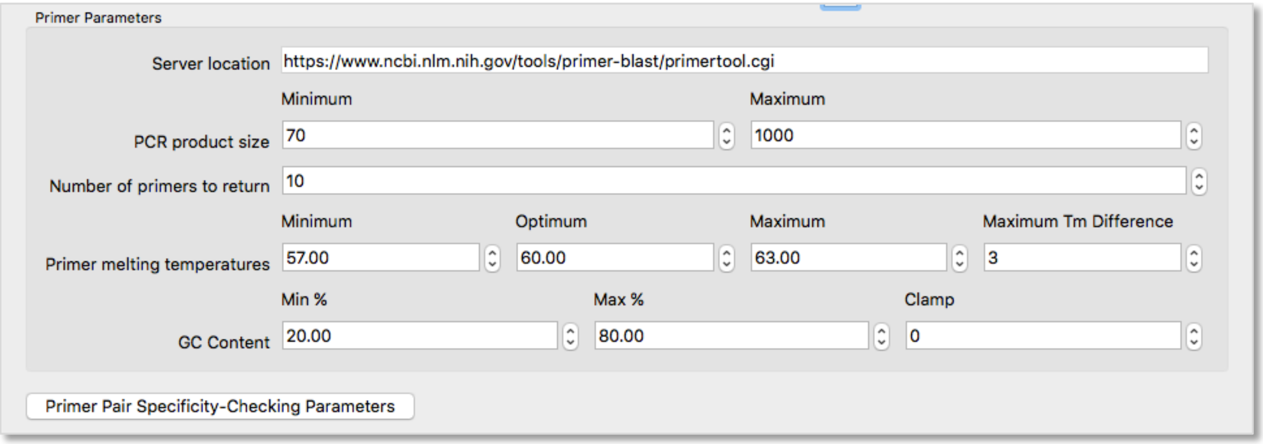
** A value of 0 indicates 'unset'.

***Note**:* The position range of the primers must not overlap.

PRIMER-BLAST OPTIONS

There are a number of options you can set besides the Primer Ranges in order to refine your primers further. Right-click on the header for the Primer-BLAST Channel and select the **Options…** menu item. The **Channel Options** dialog opens at the **Primer-BLAST** tab. The options are divided into the Primer **Parameters** and the Specificity Checking Parameters. You can ask Primer-BLAST to send back more than 10 results. You can also adjust the primer melting temperature and specify the %GC content.

Figure 25-11 Primer-BLAST primer parameters



Primer Parameters

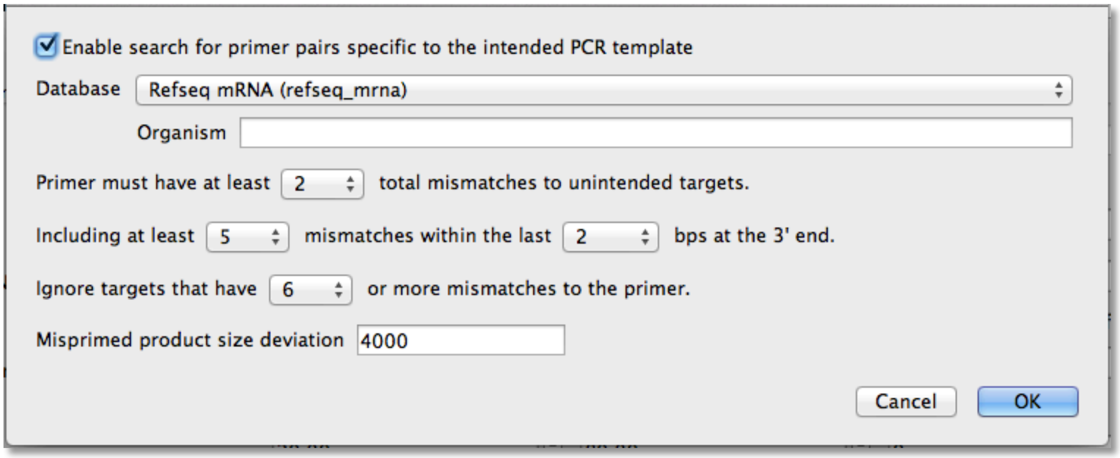
Server location: https://www.ncbi.nlm.nih.gov/tools/primer-blast/primertool.cgi

	Minimum	Optimum	Maximum	Maximum Tm Difference
PCR product size	70		1000	
Number of primers to return	10			
Primer melting temperatures	57.00	60.00	63.00	3

	Min %	Max %	Clamp
GC Content	20.00	80.00	0

To enable specificity-checking, click on the **Primer Pair Specificity-Checking Parameters** button. In the dialog that opens, click on the checkbox next to **Enable search for primer pairs specific to the intended PCR template**. Once you have made changes to the settings, click on the **OK** button to dismiss the dialog.

Figure 25-12 Primer-BLAST primer specificity parameters



☑ Enable search for primer pairs specific to the intended PCR template

Database: Refseq mRNA (refseq_mrna)

Organism:

Primer must have at least 2 total mismatches to unintended targets.

Including at least 5 mismatches within the last 2 bps at the 3' end.

Ignore targets that have 6 or more mismatches to the primer.

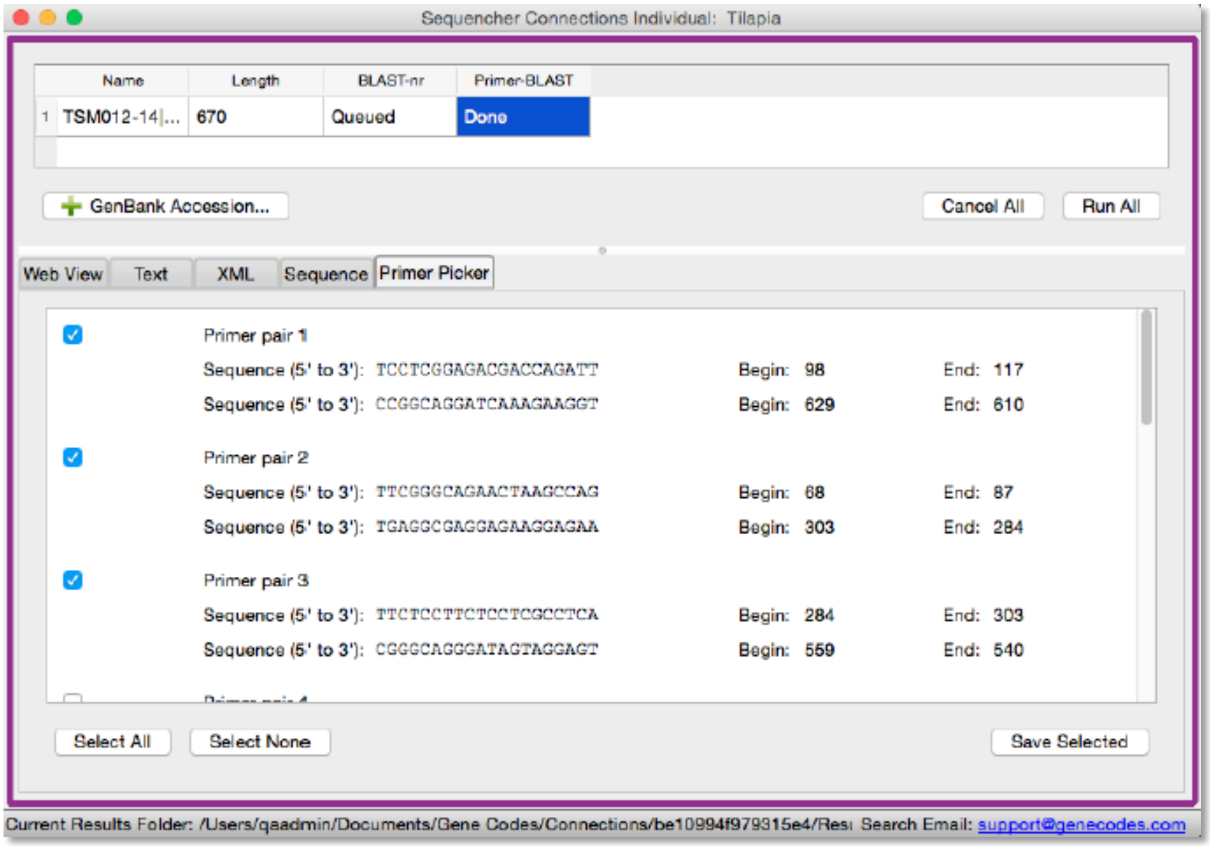
Misprimed product size deviation 4000

The results are displayed when the status of the job is shown as Done. Right-click on the name of the sequence and choose **Show Schematic**.

SAVING PRIMERS TO SEQUENCHER PROJECT

Once the Primer-BLAST analysis has completed, you can review the proposed primers by clicking on the **Text** tab. Here you will see detailed information on each primer pair. You will notice that a new tab called **Primer Picker** appears whenever you click on a cell in a Primer-BLAST analysis column. The **Primer Picker** contains a subset of the information contained in the **Text** tab. The information consists of a checkbox, the primer pair sequences, the start point for the primer on the original sequence, and the end point for the primer.

Figure 25-13 The Primer Picker tab



	Name	Length	BLAST-nr	Primer-BLAST
1	TSM012-14|...	670	Queued	Done

☑ Primer pair 1

Sequence (5' to 3'):	TCCTCGGAGACGACCAGATT	Begin: 98	End: 117
Sequence (5' to 3'):	CCGGCAGGATCAAAGAAGGT	Begin: 629	End: 610

☑ Primer pair 2

Sequence (5' to 3'):	TTCGGGCAGAACTAAGCCAG	Begin: 68	End: 87
Sequence (5' to 3'):	TGAGGCGAGGAGAAGGAGAA	Begin: 303	End: 284

☑ Primer pair 3

Sequence (5' to 3'):	TTCTCCTTCTCCTCGCCTCA	Begin: 284	End: 303
Sequence (5' to 3'):	CGGGCAGGGATAGTAGGAGT	Begin: 559	End: 540

□ Primer pair 4

Current Results Folder: /Users/qaadmin/Documents/Gene Codes/Connections/be10994f979315e4/Resı Search Email: support@genecodes.com

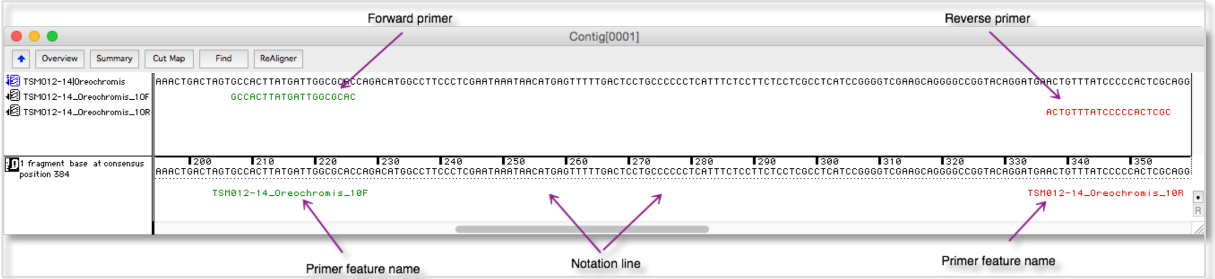
Click on the **Primer Picker** tab and select the primers you want to save to your project by clicking on the checkbox for each pair you want. If you want to save all the primer sequences to **Sequencher**, then click on the **Select All** button. If you do not like the selections you have made, you can clear all the checkmarks by clicking on the **Select None** button.

Once you have finalized your selection, click on the **Save Selected** button. This button will be greyed until you make your first primer pair selection.

Any checkboxes you select will remain checked until you close your session. The Primer-BLAST analysis itself will remain until you a) delete the Primer-BLAST column from the session, b) delete the sequence from the Session table, or c) delete the session itself. Any primers you have saved to your project are unaffected by any of these deletions.

Sequencher automatically assigns the Primer Feature key to the primer sequences. It also colors the bases of forward primers green and the reverse primers red. Once the primer sequences are in your project, you can assemble them to other sequences.

Figure 25-14 Primers aligned to a sequence

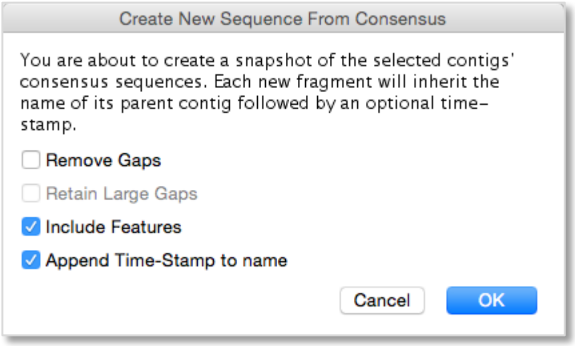


```
TSM012-14|Oreochromis              AAACTGACTAGTGCCACTTATGATTGGCGCACCAGACATGGCCTTCCCTCGAATAAATAACATGAGTTTTGACTCCTGCCCCCCTCATTTCTCCTTCTCCTCGCCTCATCCGGGGTCGAAGCAGGGGCCGGTACAGGATGAACTGTTTATCCCCCACTCGCAGG
TSM012-14_Oreochromis_10F                   GCCACTTATGATTGGCGCAC
TSM012-14_Oreochromis_10R                                                                                                                                                  ACTGTTTATCCCCCACTCGC

1 fragment base at consensus        200       210       220       230       240       250       260       270       280       290       300       310       320       330       340       350
position 384                       AAACTGACTAGTGCCACTTATGATTGGCGCACCAGACATGGCCTTCCCTCGAATAAATAACATGAGTTTTGACTCCTGCCCCCCTCATTTCTCCTTCTCCTCGCCTCATCCGGGGTCGAAGCAGGGGCCGGTACAGGATGAACTGTTTATCCCCCACTCGCAGG
                                       TSM012-14_Oreochromis_10F                                                                                                             TSM012-14_Oreochromis_10R
```

If you create a contig with primers as shown in Figure 25-14, you can create a consensus sequence that contains those primers.

Select a contig containing primers and then choose **Contig>Create New Seq From Consensus…**. A new dialog appears which allows you to set selected options. You can choose to include any GenBank style feature keys by selecting the **Include Features** checkbox.

Figure 25-15 Create New Sequence From Consensus options



Create New Sequence From Consensus

You are about to create a snapshot of the selected contigs' consensus sequences. Each new fragment will inherit the name of its parent contig followed by an optional time-stamp.

☐ Remove Gaps

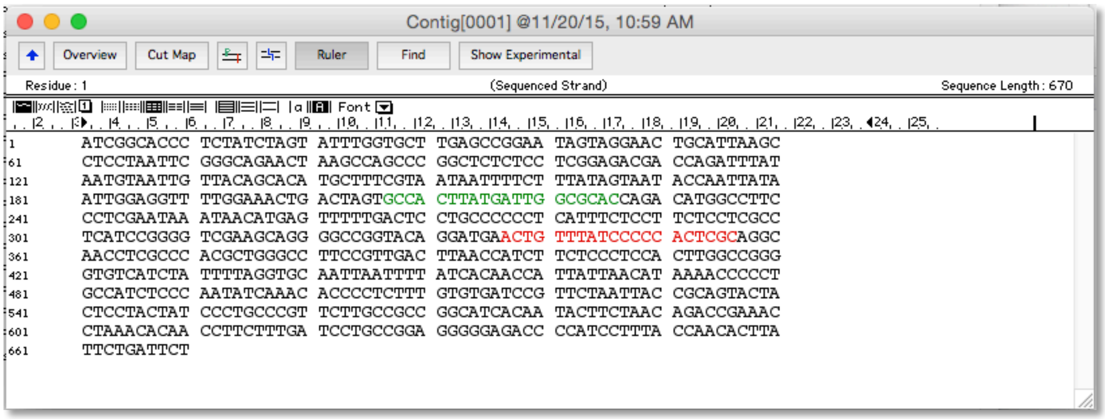
☐ Retain Large Gaps

☑ Include Features

☑ Append Time-Stamp to name

The consensus will contain the primers, but as features rather than individual sequences, as seen in the figure below.

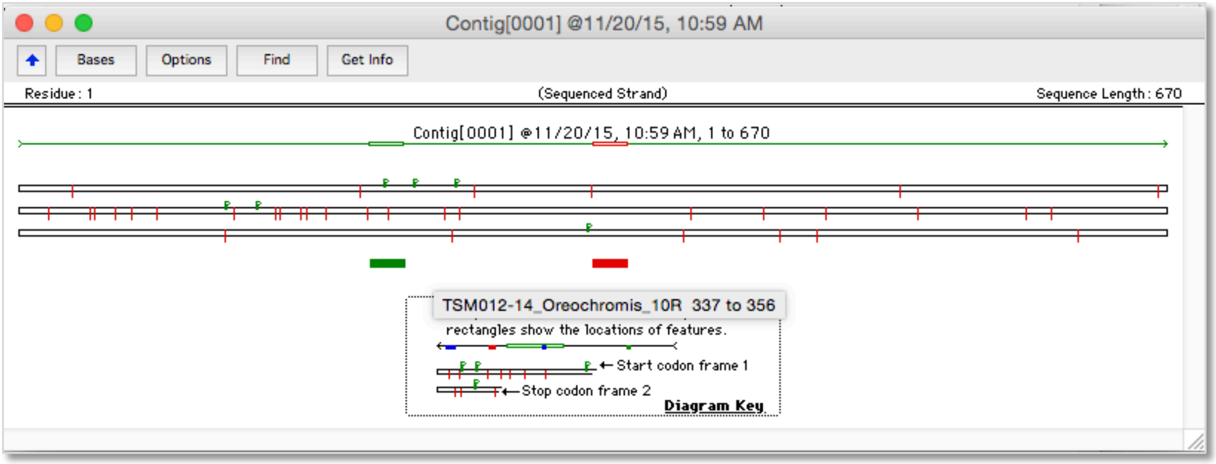
Figure 25-16 Consensus containing primers as features



```
Residue : 1                    (Sequenced Strand)                    Sequence Length : 670

1       ATCGGCACCC TCTATCTAGT ATTTGGTGCT TGAGCCGGAA TAGTAGGAAC TGCATTAAGC
61      CTCCTAATTC GGGCAGAACT AAGCCAGCCC GGCTCTCTCC TCGGAGACGA CCAGATTTAT
121     AATGTAATTG TTACAGCACA TGCTTTCGTA ATAATTTTCT TTATAGTAAT ACCAATTATA
181     ATTGGAGGTT TTGGAAACTG ACTAGTGCCA CTTATGATTG GCGCACCAGA CATGGCCTTC
241     CCTCGAATAA ATAACATGAG TTTTTGACTC CTGCCCCCCT CATTTCTCCT TCTCCTCGCC
301     TCATCCGGGG TCGAAGCAGG GGCCGGTACA GGATGAACTG TTTATCCCCC ACTCGCAGGC
361     AACCTCGCCC ACGCTGGGCC TTCCGTTGAC TTAACCATCT TCTCCCTCCA CTTGGCCGGG
421     GTGTCATCTA TTTTAGGTGC AATTAATTTT ATCACAACCA TTATTAACAT AAAACCCCCT
481     GCCATCTCCC AATATCAAAC ACCCCTCTTT GTGTGATCCG TTCTAATTAC CGCAGTACTA
541     CTCCTACTAT CCCTGCCCGT TCTTGCCGCC GGCATCACAA TACTTCTAAC AGACCGAAAC
601     CTAAACACAA CCTTCTTTGA TCCTGCCGGA GGGGGAGACC CCATCCTTTA CCAACACTTA
661     TTCTGATTCT
```

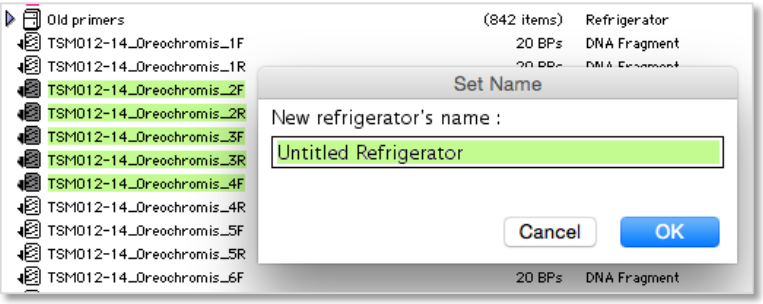
You will also see the primers as features in the **Sequence Overview**. To view the name of the feature, place the cursor over it for a few seconds. The name will appear in a tooltip.

Figure 25-17 Primer features in sequence Overview.


Contig[0001] @11/20/15, 10:59 AM
Bases
Options
Find
Get Info
Residue : 1
(Sequenced Strand)
Sequence Length : 670
Contig[0001] @11/20/15, 10:59 AM, 1 to 670
TSM012-14_Oreochromis_10R 337 to 356
rectangles show the locations of features.
Start codon frame 1
Stop codon frame 2
Diagram Key

If you want to keep a large number of primer sequences in your project, you can store them in a **Refrigerator**. This is a special folder whose contents will be saved with your project but which will not be assembled to any sequences unless first removed from the Refrigerator. You can create a Refrigerator by selecting a group of primers and choosing **Edit>Refrigerate**. You will be prompted to give your **Refrigerator** a name. Click on the **OK** button to dismiss this dialog.

Figure 25-18 Moving primers to a Refrigerator



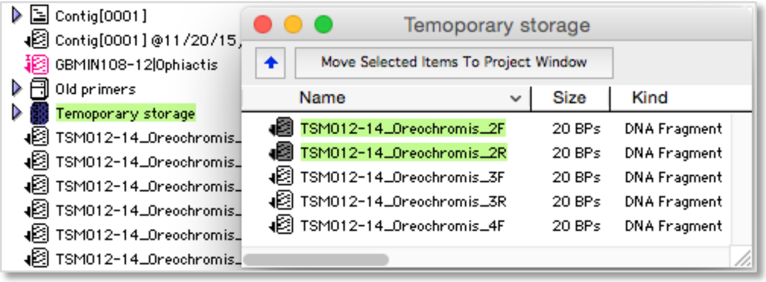
Old primers	(842 items)	Refrigerator
TSM012-14_Oreochromis_1F	20 BPs	DNA Fragment
TSM012-14_Oreochromis_1R	20 BPs	DNA Fragment
TSM012-14_Oreochromis_2F		
TSM012-14_Oreochromis_2R		
TSM012-14_Oreochromis_3F		
TSM012-14_Oreochromis_3R		
TSM012-14_Oreochromis_4F		
TSM012-14_Oreochromis_4R		
TSM012-14_Oreochromis_5F		
TSM012-14_Oreochromis_5R		
TSM012-14_Oreochromis_6F	20 BPs	DNA Fragment

Set Name

New refrigerator's name : Untitled Refrigerator

To view the contents of a Refrigerator, double-click on its icon. To move any primers to your project desktop, select the ones you want to move and click the **Move Selected Items To Project Window** button.

Figure 25-19 Moving primers to the Project Window



Contig[0001]
Contig[0001] @11/20/15,
GBMIN108-12|Ophiactis
Old primers
Temoporary storage
TSM012-14_Oreochromis_
TSM012-14_Oreochromis_
TSM012-14_Oreochromis_
TSM012-14_Oreochromis_
TSM012-14_Oreochromis_
TSM012-14_Oreochromis_

Name	Size	Kind
TSM012-14_Oreochromis_2F	20 BPs	DNA Fragment
TSM012-14_Oreochromis_2R	20 BPs	DNA Fragment
TSM012-14_Oreochromis_3F	20 BPs	DNA Fragment
TSM012-14_Oreochromis_3R	20 BPs	DNA Fragment
TSM012-14_Oreochromis_4F	20 BPs	DNA Fragment

LOCAL-BLAST - MAC

You need to have Local-BLAST installed on your computer in order to be able to utilize it. You can download the program for your computer from:

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/[5]

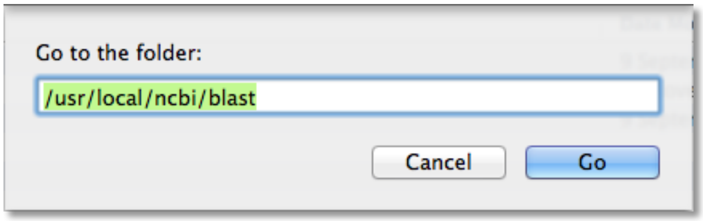
You will need to download the DMG. Once the download has completed, locate the DMG and double-click on it. The DMG will open and you will see the BLAST package installer inside. This has a file extension of pkg. Double-click on this and follow the prompts. The installer is large so be sure this is what you want to do before you start. Do not change the location of the installation.

Now you will need to create a database to query. You can use your own sequences or you can download parts of the databases that BLAST uses from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. These databases are pre-formatted and need no further attention. For more information, visit ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastdb.html.

You will need to place the databases in the correct location. From the Finder, visit the **Go** menu and choose **Go to Folder….** You will see a new dialog, type in **/usr/local/ncbi/blast** and click on the **Go** button.

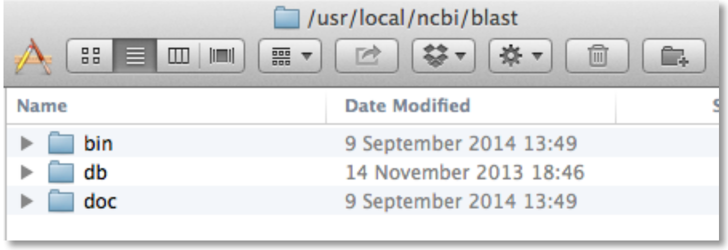
[5] At the time of writing this was version 2.2.40+

Figure 25-20 Go to the folder dialog showing blast path



You are now in the blast folder. Create a new folder here called **db**. This is where you place your database files.

Figure 25-21 Creating a db folder for BLAST



Name	Date Modified
bin	9 September 2014 13:49
db	14 November 2013 18:46
doc	9 September 2014 13:49

LOCAL-BLAST - WINDOWS

You need to have Local-BLAST installed on your computer in order to be able to utilize it. You can download the program for your computer from: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/[6]

You will need to download the EXE. Once the download has completed, locate the EXE and double-click on it and follow the prompts. The installer is large so be sure this is what you want to do before you start. Do not change the location of the installation.

Now you will need to create a database to query. You can use your own sequences or you can download parts of the databases that BLAST uses from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. These databases are pre-formatted and need no further attention. For more information, visit ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blastdb.html.

You will need to place the databases in the correct location. The location BLAST expects databases to be stored is pointed to by the environment variable “BLASTDB.” This environment variable can be viewed or set using the following procedure:

[6] At the time of writing this was version 2.2.28.

- Locate the **Control Panel** using the **Start** menu or by using **Search**.

- In the **Control Panel** window, doubl- click on the **System** icon.

- In the popup, click on the **Advanced** tab.

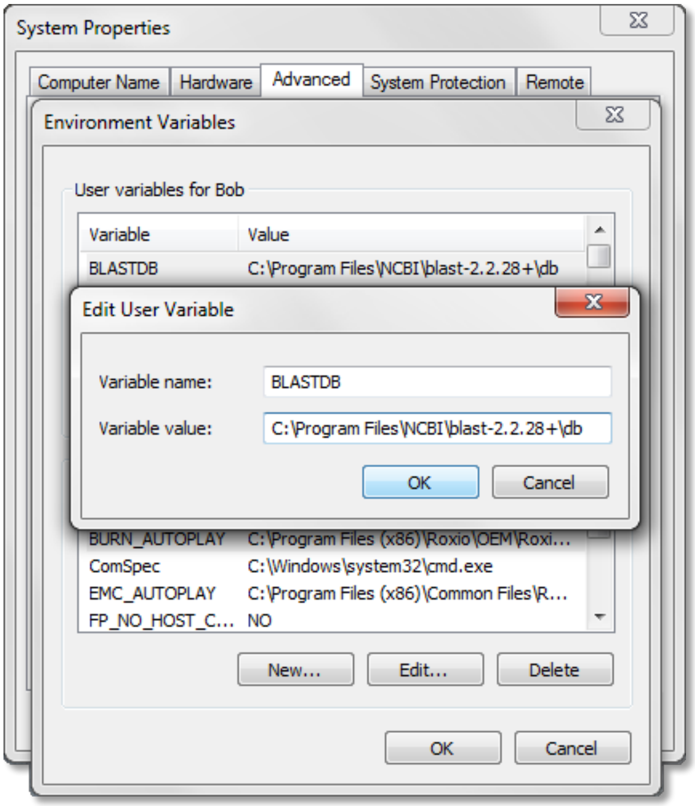
- Click on the **Environment Variables** button (if BLASTDB is already set, it may be under **User variables** for… or **System variables**).

- If BLASTDB does not yet exist, click on the **New** button under the **User variables for ...** panel (or add it as a **System** variable if you want it available for all users).

- Type the environment variable name **BLASTDB** and enter the absolute path. For example, if the databases are in the folder "db" in C:\Program Files\NCBI\blast-2.2.28+\, the path would be **C:\Program Files\NCBI\blast-2.2.28+\db**.

- Click on the **OK** button to dismiss any open dialogs.

Figure 25-22 Editing the Environment Variable



User variables for Bob

Variable	Value
BLASTDB	C:\Program Files\NCBI\blast-2.2.28+\db

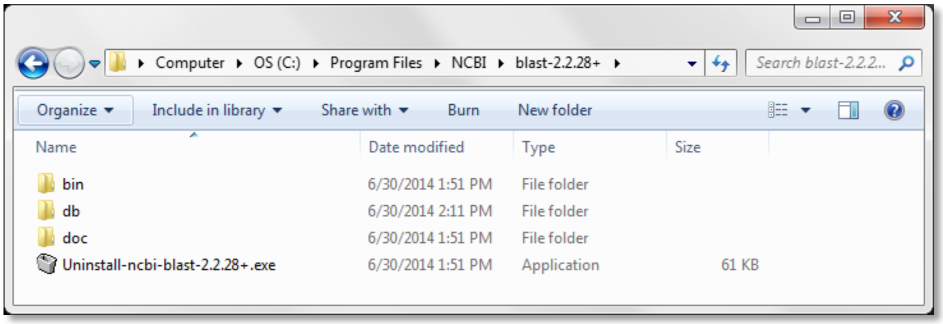
Variable name: BLASTDB

Variable value: C:\Program Files\NCBI\blast-2.2.28+\db

BURN_AUTOPLAY	C:\Program Files (x86)\Roxio\OEM\Roxi...
ComSpec	C:\Windows\system32\cmd.exe
EMC_AUTOPLAY	C:\Program Files (x86)\Common Files\R...
FP_NO_HOST_C...	NO

From a Windows Explorer window, navigate to the folder to which BLASTDB points. This is where you place your database files.

Figure 25-23 Creating a db folder for BLAST



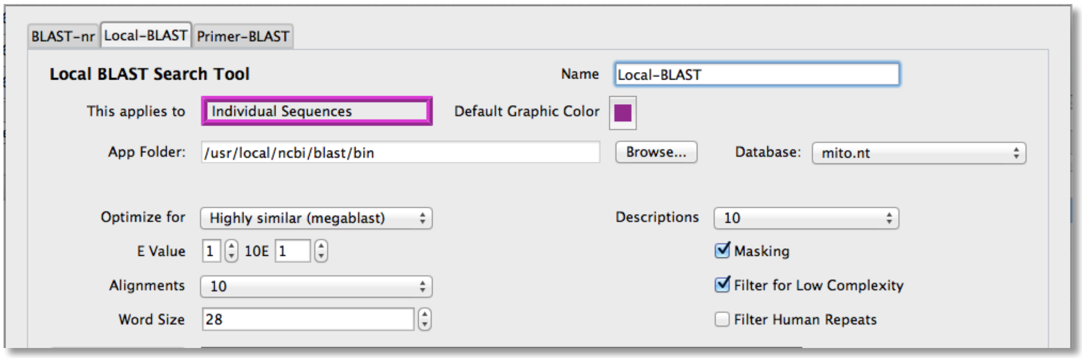
Name	Date modified	Type	Size
bin	6/30/2014 1:51 PM	File folder	
db	6/30/2014 2:11 PM	File folder	
doc	6/30/2014 1:51 PM	File folder	
Uninstall-ncbi-blast-2.2.28+.exe	6/30/2014 1:51 PM	Application	61 KB

WORKING WITH LOCAL-BLAST

Other than being installed locally and using databases that you have created, there are very few differences between BLAST and Local-BLAST options. You still have the choice between blastn and megablast. You can still mask your sequences. You can still filter for low complexity and set the E value and word size.

Note: Unlike BLAST, Local-BLAST does not set a default database for you. You must choose a database from those available by using the **Database** drop-down menu in Local-BLAST **Options**.

Figure 25-24 Local-BLAST parameters



Local BLAST Search Tool

Name: Local-BLAST

This applies to: Individual Sequences

Default Graphic Color: ■

App Folder: /usr/local/ncbi/blast/bin

Database: mito.nt

Optimize for: Highly similar (megablast)

E Value: 1 10E 1

Alignments: 10

Word Size: 28

Descriptions: 10

☑ Masking

☑ Filter for Low Complexity

□ Filter Human Repeats

CONNECTIONS SCHEMATIC

The Connection Schematic is a tool that allows you to compare the results of BLAST and Primer-BLAST runs, using different parameters, on a single sequence. This is an immensely powerful way of finding functional regions in your sequences. In this example, we are showing the results of the same sequence run with BLAST against different subsets of the GenBank collection: htgs, dbsts, human est, and mouse est. The hits are listed in a table below the diagram.

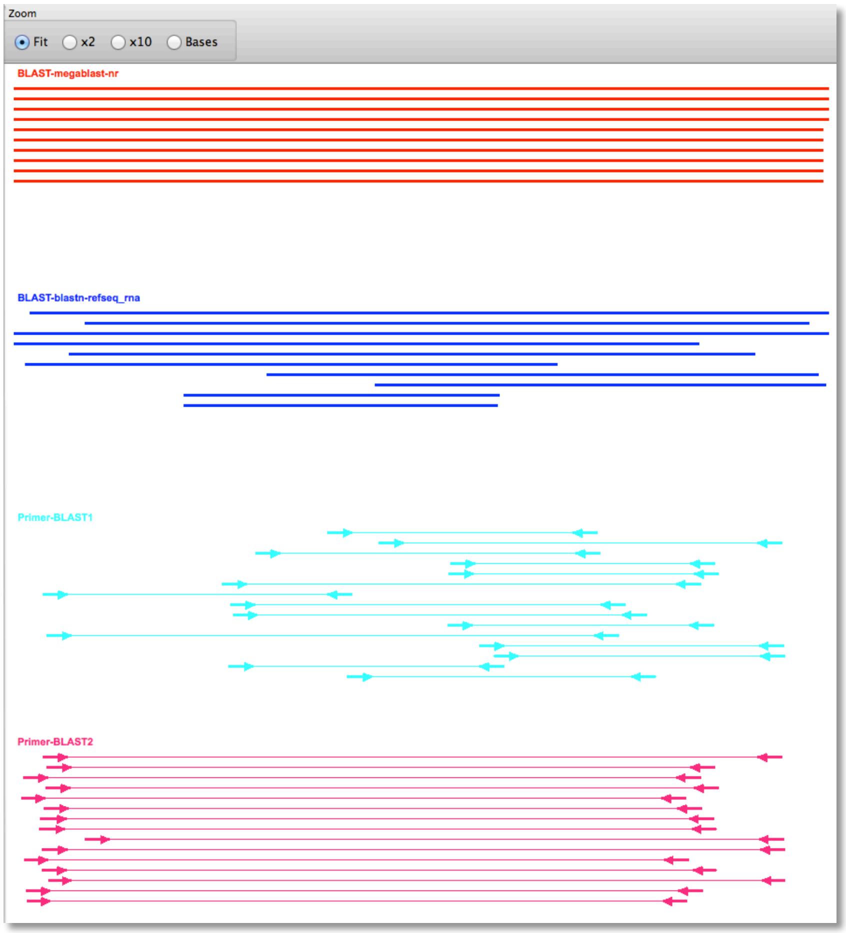
Right-click in the row header of the sequence you have results for and whose Channels you would like to summarize in a schematic. The row header has a row number in it like 1, 2, 3, 4, etc. Choose **Show Schematic**.

Depending on how much data is involved, this window may take a little time to build. To obtain more information about a particular hit, hover your cursor over the line. This will get information about that particular BLAST hit. Alternatively, you may see more detailed data in the tables below the graphic.

You will note that each Channel in the image below is displayed in a different color. This is set in the **Channel Options**. Click on the green box next to the **Default Graphic Color** text and choose a color.   By clicking on the appropriate radio button, you can zoom all the way in to view the individual bases on the Schematic.

Initially, you may prefer to use the **Web View** to examine your results using the BLAST or Primer-BLAST graphic. An important thing to remember is that BLAST and Primer-BLAST results eventually expire from the NCBI website. **Sequencher** Connections saves the results of any searches so that you can have a permanent record of your search. Moreover, the Schematic not only gives you a graphic view of your results, but also allows you to compare those results across a series of channels, dramatically enhancing the graphic overview.

Figure 25-25 The Connections Schematic showing BLAST and Primer-BLAST results


Zoom

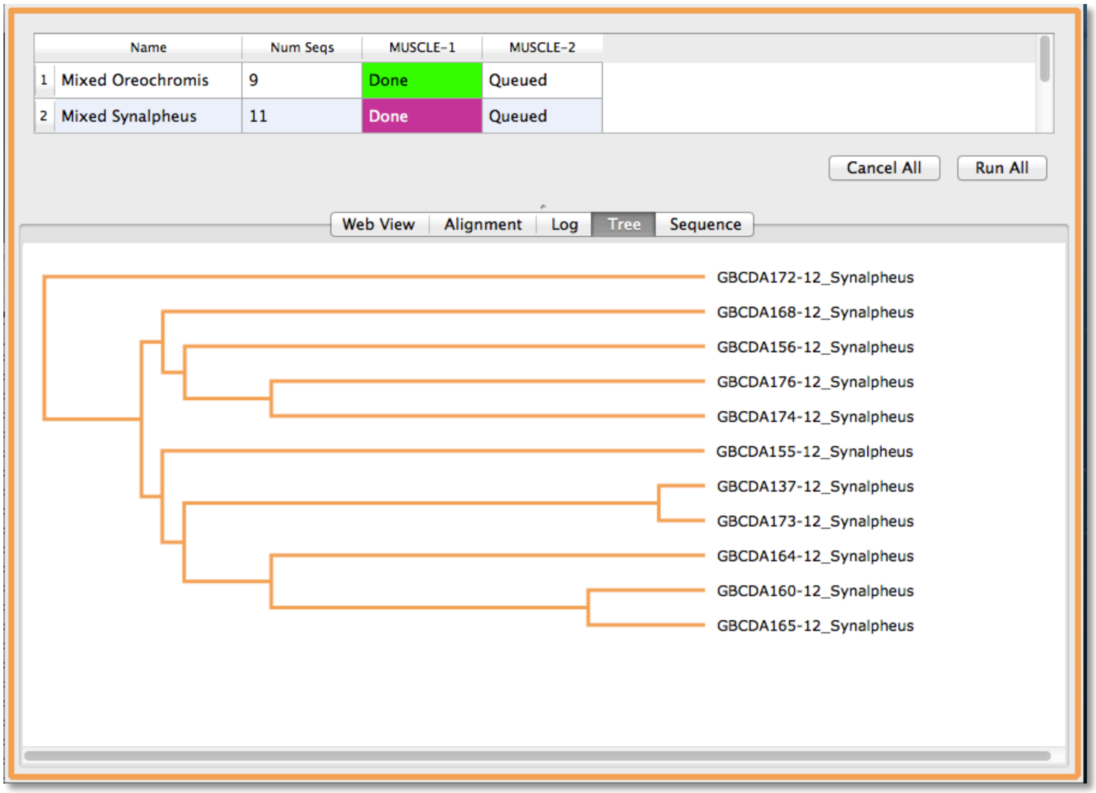
● Fit ○ x2 ○ x10 ○ Bases

SEQUENCHER CONNECTIONS, GROUP SEQUENCES AND MUSCLE

Group sequences are sets of sequences you suspect are related. One of the main ways of testing relatedness is to align all the sequences together to maximize the number of matching bases. In order to achieve this, it is often the case that many gaps will be inserted. There are several programs that are designed for this of which **ClustalW2** and **MUSCLE** are among the best known. **MUSCLE** is a multiple-sequence alignment algorithm that is claimed to achieve both better speed and accuracy than **ClustalW2**. Starting with a group of phylogenetically-related sequences in **Sequencher**, it is now possible to align those sequences and create a phylogenetic tree using the **MUSCLE** algorithm.

When working with grouped sequences, you have two options. You can either create a new session for each group or you can place related groups in the same session. In the image below, two sets of fish mitochondrial DNA data have been added to a single grouped session. The advantage of this is that you can keep related analyses together.

Figure 25-26 Group session with two groups of data

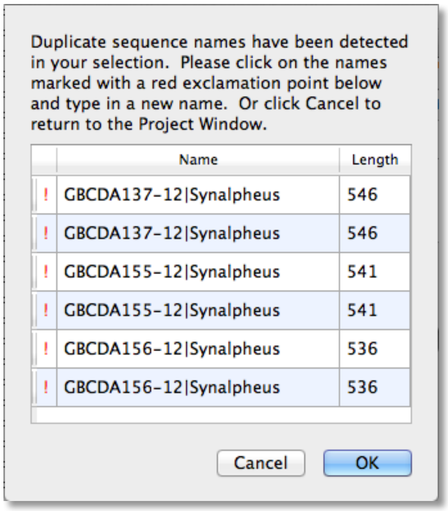


	Name	Num Seqs	MUSCLE-1	MUSCLE-2
1	Mixed Oreochromis	9	Done	Queued
2	Mixed Synalpheus	11	Done	Queued

Open a **Sequencher** project file containing your sequences. Then select some or all of the sequences. From the **Window** menu choose **Add to Connections Session...** if you are creating a grouped session for the first time. You will then have to name your session and then click on the **OK** button.

If you choose several sequences that have the same name to add to a session, **Sequencher** will warn you and ask you to change the names of those sequences.

Figure 25-27 Group Session duplicate names dialog

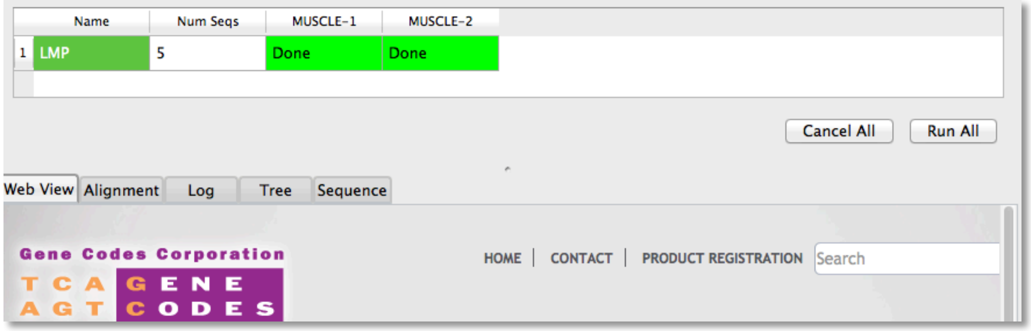


Duplicate sequence names have been detected in your selection. Please click on the names marked with a red exclamation point below and type in a new name. Or click Cancel to return to the Project Window.

	Name	Length
!	GBCDA137-12|Synalpheus	546
!	GBCDA137-12|Synalpheus	546
!	GBCDA155-12|Synalpheus	541
!	GBCDA155-12|Synalpheus	541
!	GBCDA156-12|Synalpheus	536
!	GBCDA156-12|Synalpheus	536

To edit the name, click in the **Name** field and type a new name or add an additional character or number if you want to preserve the majority of the name. As the name clashes are removed, the red exclamation points will disappear. When all duplicates have been renamed, click on the **OK** button. The session opens and the group is given a name (typically Group 1, Group 2, … Group n). Change the name by typing over the selected text in the name field.

Figure 25-28 Group session with user-specified name for group



	Name	Num Seqs	MUSCLE-1	MUSCLE-2
1	LMP	5	Done	Done

Once your session is open, you will initially see only one row in the grid bearing the name that you gave to your sequence group. You will also note that the tab views differ from those in the Individual BLAST **Sequencher Connections** dialog. You will see a **Web View** (which shows a page from the Gene Codes website), an **Alignment** view (a text-based view of the **MUSCLE** alignment), a **Log** view (view of the log file generated by the **MUSCLE** program), a **Tree** view (phylogenetic tree view), and a **Sequence** view (the unaligned sequences).

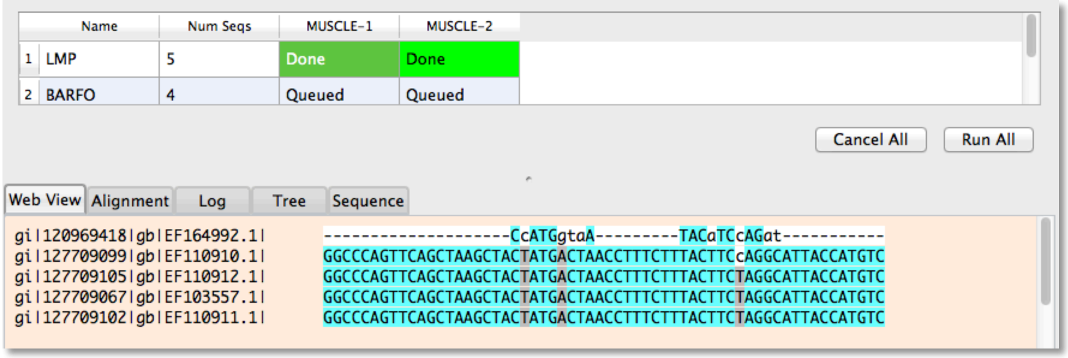
If you want to add further groups to the session, make your selection in the **Project Window**, select the **Add to Connections Session…** item on the **Window** menu, click on the **Add to an**

existing Connections Session... radio button in the **Session Launcher** dialog, choose the session you want to add the data to, and then click on the **OK** button.

Note*:* The outline colors for Individual and Group Connections sessions are different.

To run **MUSCLE** on all of the data in a session, click on the **Run All** button. If you have set up several Channels and only want to use a selected Channel, right-click in the header row for that Channel and choose the **Run on Each Group** menu item on the context menu. Once the status has changed to Done, you may view the alignment in the **Text** view or the phylogenetic tree in the **Tree** view by clicking on the appropriate tab.

Figure 25-29 Status messages in a Group session



	Name	Num Seqs	MUSCLE-1	MUSCLE-2
1	LMP	5	Done	Done
2	BARFO	4	Queued	Queued

```
gi|120969418|gb|EF164992.1|      --------------------CcATGgtaA---------TACaTCcAGat-----------
gi|127709099|gb|EF110910.1|      GGCCCAGTTCAGCTAAGCTACTATGACTAACCTTTCTTTACTTCcAGGCATTACCATGTC
gi|127709105|gb|EF110912.1|      GGCCCAGTTCAGCTAAGCTACTATGACTAACCTTTCTTTACTTCTAGGCATTACCATGTC
gi|127709067|gb|EF103557.1|      GGCCCAGTTCAGCTAAGCTACTATGACTAACCTTTCTTTACTTCTAGGCATTACCATGTC
gi|127709102|gb|EF110911.1|      GGCCCAGTTCAGCTAAGCTACTATGACTAACCTTTCTTTACTTCTAGGCATTACCATGTC
```

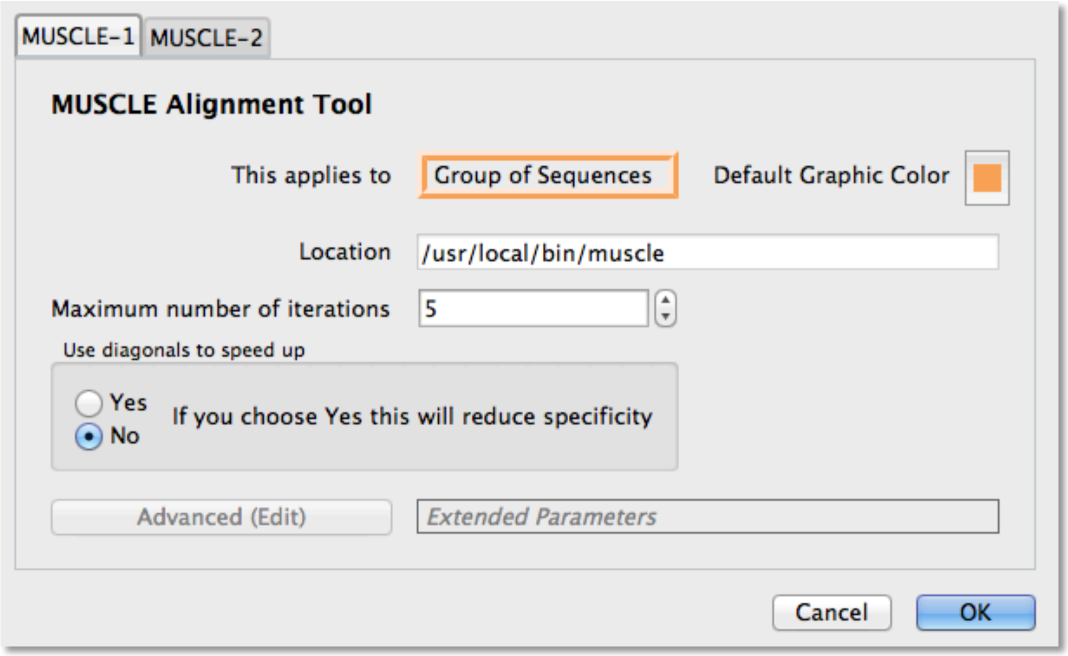
MUSCLE OPTIONS

As with the **Sequencher** Connections for individual sequences, you can also configure options for grouped sequences. Right-clicking on the **MUSCLE** column header invokes a contextual menu. Choose **Options....**

You can change the number of iterations that **MUSCLE** will go through. Each iteration will try to improve the quality of the multiple alignment. However, with very long sequences or large numbers of sequences, it may not be practicable to perform too many iterations. At this time, the maximum number of iterations you can choose is 5.

Another option you can set affects the speed of the alignment. When comparing two sequences, a shortcut method whereby short regions of high similarity are found has been implemented in **MUSCLE**. This speeds up the algorithm but does have an adverse affect on accuracy. You can turn this option on or off by choosing the appropriate radio button in the **Use diagonals to speed up** groupbox.

Figure 25-30 Options for MUSCLE algorithm



MUSCLE Alignment Tool

This applies to: Group of Sequences

Default Graphic Color: 🟧

Location: /usr/local/bin/muscle

Maximum number of iterations: 5

Use diagonals to speed up

○ Yes
● No

If you choose Yes this will reduce specificity

You can also change the color of the tree by clicking on the **Default Graphic Color** icon and selecting a replacement for the default.

SEQUENCHER CONNECTIONS MUSCLE TREE VIEW

You can view the phylogenetic tree in the **Tree** view by clicking on the appropriate tab. If you right-click within the Tree view, you will see a contextual menu that allows you to switch between the rectangular and circular rendering of the tree.

Figure 25-31 Rectangular phylogram using MUSCLE alignment

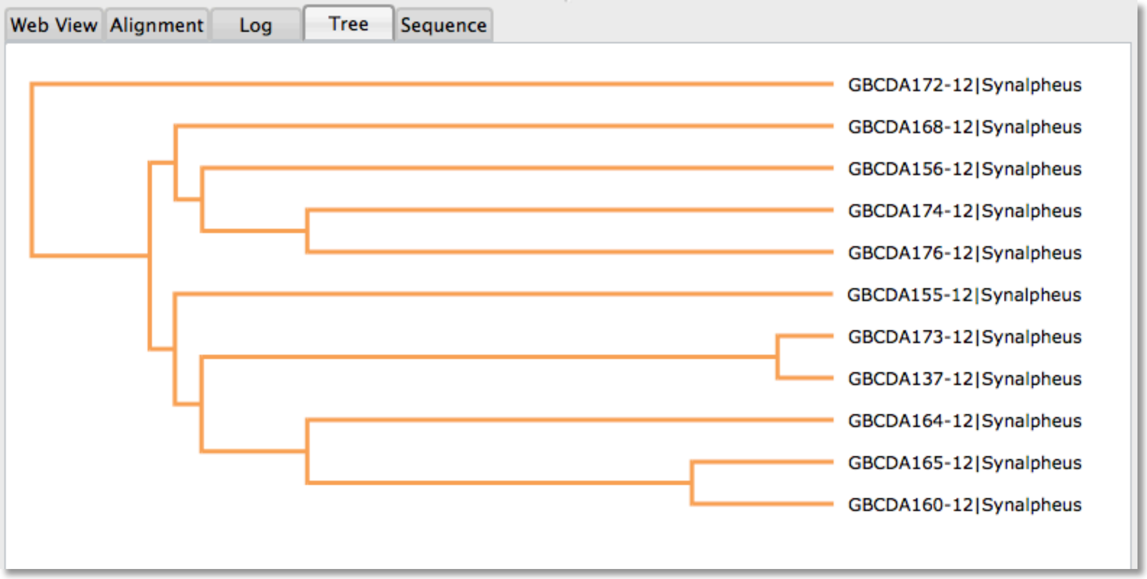
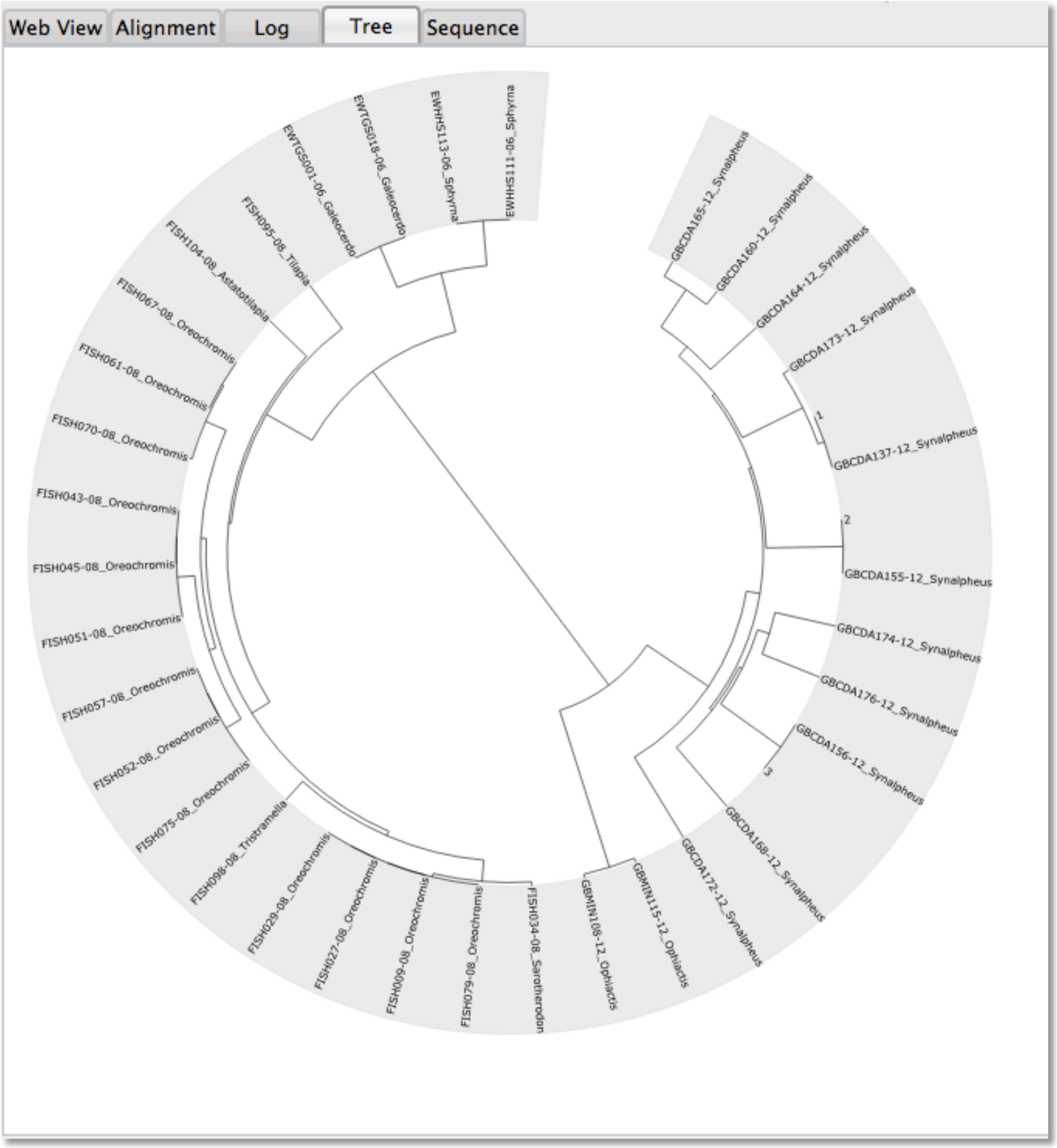


Figure 25-32 Circular phylogram using MUSCLE alignment



26. APPENDIX – KEYBOARD SHORTCUTS

WINDOWS

Menu/Window	Command	Windows
Any view	Help	F1
File menu	Open window or Open project (set in User Preferences)	Ctrl+O
	Close window	Ctrl+W
	Get Info…	Ctrl+I
	Save Project	Ctrl+S
	Print	Ctrl+P
	Exit	Ctrl+Q
Edit menu	Undo	Ctrl+Z
	Cut Selection	Ctrl+X
	Copy Selection	Ctrl+C
	Paste	Ctrl+V
	Edit Comments…	Ctrl+,
Select menu	Select All	Ctrl+A
	Select None	Ctrl+\

Menu/Window	Command	Windows
	Find bases…	Ctrl+F
	Find bases again	Ctrl+G
	Contig column	Ctrl+K
	Select Next Ambiguous base	Ctrl+N
	Select Next Contig Disagree	Ctrl+D
	Select Next Edited ase	Ctrl+E
	Select Next Low Confidence Base	Ctrl+L
	Select Next Met to Stop	Ctrl+M
Sequence menu	Mark Selection As Feature	Ctrl+‘
View menu	Display Format Ruler	Ctrl+R
	Reverse & Comp	Ctrl+4
	Sequenced Strand	Ctrl+5
View submenu	Display 1st frame translation	Ctrl+1
	Display 2nd frame translation	Ctrl+2
	Display 3rd frame translation	Ctrl+3
	Single stranded	Ctrl+-
	Double stranded	Ctrl+=

Menu/Window	Command	Windows
Window menu	Chromatogram	Ctrl+T
	User Preferences	Ctrl+U
	Switch between large icons, small icons, and list view.	Alt+click in Title line
Overview Options	Names at Left	Ctrl+Shift+left arrow key
Sequence Editor/Contig editor	Insertion point moves one base to the right.	Right arrow key
	Insertion point moves one base to the left.	Left arrow key
	Insertion point moves 3 bases to the right.	Alt+right arrow key
	Insertion point moves 3 bases to the left.	Alt+left arrow key
	Insertion point moves to 3' end of sequence.	Ctrl+right arrow key
	Insertion point moves to 5' end of sequence.	Ctrl+left arrow key
Summary view	Display Formatting Ruler.	Ctrl+Alt+R
	Frame 1	Ctrl+Alt+1
	Frame 2	Ctrl+Alt+2
	Frame 3	Ctrl+Alt+3
User defined		Ctrl+6 to Ctrl+0

MAC

Menu/Window	Command	Mac
Any view	Help	Cmd+Shift+? or Help
Project Window	Continuous selection	Click first item then Shift+Click
	Discontinuous selection	Cmd+Click each item
Sequencher menu	Quit **Sequencher**	Cmd+Q
File menu	Open window or Open project (set in User Preferences	Cmd+O
	Close window	Cmd+W
	Get Info…	Cmd+I
	Save Project	Cmd+S
	Print	Cmd+P
Edit menu	Undo	Cmd+Z
	Cut Selection	Cmd+X
	Copy Selection	Cmd+C
	Paste	Cmd+V
	Edit Comments…	Cmd+,
Select menu	Select All	Cmd+A
	Select None	Cmd+\
	Find bases…	Cmd+F
	Find bases again	Cmd+G

Menu/Window	Command	Mac
	Contig column	Cmd+K
	Select Next Ambiguous base	Cmd+N
	Select Next Contig Disagree	Cmd+D
	Select Next Edited Base	Cmd+E
	Select Next Low Confidence Base	Cmd+L
	Select Next Met to Stop	Cmd+M
Sequence menu	Mark Selection As Feature	Cmd+‘
	Delete Bases Fill Void From Right	Delete
	Delete Bases Fill Void From Left	Alt+Delete
	Insert Gaps & Move Bases Right	Tab
	Insert Gaps & Move Bases Left	Alt+Tab
View menu	Display Format Ruler	Cmd+R
	Reverse & Comp	Cmd+4
	Sequenced Strand	Cmd+5
View submenu	Display 1st frame translation	Cmd+1
	Display 2nd frame translation	Cmd+2
	Display 3rd frame translation	Cmd+3
	Single stranded	Cmd+-
	Double stranded	Cmd+=

Menu/Window	Command	Mac
Window menu	Chromatogram	Cmd+T
	User Preferences	Cmd+U
	Switch between large icons, small icons and list view.	Alt+click in Title line
Overview Options	Names at Left	Ctrl+Alt+left arrow
	Condensed Base Numbers	Ctrl+Alt+C
	Diagram Key	Ctrl+Alt+K
	Fragment Labels	Ctrl+Alt+L
	Fragment Names	Ctrl+Alt+N
	Fragment Positions	Ctrl+Alt+P
	Start & Stop Codons	Ctrl+Alt+S
	Base Numbers at Transitions	Ctrl+Alt+T
	Scale to Window	Ctrl+Alt+W
	Base numbers every X Bases	Ctrl+Alt+X
Sequence Editor/Contig editor	Insertion point moves one base to the right	Right arrow key
	Insertion point moves one base to the left	Left+arrow key
	Insertion point moves 3 bases to the right	Alt+right arrow key
	Insertion point moves 3 bases to the left	Alt+left arrow key

Menu/Window	Command	Mac
	Insertion point moves 9 bases to the right	Ctrl+Alt+right arrow key
	Insertion point moves 9 bases to the left	Ctrl+Alt+left arrow key
	Insertion point moves to 3' end of sequence	Cmd+right arrow key
	Insertion point moves to 5' end of sequence	Cmd+left arrow key
	Continue selection to 5' end	Cmd+[
	Continue selection to 3' end	Cmd+]
Summary view	Display/Hide Bullets and Pluses	Ctrl+Alt+B
	Display/Hide Consensus Sequence	Ctrl+Alt+C
	Display/Hide & Dashes	Ctrl+Alt+D
	Display/Hide Icons	Ctrl+Alt+I
	Display/Hide Fragment Sequences	Ctrl+Alt+F
	Display/Hide Matching bases as dashes	Ctrl+Alt+M
	Display Formatting Ruler	Cmd+R
	Frame 1	Cmd+1
	Frame 2	Cmd+2
	Frame 3	Cmd+3
Restriction map display options	1 cutter (unique cutter)	Ctrl+Alt+1

Menu/Window	Command	Mac
	2 cutters	Ctrl+Alt +
	3 cutters	Ctrl+Alt+3
	4 & more cutters	Ctrl+Alt+4
	All cutters	Ctrl+Alt+A
	Cut Positions	Ctrl+Alt+C
	Fragment Sizes	Ctrl+Alt+F
	Cut Positions	Ctrl+Alt+C
	Fragment Sizes	Ctrl+Alt+F
	Multiple Lines	Ctrl+Alt+M
	List selected enzyme names	Ctrl+Alt+N
	Single Lines	Ctrl+Alt+S
	Text	Ctrl+Alt+T
	Scale to Window	Ctrl+Alt+W
User defined		Cmd+6 to Cmd+0

27. APPENDIX – ADVANCED EXPRESSIONS

A regular expression is a special way of describing a search pattern using text strings according to certain syntax rules. **Sequencher** uses regular expressions to help you break down your sequence-naming schema in order to define your Assembly Handles.

Table 27-1 Regular Expression special characters

Character	Description
.	Represents any character
[A-Za-z]	Represents any letter
[A-Z].	Represents any capital letter
[a-z].	Represents any lowercase letter
\d	Represents any digit
[0-9]	Represents any digit
*****	A character or special character which is followed by ***** matches *zero* or *more* instances of the character
+	A character or special character which is followed by **+** matches *one* or *more* instances of the character
?	A character or special character which is followed by **?** matches *zero* or *one* instances of the character
{N}	A character or special character followed by **{**N**}**, where "N" = a whole number, matches that number of instances of the preceding character
[ ]	The contents of square brackets describe a list of matching items, regardless of order.
^	Use this in square brackets to *exclude* the other items contained within the brackets.
\	The \ is known as the escape character and it allows the character to be interpreted literally. It can also give a standard character a special meaning as in **\d.**
( )	Matches whatever you type within the parentheses
|	The | acts as an "or".

You may require a more complex delimiter than those provided in the **Name Delimiters** drop-down menu. You can use regular expressions to define these more complex delimiters.

Table 27-2 Examples of delimiter expressions

Delimiter Expression	Sequence Name	Handle 1	Handle 2	Handle3	Handle4	Handle5
%	Gel?%A%T7?%dog	Gel?	A	T7?	dog	
\?	Gel?%A%T7?%dog	Gel	%A%T7	%dog		
\?%	Gel?%A%T7?%dog	Gel	A%T7	dog		
\?|%	Gel?%A%T7?%dog	Gel		A	T7	dog

If your sequence name is complex, you may need to write a regular expression that describes the delimiters and the Assembly Handles. In the example below, the text between the parentheses defines the Assembly Handles. Anything not enclosed by parentheses will be treated as a delimiter.

(Assembly Handle 1**)**Delimiter**(**Assembly Handle 2**)**Delimiter**(**Assembly Handle 3**)**

You would use this kind of expression with the **Expression is a delimiter** box unchecked.

Regular expressions are powerful tools. There are many resources on the Internet that describe how to use regular expressions. Here is one to help you get started:

www.zytrax.com/tech/web/regex.htm

28. APPENDIX - IUPAC-IUB STANDARD CODES

Code	Base(s)	Meaning
A	A	Adenosine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil (RNA)
R	A or G	Purine
Y	C or T (or U)	Pyrimidine
K	G or T	Keto
M	A or C	Amino
S	G or C	Strong
W	A or T (or U)	Weak
B	C, G, T (or U)	Not A
D	G, A, T (or U)	Not C
H	A, C, T (or U)	Not G
V	A, C, G	Not T (or U)
N	A, C, G, T (or U)	(Any)

29. APPENDIX – DEFAULT FEATURE STYLES

The following table displays the default settings for the colors and styles of new features. New features are those you create in an existing sequence or those described in an imported sequence with a GenBank style Feature Table.

You can change the default settings in the Feature, Motif preference pane. (For more information, refer to Chapter 23 “Customizing Sequencher and User Preferences”.

Feature	Display	Color	Style	Second Strand
A. misc_feature				
1. misc_difference	√	Red	Underlined	
a) conflict	√	Red		
b) unsure	√	Red		
c) old_sequence	√	Red		
d) variation	√	Red		
e) modified_base	√	Red		
f) mutation	√	Red		
g) allele	√	Red		
2. gene				

Feature	Display	Color	Style	Second Strand
3. misc_signal				
a) promoter				
1) CAAT_signal				
2) TATA_signal				
3) -35_signal				
4) -10_signal				
5) GC_signal				
b) RBS				
c) polyA_signal				
d) enhancer				
e) attenuator				
f) terminator				
g) rep_origin				
h) oriT				

Feature	Display	Color	Style	Second Strand
4. misc_RNA				
a) prim_transcript				
1) precursor_RNA				
a) mRNA				
b) 5'clip				
c) 3'clip				
d) 5'UTR				
e) 3'UTR				
f) exon	√	Blue		
g) CDS	√	Blue		Protein
1) sig_peptide				
2) transit_peptide				
3) mat_peptide				
h) intron	√	Cyan	Invert case	
i) polyA_site				
j) rRNA				
k) tRNA				
l) scRNA				
m) snRNA				
n) snoRNA				

Feature	Display	Color	Style	Second Strand
5. Immunogobulin related **a) C_region** **b) D_segment** **c) J_segment** **d) N_region** **e) S_region** **f) V_region** **g) V_segment**				
6. repeat_region **a) repeat_unit** **b) LTR** **c) satellite** **d) transposon**				
7. misc_binding **a) primer_bind** **b) protein_bind** **c) STS** **d) primer**				
9. misc_structure **a) stem_loop** **b) D-loop**				

Feature	Display	Color	Style	Second Strand
10. gap				
11. operon				
12. source				
-				
Unrecognized*				

**Note* – The Unrecognized key is at the same level at misc_feature.

30. APPENDIX - FEATURE KEYS & QUALIFIERS

GENBANK FEATURE TABLES

GenBank Feature Tables provide a way of describing and locating features in a DNA sequence using internationally agreed-upon layout and vocabulary. Although these tables look complex, all tables are comprised of three basic elements which are described in the table below.

Table 30-1 Elements in a feature table

Element	Description
Feature Key	A single word or abbreviation which indicates a functional grouping
Location	The instruction for where to find the feature in the sequence
Qualifier	Additional information about the feature

Feature tables may look simple or complex depending on the number of annotations in the table. The following figure shows a typical entry.

Figure 30-1 A simple feature table entry

```
Key       Location/Qualifiers source   1..1509

/organism="Mus musculus"

/strain="CD1"
```

In the example above, the feature key is “source” and the location is represented by a range, “1…1509”. A location can be a single base, a range of bases, a single base within a range, etc. The qualifier, which can be free text, controlled vocabulary, citation or references numbers, sequences, or user-defined feature labels is separated from the location by a forward slash.

For a more detailed discussion of table entries and their standard typography, see http://www.ncbi.nlm.nih.gov/collab/FT/index.html.

The table below contains a listing of the standard feature keys which have been implemented in **Sequencher**.

Table 30-2 Feature keys groups and hierarchies

Feature	Specificity or detail	Functional grouping
A. misc_feature		
1. misc_difference	a) conflict b) unsure c) old_sequence d) variation	**Difference and change features**
2. gene		
3. misc_signal	a) promoter 1) CAAT_signal 2) TATA_signal 3) -35_signal 4) -10_signal 5) GC_signal b) RBS	**Expression signal features**

Feature	Specificity or detail	Functional grouping
4. misc_RNA	a) prim_transcript 1) precursor_RNA a) mRNA b) 5'clip c) 3'clip d) 5'UTR e) 3'UTR f) exon g) CDS 1) sig_peptide	**Transcript features**
5. Immunogobulin related	a) C_region b) D_segment c) J_segment d) N_region	
6. repeat_region	a) repeat_unit b) LTR c) satellite	**Repeat features**
7. misc_binding	a) primer_bind b) protein_bind	**Binding features**
8. misc_recomb	a) iDNA	**Recombination features**
9. misc_structure	a) stem_loop b) D-loop	**Structure features**
10. gap		
11. operon		
12. source		

FEATURE TABLE QUALIFIERS

Genbank Feature Qualifiers convey many different types of information. To accommodate this breadth there are several value formats.

Table 30-3 Feature Qualifiers

Qualifier Type	Description
Free text	Free text qualifiers are usually descriptive phrases enclosed in double quotation marks.
Controlled vocabulary or enumerated values	Controlled vocabulary or enumerated values qualifiers are selections from a controlled vocabulary list and are entered without quotation marks. Examples of controlled vocabularies can be found in the Appendices to the GenBank Feature Table documentation (http://www.ncbi.nlm.nih.gov/collab/FT/index.html).
Citation or reference numbers	A citation or reference number is enclosed in square brackets to distinguish it from other numbers.
Sequences	Since 1998, it has not been acceptable to use a literal sequence of bases as the Sequence Qualifier.
Feature labels	The feature label qualifier supports clarity in reporting by making sure your references are unambiguous. An example of a feature label would be alcA (see example below).

The figure below shows a typical example of a feature table entry with qualifiers. In this example, the feature key is CDS. The location is expressed as two numbers. The qualifiers tell the reader that the gene is alcA, the product of the gene is alcohol dehydrogenase, and that the evidence is experimental.

Figure 30-2 Qualifier examples

```
Key   Location/Qualifiers

CDS   6006..>6253

/gene="alcA"

/experiment="experimental evidence, no additional details
recorded"

/note="ADHI; ethanol regulon"

/codon_start=1

/product="alcohol dehydrogenase I"
```

Information on the Feature Keys and Tables and Qualifiers has been taken from the official GenBank documentation which can be obtained from the following URL: http://www.ncbi.nlm.nih.gov/collab/FT/index.html.

31. APPENDIX - DEFAULT FEATURE QUALIFIERS

In this table you will find the default Feature Qualifier for each Feature Key as implemented in **Sequencher**.

Feature Key	Default Feature Qualifier
Sequencher *	
A. misc_feature	misc feature [note]
1. misc_difference	misc difference [note]
a) conflict	conflict [compare]
b) unsure	unsure difference
c) old_sequence	old sequence [compare]
d) variation	variation [note]
e) modified_base	[gene] modified base[mod_base]
Allele **	allele
Mutation **	mutation
2. gene	[gene] gene
3. misc_signal	[gene] misc signal
a) promoter	[gene] promoter
1) CAAT_signal	[gene] CAAT signal

Feature Key	Default Feature Qualifier
2) TATA_signal	[gene] TATA signal
3) -35_signal	[gene] -35 signal
4) -10_signal	[gene] -10 signal
5) GC_signal	[gene] GC signal
b) RBS	[gene] RBS
c) polyA_signal	[gene] polyA signal
d) enhancer	[gene] enhancer
e) attenuator	[gene] attenuator
f) terminator	[gene] terminator
g) rep_origin	[gene] replication origin
h) oriT	[gene] oriT
4. misc_RNA	[gene] misc RNA
a) prim_transcript	[gene] primary transcript
1) precursor_RNA	[gene] precursor RNA
a) mRNA	[gene] mRNA
b) 5'clip	[gene] 5'clip

Feature Key	Default Feature Qualifier
c) 3'clip	[gene] 3'clip
d) 5'UTR	[gene] 5'UTR
e) 3'UTR	[gene] 3'UTR
f) exon	[gene] exon
g) CDS	[gene] CDS
1) sig_peptide	[gene] signal peptide
2) transit_peptide	[gene] transit peptide
3) mat_peptide	[gene] mature peptide
h) intron	[gene] intron
i) polyA_site	[gene] polyA site
j) rRNA	[gene] rRNA
k) tRNA	[gene] tRNA
l) scRNA	[gene] scRNA
m) snRNA	[gene] snRNA
n) snoRNA	[gene] snoRNA

Feature Key	Default Feature Qualifier
5. Immunogobulin related	Immunogobulin related
a) C_region	[gene] C region
b) D_segment	[gene] D segment
c) J_segment	[gene] J segment
d) N_region	[gene] N region
e) S_region	[gene] S region
f) V_region	[gene] V region
g) V_segment	[gene] V segment
6. repeat_region	[gene] [rpt_type] repeat region [rpt_family]
a) repeat_unit	[gene] [rpt_type] repeat unit
b) LTR	[gene] LTR
c) satellite	[gene] [rpt_type] satellite
**Transposon ** **	transposon
7. misc_binding	misc binding of [bound_moiety]

Feature Key	Default Feature Qualifier
a) primer_bind	[gene] primer binding site
b) STS	STS [locus_tag]
c) protein_bind	[bound_moiety] protein binding site
Primer**	primer
8. misc_recomb	Misc recombination [gene] [map]
a) iDNA	[gene] iDNA
9. misc_structure	misc structure [note]
a) stem_loop	[gene] stem loop
b) D-loop	[gene] D-loop
10. gap	gap of [estimated_length]
11. operon	[operon] operon
Source	[organism]
Unrecognized ***	

* The **Sequencher** Feature Key is for personal annotation.

** Legacy keys.

*** This is used for Feature Keys which are unknown to **Sequencher** so that they can be imported in to a project.

32. GLOSSARY

Name	Description
Administrator	Also known as Admin or admin user. A user account that typically has privileges that allow the user to change system settings and install software.
Agent	A region in some dialogs (and also in the Contig Editor) that contains information in text form.
ASCII	American Standard Code for Information Interchange; the most widely accepted coding system that permits letters, numbers, and punctuation to be represented in binary computer code.
Assemble by Name	A feature that uses portions of a sequence's name to determine which other sequences it will be compared against to construct a contig.
Base caller	Program that analyses fluorescent trace intensities from an automated sequencer and assigns bases together with a probability of error.
BLAST	Basic Local Alignment Search Tool. Developed by NCBI. This algorithm will identify sequences in a database that align to the query sequence. Most searches are conducted over the internet using the BLAST servers at NCBI. The program can also be installed on a desktop computer in which case it is called Local-BLAST.

Name	Description
Button bar	The tool region in a **Sequencher** window. The button bar is located just below the title bar.
CAF	A format for describing sequence assemblies
Channel	The channel in a Connections Session contains a single program and associated settings and is represented in the Session by a column.
Checkbox	A control that allows you to toggle a parameter. Click on the box to turn it on or off. When a checkbox is active, it is displayed with an "X" in the box.
Chromatogram	Also known as a sequencing trace or electropherogram. This is the plot of results from electrophoresis of a DNA sequencing reaction. The results are visualized as traces of four separate colors where each peak represents a base.
Condition	In RNA-Seq, a condition is a single sample, tissue, or treatment on a single sample or tissue. RNA-Seq experiments usually contain at least two conditions.
Confidence score	A number associated with each base call and which defines the likelihood that a base call is incorrect. The most common scale is from 1-60, where "60" represents a $1/10^6$ chance of a wrong call and 20 represents a $1/10^2$ chance. See also Quality score.
Contig	Contiguous alignment of a set of sequences to make up the sequence of bases from a longer piece of DNA.
Cursor	The indicator on a computer screen that shows the currently active location. Cursors have various shapes, indicating different purposes and capabilities.
Data source	A data source is a sequence that will be analysed in a Connections Session.
Dialog	Any of several special windows that the operating system or **Sequencher** displays when it needs to alert you to something you need to know, or needs to ask you a question.

Name	Description
Delimiter	Usually a character such as / or – which separates elements of a sequence name
Elevator button	A very small scroller that allows you to increase or decrease a numerical value by using the mouse. An elevator button has an up arrow and a down arrow.
Enzyme	In this manual, enzyme always refers to a restriction endonuclease.
Exemplar	Chosen or primary sequence. Literal meaning typical example or excellent model.
External Data Browser	A tool for managing NGS, RNA-Seq, and MSA (Multiple-Sequence Alignment) run folders. With this tool, you may open the folders, delete them, view the log files, view alignments in Tablet, and annotate the run.
External Data Home	The location where your run folders are saved. The default location is your Documents folder.
External Tools	Programs that can be used by **Sequencher** to perform analysis. Specifically, the External Tools are BWA, GSNAP, Maq, Velvet, MUSCLE, Tablet and which are installed using a single installer. The RNA-Seq tools are also an example of external tools.
Field	A rectangle in which the user can (or in some cases must) enter names or other information.
Gap	An insertion or deletion between two or more aligned sequences. This may be one or more bases long.
Handle	The portion of a sequence name between two consecutive delimiters. Also known as Assembly handle.
Large Gap	An insertion or deletion consisting of 10 or more bases between two or more aligned sequences.

Name	Description
Modal dialog	A dialog that requires a response before you can continue with regular actions in the application.
Motif	In **Sequencher**, this is a short subsequence of 50 bases or less.
MSF	Multiple Sequence Format. A format used by multiple-sequence alignment programs.
Multiplex ID	A system for tagging reads using short sequences of DNA. Used to distinguish samples in the same sequencing reaction. **Sequencher** separates the samples using these tags prior to aligning them.
NGS	Also known as Next-Generation Sequencing. This term covers a number of technologies that enable high-throughput sequencing.
NEXUS/Paup	Format used by phylogenetic and cladistic software.
Non-stranded	In GSNAP, this refers to protocols where reads that align 5’ to 3’ on either strand of the genome and their reverse complements are considered.
Project	A project contains sequences, information on how these sequences fit together to form larger pieces, and parameter information, specified by the user, that controls alignment operations.
Project window	The window in which **Sequencher** displays the Contigs and unincorporated Sequences of a project. Items in **the Project Window** can be shown as icons, small icons, or as a sorted list.
Quality score	A number associated with each base call and which defines the likelihood that a base call is incorrect. See also Confidence score.
Radio button	A control that allows you to choose among alternate possibilities. Radio buttons always appear in sets of at least two. Click one of the buttons to select it. A button that is selected is shown with a filled circle; buttons that are not selected are shown with open circles.

Name	Description
Reference sequence	A sequence used as a prototype or benchmark sequence. If a sequence has been designated as Reference within **Sequencher**, it has special properties.
Refrigerator	A special folder in which to store subsets of sequences which need to remain in your project.
Replicate	Duplicates of samples used in RNA-Seq to control variability. There are two kinds of replicates – biological and technical. Biological replicates are separate samples that are then processed independently. Technical replicates use the same sample but perform the technical steps separately.
RNA-Seq	This term applies to the use of NGS sequencing of transcriptomes. RNA-Seq provides a quantifiable snapshot of the expressed RNA in the sample at a given time.
RNA editing	Conversion of Adenosine to Inosine by the ADAR gene. This can be detected by NGS sequencing as an Adenosine to Guanine change when compared to a Reference Sequence.
SCF	Standard Chromatogram Format. A format used to store analyzed data or fragment data for a single sample.
Schematic	Visualization of BLAST, Local-BLAST, and Primer-BLAST results. The Schematic presents a consolidated view across all analyses for an individual sequence.
Secondary peak	A chromatogram peak whose height may be less than that of the primary peak and whose presence may indicate that a heterozygote exists at that position.
Sequence	A small piece of DNA, decoded into its component bases, in order, represented as text. Sometimes in reverse order, depending upon the vagaries of the sequencing procedure.
Sequence Editor	The window used for editing DNA sequences. This editor looks much like a word processor and supports the standard text editing commands: Cut, Copy, Paste, and Undo.

Name	Description
Session	Term used to describe a collection of channels (analyses) and data sources (sequences) in **Sequencher** Connections. There are two types of sessions – individual or grouped. In Individual Sessions, the analysis operates on single sequences although it can be applied to many sequences in parallel. In a Grouped Session, a single analyses operates on several sequences such as in a multiple alignment.
Slider	An on-screen control that simulates the action of a slide control. This is used as a control for increasing/decreasing a value.
Stranded	In GSNAP, this refers to protocols where only reads that align 5' to 3' on either strand of the genome are considered.
Title bar	A text region at the top of a window generally describing the function of the window or the name of the document displayed in the window.
Variant Calling	Identifying single bases that differ when compared to a Reference Sequence. Unlike Sanger sequencing where the traces may be examined, in NGS, reliance is placed on the algorithm used to distinguish between a true variant and technical problems with the sequence.
VecBase	A cloning vector sequence database prepared by Friedhelm Pfeiffer, in collaboration with William Gilbert. A file with this name, in VecBase (CODATA) format. Part of the **Sequencher** distribution.
Volcano plot	Specialised form of scatter plot. Plots –log10 (p-value) against log2fold change in expression.