

SEQUENCHER®

Tutorial for Windows and Macintosh

Comparative Sequencing

© 2017 Gene Codes Corporation

Gene Codes Corporation

T C A G E N E
A G T C O D E S

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

Comparative Sequencing

Getting started	3
The Reference Sequence	3
Marking features on a Reference Sequence	4
Assembling your data.....	5
Using the Summary View to explore your sequences	5
Using a Variance Table to analyze sequence differences.....	7
Focusing on Regions of Interest in the Reference Sequence	7
Removing unwanted data from the table	9
Making a Report	10
Saving your report	11
Conclusion	11

Comparative Sequencing

Comparative sequencing is a comprehensive term used to describe many types of sequencing problems. In some instances, you focus on the similarities between samples, sequences, strains or species and, in other instances, you focus on the differences between them. In this tutorial, you will analyse a group of 116 bacterial sequences. You will first identify the differences between the individual samples. Then you will create a Population report that provides a description of the frequency of the differences in the sample population.

The differences between similar sequences may represent SNPs, polymorphisms, mutations, or bases that require editing in order to be resolved. **Sequencher** provides a number of tools and functions that help you find the differences between selected sequences. It also enables you to summarize all of the differences between each sequence in your project and a common Reference Sequence using **Sequencher's Variance Table**.

GETTING STARTED

In this tutorial, you will use the **Summary** view and the **Variance Table** to identify and report on differences between sample sequences. First, you need to open a project.

- Launch **Sequencher**.
- Go to the **File** menu and select **Import > Sequencher Project...**
- Navigate to the **Sample Data** folder inside the **Sequencher** application folder.
- Choose the project called "**Comparative Sequencing**" and select **Open**.

The project contains 116 bacterial sequences. One sequence will be used as a reference sequence. The remaining sequences are samples for your analysis. The samples represent a region of the gene coding for recombinase A from different strains of *Haemophilus influenzae*. You will be looking for differences between the sequences. You will then group samples with same set of differences.

The samples in this project have already been trimmed and edited. They are ready for the next step in your analysis.

THE REFERENCE SEQUENCE

You are now ready to mark a sequence to serve as the Reference Sequence and annotate it with features. The Reference Sequence function sets the base numbering and the orientation of the contig you are about to assemble. This allows you to reference a variant in relation to a fixed position. The Reference Sequence acts as a standard against which the sample sequences are compared. The Reference Sequence does NOT contribute to the consensus sequence.

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

You will notice that the icon of the Reference Sequence now contains an R. The sequence is now protected from editing. If you try to edit the Reference Sequence a warning window will appear.



Reference Sequence

MARKING FEATURES ON A REFERENCE SEQUENCE

You can annotate your Reference Sequence with features representing regions with biological functions. There are two ways to annotate your sequence. The first method is simply to select some bases and use the **Mark Selection As Feature** from the **Sequence** menu. The example you are about to work through demonstrates the use of the **Feature Editor**. This is the tool to use when you have several features to annotate.

- Double-click on the **Reference Sequence** icon to open its **Sequence Editor**.
- Choose the **Sequence** menu and select **Edit Features...**
- Click on the **Add** button.
- From the **Feature Key:** list, choose **CDS**.
- Append **RecA** to CDS in the **Feature Name:** input field.
- Type the following numbers into the **Feature Location** input fields: **1, 425**.

The screenshot shows the 'Feature Editor' dialog box. It has a 'Feature Key:' dropdown menu set to 'CDS'. The 'Feature Name:' text box contains 'CDS RecA'. Below this is a list box containing 'CDS RecA'. To the right, there is a checked checkbox for 'Display Feature in Editors'. Below that, the 'Feature Location:' is set to '1 : 425'. There is a 'Which strand?' section with an unchecked 'Complement' checkbox. A 'Display Style' dropdown menu is also present. At the bottom left are 'Add' and 'Remove' buttons. At the bottom right is a 'Style Preview' box showing the sequence '91 XXXXCGGATACTTGCAAGCXXXX' with 'ArgIleLeuAlaSer' highlighted below it.

Repeat these steps to add a second feature:

- Click on the **Add** button.
- From the **Feature Key:** drop-down menu, choose **misc_feature**.
- Replace **misc feature** with **ATP Site** in the **Feature Name** input field.

- Check on **Display Feature in Editors**.
- Type the following numbers into the **Feature Location** input fields: 207, 225.
- From the **Display Style** drop-down menu, select **Cyan**.
- Click on the **Done** button.
- Close the Reference Sequence **Sequence Editor** window.

Note: If you prefer that feature keys be listed alphabetically rather than hierarchically in the **Feature Key** drop-down menus, you can set that preference in **User Preferences** via the **Window** menu. Choose the **Feature, Motif** section of user preferences, click on the button labeled **Define Feature Key Default Styles...**, and then click on the **Alphabetical** radio button towards the bottom of the dialog.

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. If you are dealing with sequences downloaded from GenBank, files from automated sequencers or files from colleagues, you can assemble them into one contig.

The data you are using is finished sequence. You do not need to trim off any low quality data. You will use the default settings for the **Assembly Parameters**.

- Choose **Select > Select All** and click on the **Assemble to Reference** button.
- Click the **Close** button to dismiss the **Assembly Completed** dialog.

When you are exploring your data and looking for differences it is better to use **Consensus Inclusively** since this will expose any differences between sequences. If you do not have this consensus calculation set you will need to do so now.

- Choose **Contig > Consensus Inclusively**. Note the caution window.
- Double-click on the contig icon to open the **Overview**.

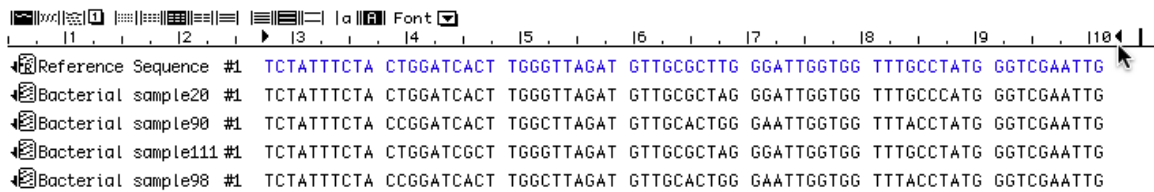
USING THE SUMMARY VIEW TO EXPLORE YOUR SEQUENCES

The **Summary View** is a report that can demonstrate the similarities and the differences between assembled sequences. You can control the display of this report in order to optimize your view of the data.

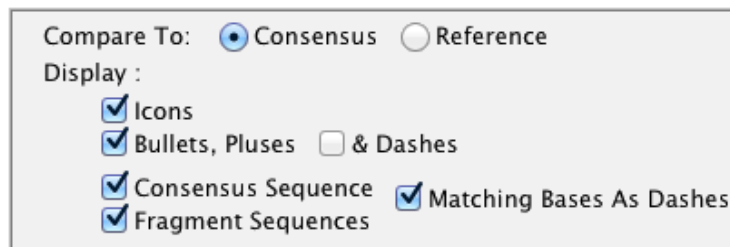
- From the **Contig Editor** button bar, click on the **Summary** button.
- Click on the **Ruler** button.
- From the **File** menu, select **Page** or **Print Setup** and choose the **Landscape** (or appropriate) setting.
- Click on the **OK** button to close the **Page Setup** window.

You can maximize the amount of data per line by using a smaller font size and a wider line setting.

- Choose font size **10** from the **Font** drop-down menu.
- Click on the right hand margin marker and move it as far right as it will move.

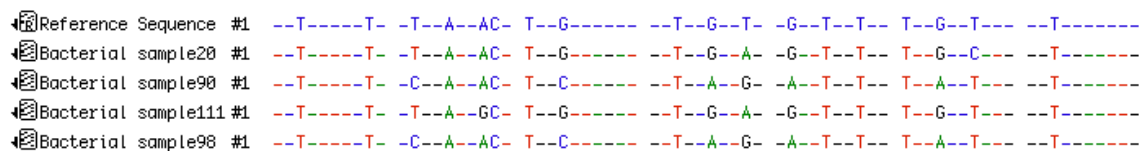


- Choose **View > Display Color Bases** and ensure that this menu item is checked.
- Click on the **Options** button.
- Match the **Display** settings to the following image:



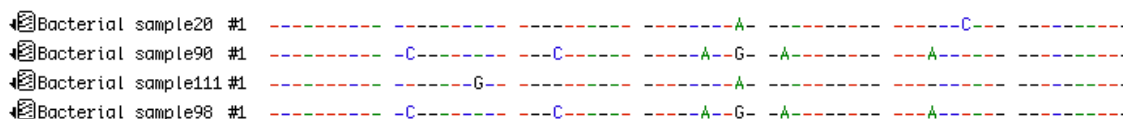
- Click on the **Compare To: Consensus** radio button.
- Click **OK** to close the **Options** window.

The report now shows only the disagreements. You can toggle between the two displays by checking or un-checking the **Matching Bases As Dashes** box in the **Options** section. Similarly, you can display the protein translations for the samples in the assembly.



Sequencher allows you to look at the differences compared to your Reference Sequence or the Consensus.

- Click on the **Options** button.
- Click on the **Compare To: Reference** radio button.
- Click **OK**.



The **Summary View** gives you a very specific view where the sequences are arrayed in lines so you can review either the similarities or the differences between the sequences in your comparative study.

Note the difference between the two **Compare To:** views. **Compare to Consensus** displays every base at a position where a difference exists. **Compare to Reference** only displays a base where a sequence differs from the Reference.

USING A VARIANCE TABLE TO ANALYZE SEQUENCE DIFFERENCES

The **Variance Table** provides an overview of the differences in your data relative to a selected primary sequence or exemplar. You are now ready to create a **Variance Table** with your data and identify differences.

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

Sequencher generates the **Variance Table**. Your initial view of the **Variance** b displays the differences for the 116 sequences. Each column represents a sample and each row is ordered according to the positions of the differences relative to the Reference Sequence.

Description: 116 sequences compared to Reference Reference Sequence of contig Contig[0066].
 Comparison Range: Unfiltered
 Base Positions: 1..425
 Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

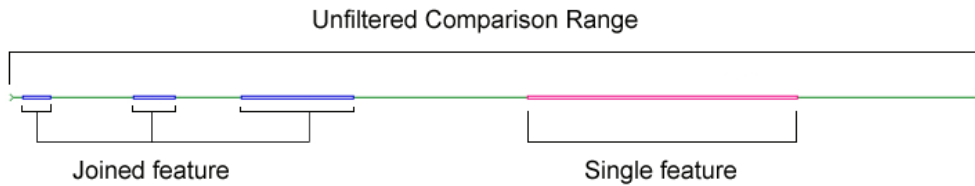
Reference		Refer...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Total
3	T									2
9	T									2
12	T			C		C			C	28
15	A									2
18	A				G					3
19	C									2
21	T									2
24	G			C		C			C	53
33	T									5
36	G			A		A			A	69
39	T		A	G	A	G	A	A	G	103
+	Total	0	5	12	6	20	3	1	12	1206

The Total cell in the bottom right corner shows that there are a total of 1206 variants listed in the table.

FOCUSING ON REGIONS OF INTEREST IN THE REFERENCE SEQUENCE

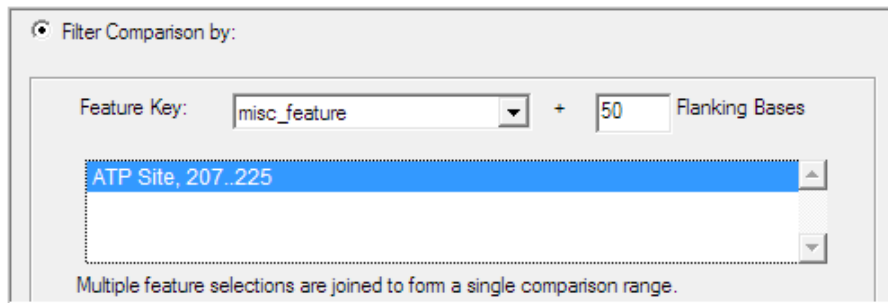
In some cases, you may want to explore the entire length of the Reference Sequence. However, if your Reference Sequence contains features such as an exon or a CDS, you can direct the **Variance Table** to focus on just these features. Sequences from GenBank are annotated using Feature Keys, a (a standardized method of referring to biologically important regions). GenBank annotations are listed in a Feature Table that is read by **Sequencher** when the sequence is imported into a project. You created GenBank annotations on the Reference Sequence at the beginning of this tutorial.

The default **Comparison Range** is defined by the entire Reference or exemplar sequence. In the example in this tutorial, the **Comparison Range** is defined by the Reference Sequence from bases 1 to 425.



You can restrict the **Comparison Range** by choosing one of the **Feature Keys** used to annotate the sequence.

- 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 **misc_feature**.
- Type **50** into the **Flanking Bases** number box.
- Click on the **OK** button to dismiss the **Comparison Range** dialog.



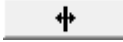
The **Variance Table** is redrawn. Notice that the only variants in the table are now within the feature you selected and the flanking bases. The **Total** cell at the bottom right of the table shows that there are now only 210 variants listed in the table.

The **Variance Table** is linked to the underlying contig. If your data contained chromatograms you would be able to view this together with the **Contig Editor** in the **Review** mode. You can read more about this in the manual and other tutorials in this series.

Description: 116 sequences compared to Reference Reference Sequence of contig Contig[0066].
 Comparison Range: Filtered by ATP Site plus 50 flanking bases
 Base Positions: 157..275
 Display Options: Large gap insertions (10 or more bases) included. Matches to ambiguous reference positions excluded.

Reference	Refer...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Bacter...	Total	
159	A								6	
168	T								6	
171	A								8	
177	C								2	
180	A								9	
183	T								2	
186	T				C				21	
198	A		T					T	7	
199	G								2	
207	T				A				6	
210	T								8	
✚	Total	0	0	1	0	5	0	0	1	210

The names of the sequences are quite long and 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.

- Click once again on the resize icon button  to return to the original display.

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 differences compared to the Reference Sequence are now grouped together at the left hand end of the table. Some samples have no differences. You may remove them from the table.

- Locate the first column that contains no differences. Click on the header of this column.
- Scroll the table to the right.
- **Shift+click** on the header of the last column in the table. You have now selected a range of columns.
- Choose **Edit > Remove From Table**.

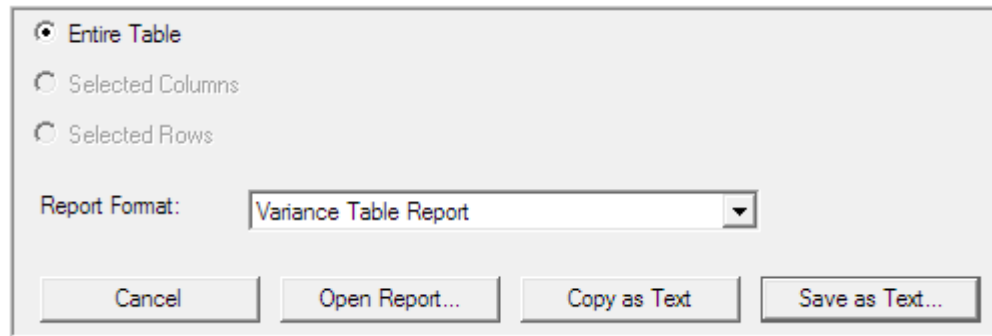
This step removes data only from the table, not from the underlying contig.

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 export 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: **Variance Table Report**, **Individual Variance Reports**, **Variance Detail Report**, and **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 **Individual Variance Reports** are also available to either **Copy as Text** or **Save as Text...** if you want to export your data.

The **Population Report** is a unique report that groups together samples with identical sets of differences. You can use this report in a number of ways for instance if you are looking for sequences with the same alleles.

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

Note that if you are following this tutorial using the Viewer demo version of **Sequencher**, you cannot generate a report but excerpts from the report follow.

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

The **Population Report** groups the samples from the **Variance Table** that share the same pattern of variants. The different sections of the **Population Report** provide you with separate summaries of your selected data. One section of the **Population Report** tells you how many variants there are in the group and how many samples share that set of variants. The subsequent section describes the frequency of a variant as a percentage of the total number of samples. The first example of a unique sequence (alphabetically) is the sample that defines the name of the group. For instance, in the excerpt from the **Population Report** below, samples

that share the same differences from the reference as Bacterial Sample 62 are placed in a group called "Bacterial sample62 - like".

- Scroll down the Population Report to view the data for each Bacterial sample62 - Like.

Bacterial sample62 - Like		
Frequency	2.94%	
Variants	20	12 C, 24 C, 36 A, 39 G, 42 A, 54 A, 78 C, 84 A, 96 T, 135 A, 186 C, 252 A, 258 A, 270 A, 288 T, 291 T, 294 C, 366 G, 390 G, 411 A
Samples	2	Bacterial sample62, Bacterial sample63

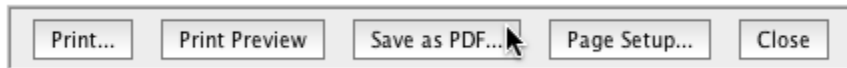
Nonparticipating data include any samples that do not span the length of the chosen **Comparison Range**, in this case defined by the **Feature Key** selected earlier.

Participating Data: 33 population groups consisting of 68 samples
Nonparticipating Data: 0 samples dropped from the report due to incomplete comparison range coverage

SAVING YOUR REPORT

You can save this report as a PDF if you want to archive your results. Perform the following steps:

- Click on the **Save as PDF...** button in the **Population Report** button bar.



- Select a location and file name from the **Save PDF File** dialog.
- 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 Summary View** and **Variance Table** immediately identified the differences from sample sequences assembled to a Reference Sequence of 425 bases in length. You learned how to mark **Features** on your sequences. You used the **Summary Report** to look at regions of similarity and regions where differences appeared. You used some of the **Summary View Options** to control the display. With a few keystrokes you were able to generate a **Variance Table**. You defined a sub-region based on a feature in your Reference Sequence. You then displayed this more focused region. The **Population Report** helped you to sort your samples into population groups that share the same set of variants.

If you want to learn more about these features, see the associated tutorials and the **Sequencher** manual.