SEQUENCHER®

Tutorial for Windows and Macintosh

SNP Hunting

Gene Codes Corporation
525 Avis Drive, Ann Arbor, MI 48108 USA
1.800.497.4939 (USA) +1.734.769.7249 (elsewhere)
+1.734.769.7074 (fax)
www.genecodes.com gcinfo@genecodes.com

SNP Hunting

SNP Hunting

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

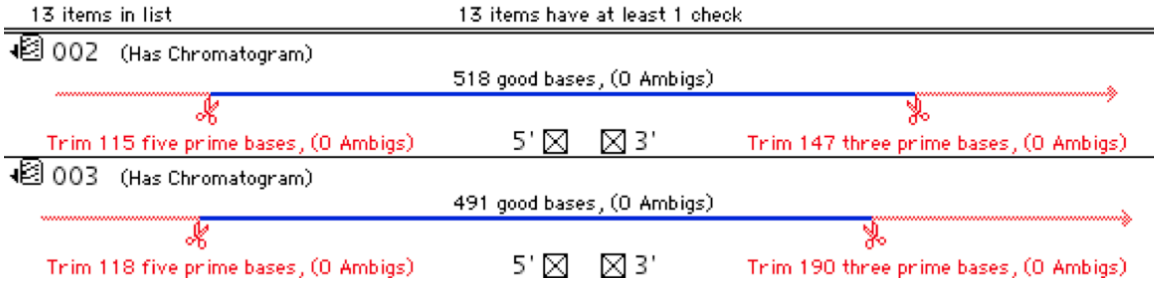
The SNP project contains 14 sequences. In addition to a reference sequence, the sequences include auto-sequencing data for one exon from different individuals.

TRIMMING LOW QUALITY DATA

The first step in analyzing data generated by an automated sequencer is trimming the data. Files from automated sequencers frequently have low quality data at the ends of the sequence. Low quality data may include miscalled bases. This can affect the outcome of the proposed assembly and so should be removed.

- On import, all of your sequences should be selected. If they are not, choose **Select** > **Select All**.
- Use **Apple+Click** on a Mac or **Ctrl+Click** on a PC to deselect the reference sequence.
- 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. The proposed trim site is indicated by a scissors icon.



13 items in list 13 items have at least 1 check

002 (Has Chromatogram)

518 good bases, (0 Ambigs)

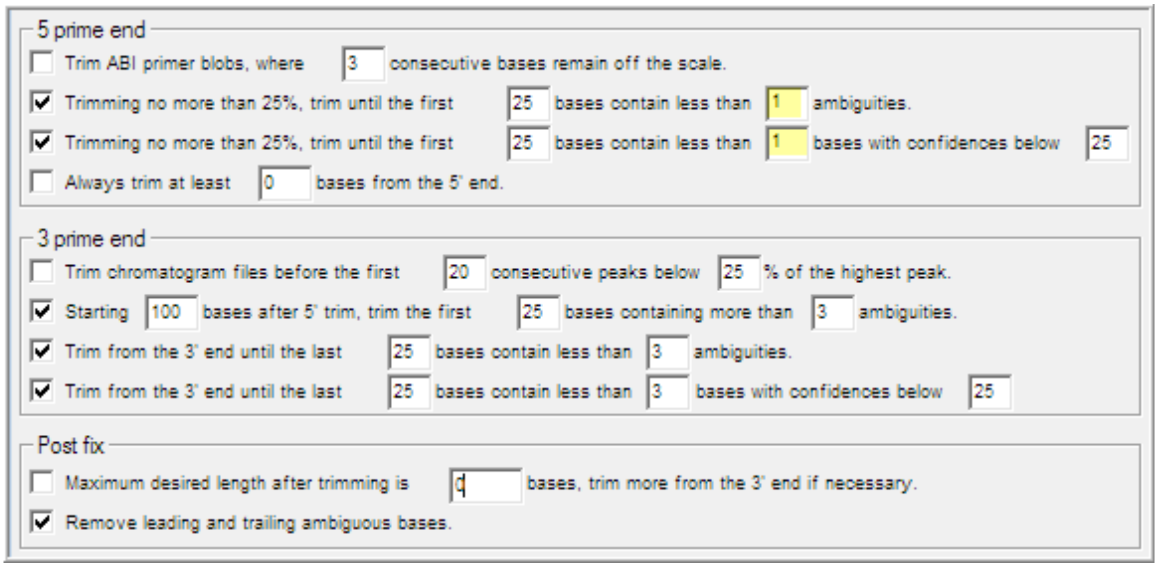
Trim 115 five prime bases, (0 Ambigs) 5' ☒ ☒ 3' Trim 147 three prime bases, (0 Ambigs)

003 (Has Chromatogram)

491 good bases, (0 Ambigs)

Trim 118 five prime bases, (0 Ambigs) 5' ☒ ☒ 3' Trim 190 three prime bases, (0 Ambigs)

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



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 1 ambiguities.
- ☑ Trimming no more than 25%, trim until the first 25 bases contain less than 1 bases with confidences below 25
- □ Always trim at least 0 bases from the 5' end.

3 prime end

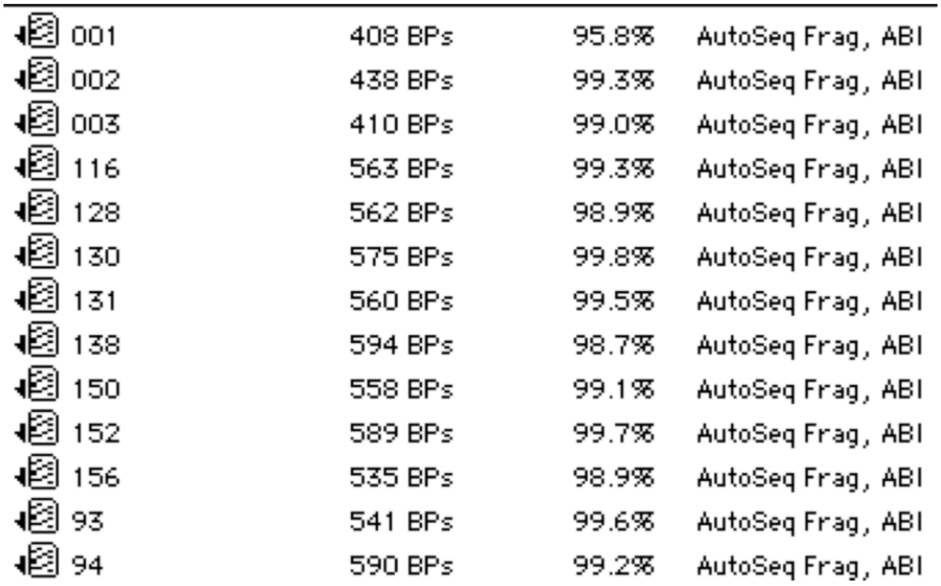
- □ Trim chromatogram files before the first 20 consecutive peaks below 25 % of the highest peak.
- ☑ Starting 100 bases after 5' trim, trim the first 25 bases containing more than 3 ambiguities.
- ☑ Trim from the 3' end until the last 25 bases contain less than 3 ambiguities.
- ☑ Trim from the 3' end 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.

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

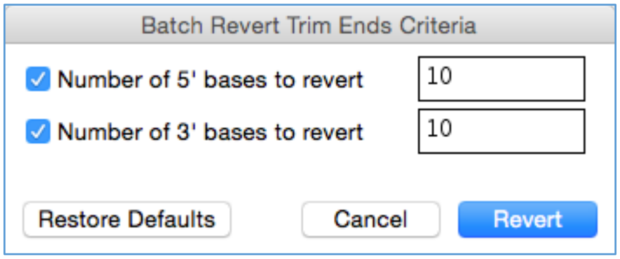
Notice the improvement of the values in the **Quality** column in the **Project Window**.



001	408 BPs	95.8%	AutoSeq Frag, ABI
002	438 BPs	99.3%	AutoSeq Frag, ABI
003	410 BPs	99.0%	AutoSeq Frag, ABI
116	563 BPs	99.3%	AutoSeq Frag, ABI
128	562 BPs	98.9%	AutoSeq Frag, ABI
130	575 BPs	99.8%	AutoSeq Frag, ABI
131	560 BPs	99.5%	AutoSeq Frag, ABI
138	594 BPs	98.7%	AutoSeq Frag, ABI
150	558 BPs	99.1%	AutoSeq Frag, ABI
152	589 BPs	99.7%	AutoSeq Frag, ABI
156	535 BPs	98.9%	AutoSeq Frag, ABI
93	541 BPs	99.6%	AutoSeq Frag, ABI
94	590 BPs	99.2%	AutoSeq Frag, ABI

Batch Reverting Trimmed Ends

After you have made the initial trim of your sequences in the **Project Window**, you may feel that the trim criteria were too stringent. You can revert all of the trim or a portion of the trim for the current sequence selection by going to the **Sequence** menu, selecting the **Batch Revert Trim Ends…** menu item, and entering the number of 5′ bases to revert and/or the number of 3′ bases to revert.



Batch Revert Trim Ends Criteria

☑ Number of 5' bases to revert 10

☑ Number of 3' bases to revert 10

The Reference Sequence

You are now ready to mark a sequence to serve as the Reference Sequence. The Reference Sequence has certain properties that 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–an especially important fact in situations where you only have a few sequences in the contig.

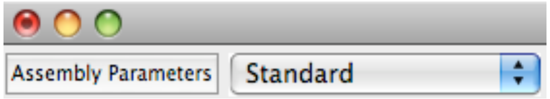
- Click on the sequence called **NC_000022** to select it.
- Choose the **Sequence** menu and ensure that **Reference Sequence** menu item is checked.

The icon of the Reference Sequence should now contain an R and is now protected from editing.

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 to create your contig.

- Choose **Select > Select All**.
- Ensure **Assembly Mode** is set to **Standar**d, like in the following picture:



Assembly Parameters: Standard

- Click on the **Assemble to Reference** button.
- Click on the **Close** button to dismiss the Assembly Completed dialog.
- Choose **Contig > Consensus Inclusively**. Note the caution window.

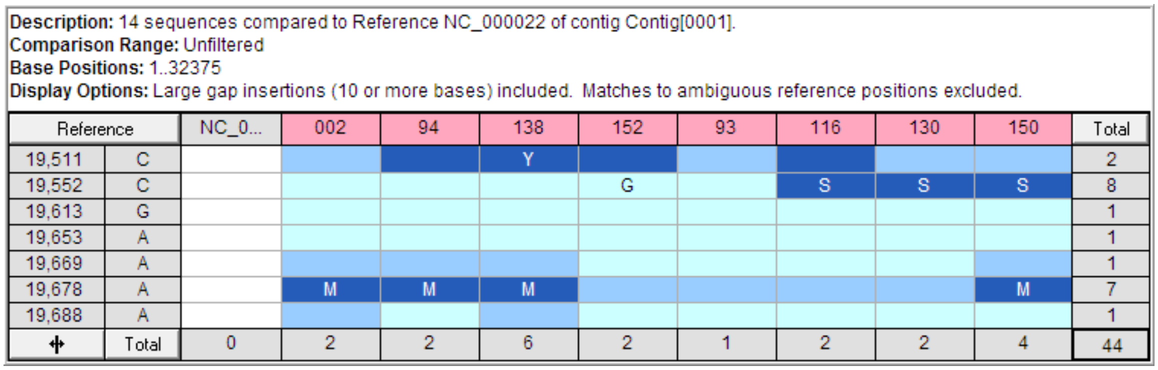
Creating a Variance Table to locate SNPs

You are now ready to create a **Variance Table** with your trimmed data and identify candidate SNPs.

The **Variance Table** provides an overview of the differences within the same contig relative to a selected primary sequence or exemplar. These differences will be your candidate SNPs. You can choose which sequence you want to use for your exemplar. Your choices are the Reference Sequence, the Consensus Sequence, or the Top Sequence (from **Sequence > Compare Bases To**).

In this tutorial, you will use the Reference Sequence as the primary sequence.

- From the **Project Window**, ensure that the contig is selected.
- From the **Sequence** menu, select the command **Compare Bases To > Reference Sequence**.



Description: 14 sequences compared to Reference NC_000022 of contig Contig[0001].
Comparison Range: Unfiltered
Base Positions: 1..32375
Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

Reference		NC_0...	002	94	138	152	93	116	130	150	Total
19,511	C				Y						2
19,552	C					G		S	S	S	8
19,613	G										1
19,653	A										1
19,669	A										1
19,678	A		M	M	M					M	7
19,688	A										1
✚	Total	0	2	2	6	2	1	2	2	4	44

Sequencher generates the **Variance Table**. Your initial view of the **Variance Table** displays the candidate SNPs for the 14 sequences in the contig. Columns are in the order that sequences appear in the contig, and rows are ordered according to the positions of the differences relative to the Reference Sequence.

Long sequence names are truncated at the default column width.

- Click once on the resize ⇹ icon button next to the bottom left Total button.

The table expands all of the columns to display the full name for each sequence.

- Click once more on the resize ⇹ icon button.

The column widths in the table change to display only a single base's width. This enables you to see more data in your viewing window. Clicking once again on the resize ⇹ icon button will return to the original display.

One sequence has a gray header. This is the Reference Sequence. You will notice that each of the remaining column headings is shaded in pink. This indicates these sequences do not cover the full comparison range of the Reference Sequence. Notice that several 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 blue shading in some cells corresponds to one of the three base call Confidence Ranges (Low, Medium or High). Darker shading indicates the low Confidence Range. You alter the values that set the boundaries for the Confidence Ranges in the Preferences.

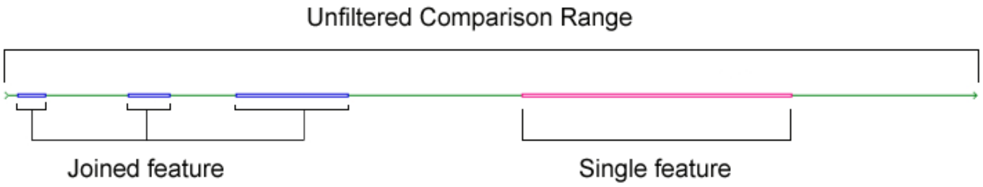
The Total cell in the bottom right corner shows that there are a total of 44 variants listed in the table.

- Move your cursor to the line that divides the columns between each sequence name until the arrow icon becomes the resize column icon.
- Drag the column divider for any individual cell and note that the column resizes to your chosen size. Or double-click to resize to the full width of the name.

Focusing on regions of interest on the Reference Sequence

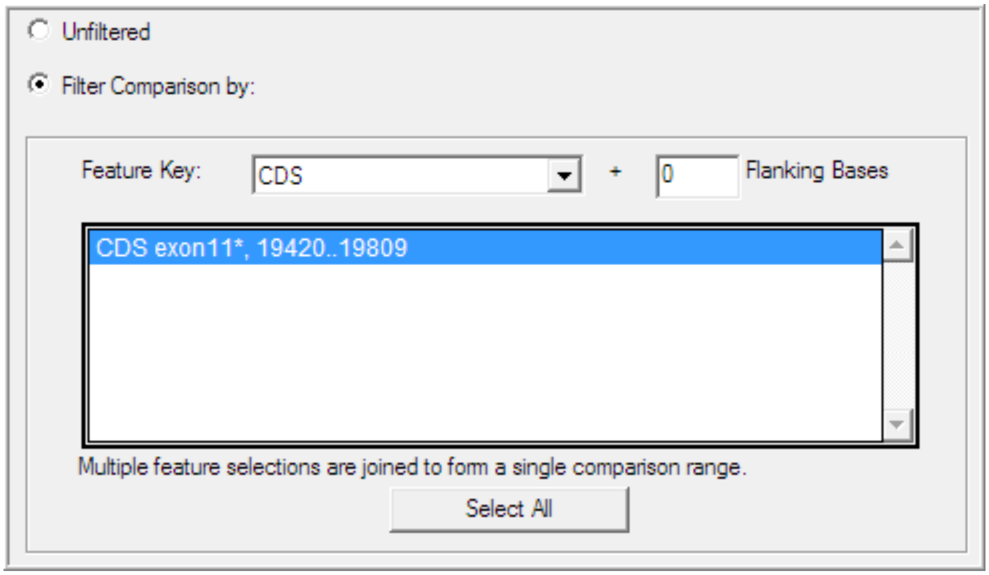
In some cases, you want to explore the entire sequence. However, if your Reference Sequence contains features such as an exon or a CDS, you can direct the **Variance Table** to focus on these features. This will reduce the amount of data you have to review. Sequences from GenBank are annotated with features 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 numbered in the primary or exemplar sequence. In the example in this tutorial, the Comparison Range is defined by the Reference Sequence from bases 1 to 32,375. This is the unfiltered Comparison Range.



You can restrict the Comparison Range by choosing the Feature used to annotate the Reference Sequence. You will use the CDS Exon 11* Feature in this tutorial.

- Go to the **Variance Table** 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 **CDS Exon 11*, 19420..19809** feature from the list box.
- Click on the **OK** button to dismiss the **Comparison Range** dialog.



○ Unfiltered

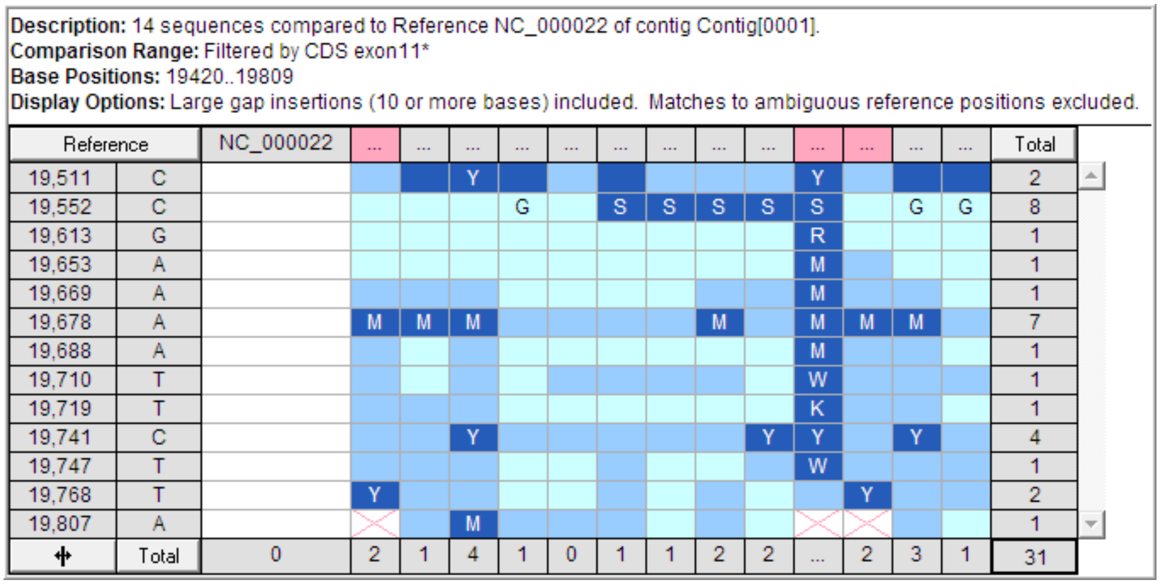
● Filter Comparison by:

Feature Key: CDS + 0 Flanking Bases

CDS exon11*, 19420..19809

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

The **Variance Table** is redrawn and the only variants in the report are now within the selected feature. The Total cell at the bottom right of the table shows that there are now only 31 variants listed in the table.



Description: 14 sequences compared to Reference NC_000022 of contig Contig[0001].
Comparison Range: Filtered by CDS exon11*
Base Positions: 19420..19809
Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

Reference		NC_000022	...	...	...	...	...	...	...	...	...	...	...	...	...	Total
19,511	C				Y							Y				2
19,552	C					G		S	S	S	S	S		G	G	8
19,613	G											R				1
19,653	A											M				1
19,669	A											M				1
19,678	A		M	M	M					M		M	M	M		7
19,688	A											M				1
19,710	T											W				1
19,719	T											K				1
19,741	C				Y						Y	Y		Y		4
19,747	T											W				1
19,768	T		Y										Y			2
19,807	A		☒		M							☒	☒			1
✥	Total	0	2	1	4	1	0	1	1	2	2	...	2	3	1	31

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**.

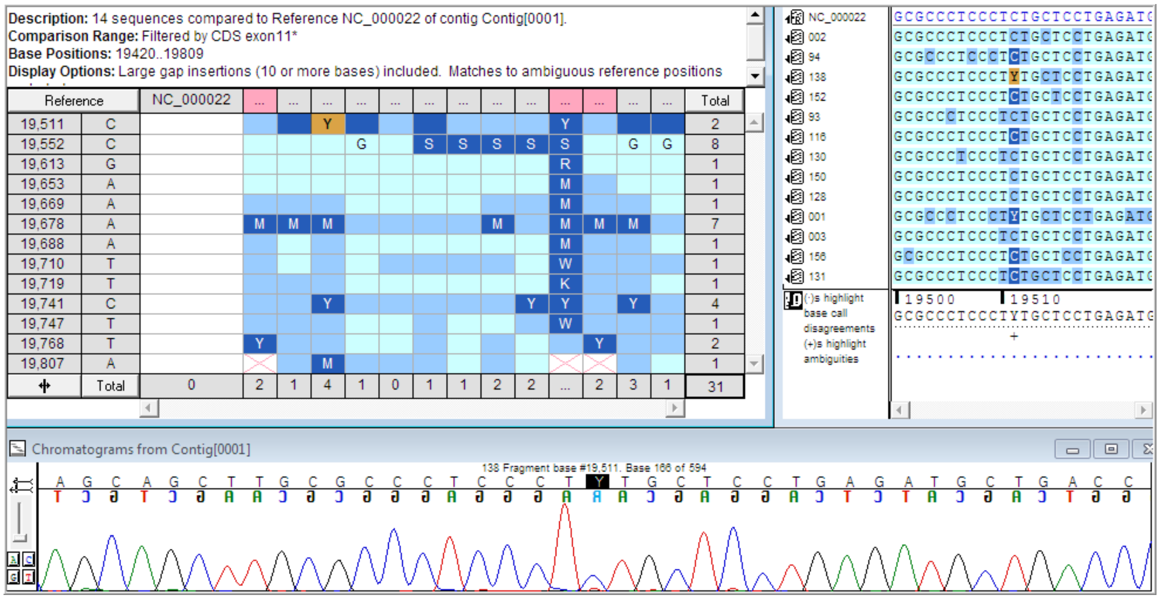
You can move quickly from one difference in the table to the next using a keyboard shortcut.

- Locate base position 19,511 and place your cursor in the first cell 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**.

- Go to the button bar and click on the **Review** button.
- Arrange the windows to suit your viewing preference.

The base call in this position in Sample 138 is a "Y". There is a black + below the consensus sequence line indicating that there is an ambiguity here. You will notice that the background color of the cell containing the "Y" is dark blue indicating that the base call is low quality, which is typical for ambiguous base calls. The next step is to check the chromatogram underlying this base call.



Description: 14 sequences compared to Reference NC_000022 of contig Contig[0001].
Comparison Range: Filtered by CDS exon11*
Base Positions: 19420..19809
Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions

Reference		NC_000022	...	...	...	...	...	...	...	...	...	...	...	...	...	Total
19,511	C				Y							Y				2
19,552	C					G		S	S	S	S	S		G	G	8
19,613	G											R				1
19,653	A											M				1
19,669	A											M				1
19,678	A		M	M	M					M		M	M	M		7
19,688	A											M				1
19,710	T											W				1
19,719	T											K				1
19,741	C				Y						Y	Y		Y		4
19,747	T											W				1
19,768	T		Y										Y			2
19,807	A				M											1
✣	Total	0	2	1	4	1	0	1	1	2	2	...	2	3	1	31

NC_000022	GCGCCCTCCCTCTGCTCCTGAGATG
002	GCGCCCTCCCTCTGCTCCTGAGATG
94	GCGCCCTCCCTCTGCTCCTGAGATG
138	GCGCCCTCCCTYTGCTCCTGAGATG
152	GCGCCCTCCCTCTGCTCCTGAGATG
93	GCGCCCTCCCTCTGCTCCTGAGATG
116	GCGCCCTCCCTCTGCTCCTGAGATG
130	GCGCCCTCCCTCTGCTCCTGAGATG
150	GCGCCCTCCCTCTGCTCCTGAGATG
128	GCGCCCTCCCTCTGCTCCTGAGATG
001	GCGCCCTCCCTYTGCTCCTGAGATG
003	GCGCCCTCCCTCTGCTCCTGAGATG
156	GCGCCCTCCCTCTGCTCCTGAGATG
131	GCGCCCTCCCTCTGCTCCTGAGATG
(·)s highlight base call disagreements (+)s highlight ambiguities	19500 19510 GCGCCCTCCCTYTGCTCCTGAGATG +

Chromatograms from Contig[0001]

138 Fragment base #19,511. Base 166 of 594

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

You will now see the contig bases and chromatogram data for the sample. The base call is a "Y". This base call is not well supported by the underlying chromatogram.

- Use the right arrow key on your keyboard to view other samples at this position.

All of the samples at this position have a weak "C" peak. Underneath the "C" is a small "T" peak. This "T" is probably noise. It is not likely that there is a SNP at this position.

EDITING DATA FROM THE VARIANCE TABLE

Although the **Variance Table** is used primarily for display and review of your data, you can make use of its links to the underlying data in another way.

- Click once on the resize ✣ icon button next to the bottom left Total button to see the sample names again.
- Double-click in the **Variance Table** in sample **138** at the "Y" in base position **19,511**.

You will now see the contig bases and chromatogram data for the selected sample cell. The base call is a "Y", and you can see that there are two small peaks (relative to the surrounding data) at this position.

You decide to edit the "Y" base call, because the trace data in this sample matches "C" calls in the majority of the other samples at this position.

- Type a **C** into the cell.

The base now matches the Reference Sequence. Since you have removed the conflicting base, the cell is empty. Any edits you make are reflected immediately due to the dynamic link between the windows. They will also be updated in the Totals at the left and bottom of the Table.

- In the **Variance Table**, find the second "Y" call at position 19,511 in sample 001, and edit the call to a "**C**".
- Click on the **Refresh** button in the **Variance Table** button bar.

Note that the row for position 19,511 is no longer included in the table, because there are no longer samples with differences at this position.

- Click the **Review** button to exit Review Mode and 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 further with your analysis. In this example, you will remove the samples that do not have any 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 candidate SNPs are now grouped together at the left hand end of the table.

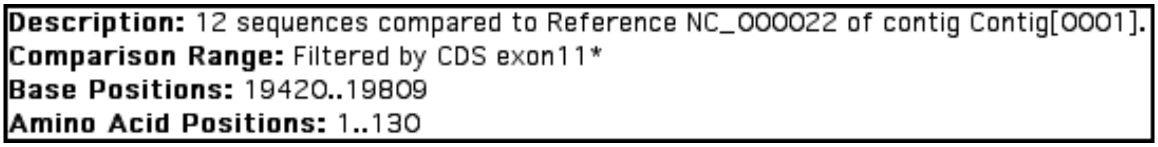
- **Shift-click** in the column headers of the last two columns with no variants, sample **93** and the **reference sequence**.
- Choose **Edit > Remove from Table**.

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

Viewing the impact of candidate SNPs on translated regions

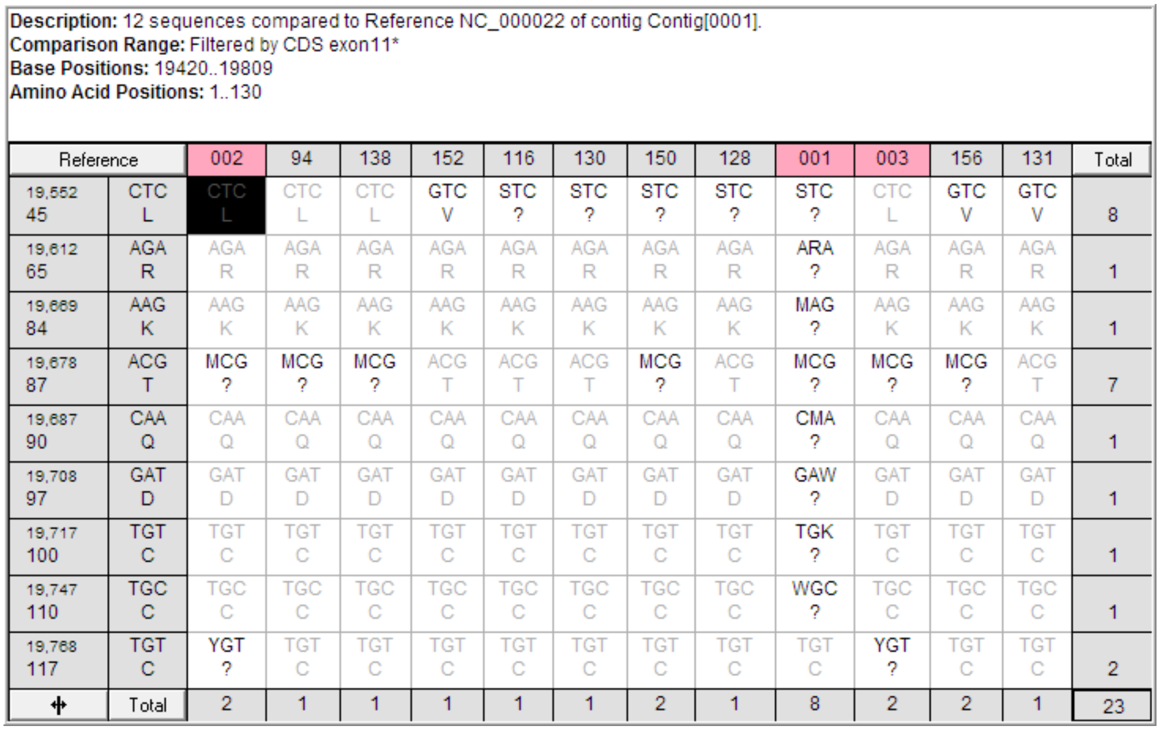
It is possible to view the coding impact of any difference in the **Variance Table**. The **Translated Variance Table** is the sister table to the **Variance Table**. This directs your attention to the putative effects a base difference has on a codon and its associated amino acid. This is most informative when your Reference Sequence has a feature that can be translated.

- Click on the **Translation** button on the **Variance Table** button bar. Note that this Sister Table shares the same CDS filter you chose for the Bases table's Comparison Range.



Description: 12 sequences compared to Reference NC_000022 of contig Contig[0001].
Comparison Range: Filtered by CDS exon11*
Base Positions: 19420..19809
Amino Acid Positions: 1..130

The **Translated Variance Table** shares many features with the **Variance Table**. You can move around from cell to cell using arrow keys. You can move across gaps using the Option/Alt + arrow key combination. The pink headers and Xs have the same function. You can also view the underlying base calls and trace data by using the Review mode.



Description: 12 sequences compared to Reference NC_000022 of contig Contig[0001].
Comparison Range: Filtered by CDS exon11*
Base Positions: 19420..19809
Amino Acid Positions: 1..130

Reference		002	94	138	152	116	130	150	128	001	003	156	131	Total
19,552 45	CTC L	CTC L	CTC L	CTC L	GTC V	STC ?	STC ?	STC ?	STC ?	STC ?	CTC L	GTC V	GTC V	8
19,612 65	AGA R	AGA R	AGA R	AGA R	AGA R	AGA R	AGA R	AGA R	AGA R	ARA ?	AGA R	AGA R	AGA R	1
19,669 84	AAG K	AAG K	AAG K	AAG K	AAG K	AAG K	AAG K	AAG K	AAG K	MAG ?	AAG K	AAG K	AAG K	1
19,678 87	ACG T	MCG ?	MCG ?	MCG ?	ACG T	ACG T	ACG T	MCG ?	ACG T	MCG ?	MCG ?	MCG ?	ACG T	7
19,687 90	CAA Q	CAA Q	CAA Q	CAA Q	CAA Q	CAA Q	CAA Q	CAA Q	CAA Q	CMA ?	CAA Q	CAA Q	CAA Q	1
19,708 97	GAT D	GAT D	GAT D	GAT D	GAT D	GAT D	GAT D	GAT D	GAT D	GAW ?	GAT D	GAT D	GAT D	1
19,717 100	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGK ?	TGT C	TGT C	TGT C	1
19,747 110	TGC C	TGC C	TGC C	TGC C	TGC C	TGC C	TGC C	TGC C	TGC C	WGC ?	TGC C	TGC C	TGC C	1
19,768 117	TGT C	YGT ?	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	TGT C	YGT ?	TGT C	TGT C	2
✥	Total	2	1	1	1	1	1	2	1	8	2	2	1	23

Notice the difference in numbering between the two tables. The **Variance Table** reports the actual base position of the candidate SNP. For example, the **Variance Table** reports a SNP at position 19,613. The **Translated Variance Table** reports the position of the first base in the codon containing the SNP, which is in this case 19,612. The numbering for both tables is relative to the Reference Sequence.

If you want to learn more about the **Translated Variance Table**, refer to the manual and the associated tutorial.

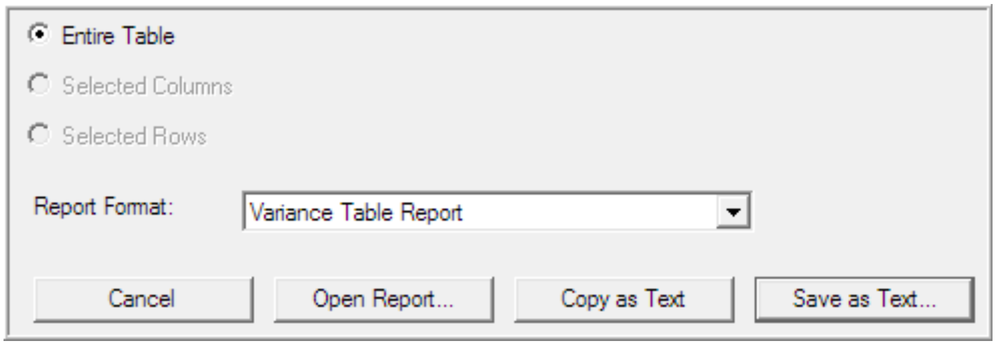
Making a Report

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. It can be exported as individual column reports that reflect the original comparison sequences or you can export selected rows or selected columns. You will now create a Report as if you required it for printing.

- Click on the **Comparison Range** button on the **Compare Bases to Reference Sequence** button bar.
- Check the **Filter Comparison by:** radio button.

- Click on the **Feature Key:** drop-down menu and choose **gene**.
- Select the feature **HPS4 gene, 1..32375**.
- Click on the **OK** button to dismiss the **Comparison Range** dialog.
- Click on the **Reports** button on the **Variance Table** button bar.

Sequencher will bring up the following Report dialog:



● Entire Table

○ Selected Columns

○ Selected Rows

Report Format: Variance Table Report

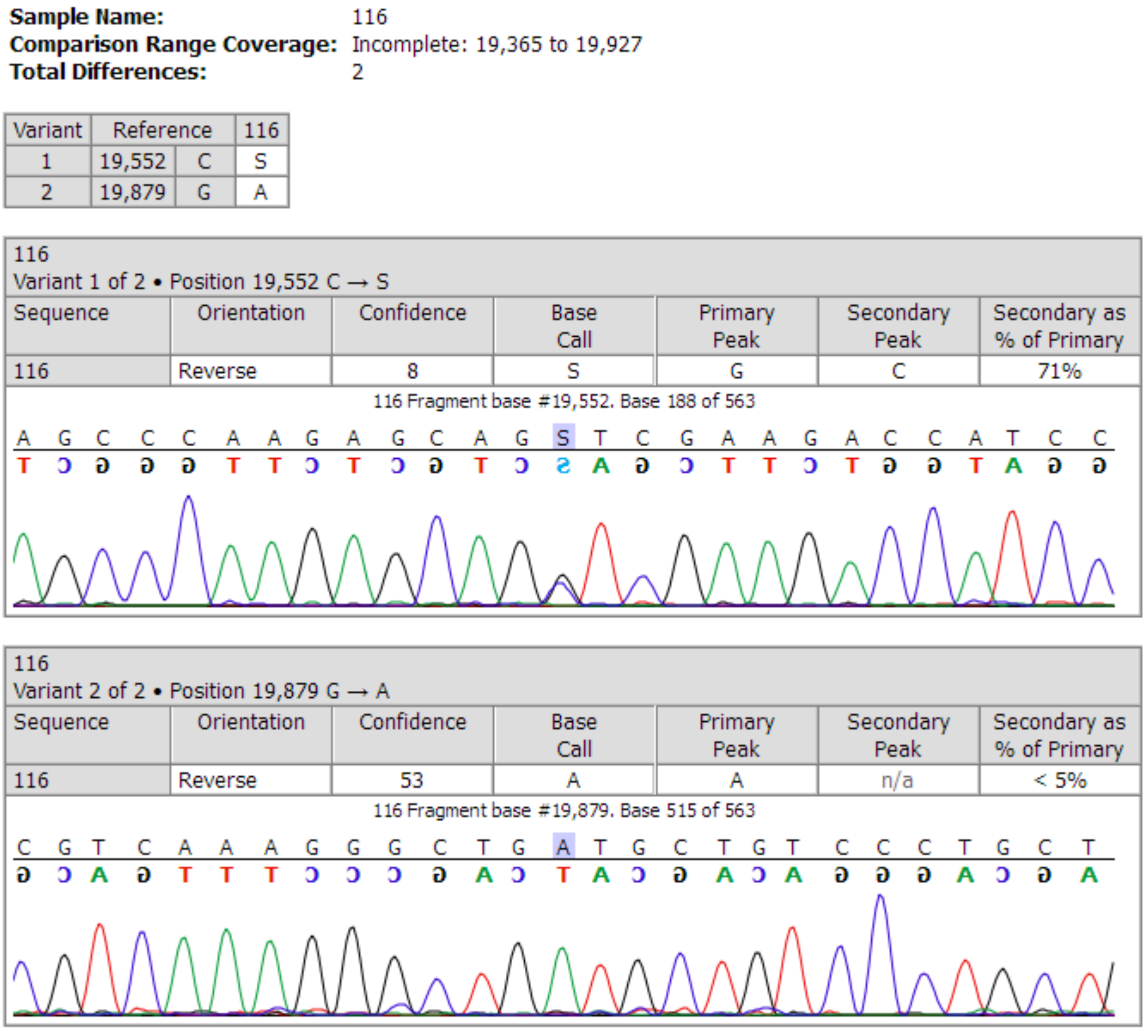
The Report Format: drop-down menu provides four different report options: the Variance Table Report, Individual Variance Reports, the Variance Detail Report, and the Population Report. The **Open Report...** command displays a view of each report, which you can either print or save as a PDF (Portable 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.

The Variance Detail Report contains a wealth of information pertaining to each variant. It will contain the exact Confidence value, the base call, an extract of the trace (this is called a tracelet) around the variant, as well as information on any potential secondary peak.

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

*Note: The Reporting functions are not accessible when you are running the Viewer version of **Sequencher**.*

- Scroll down the Variance Detail Report to view the data for sample 116.



Sample Name: 116
Comparison Range Coverage: Incomplete: 19,365 to 19,927
Total Differences: 2

Variant	Reference		116
1	19,552	C	S
2	19,879	G	A

116
Variant 1 of 2 • Position 19,552 C → S

Sequence	Orientation	Confidence	Base Call	Primary Peak	Secondary Peak	Secondary as % of Primary
116	Reverse	8	S	G	C	71%

116 Fragment base #19,552. Base 188 of 563

A G C C C A A G A G C A G S T C G A A G A C C A T C C
T ɔ ә ә ә T T ɔ T ɔ ә T ɔ ƨ A ә ɔ T T ɔ T ә ә T A ә ә

116
Variant 2 of 2 • Position 19,879 G → A

Sequence	Orientation	Confidence	Base Call	Primary Peak	Secondary Peak	Secondary as % of Primary
116	Reverse	53	A	A	n/a	< 5%

116 Fragment base #19,879. Base 515 of 563

C G T C A A A G G G C T G A T G C T G T C C C T G C T
ә ɔ A ә T T T ɔ ɔ ɔ ә A ɔ T A ɔ ә A ɔ A ә ә ә A ɔ ә A

Sequencher reports the orientation, confidence score, and base call for the variants in this sample. The table also provides information on the exact ratio of the primary and secondary peaks in such a way that you can determine whether these are true heterozygotes or artifacts of the sequencing process. Snapshots of the supporting traces are included with the report.

You will notice that for position 19,879 the Confidence Value is 53. You will also notice that there is no base call for the secondary peak. The secondary peak is less than five percent of the primary peak height. This provides strong evidence in support of the pure base call.

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.
- Click on the **Save** button to dismiss the dialog.
- 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 16,936 bases in length. With a few keystrokes, you were able to define the Comparison Range so that you could further narrow your results to just the differences you wanted to see. The Review Mode and Variance Detail Report supported your validation of the candidate SNPs. You were able to make edits from the **Variance Table**. You were able to review the underlying trace data and obtain the Confidence Values for your candidates and confirm them efficiently.