SEQUENCHER™

Tutorial for Windows and Macintosh

More on SNP Hunting

© 2007 Gene Codes Corporation

Gene Codes Corporation

TCA GENE
AGT CODES

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

More on SNP Hunting

Getting started 3

Trimming low quality data 3

The Reference Sequence 4

Changing feature styles on the Reference Sequence 5

Assembling your data 6

Creating a Variance Table to locate SNPs 8

Focusing on regions of interest on the Reference Sequence 9

Reviewing the data 11

Removing unwanted data from the table 13

Making a Report 14

Saving your report 15

Conclusion 16

More on SNP Hunting

Sequencher's Variance Table gives you a summary analysis of your data that focuses on differences between sequences. The differences between two similar sequences may represent SNPs, polymorphisms, mutations, or just bases that require editing in order to be resolved. A Variance Table can compare two selected sequences or summarize all of the differences between each consensus sequence and a common Reference Sequence.

You can use the Variance Table to validate your data. Each cell in the Variance Table is linked to the data used to generate the base call. The sorting tools in the Variance Table make it easy to find novel SNPs or to identify regions prone to base calling errors. You can create and export a variety of reports based on the Variance Table.

GETTING STARTED

In this tutorial, you will use the Variance Table to identify and report on candidate SNPs. You will first need to open a project and trim the data within it.

- Launch Sequencher.
- Go to the **File** menu and select **Import > Sequencher Project…**
- Navigate to the Sample Data folder inside the Sequencher application folder.
- Choose the **More on SNP Hunting** project and select **Open**.

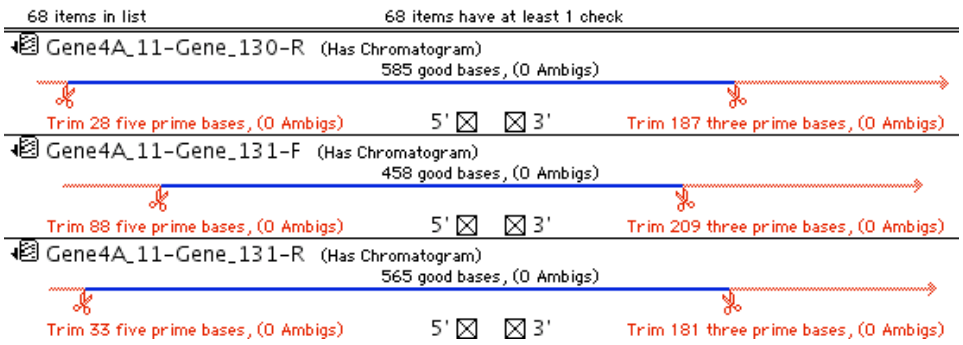
The SNP project contains 69 sequences. In addition to a Reference Sequence, there are 68 sequences composed of auto-sequencing data spanning different exons along a stretch of genomic DNA.

TRIMMING LOW QUALITY DATA

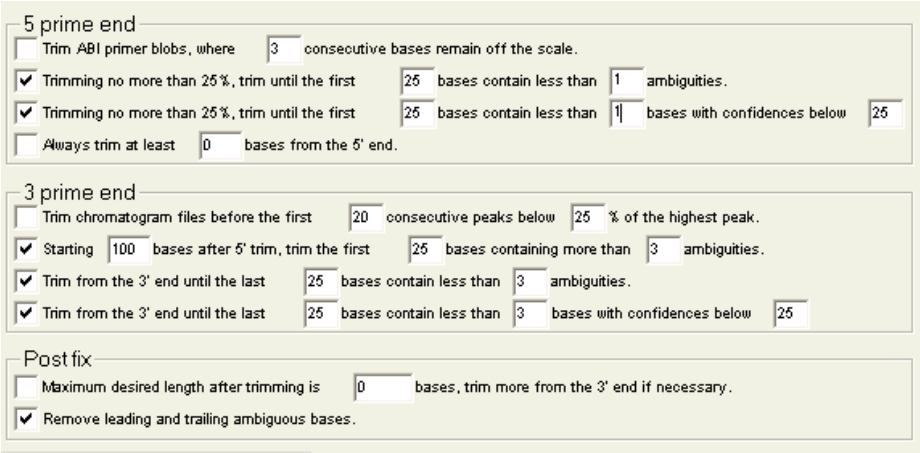
The first step in analyzing data generated by an automated sequencer is trimming the data. Automated sequencers create data where the ends of the sequence are low quality. Low quality data may include miscalled bases because this can affect the outcome of the proposed assembly, it needs to be removed.

- All of the items in your project should be selected on import, if not choose **Select > Select All**.
- Deselect the sequence called Reference with 3 exons, by clicking on it while holding the Apple (Mac) or Ctrl (PC) key.
- From the **Sequence** menu, select **Trim Ends…**

Sequencher will open the **Ends Trimming** window and display a graphic representation of the proposed trim for each of the sequences. You will see that each line has a red region at either end. A scissors icon indicates each proposed trim site.

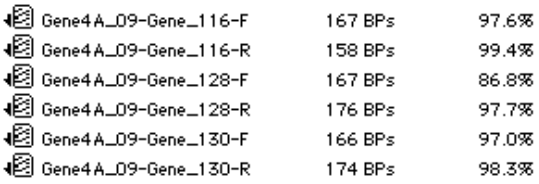


- Go to the button bar.
- Click on the **Change Trim Criteria** button.
- The **Ends Trimming Criteria** window will appear. Change the settings to match the following selections:



- Click **OK** to close the window.
- Go to the button bar and click the **Trim Checked Items** button.
- A caution window appears. Click the **Trim** button.
- Now only blue lines remain.
- Close the **Ends Trimming** window.

Note the improvement of the values in the **Quality** column in the Project Window.



The Reference Sequence

You are now ready to mark a sequence to serve as the Reference Sequence. The Reference has certain properties which are useful for characterizing SNPs. The Reference Sequence will set the base numbering and the orientation of the contig you are about to assemble. This allows you to reference a SNP in relation to a standardized position. In addition, the Reference Sequence does *not* contribute to the consensus sequence. Having a Reference Sequence does not skew the consensus sequence calculation, which is an especially important fact in situations where you only have a few sequences in the contig.

- Click the sequence called Reference with 3 exons to select it.
- Choose **Sequence > Reference Sequence.**

You will notice that the icon of the Reference Sequence now contains an R. The sequence is now protected from editing.

Reference with 3 exons

- Double click on the Reference Sequence icon to open its sequence editor.

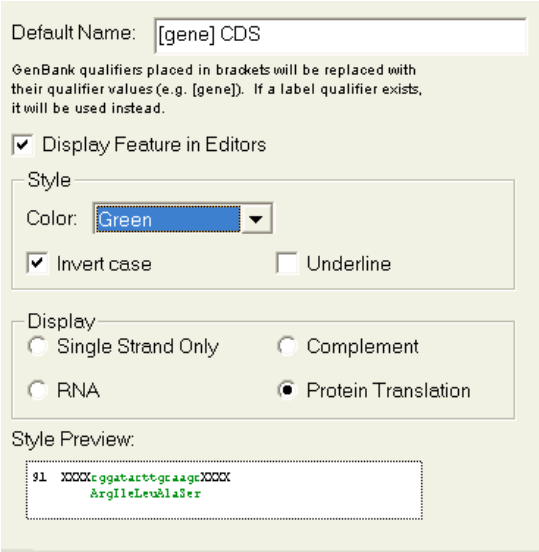
Notice that the numbering starts at 14,587.

CHANGING FEATURE STYLES ON THE REFERENCE SEQUENCE

Sequencher provides you with a set of powerful tools for adding features to your Reference Sequence or changing its display. You can add feature annotations that are compatible with GenBank or you can add your own annotations. In both instances you can choose the text display style and color. You can also control whether a specific set of features will be displayed or not in the editors or Overview.

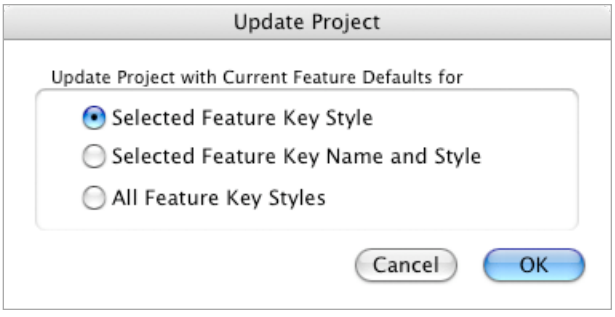
In this section of the tutorial you will change the style for all the features with a CDS feature key.

- Choose **Window > User Preferences…**
- Choose the **Display** tab.
- Click on **Feature, Motif**.
- Click on the **Define Feature Key Default Styles…** button.
- Scroll down in the **Feature Key:** pane until you find **CDS**. Select **CDS**.
- Change the color to Green. Click on **Invert Case**.

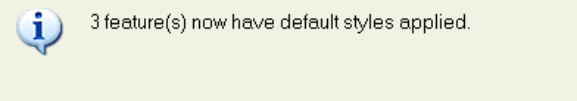


Note that the **Default Name** is **[gene] CDS**. The bracketed word defines an additional qualifier value from a GenBank feature that will be used to identify this feature in Sequencher. The name of these features will be HPS4 CDS [#], where HPS4 is the gene, CDS is the following text, and the bracketed number is the number of the exon in the original joined feature.

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



- Dismiss the window confirming that the new styles have been applied by clicking **OK**.



- Click **Done**. Close the **User Preferences** window.

Scrolling through the Sequence Editor window, you will notice that three segments of the sequence are green and in lower case. The red underlined bases, also imported features, are known variants in the HPS4 gene.

- Close the Sequence Editor.

Follow these steps in User Preferences whenever you want to change the style of all features that share a specific feature key. If you want to add a single feature or change its style, use the **Feature Editor**. You will learn how to do this later in the tutorial.

ASSEMBLING YOUR DATA

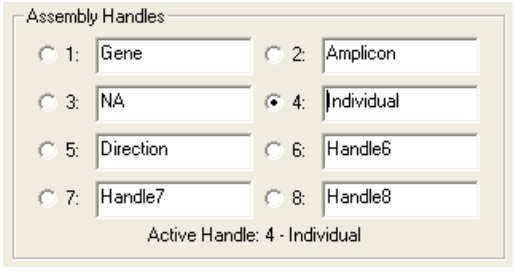
Sequencher provides several algorithms for data assembly. Each algorithm has been devised for a specific purpose and contains parameters you can control. In this tutorial you will use the default **Assembly Parameters** together with the **Assemble by Name** function to create your contig.

Assemble by Name is a powerful function which uses the information you have incorporated into your sequence names to decide which sequences should be considered for assembly in the same contig. In this way you can take multiple samples from multiple sources. Then you can assemble them into distinct contigs without manually selecting them first.

- Click on the **Assembly Parameters** button in the Project Window button bar.
- In the **Assemble by Name** pane ensure that the **Enabled** check box is selected.
- Click the **Names Settings…** button.

The sequence names in this tutorial are composed of text or numbers divided by characters such as – (dash) and _ (underscore). The numbers and text are referred to as **Assembly Handles**. Characters such as dashes or underscores are called **Name Delimiters**.

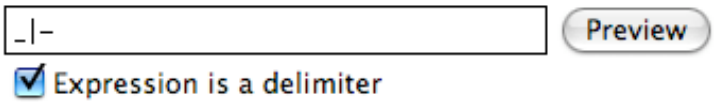
- Change the **Assembly Handles** names to match those in the image below.



Now you need to set up the **Name Delimiters**. In most instances you can select the appropriate character from the **Name Delimiter** drop down menu. Sometimes, however, the delimiters are more complex and can consist of more than one character or a mixture of characters. In these cases you need to set up an **Advanced Expression**. For instance in this

example you have two different delimiters. As an Advanced Expression, you would first select the check box defining your expression as a delimiter. Then you would type both the _ (underscore) and the – (dash) separated by the | (vertical bar). The vertical bar means "or" in a regular expression, so Sequencher will recognize both characters as delimiters.

- Choose **Advanced Expression…** from the **Name Delimiters** drop down menu.
- Click the **Define…** button.
- Ensure that the check box, **Expression is a delimiter**, is checked.
- Type _|- to match the **Advanced Expression** from the image below.



You can check the expression you have written by using the **Preview** function. It should match these preview results.

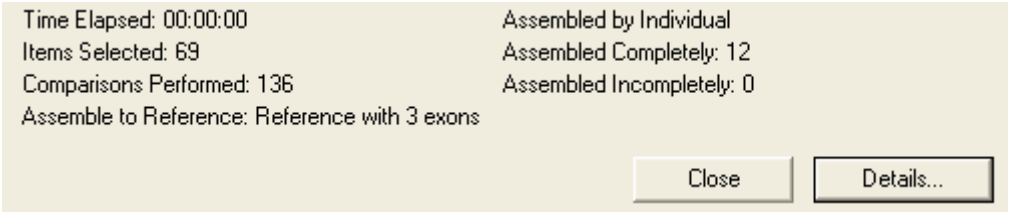
Gene	Amplicon	NA	Individual	Direction
Gene4A	09	Gene	116	F
Gene4A	09	Gene	116	R
Gene4A	09	Gene	128	F
Gene4A	09	Gene	128	R

- Click the **OK…** button to dismiss the **Advanced Expression for Name Parsing** window.
- Click on the fourth radio button. This should be labeled Individual.
- Ensure that you can see Active Handle: 4 – Individual at the bottom of the **Assembly Handles** pane.
- You may now dismiss the **Assemble by Name Settings** window by clicking the **OK** button.
- Click **OK** to close the **Assembly Parameters** window.

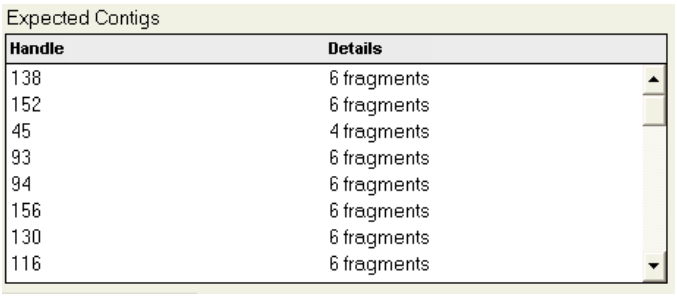
When you return to the project window, you will notice a few changes. You have a new column called **Handle**, which lists the **Individual** Assembly Handle appropriate for each sequence. The title of the assembly buttons has also changed indicating that **Assemble by Name** is now enabled. To revert to the standard assembly commands, you can return to the Assembly Parameters dialog and uncheck the **Assemble by Name Enabled** check box, or you can just toggle **Assemble by Name** off and on from the **AbN** button on your project window. For now, leave **Assemble by Name** enabled.

You are now ready to perform the assembly.

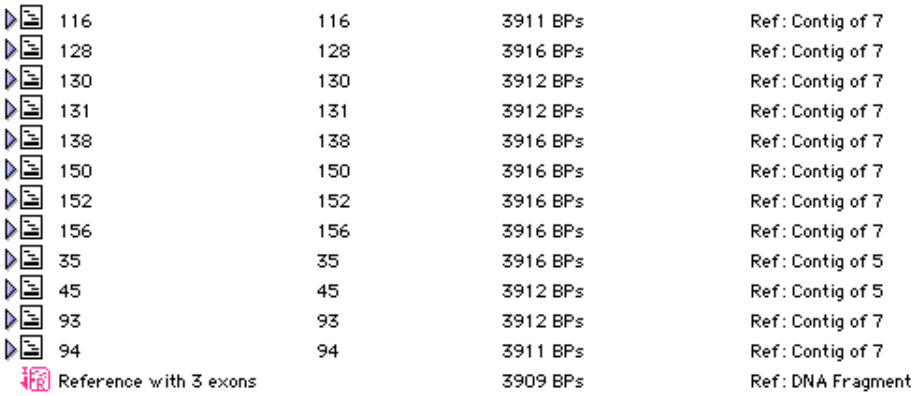
- Choose **Select > Select All** and click on the **To Reference by Name** button.
- Dismiss the **Assembly Preview** dialog window by clicking the **Assemble** button.
- Click the **Close** button to dismiss the **Assembly Completed** window.



Did you notice that the Assembly Preview dialog window indicated the expected number of fragments in each contig?



The actual number of sequences that will be assembled into a particular contig depends on the factors that normally affect a sequence assembly, such as the presence of a matching overlap. You should now have 12 contigs, one for each set of data. Each contig contains a copy of the Reference and the original Reference is still available in the project window.



Since you will be constructing a Variance Table using consensus sequences it is important to make sure you have the correct consensus calculation selected before you start. The consensus calculation you choose will affect the results of the Variance Table.

The **Consensus by Plurality** calculation uses a majority vote rule. **Consensus Inclusively** determines the consensus by the smallest category of ambiguity that covers all the available data.

- Choose **Contig > Consensus Inclusively**.

Creating a Variance Table to locate SNPs

The Variance Table provides an overview of the data within the same contig relative to a selected primary sequence or exemplar. You are now ready to create a Variance Table with your trimmed data and identify candidate SNPs. Since you have data from a number of different individuals and each has been assembled to the same Reference Sequence, you will construct a Variance Table using their consensus sequences.

- From the Project Window ensure that each contig is selected.
- Choose **Contig > Compare Consensus To Reference**.

Sequencher generates the Variance Table. Your initial view of the Variance Table displays the candidate SNPs for the 12 consensus sequences. Each column represents a contig. The rows are ordered according to the positions of the differences relative to the Reference Sequence.

Description: 12 consensus sequences compared to Reference Reference with 3 exons.
Comparison Range: Unfiltered
Base Positions: 14587..18495
Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

Reference		116	128	130	131	138	150	152	156	35	45	93	94	Total
15,395	G		C			C	C	C	C	C	C	C		8
15,400	T		A	A	A	A	A	A	A	A	A	A		10
15,401.1	:		T	T	T	T	T	T	T	T	T	T		10
15,403	C		S							S		S		3
15,517	A	g	g	g	g	g	g	g	g	r	g	r	g	12
15,564	C	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	10
15,576	G	T	T	T	T	T	T	T	T	K	T	K	T	12
15,579	C				Y									1
15,612	G		W					R	R					3
15,615	G							R						1
15,624	G		R											1
15,644.1	:		G	G	G	G	G	G	G	G	G	G		10
15,644.2	:		G	G	G	G	G	G	G	G	G	G		10
✣	Total	26	33	19	23	32	27	35	28	20	25	27	21	316

You will notice that each of the column headings is shaded in pink. This indicates that these sequences do not cover the full comparison range (the full length of the Reference Sequence). Notice that some of the cells contain a pink X. This indicates that this sequence does not have a base for the equivalent position in the Reference Sequence.

The **Total** cell in the bottom right corner shows that there are a total of 316 variants listed in the table. Notice that some of the bases in the table are in black and some are colored. If you scroll down the table you will see that some are red and some are green. Bases in green lie within a CDS, bases in red lie within a known variant.

You can obtain a listing of all the features marked on the Reference Sequence.

- From the Project Window select sequence **Reference with 3 Exons**.
- From the **Sequence** menu, select the command **Feature Listing**.
- Examine the features by scrolling through the window.

Notice that the name and location of each feature is listed. You will also see Feature Qualifier information and the display style for each feature. This provides more detail about the feature, such as the nature of a given variant.

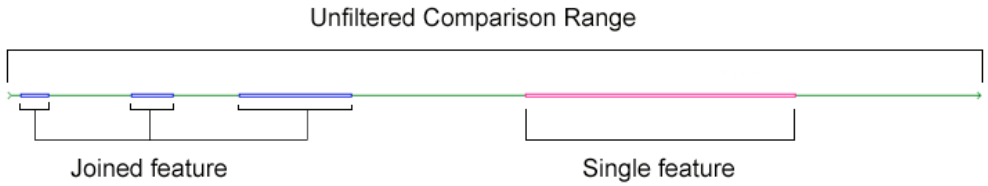
- Close the window.

You can turn off the display of features in the Variance Table by unselecting the **Display Features** command in the **View** menu.

Focusing on regions of interest on the Reference Sequence

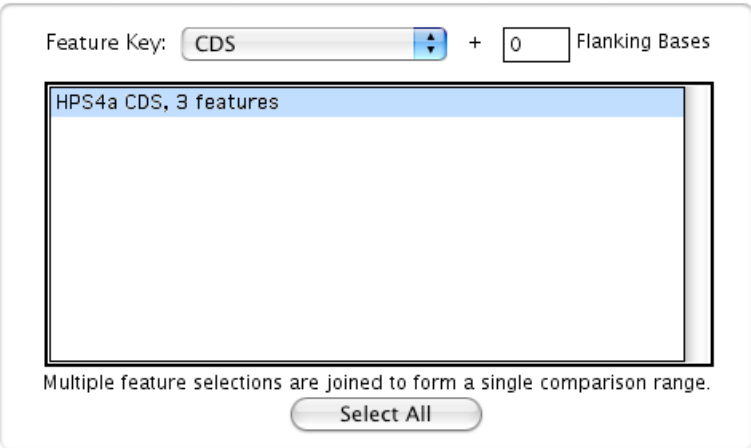
In some cases you want to explore the entire length of the Reference. However, if your Reference contains features such as exons or CDSes you can direct the Variance Table to focus on these features. This will reduce the amount of data you have to review. Sequences from GenBank are annotated with features using **Feature Keys** (a standardized method of referring to biologically important regions). The annotations are listed in a Feature Table that is read by Sequencher when the sequence is imported into a project.

The **Comparison Range** is defined by the bases as numbered in the primary or exemplar sequence. In the example in this tutorial, the Reference Sequence bases 14,587 to 18,495 define the unfiltered **Comparison Range**.



You can restrict the **Comparison Range** by choosing one or more of the features used to annotate the sequence.

- Return to the Variance Table.
- Go to the button bar and click on the **Comparison Range** button.
- Check the **Filter Comparison by:** radio button.
- Click on the **Feature Key:** drop down menu and choose **CDS.**
- Select the feature **HPS4a CDS, 3 features**.
- Click on the **OK** button to dismiss the **Comparison Range** dialog window.



The Variance Table is redrawn. Notice that the only variants in the table are now within a CDS. The **Total** cell at the bottom right of the table shows that there are now only 50 variants listed in the table. If, in some instance, you are interested in sequences flanking the feature, you would type a number in the **Flanking Bases** text box.

Description: 12 consensus sequences compared to Reference Reference Sequence.
Comparison Range: Filtered on CDS plus 0 flanking bases
Base Positions: 15501..15537 16212..16308 16937..17846
Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

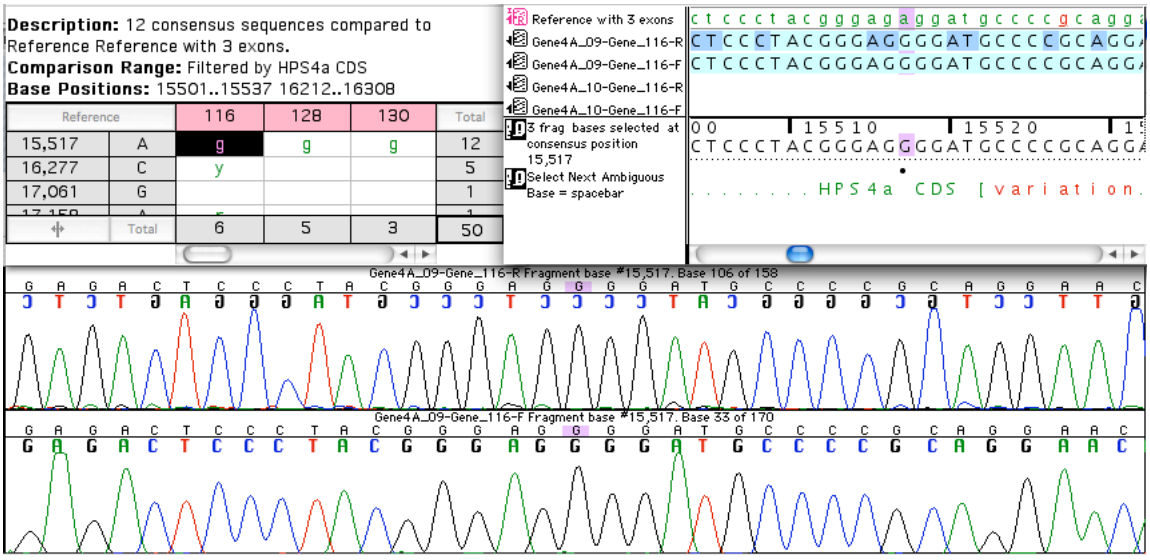
Reference		128	130	131	138	150	152	156	116	35	45	93	94	Total
15,517	A	g	g	g	g	g	g	g	g	r	g	r	g	12
16,277	C					y	y	y	y		y			5
17,061	G			r										1
17,158	A								r					1
17,229	G											r		1
17,353	G								k					1
17,419	C				y									1
17,460	C	s	s	g		s	g	g	s					7
17,586	A				m	m		m					m	4
17,649	C	y			y			y						3
17,715	A				m									1
17,771	C	m			m	m								3
17,787	G	a	a	a	a	a	a	a	a			r	a	10
✣	Total	5	3	4	7	6	4	6	6	1	2	3	3	50

REVIEWING THE DATA

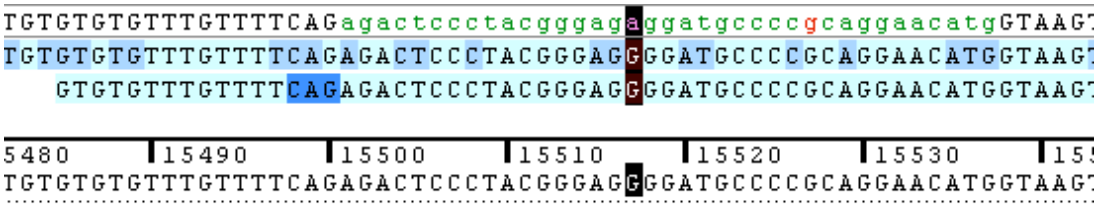
The **Review** mode of the Variance Table lets you use the table of differences to navigate to areas of interest and explore the underlying data. When you click on the **Review** button in the button bar, or when you double-click on a cell in the table, Sequencher opens the Contig Editor and Contig Chromatogram windows. The data displayed in each of the windows updates to reflect your selection in the Variance Table.

- Place your cursor in the cell in Reference row 15,517 and Individual column 116.
- Go to the button bar and click on the **Review** button.
- Arrange the windows to suit your viewing preference.

You can now see the chromatograms forming the underlying data for this consensus base call. There are two overlapping sequences in the Contig Editor. The consensus base and its component bases are highlighted. Both base calls have a pale blue background indicating that they have a confidence score falling within the High Range (40 – 60). Settings for the confidence score ranges can be altered in the User Preferences Confidence pane.



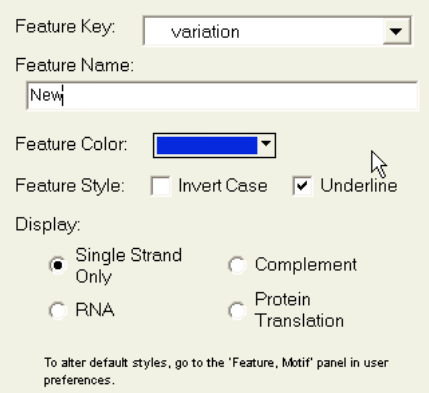
On closer examination you will notice the chromatogram data from two sequences are in opposite orientations. Look at the upper chromatogram. You can see the original base calls, below the current base calls in color, are displayed in their mirror image. This visual cue tells you that this sequence is in the reverse orientation relative to the Reference Sequence. Both chromatograms support the base call. The candidate SNP has been confirmed.



You can also see that the base falls within a green CDS region. There are no previously known variants (marked in red) at this position. To mark this base as a candidate SNP use the Variant Feature Key.

- Click on the consensus base to ensure it is highlighted..

- Choose **Sequence > Mark Selection as Feature**.
- Click on the **Feature Key:** drop down menu and choose variation.
- Change the **Feature Name:** New
- Set the **Feature Color:** to Blue.
- Click the **Underline** check box if it is not already checked.
- Click the **OK** button to dismiss the dialog window.



You will see the sample sequences' bases are now blue and underlined and that the word New has appeared below the Consensus line.

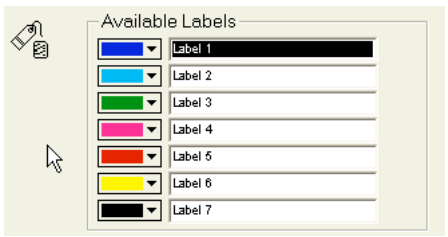
You can move quickly from one difference in the table to the next using a keyboard shortcut.

- Locate row base position 17,061 and place your cursor in the first column of the Variance Table at this position.
- While holding down the **Option/Alt** key, press the **right arrow** key.

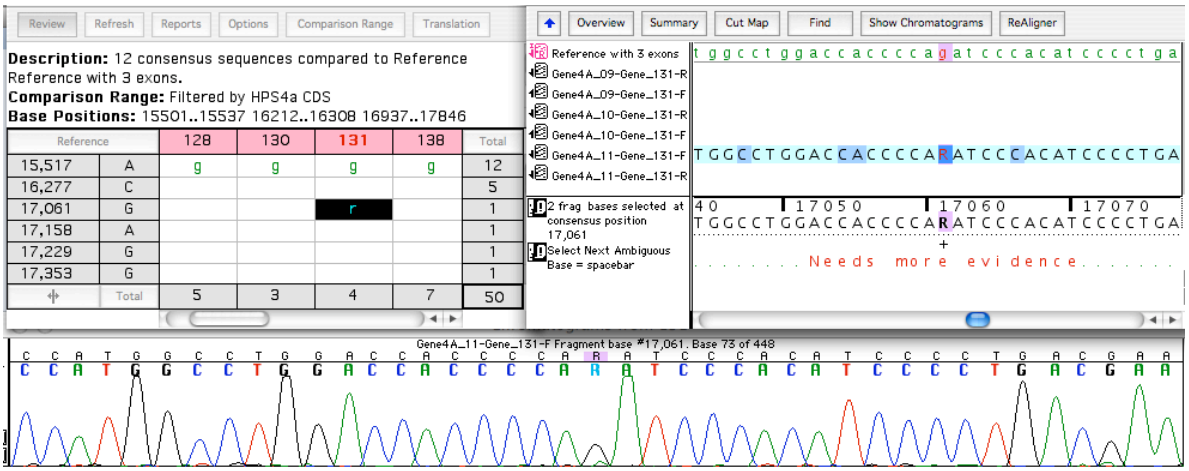
The **Option/Alt** key plus arrow key combination moves your cursor to the next difference in the Variance Table. Notice that the display in the Contig Editor and the Contig Chromatogram have updated to match the new position. Since there is only one chromatogram underlying this base call you may wish to reserve judgment until you have additional supporting evidence. You can still annotate it but this time use the Sequencher Feature Key to indicate that this is not yet a confirmed variation.

- With your cursor still in the Variance Table, choose **Sequence > Mark Selection as Feature**.
- Click on the **Feature Key** drop down menu and choose **Sequencher** from the top of the list.
- Insert "Needs more evidence" into the **Feature Name:** text box.
- Set the **Feature Color:** to Red.
- Click the **OK** button to dismiss the dialog window.

In order to indicate that you have finished your review of this sample you can change its label. This will change the text color of the header, giving you a visual cue about the status of the sample. You can set up names for labels in the **User Preferences**.



- Locate contig 131 and click in the header.
- Choose the **Edit** menu and select **Label > Label 5**.



The text in the header has changed to red. You can mark the rest of the contigs in a similar fashion using different labels.

- Close the Contig Editor and Contig Chromatogram windows.

REMOVING UNWANTED DATA FROM THE TABLE

There may be instances when you wish to remove data from your table before proceeding with your analysis. In this example you will remove the samples that do not have many variants.

First sort the table so that all the sample sequences containing variants are grouped together.

- Click on the **Total** button at the bottom left of the table.

The samples with the most candidate SNPs are now grouped together at the left hand end of the table. Sample 45 and Sample 35 only have a few variants. The numerous pink Xs indicate that they have incomplete coverage over significant areas of the gene. You therefore wish to remove them from the table.

- Select the columns for individual samples 45 and 35 by clicking on 45 and then shift clicking on 35.

Reference		138	156	150	116	128	152	131	94	93	130	45	35	Total
15,517	A	g	g	g	g	g	g	g	g	r	g	g	r	12
16,277	C		y	y	y		y					y		5
17,061	G							r						1
17,158	A				r									1
17,229	G									r				1
17,353	G				k									1
17,419	C	y												1
	Total	7	6	6	6	5	4	4	3	3	3	2	1	50

- Choose **Edit >Remove from Table**.

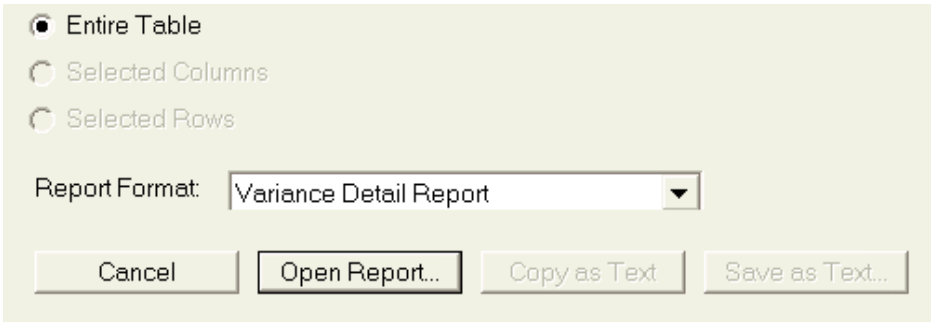
This step removes data only from the table and not from the underlying contig.

MAKING A REPORT

Now that you have reviewed some of your results in the Variance Table, you can create a report and print or export it. Sequencher provides a number of report formats. The entire table can be exported as a single entity. You can export it as individual column reports that reflect the original comparison sequences or you can report on selected rows or selected columns. You will now create a Report as if you required it for printing.

- Click on the **Reports** button on the button bar.

Sequencher will bring up the following Report dialog.



The drop down menu provides four different report options: the Variance Table Report, Individual Variance Reports, the Variance Detail Report, and the Population Report. The **Open Report…** command displays a view of each report, which you can either print or save as a PDF (Portable Document Format). The Variance Table and the Individual Variance Reports are also available to either **Copy as Text** or **Save as Text** if you want to export your data.

- Choose **Variance Detail Report** from the Report Format drop down menu.
- Click on the **Open Report…** button.

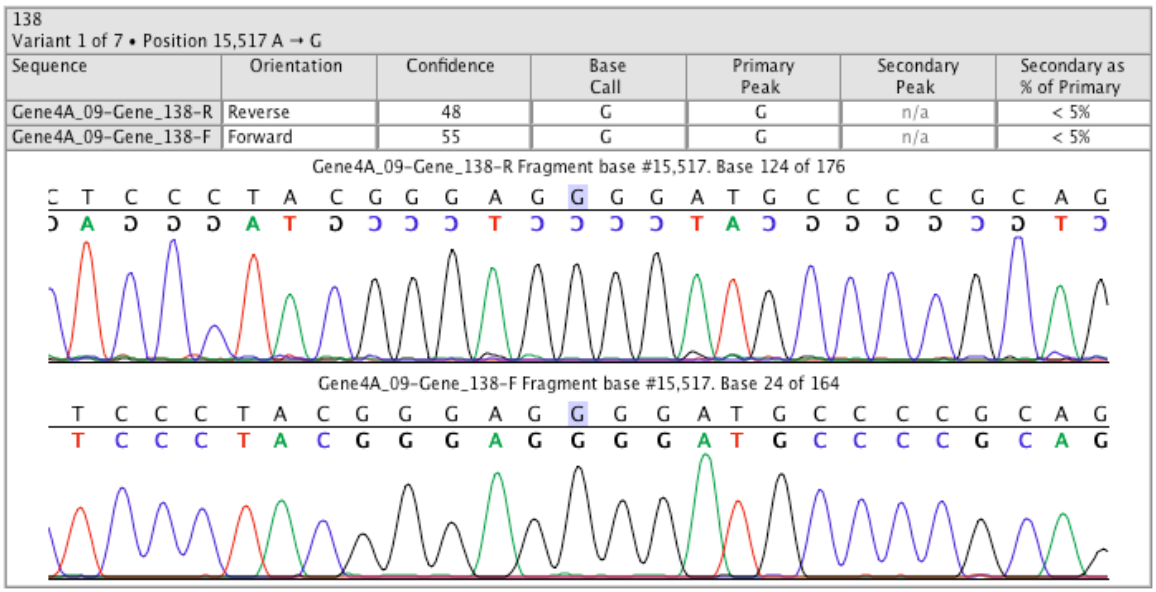
The Variance Detail Report contains a wealth of information pertaining to each variant. Sequencher reports the orientation, confidence score, and base call. The table also provides information on the exact ratio of the primary and secondary peaks in such a way that you can determine whether you are seeing a true heterozygote or an artifact of the sequencing process. Snapshots of the supporting traces (called tracelets) around the variant are included with the report.

Note: The Reporting functions are not accessible when you are running a copy of Sequencher in demo mode or using the special demonstration version of Sequencher. However, you can view an excerpt of the Variance Detail Report below.

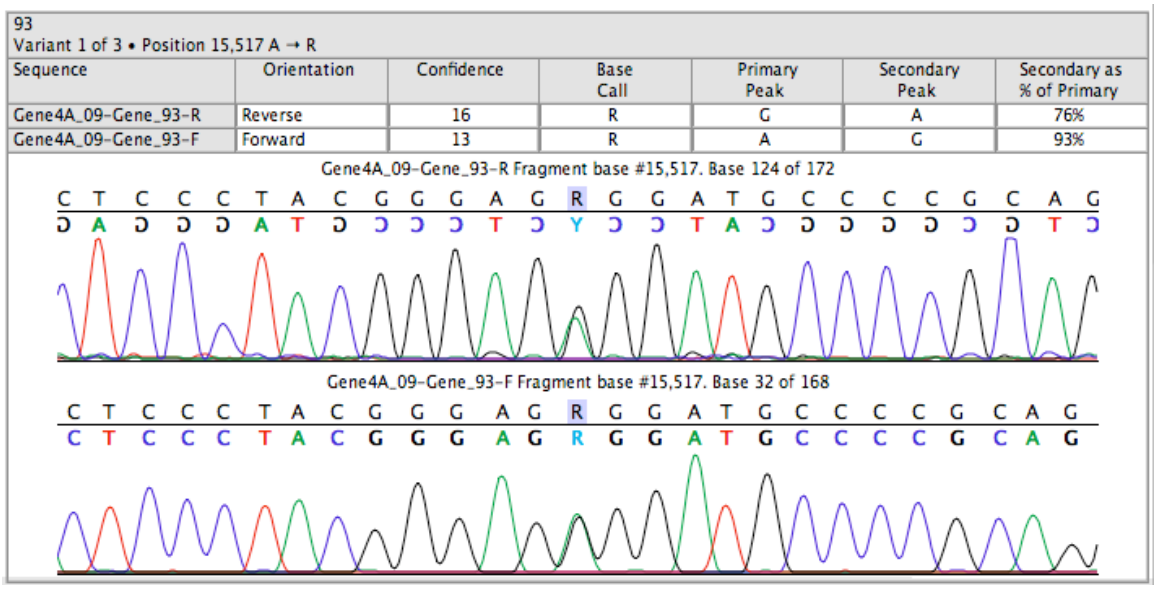
- Scroll down the **Variance Detail Report** to view the data for each sample.

The Variance Detail Report goes into much more detail than the Variance Table. Look at the section of the report dealing with sample 138, and sample 93 at base position 15,517.

For Individual sample 138 below, you will notice that the confidence scores are high for both supporting sequences, 55 for the forward orientation and 48 for the reverse. You will also notice, as you look to the columns on the right, that there is no base call for the Secondary Peak. In this sample's forward and reverse sequences the Secondary Peak is less than five percent of the Primary Peak height, and so does not meet the threshold for calling a Secondary Peak. This provides strong evidence in support of the "G" base call.

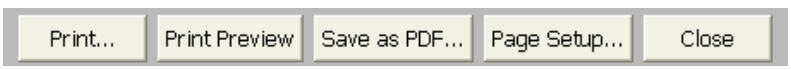


In contrast, there is a clear heterozygote for Individual sample 93 at the same base position in the report. The base call is an "R". In the forward orientation, the report tells you that there is a secondary peak at this location. It is 93% of the height of the primary peak. In addition, the report tells you that the primary peak is an "A" and the secondary peak is a "G". There is also strong evidence for the "R" heterozygote in the reverse orientation; its secondary peak is 76% of the primary. The evidence supporting a heterozygote at position 15,517 in sample 93 is even stronger when contrasted with the information for the clean base call in the same position in sample 138.



SAVING YOUR REPORT

You can save this report as a PDF if you want to archive your results.



- Click on the **Save as PDF…** button.
- Select a location and file name from the **Save as PDF File** dialog window.
- Click on the **Save** button to dismiss the window.
- Close the project without saving.
- **Quit** Sequencher.

CONCLUSION

In the example you've used in this tutorial, Sequencher's **Variance Table** immediately identified the candidate SNPs from sample sequences assembled to a Reference Sequence of 3,909 bases in length. With a few keystrokes you were able to define the Comparison Range so that you could further narrow your results to the differences that you wanted to see. You learned how to mark Features on your sequences. You learned how to label contigs providing visual cues for their status. The Review Mode and Variance Detail Report assisted in your validation of the candidate SNPs. You were able to review the underlying trace data and obtain the confidence scores for your candidates and confirm them efficiently.

To learn more about the Reference Sequence and the Variance Table see the manual or other tutorials in this series.