SEQUENCHER™

Tutorial for Windows and Macintosh

Reference Sequence

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

Reference Sequence

Reference Sequence

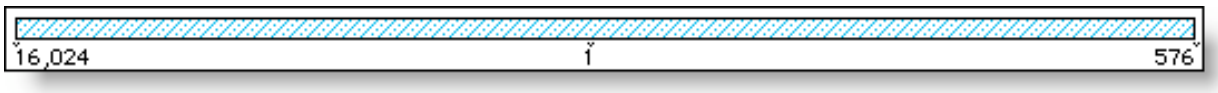
There are many cases where you might want to align your data to a Reference. In the Demo Tour Guide, we demonstrated how to use the Reference Sequence function in Sequencher to check a clone. In this case, we will align sequencing data from the HV1 and HV2 hyper variable regions of the human mitochondrial genome to the Cambridge Reference Sequence. This is a standard practice in forensic labs where mitochondrial sequence facilitates human identification.

The Reference is particularly useful for characterizing SNPs in aligned sequences. The Reference Sequence controls the base numbering and the orientation of the overall contig. By providing consistent base numbering, the Reference Sequence allows you to reference a SNP by a given base position without concern for the effect of upstream insertions or deletions. Unlike other fragments in a contig, the Reference Sequence does not contribute to the consensus calculation, and it is immune to edits in the consensus.

This tutorial does not describe all the features for mitochondrial analysis, so if you would like more information on SEQUENCHER's capabilities for forensic or reference-comparison work, please follow this tutorial with the Mitochondrial DNA Typing Tutorial.

Import and Configure Reference Sequence

The HV1 and HV2 regions of the mitochondrial genome lie next to each other, but across from the origin, base 1, of the genome. The numbering of the region of interest from 5′ to 3′ starts counting up from base position 16,024 until it reaches the end of the genome, 16,569. At this point the numbering starts again at 1. In Sequencher you can create a Reference Sequence that mimics the numbering in the control region of the mitochondrial genome.



- Launch Sequencher.
- Choose **New Project** from the **File** menu to open a new, empty project.
- From the **File** menu, select **Import > Sequences**…
- Navigate to the **Sequencher > Sample Data > Mitochondrial Sequences** folder.
- Select the Cambridge Reference Sequence.
- Click **Open**.

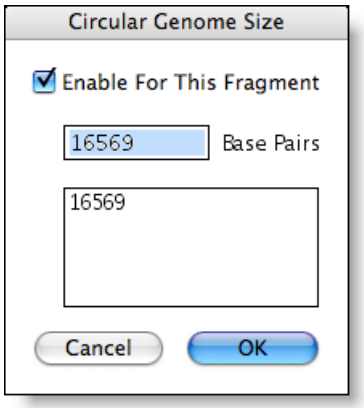
You have now imported a sequence into your project. The bases match those of the **Cambridge Reference**, but the numbering does not.

- From the **Sequence** menu, choose the **Reference Sequence** command to give the **Cambridge Reference** reference properties.

The **Reference Sequence** menu command acts as a toggle switch, enabling you to endow or remove Reference Sequence properties from a selected sequence. Note the "**R**" in the icon of this sequence and the prefix "**Ref**" displayed in the **Kind** column of the Project window. These distinctions tell you that this sequence has the special

properties associated with a Reference Sequence. These properties, which are demonstrated in this tutorial, include controlling the numbering of a contig, immunity to editing, ability to direct assembly, and special comparison features. Also, you can only assign circular genome size to a Reference Sequence.

- From the **Sequence** menu, choose **Set Circular Genome Size**… .
- Click on the box **Enable For This Fragment**.



Sequencher automatically displays the genome size of the human mitochondria, 16,569, and the last five genomes that you entered. The most recent is selected.

- Click OK.

You are now ready to modify the numbering of the **Cambridge Reference**, so that it corresponds to the published sequence.

- Double-click on Cambridge_Reference sequence.
- Place your cursor on base position 1.
- From the **Sequence** menu, choose **Set Base Number > As Base Number…**
- Enter 16024.
- Click **OK**.

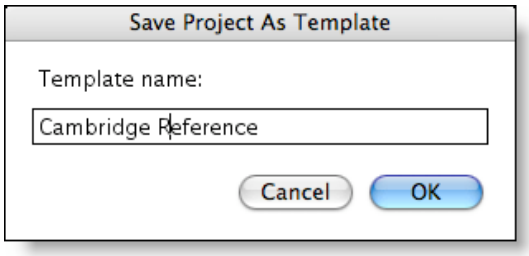
Note that the numbering of this sequence now starts at 16024, goes up to 16569, then restarts numbering again at position 1.

- Close the **Cambridge_Reference** window by clicking on its close box.

CREATE A PROJECT TEMPLATE

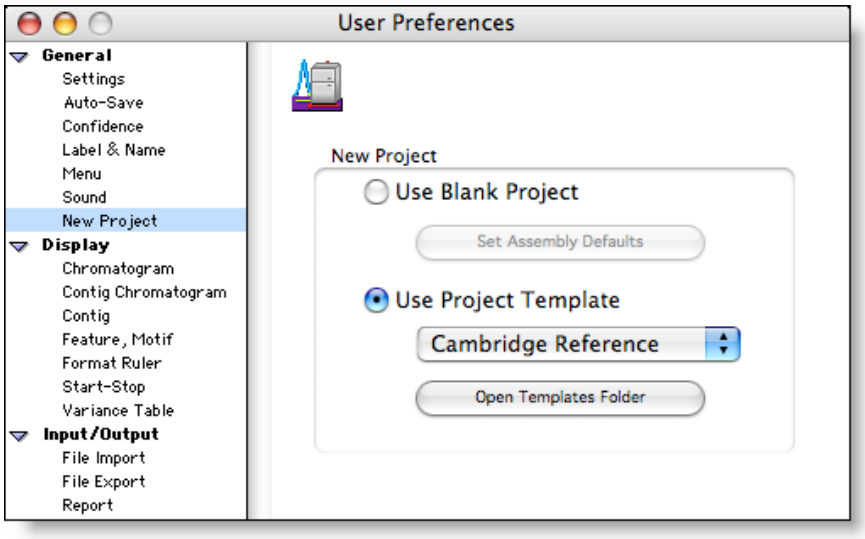
If you are using a Reference Sequence in one project, it is likely that you may need the same Reference Sequence in another project. You can save a project as a **Project Template**. The **Cambridge Reference** sequence is the basis for comparison in reporting mitotyping results, so it will be useful to save the sequence as we have just modified it as a template.

- From the **File** menu, choose **Save Project As Template…** .
- Give the Project Template the name **Cambridge Reference**.



- Click OK.

Until you delete the Project Template, you can access it through the **File** > **New Project From Template** command or from the **File** > **Import** > **From Template** command. In the **User Preference**, **New Project** pane, you can also make a Project Template the default template to open when starting a new project. This can save you time if you always start a project with the same Reference.



IMPORT AND ORGANIZE YOUR DATA

- From the **File** menu, select **Import** and from the sub-menu, select **Folder of Sequences…**
- Select the **Sequencher/Sample Data/Mitochondrial Sequences/HV1 and HV2** folder.
- Click **OK**.

Sequencher will alert you that you are about to import 9 files.

- Click the **Import All Files in Folder** button.

This will load nine AutoSeq fragments. Note that the names of the sequences are selected. Each column of the Project window describes specific details about the imported sequences such as size, quality, and kind. You can also modify the look of the Project window.

- From the **View** menu, choose **Project Window Columns > Label**.

This removes the **Label** column from the Project window until you execute this command again. You can also configure the Project window columns to include the Sample names of your data. This imported data does not have a separate sample name, because, as you can see in the Kind column, it is based on SCF files. However, most ABI files do have a Sample Name.

Because of preprocessing by an upstream software program, the AutoSeq data that you imported does not require trimming. However, most sequencing reactions do. For more information on setting effective trim criteria, see the Demo Tour Guide, Trimming Sequence, and the Quality Scores tutorials.

Set Assembly Parameters and Assemble

The Assembly Parameters allow you to control the limits of your assembly. You can define the minimum % match and the minimum number of overlapping bases required to build a contig. There are also additional criteria that you can set to define how to optimize the placement of gaps and how to select and name your contigs. The default values for assembly will work under most circumstances, but read the Assembly Strategies Tutorial for more information. The following will focus on the use of ReAligner and Assemble by Name.

- Click on the **Assembly Parameters** button.
- Check **Use ReAligner**, if it is not already checked, to optimize gap placement and check **Prefer 3′ Gap Placement** to gather gaps to the right-most position.

Sequencher allows you to choose how you want to position gaps created by small inserts and double-called bases.

5′ Preferred	Non-optimized	3′ Preferred
a:ggatc	a:ggatc	agg:atc
a:ggatc	agg:atc	agg:atc
a:ggatc	ag:gatc	agg:atc
agggatc	agggatc	agggatc

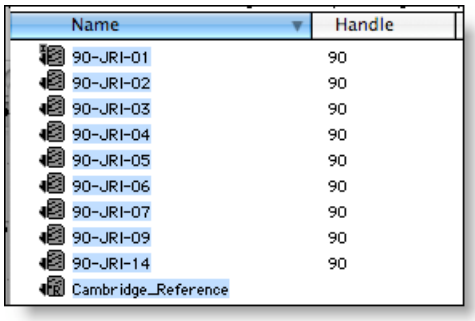
- Click on the **Name Settings…** button in the bottom section on **Assemble By Name**.

This opens the dialog that allows you to separate the information contained in the names of your imported sequences into useful parts that Sequencher calls Assembly Handles. The sequences in the Project window have two sets of information separated by a dash, so we will change the delimiter to a dash and use Handle 1 to name and select our contigs.

- Scroll down the list of delimiters and select the Dash – name delimiter.
- Click on the **Handle 1** Assembly Handle.

- Click **OK** to close the Assemble by Name Settings dialog.
- Turn on Assemble by Name by clicking in the **Enabled** box in Assembly Parameters.
- Click **OK** to close the **Assembly Parameters** window and return to the Project window.

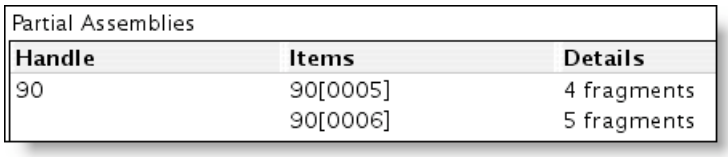
Notice that there is now a new column in the Project window, **Handle**, and across from each of the imported sequences is the portion of the sequence name that is followed by a dash. Also note that the names of the buttons on the Project window have changed to reflect the new status of the Assemble command. The **AbN** button on the Project window provides a shortcut to turn Assemble by Name off or on. For more information on this feature, read the Assemble by Name tutorial.



- Select all of the fragments by dragging a box around them or choosing **Select All** from the **Select** menu.
- From the Project window, click on the button **Auto Assemble by Name**.
- Review the **Assembly Preview** window and then click **Assemble**.

After Assembly, Sequencher will present an Assembly Completed dialog.

- Click on the **Details**… button.



The result of the assembly is two partial assemblies, which we expect. There is no overlap between the regions of HV1 and HV2 sequenced in this project.

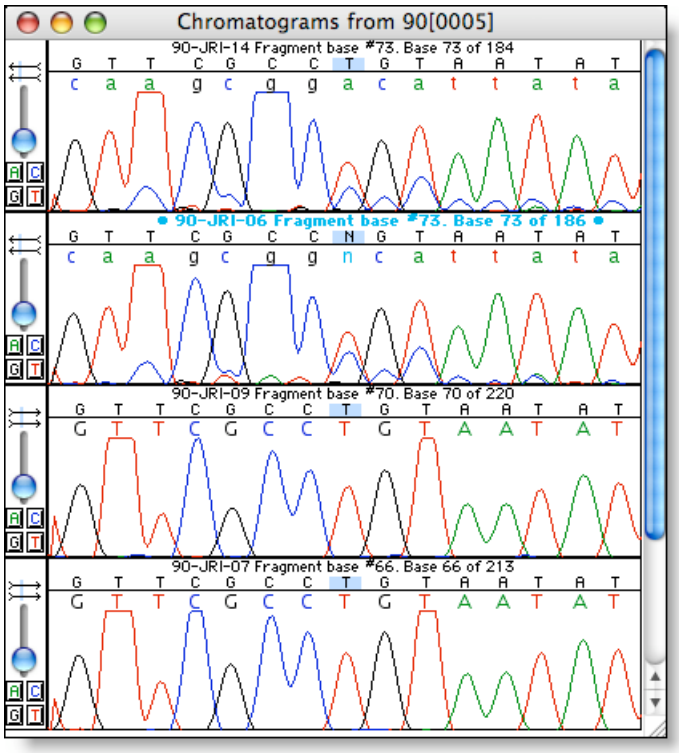
- Click on the **Close** button to return to the Project window.
- Open a Contig Editor by double-clicking on the contig icon **90[0005]**.

In the **Overview** of the contig, you will notice two forward and two reverse sequences.

- Click on the **Bases** button.
- Select the first base in the consensus sequence.
- Click on the **Show Chromatograms** button to display all of the trace data involved in calling this consensus base position.
- From the **Select** menu, choose **Next Ambiguous Base**.

The first ambiguous base is at consensus position 73. Sequence 90-JRI-06, which is in the reverse orientation, has an N at this position while the other sequences have a T. Also note the dark blue confidence shading of the equivalent base in the reverse orientation, sequence 90-JRI-14. This indicates that the quality score for this sequence is very low.

A look at the chromatogram data shows that there is a lot of "C" background noise in this region in the reverse sequences, but the forward sequences clearly support a "T" call.



- Keep your selection in the consensus line at base position 73, and type "T".
- With your cursor still in the consensus, press the space bar to again invoke the **Select** > **Next Ambiguous Base** command.

Sequencher alerts you that there are no more ambiguities in the data.

- Repeat the editing process with contig 90[0006].

These data have four forward sequences and one reverse sequence. In editing, one must be careful not to give too much credence to supporting data sequenced in the same orientation. When making the consensus call, it is best to use the one best forward and the one best reverse reactions. Typically sequences repeated in the same orientation contain the same errors. Contig 90[0006] requires edits at positions 73, 85, and 141.

- Close both the Contig and the Chromatogram Editor windows.

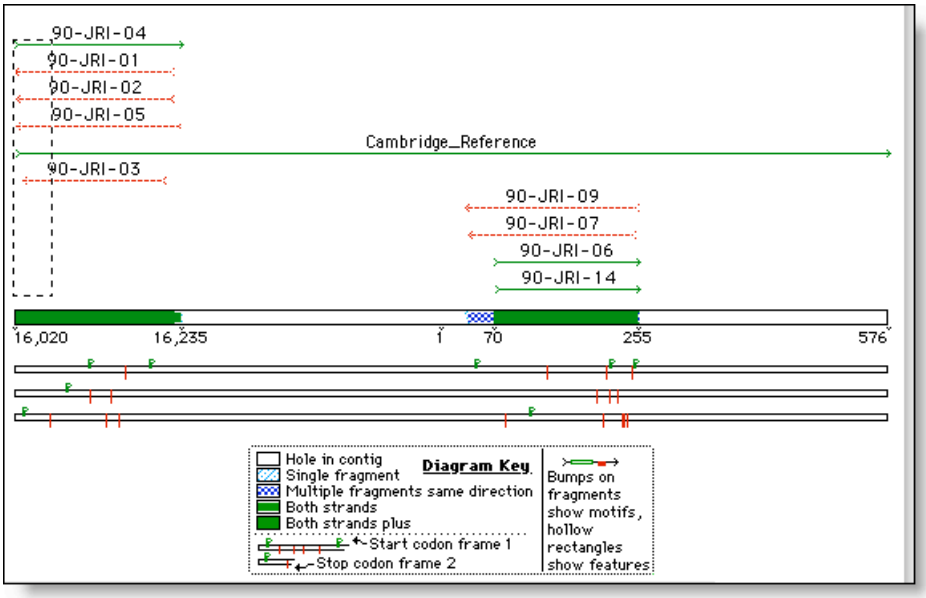
Assemble to Reference by Name

The **To Reference by Name** command allows all samples to align to a single Reference, regardless of inconsistencies between the individual fragments. 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** assembly option.

- Select all.
- Click on the **To Reference by Name** button.
- Click **Assemble** on the **Assembly Preview** window.
- Click **Close** after reading the Assembly Completed dialog.

After assembly, you should have only two items in your Project window, the **Cambridge_Reference**, because it is not consumed in the assembly, and the contig 90. The complete contig is named with the Assembly Handle alone. The distinguishing suffixes, 0005 and 0006, are no longer needed.

- Double-click on the icon of contig 90.



The Overview shows the layout of the contig relative to the Cambridge Reference sequence. Also note the numbering of the contig matches the numbering of the Reference Sequence and the information displayed in the coverage map ignores coverage contributed exclusively by the Reference.

- Click on the **Bases** button.

You can see that the Reference Sequence still has the **R** in the icon and a gray border surrounds the bases. You can move the Reference Sequence to the top or bottom of the contig by clicking on its name and dragging.

The Reference Sequence has several advantages. It allows you to focus on your region of interest and it provides a consistent framework on which you can describe your project. In addition, the Reference Sequence is unobtrusive. It does not contribute to the consensus and it protects you from making accidental changes to the Reference while editing in the consensus. Gaps in the Reference Sequence are given decimal numbers, thereby preserving the numbering of base positions.

- **Close** the Contig Editor window.

Analyze Variations and Generate Reports

Since you have already edited the sample sequence, 90, you can now view a compact table of the genetic variations that characterize this individual.

- From the Project window, select contig **90**.
- From the **Sequence** menu, choose **Compare Bases To** > **Reference Sequence**.

- In the bottom left corner of the Variance Table, click on the open elevator icon  to optimize the column width for the display of the full name of the sequences.

In effect, we are asking how each of the sequences in the contig compares to the Reference. The results are displayed in an interactive Variance Table.

The left hand side of the table shows that there are no differences between the contig 90 and the Reference Sequence in HV1.

Reference		90-JRI-04	90-JRI-01	90-JRI-02	90-JRI-05	Cambridge_Reference	90-JRI-03	Total
73	A							4
195	T							4
	Total	0	0	0	0	0	0	8

Scroll the table to the right to see that there are two well supported differences to the Reference Sequence in the HV2 region, at reference base positions 73 and 195.

Reference		Cambridge_Reference	90-JRI-03	90-JRI-09	90-JRI-07	90-JRI-06	90-JRI-14	Total
73	A			G	G	G	G	4
195	T			C	C	C	C	4
	Total	0	0	2	2	2	2	8

The quality of some of the data supporting a difference at base position 73 is weak, as noted by the darker blue shading.

- Double click on the cell that reports a G at position 73 in sequence 90-JRI-14.

Sequencher opens the Contig and Chromatogram editors that support that base call. This is what the Variance Table looks like in Review mode.

- Use the right arrow key on your keyboard to navigate to each of the G calls at position 73.
- Close all of the windows when you are confident of the base calls reported as differences.

You can also view your data by just comparing the consensus to the Reference Sequence.

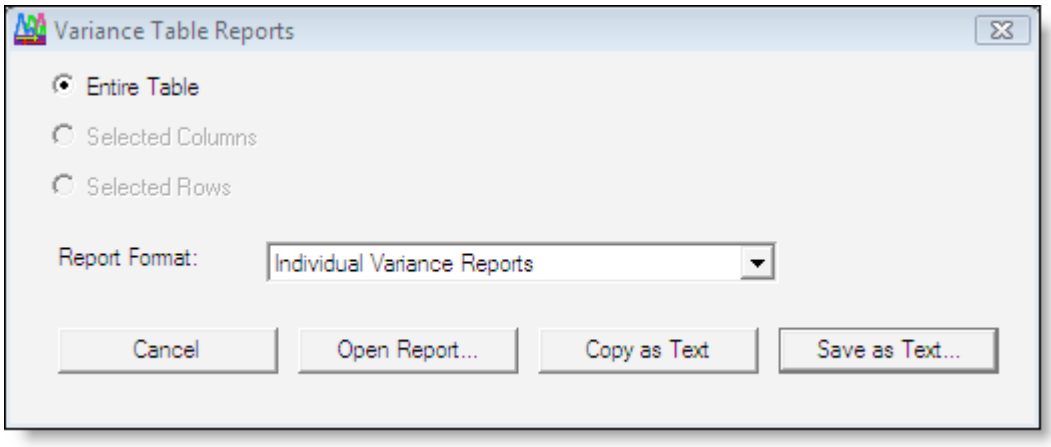
- In the Project window, choose the command **Select** > **Select All**.
- From the **Contig** menu, choose the command **Compare Consensus to Reference**.

The Variance Table that you create now compares the consensus of all of your sequences to your Reference. Each cell in the table is linked to all of the data that support that consensus call.

- Double-click on the G at reference position 73.

This puts you in the Review mode, where you can see all of your data and you can also create a report of your results.

- Click on the **Reports** button at the top of the Variance Table window.
- Change the format selection from Variance Table Report to Individual Variance Reports.



- Click on the **Save as Text...** button.
- Choose a location to store your difference report and select **OK**.

The exported report looks like the following:

2 differences between
Cambridge_Reference and 90
06/12/2006 12:49

Pos	Ref	Con	Required Edit
73	A	G	Change base
195	T	C	Change base

This tutorial provides generic information for the analysis of a sequence to a Reference. For more information you can try the Mitochondrial DNA Typing Tutorial, which focuses on specific tools built into the Forensic version of Sequencher. These tools help to further validate and report differences in mitochondrial sequence. You can learn more about tools for batch processing and automation by referencing other tutorials in this series as well as the User Manual.

- Close the project from the **File** menu by choosing the **Close Project** command.