SEQUENCHER®

Tutorial for Windows and Macintosh

Quality Scores

Gene Codes Corporation

T C A **GENE**
A G T **CODES**

Quality Scores

The process of generating DNA sequencing data involves many steps. Part of the process involves recording the intensity of fluorescence as a nucleotide-associated dye passes a laser. A subsequent step requires analysis software to convert these recordings into base calls. The ability of base calling software to accurately interpret raw traces varies with the quality of the data. The most common early base callers called an ambiguous base, or an "N," whenever the quality of the raw data failed to yield a clear signal for A, C, G, or T. Advances in base callers led to the evolution of the quality or confidence score.

Quality scores are numeric values associated with each base call that define how likely it is that a base call is incorrect. The most common scale is from 1-60, where "60" represents a $1/10^6$ chance of a wrong call, "50" a $1/10^5$ chance, 40 a $1/10^4$ chance, etc. Depending on the program used to generate the confidence value, the quality score may be based on peak height, the presence of more than one peak, and/or the spacing between the peaks.

Many instruments sold today come with base callers that generate confidence scores as an integral part of the chromatogram file. These include Applied Biosystems's KB base caller, MJ Research's SCF files, and the ESD files generated by the Beckman CEQ and the Amersham/GE MegaBace. Some labs use Phred, TraceTuner, or other third-party programs to generate confidence values in separate documents that are associated with the chromatogram files. Various Base Calling software programs generate a variety of different file formats. To import sequences into **Sequencher** with all of the associated data as one sequence, you may import a FASTQ or PHD file or you may import the individual sequence/qual/trace files. The key to associating the data within **Sequencher** is the naming of the file. If you have any specific difficulties with your file format, please feel free to ask Gene Codes Technical Support for assistance.

Sequencher is capable of working with confidence values from all of the above files types. **Sequencher** imports and displays confidence ranges in three user-defined ranges: low, medium, and high. **Sequencher** trims data based on confidence scores, and automatically navigates to regions of low confidence. You can view summary confidence information about your sequences in the **Project Window**, the **Sequence Editor**, and in the **Get Info** dialog.

Sequencher displays the confidence scores in three user-defined colors so that display of confidence will not interfere with your display of base calls.
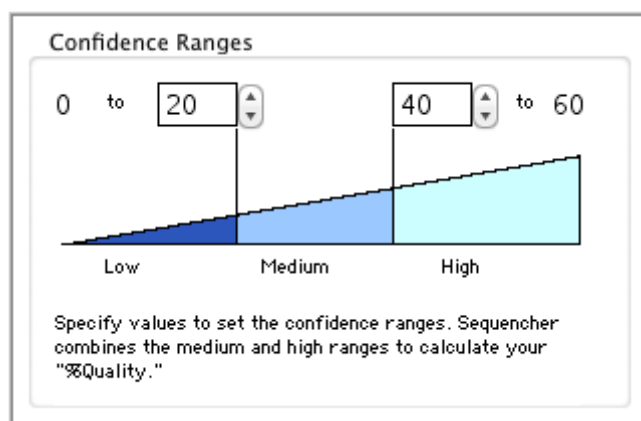
QUALITY SCORES

- Launch **Sequencer**.
- From the **File** menu, choose **Import > Sequencer Project…** and load **QualScore Project.SPF**.

Sequencer will load nine auto-seq fragments into your **Project Window**. The raw data was base called with Applied Biosystems' analysis version KB1.1, which generates confidence scores for each base call. Next to the name of each of the ABI sequences loaded into your **Project Window** is a value for % Quality. This number defines the percent of quality bases for each of the sequences. Quality - six, for instance, has a % Quality of 51.3– 48.7 % of its bases are low quality.

Name	Size	Quality
Quality – eight	652 BPs	93.1%
Quality – five	758 BPs	94.9%
Quality – four	794 BPs	91.9%
Quality – nine	796 BPs	91.1%
Quality – one	798 BPs	91.7%
Quality – seven	791 BPs	92.2%
Quality – six	805 BPs	51.3%
Quality – three	796 BPs	92.8%
Quality – two	798 BPs	91.7%

- Under the **Window** menu, select **User Preferences…**.
- One of the options in the **General** category is called **Confidence**. Select that topic.
- Set the confidence ranges as shown in this figure:

Confidence Ranges

0 to [20]   [40] to 60

Low   Medium   High

Specify values to set the confidence ranges. Sequencer combines the medium and high ranges to calculate your "%Quality."

With these settings, base calls with scores between 0 and 20 will be displayed with a cobalt blue background; those with scores between 21 and 39 will be displayed in cadet blue, and those 40 and above will be displayed in aqua. Bases that you edit by hand (and are therefore sure of) will be displayed on a white background. Your base edits are treated as high quality in all calculations.

- Exit **User Preferences** by closing the window.

- From the bottom of the **View** menu, make sure that **Display Base Confidences** is checked.
- From the **Select** menu choose **Select None**.
- Double-click on Quality - five to open its **Sequence Editor** window.
- Select a range of bases that includes some low quality bases.

You can see the lower quality bases in darker colors. Note the number and percentage of low quality bases in your selection appears above the bases in the sequence editor.

A schematic representation of each base's confidence score may also be displayed above each base. You may choose a maximum score setting based upon the scale used in your data. The most common scale has a maximum of 60. Use a maximum of 100 for ESD data, or consensus sequences where you may expect to see higher scores.

- From the **View** menu, choose **Confidence Histogram (max 60)**.

TRIM SEQUENCES BASED ON QUALITY SCORES

The quality scores can be used to trim the data before assembly.

- Close the **Sequence Editor** window.
- Select all nine sequences by choosing **Select All** from the **Select** menu.
- From the **Sequence** menu, select **Trim Ends…**
- Click on the **Change Trim Criteria** button at the top of the window.
- Deselect all options with the exception of the following:

- Click **OK** in the **Ends Trimming Criteria** window.
- Click on the **Trim Checked Items** button.
- When you are asked to confirm your decision, click **Trim**.
- Close the **Ends Trimming** window to return to the **Project Window**.

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.

## ASSEMBLE DATA AND CALCULATE CONSENSUS BY CONFIDENCE

You are now ready to view the data in an assembly.

- Select all 9 items in the **Project Window**. (They should be selected already.)
- Click on the **Assemble Automatically** button at the top of the **Project Window**. When the assembly completes, click **Close.**
- Double-click on the Contig icon to see the Overview of the sequence assembly.
- Click on the **Bases** button to see the assembled sequences.
- From the **Select** menu, choose **Next Contig Disagree** to find the first discrepancy between the overlapping fragments.

You are probably looking at base 126 where one, low-quality base call is a 'T' but the overlapping fragments all have a medium quality 'C' at that position.

- From the **Contig** menu, choose **Consensus by Confidence** to recalculate the consensus using the best quality data at each position. Note that the 'C' in the consensus at 126 is displayed as low confidence because of the discrepant T in the Unsequenced strand Quality – eight sequence.
- To review the trace data, click on the **Show Chromatograms** button at the top of the window.

All of the trace data have evidence of both the 'C' and 'T' base. This is a location where there is a heterozygous base, both a 'C' and a 'T', or a 'Y'. In general, most base callers assign low quality scores to heterozygotes.
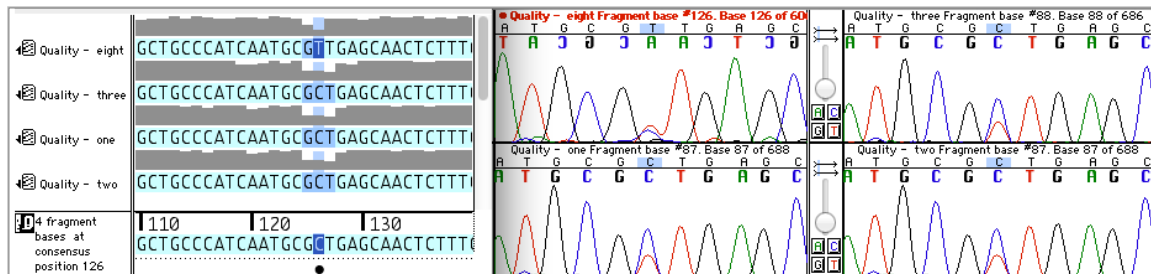
- With the consensus base still selected, type a '**Y**'.

Not all bases that require editing will be flagged with a disagreement. For example, mixed bases that may have been called an "N" in a previous base caller will now be called a low quality A, C, G, or T. Also, consensus sequence based on single coverage will never create a disagreement unless you have a Reference Sequence. In order to examine the contig with greater scrutiny, you may want to look at all of the low confidence bases using the **Select > Next Low Confidence Base** menu command.

The **Select > Next Ambiguous Base** menu item is also sensitive to the presence of confidence scores. This feature enables you to navigate from one position in the consensus where the base call is ambiguous to the next. **Sequencher** defines the following as ambiguous consensus bases:

- ➢ Where there is a disagreement between contributing bases
- ➢ Where one of the contributing bases is itself an ambiguous base call
- ➢ Where there is a gap in the contributing sequence
- ➢ Where the quality of the only sequence(s) contributing to the consensus falls in the low range as defined in User Preferences

- Close the project without saving and exit **Sequencher** if desired.