APPLICATION INFORMATION

CEQ Series

CALL SCORES AND QUALITY VALUES: TWO MEASURES OF QUALITY PRODUCED BY THE CEQ™ GENETIC ANALYSIS SYSTEMS

Roger Winer, George Yen, and Jennifer Huang
Beckman Coulter, Inc.

Introduction

Data produced by automated DNA sequencing instruments are used for a variety of genetic analysis applications including genome sequence assembly, polymorphism studies, expressed sequence tag (EST) analysis, identity and paternity testing, phylogenetic analysis, and others. Along with the nucleotide sequence data (base-call) produced by the sequencer, it is helpful to have an objective assessment of the accuracy of the reported sequence. The CEQ™ analysis software produces two different measures of sequence accuracy, each meant to address a different set of applications. The discussion that follows describes the technique used to derive the measures assigned by the CEQ and suggests ways to take advantage of the additional information these measures provide. Advantages include aiding in Editing, Trimming, selecting primers, alignment of an assembly, and increasing the BLAST score for true positives and decreasing the BLAST score for false positives.

History and Definition of Quality Values

In order to cope with the volume of data involved in large-scale sequencing projects, software tools have been developed to help manage and automate much of the sequence assembly process. The typical sequencing project comprises four main activities: 1) Breaking down the DNA for which the sequence is to be determined into smaller, overlapping pieces (library construction); 2) Acquisition of primary read data (sequence data from the samples in the library, usually 500-1000 nt/sample); 3) Piecing the primary read data into longer stretches (contig assembly) by determining which parts of the read data overlap; and 4) Finishing, which involves closing gaps between contigs and shoring up the ambiguous portions of the contig sequences that have been determined. Base-calling errors in the primary read data complicate the contig assembly process. If not accounted for, higher error rates at the beginnings and ends of reads may prevent overlapping reads from being assembled into the same contig. When too much disagreement between read sequences exists, the assembly software will not join the reads into the same contig. Figure 1 shows an example of data from overlapping reads that might be misinterpreted by sequence assembly software.

Several methods have been employed to prevent assigning reads that truly overlap from being assigned to more than one contig erroneously. In general, these methods can be divided into two categories: trimming and weighting.

Trimming methods discard portions of sequence data from each read prior to the assembly process. Criteria for trimming vary and may include: 1) Trimming the ends of the reads from the point at which a certain number of Ns are found within a window of some number of bases; 2) Trimming a fixed number of bases from either end of each read (e.g., trimming the first 25 and last 100 bases); 3) Trimming all data that falls outside some base number range (e.g., keeping only data between base numbers 25 and 500). Each of the methods has the disadvantage that some valuable information

```
Base # 750:   …ATGTGACCCGGGGTCCACATGGGAAATTTAAC
Base # 5:     …AAATGACC-GGG-TC-ACATGG-AA-TTTAACTGTTGCACAC…
Base # 150:   …ATGTGACC-GG--TC-ACATGG-AA-TTTAACTGTTGCACAC…
Consensus:    …ATGTGACC-GG--TC-ACATGG-AA-TTTAACTGTTGCACAC…
```

Figure 1. *Alignment of stretches of sequence data from three reads showing base number of first represented base for each read. Lower-quality portions of read data are shown in italics. The actual sequence is shown on the Consensus line.*

from each read is discarded, leading to less data from each read being incorporated into the contig assembly process. This in turn leads to a greater number of reads being required to assemble a contig of a given length. In addition, the last two methods make assumptions about the accuracy of data in specific portions of reads that will not be true for all reads in a project.

Weighting methods allow all data from the reads to enter the assembly process. Weighting factors are then used to indicate the portions of each read where the data is less important in deriving the true sequence. Different types of weighting methods include: 1) Fixed weighting methods and 2) Methods employing an estimate of the confidence of each called base. An example of a fixed weighting method is the trapezoidal weighting technique, where the importance of data gradually increases within a read to some plateau and then falls gradually at the end of the run (Figure 2).

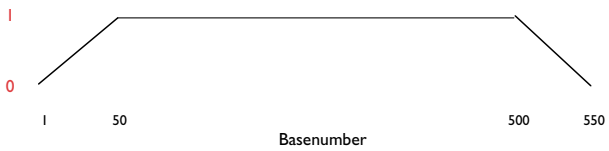

Figure 2. *Trapezoidal Weighting method assigns a weighting factor that varies between 0 and 1 to the importance of each called base. The weight assigned depends on the base number of the called base.*

Methods using estimates of the confidence of each called base include several discussed by Dear and Staden[1], Lawrence and Solovyev[2], Berno[3], Ewing and Green[4], and others. Dear and Staden proposed using a technique where four probability estimates for each base position would be used: the probability of a base being an "A," the probability a base being a "T," the probability of a base being a "C," and the probability of a base being a "G." Others have proposed using multiple confidence estimates such as the probability that the identity of a called base is correct and the probability that a base should really be called at that position. Still others have proposed assigning probability estimates for each type of error that can occur at each position in a base-call: delete (no call), insert, miscall.

The method used by Ewing, Green, and Gordon (in the Phred/Phrap/Consed suite of software tools) uses a single measure of confidence for each called base in the primary read data. The measure used is the Quality Value (QV) and relates to the estimated error rate, *p*, according to the equation:

$$QV=-10 \times \log_{10}(p)$$

Quality Values are assigned by the program Phred based on a calibration process where certain characteristics (indicators) of the trace data from automated sequencers is measured at or near each called base. The observed error rates of groups of bases are determined and, using a technique described by Ewing and Green, a lookup table is produced relating threshold values for each of the indicators to an estimated probability of error. Millions of bases of aligned sequence are required for the calibration process along with high-quality consensus sequence. The indicators used include:

- **Peak Spacing ratio:** the ratio of the largest peak-to-peak spacing in a seven-base window to the smallest peak-to-peak spacing in the same window;
- **Uncalled/called ratio:** The height ratio of the highest uncalled peak to the lowest called peak in a seven-peak window around the called base;
- **Uncalled/called ratio:** same as above, but in a three-peak window
- **Distance to Unresolved base:** Number of bases between the called base and the nearest unresolved base.

Several properties of QVs make them particularly useful for sequence assembly projects:

- **Accuracy:** the **validity** of QVs depends on the estimated error rates of groups of bases corresponding well to the observed error rates for the same groups.
- **Discriminating Power[4]:** the **utility** of QVs depends on the ability of the error rate estimates to discriminate between various qualities of data. For example, an estimated error rate of 5%

might apply to all bases taken together in a sequencing project. However, such an estimate would not be very useful for determining which bases in a read are more reliable than others. Therefore, the ability to discriminate between a wider range of error rates with greater resolution increases the utility of the QVs.

- Log-based estimate of error rate: This allows QVs for bases from independent reads (e.g., reads from opposite strands, different chemistries, etc.) to be added at a particular position to give an estimate of the error rate for the consensus at that position.

Taken together, the estimated error rates for all bases in a consensus can provide an estimate of the total number of errors one might expect to find in the sequence.

Implementation of Call Scores and Quality Values on the CEQ™

Two estimates of error rates for called bases are available on the CEQ sequencers: Call Scores and Quality Values. Both estimates have all the properties of QVs produced by Phred, but they are tailored for the sequences called by the CEQ. Over the past several releases of CEQ software, several changes have been made to improve the accuracy of the base-calling. These changes have resulted in decreased error rates. Since the estimated error rates have not changed, discrepancies between predicted and actual error rates have grown. The greatest increases in base-calling accuracy have occurred in the latter portions of the reads. As suggested in a paper by Richterich[5], predicted error rates can be compared to observed error rates at each base position over multiple reads in a sequencing project. The discrepancy between predicted and observed error rates for data produced by CEQ software version 4.2 are shown for about 9 million bases of data in Figure 3.

Since actual error rates will vary among different base-callers (e.g., Phred and the CEQ Sequence Analysis Software), the estimated error rate for a given set of bases may differ as well. Several calibrations were performed using over 9 million bases of data called by the CEQ which could be aligned against known sequences. Several combinations of indicators were tested including the ones used by Phred. A discriminant analysis performed by Beckman Coulter's Math Group[6] was used to identify several indicators that are well-correlated with error rates.

Call Scores vs. Quality Values

The two measures of quality produced by the CEQ are Call Scores (CS) and Quality Values (QV). Both provide an accurate estimate of the error rates of groups of bases. It has been suggested by Ewing, et al.[7], that "[improved base-calling] accuracy in the lower quality part of the trace would be useful in single-read applications." Toward this end, Beckman Coulter has developed Call Scores as an aid to editing and selecting portions of single-pass sequencing results that will provide the greatest utility, while QVs are most useful in identifying the highest-quality portions of reads for use in sequence assembly and multiple sequence applications. ***The main difference between the two measures is that Quality Values discriminate better between various levels of high-quality data, and Call Scores discriminate better between various levels of low-quality data.***

When reviewing the results of calibration tables obtained using different sets of indicators, several criteria were assessed for each calibration including:

- The minimum and maximum estimated error rates;
- The percentage of bases grouped into the lowest and highest quality levels (quality levels with error rates greater than 1% and those with error rates lower than 1%);
- The number of different quality levels assigned at the high and low end of the error rate spectrum and the evenness of the distribution of those levels;
- The reproducibility of the accuracy of calibrations performed using the same indicators but with smaller subsets of data (cross-validation);

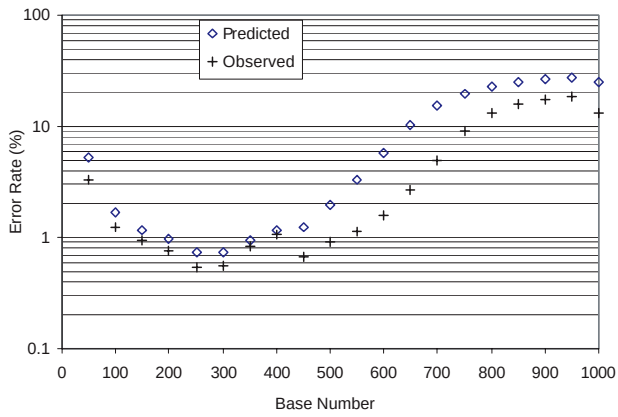

Figure 3. *Predicted vs. observed error rates for positions grouped by 50 base numbers (using QV estimates from version 4.2 software for the CEQ).*

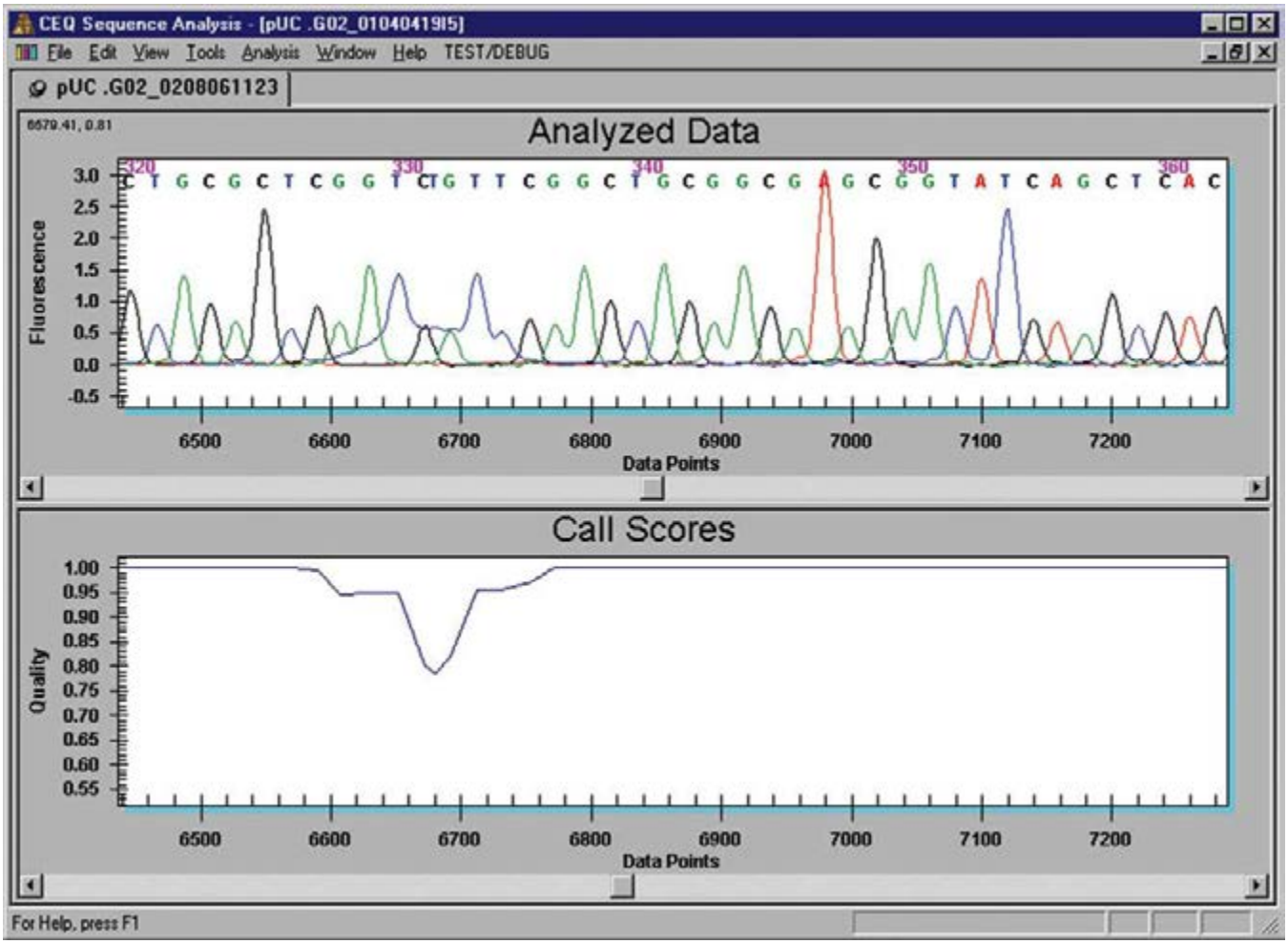

Figure 4A. *Call Scores viewed in linear scale.*

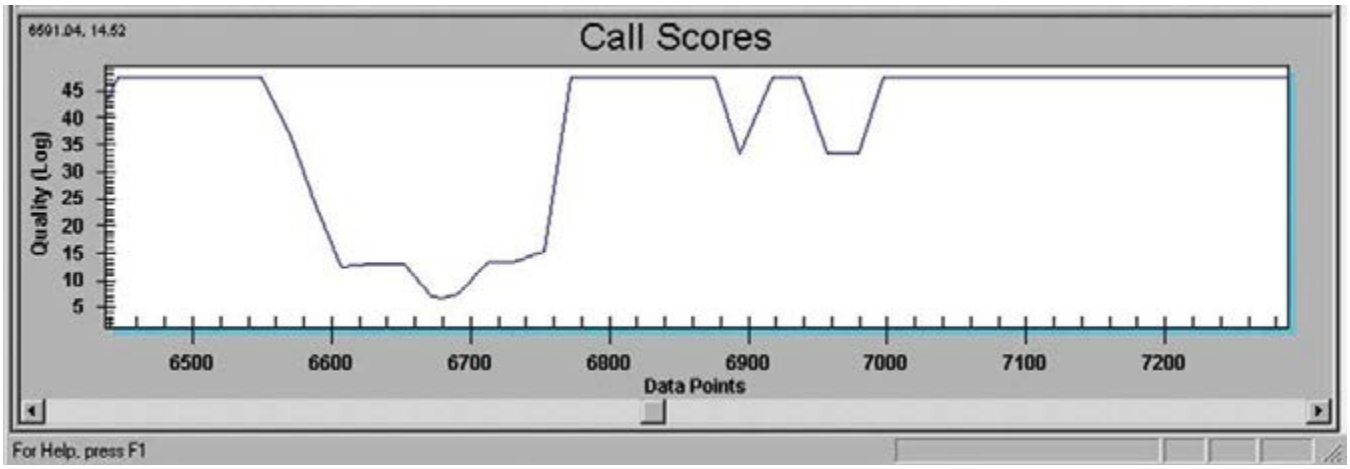

Figure 4B. *Call Scores viewed in log scale.*

- The differences between predicted error rates and observed error rates when viewed for each base position (Binned by Base Number);
- The differences between predicted error rates and observed error rates when viewed for bases with different levels of estimated quality (Binned by QV).

The first three criteria pertain to the **Discriminating Power** of the calibrations, while the last three pertain to the **validity** of the calibrations. The calibrations that showed the best accuracy and reproducibility of accuracy were examined further. The one that showed the greatest discriminating power on the low end of quality was selected as the **Call Score calibration.** The one that showed

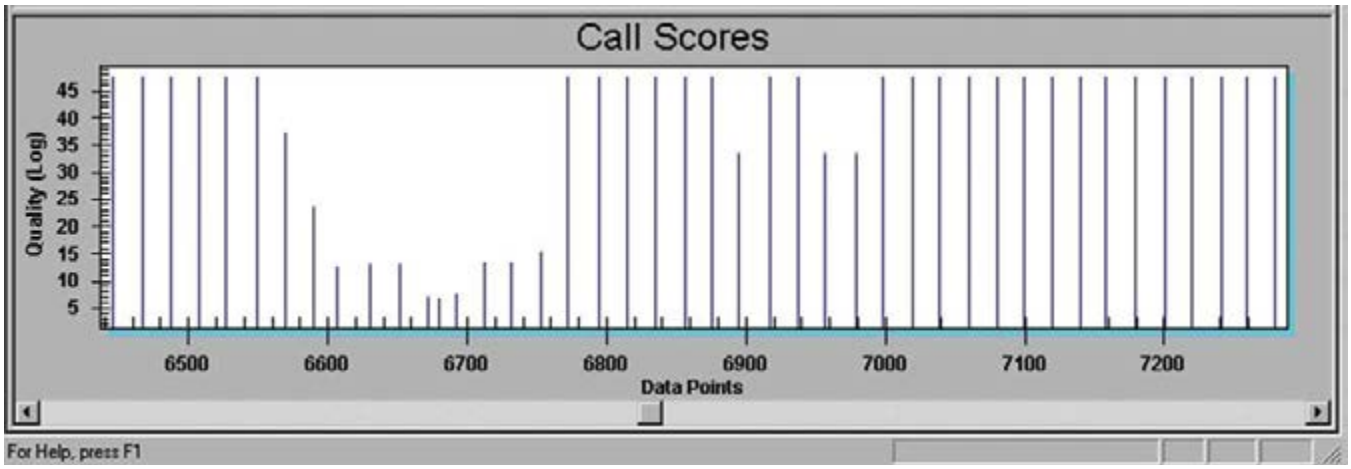

Figure 4C. *Call Scores viewed in log scale as a bar plot.*

the greatest discriminating power on the high end of quality was selected as the **Quality Values calibration.**

Log vs. Linear Scale

Both CSs and QVs may be viewed in either Log or Linear scale in the CEQ™ Analysis Software (Figures 4A and B).

In general, it is more useful to use the Log Scale view when looking for or examining higher-quality data. The Linear Scale view isn't as useful for discriminating between various levels of high-quality data. On the other hand, the Linear scale is often useful when looking for lower-quality data that might benefit from editing by the user.

Users may wish to view the QV or CS data as a bar plot as shown in Figure 4C. This has the advantage of making it easy to find peak-spacing anomalies within the Quality Parameters View.

Trace Characteristics Used in Calibration Lookup Tables

The metrics (indicators) used to lookup the associated CS or QV for each called base are listed in Table 1. Both calibration tables use five indicators to classify called bases into different Quality "bins." Each "bin" has an estimated probability of error associated with it and is bounded by values for each of the five indicators. The calibrations share three of their indicators. The indicators that differ between the two help provide added discriminating power at one end of the quality spectrum or the other. The "Prob Ratio" and "Peak Score Ratio" are metrics that use quantities estimated by the base-caller; the other indicators measure characteristics of the analyzed trace data.

Accuracies of New Calibrations

The accuracies of the new CS and QV calibrations are shown in Figures 5A-D.

Using Call Scores as an Aid to Editing

Call Scores are particularly well suited to helping identify parts of reads that can benefit from visual inspection and manual editing. This is true since they have more power to discriminate between various levels of error-prone sequence. Viewing the Call Scores, Base Sequence Text, and Trace Data for a read on a single screen facilitates the editing process. This is particularly useful when working with single-pass sequence from projects using enti-

Table 1. Metrics Used to Estimate CS and QV Values

Call Scores	***Quality Values***
Base Spacing Consistency	Base Spacing Consistency
Distance to Nearest Unresolved Base	Distance to Nearest Unresolved Base
"Peak Score Ratio:" Highest Score of Uncalled Peak/Lowest Score of Called Bases in Seven-Base Window	"Peak Score Ratio:" Highest Score of Uncalled Peak/Lowest Score of Called Bases in Seven-Base Window
Height Ratio of Highest Uncalled Peak to Lowest Called Base in a Three-Base Window	Height Ratio of Highest Uncalled Peak to Lowest Called Base in a Seven-Base Window
Peak/Shoulder/Interpolated: Whether Called Base is a Peak, a Shoulder, or Neither	"Prob Ratio:" Ratio of the Probability for the Most Likely Alternative Call to the Probability of the Current Call Being Correct

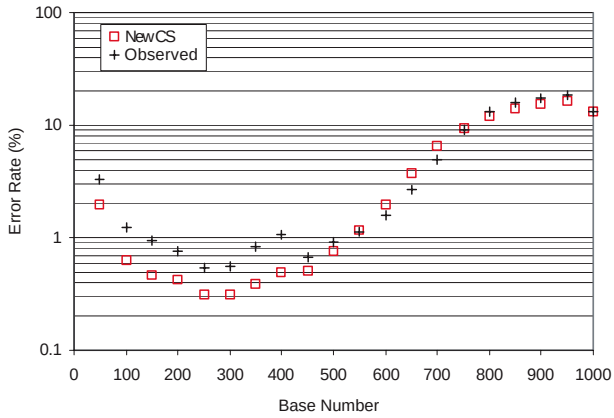

Figure 5A. *Comparison of predicted vs. observed error rates for new Call Scores (by base position).*

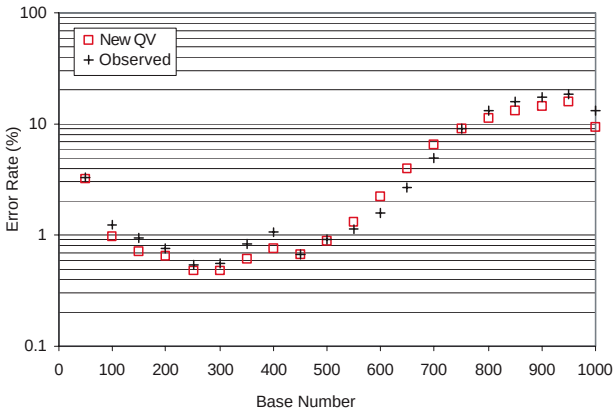

Figure 5B. *Comparison of predicted vs. observed error rates for new Quality Values (by base position).*

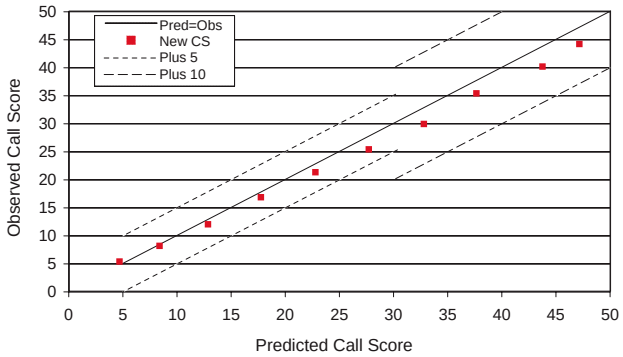

Figure 5C. *Comparison of predicted vs. observed error rates for new Call Scores (by quality).*

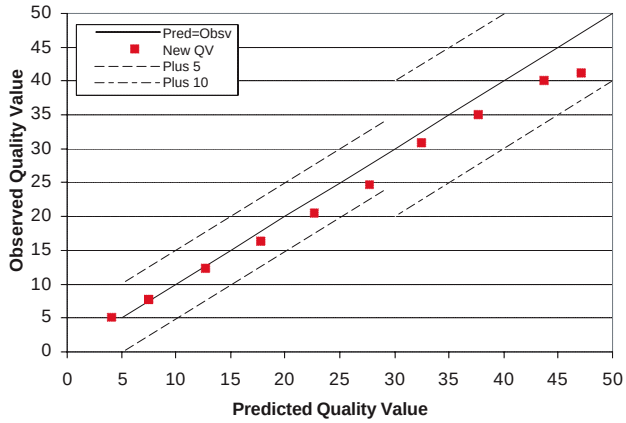

Figure 5D. *Comparison of predicted vs. observed error rates for new Quality Values (by quality).*

ties such as Expressed Sequence Tags (ESTs). The user may choose a level of accuracy below which Call Scores for error-prone bases will be visible. In general, it is best to set the Call Threshold to a very low or 0 value when processing the data. This has the advantage that fewer edits will be necessary since only the incorrect bases will need to be changed instead of all "Ns." Usually the goal of manual editing is to obtain a very long, highly accurate sequence or to obtain highly accurate sequence in particular regions of interest. In Figure 6, the Call Score View has been "zoomed-in" to a level that only displays Call Scores associated with bases having less than a 0.95 probability of being correct. Regions of lower quality (generally internal to the lower quality ends of the read) are inspected to obtain a highly accurate, long read.

In the example above, 25 bases out of 63 bases had Call Scores with accuracy probabilities below 0.95. There were 7 base-calling errors among these 25 bases which were easily identified and corrected. BLAST scores (Expectation Values) for the 63-base subsequence before and after editing were 2 E-5 and 7 E-23, respectively. Using Call Scores as a navigation aid greatly simplifies the task of editing the called sequence.

In addition to helping navigate between low-quality regions, Call Scores can provide an estimate of the number of errors one might expect to find within a region. One can View the **Base Sequence Toolbar**, select (highlight) a group of bases within the Base Sequence Text View, and obtain an estimated error rate for the selected bases. Figure 7 shows the Base Sequence Toolbar.

In Figure 7, 100 bases have been selected and they show a Call Quality Score of 0.98. This means they have roughly a 98% chance of being correct. For this 100-base region, roughly two errors can be expected.

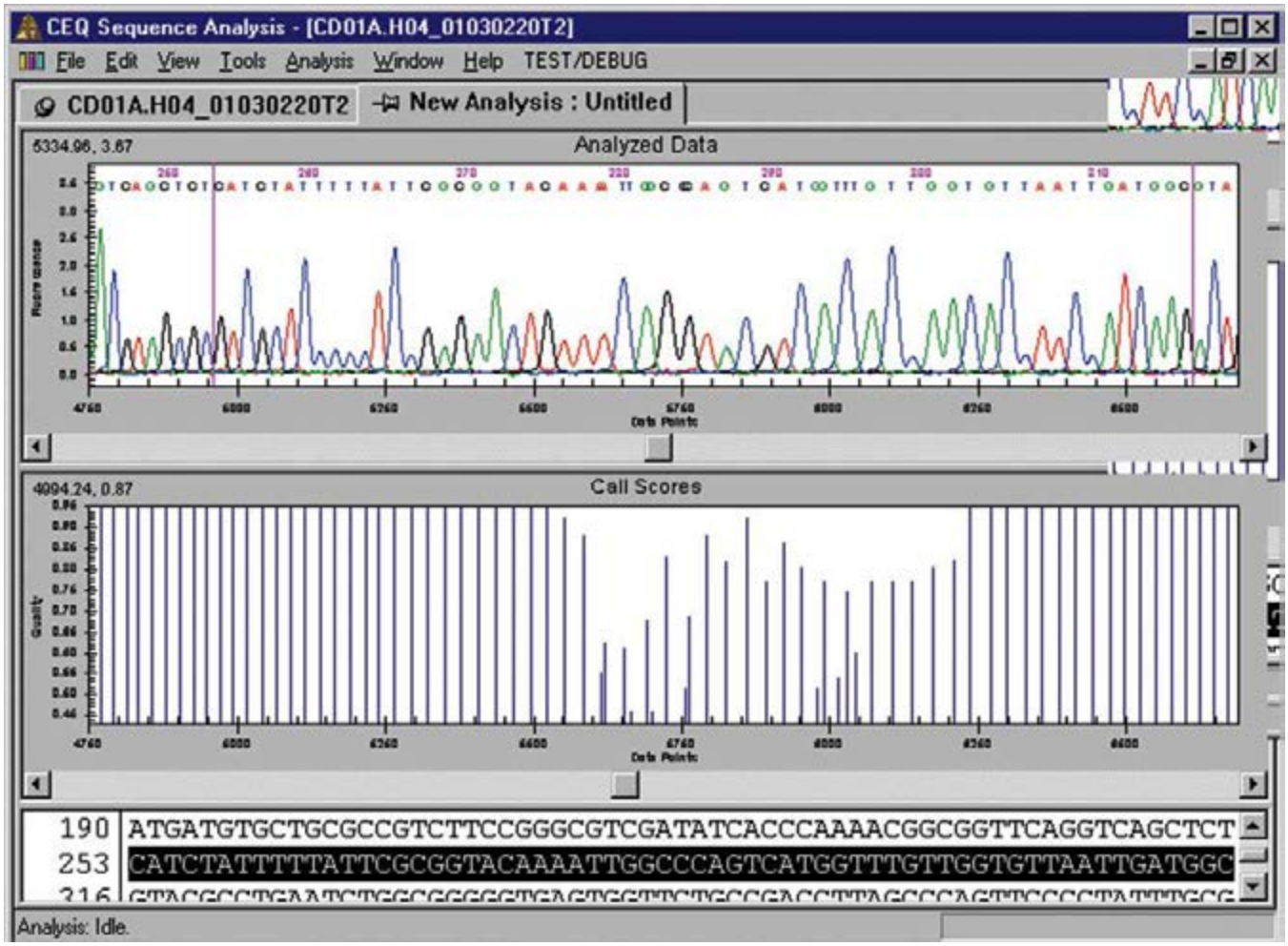

Figure 6. *Screen showing Call Scores, Base Sequence Text, and Trace Data for approximately 63 bases of data.*

Using Quality Values for Trimming

The use of Quality Values "in conjunction with appropriate assembly software can improve the accuracy and completeness of assembly by allowing better discrimination of repeats and by making it possible to use full read lengths; permit a more accurate consensus sequence to be derived; and provide an objective criterion for finishing… ."[4]" In the absence of such assembly software, it is often advantageous to trim the read ends. Most trimming software requires the user to set some fixed criteria (e.g., trim the first 50 bases and last 100 bases of each read) or to set criteria based on the number of Ns within a fixed window (e.g., three Ns within a 25-base window). Toward the goal of using as much sequence as possible from each read, trimming using Quality Values is preferable to the trimming methods described above.

In the example in Figures 8A-C, Trimming using the criterion of three Ns within a 25-base window results in a readlength of 663 bases (Figure 8C). Trimming using Quality Values results in a readlength of 768 bases (Figure 8A). Trimming with Quality Values was performed by looking for the Highest Scoring Read with error probabilities of <20% [$7 \cong -10 \times \log_{10}(0.2)$, was subtracted from each QV; QVs of all possible substrings were then summed and the length of the maximal scoring substring was taken].

Picking Primers Using Call Scores

Call Scores (or QVs) can be viewed to select regions of reads that are likely to be accurate. These regions can then be used to pick primers for further experiments. This may be particularly useful for polymorphism studies where indels (insertions/deletions) are involved or to resolve samples where multiple templates were amplified during the sequencing reaction. By choosing a primer that incorporates high-quality bases up to an ambiguous region and one or two possible bases into the ambiguous region, the template for a single sequence can then be amplified and its sequence determined. Primers can then be designed that incorporate the same high-quality bases followed by other possible one or two base combinations to sequence other templates that may be present.

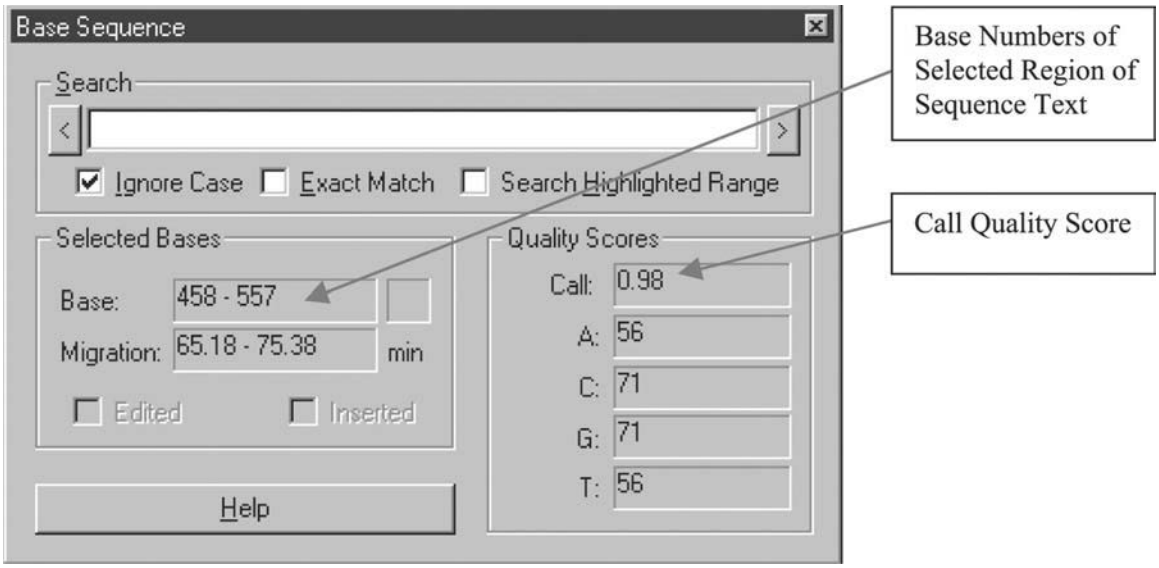

Figure 7. *Base Sequence Toolbar displays the "Call Quality Score" for a region of 100 bases.*

Using Call Scores Prior to Searching or Using BLAST

As detailed in the section titled "Using Call Scores as an Aid to Editing," editing a single-pass sequence prior to performing a BLAST can increase the search score. This can make it easier to determine the most likely identity of the template from which the sequence was determined. It can also prevent falsely identified hits from giving an erroneously high score. If a quality-weighted search program is available, editing the sequence prior to searching becomes unnecessary since bases with high error probabilities will be de-weighted when computing the score of search results.

Using Quality Values as an Aid to Alignment and Assembly

The value of using Quality Values when performing alignments and assemblies has been discussed throughout this document. Programs such as Phrap/Consed and others make use of one or another estimation of quality to improve the accuracy of consensus calling and to improve the efficiency of assembly by using full-sequence readlengths. An example of consensi determined both with and without Quality Values is illustrated in Figure 9. In the first view, the consensus as determined by Sequencher[8] is displayed. The second view shows the consensus from the same constituent sequences as determined by Phrap along with the Quality Values of the resulting consensus.

GTGCCAAGCTTAATCCTCACGAGCATCCTGTTCTGCACTCTGACCAGGGATGGCAGTATCGTATGAGAA-
GATATCAAAATATCCTTAAAGAACATGGTATTAAACAAAGCATGTCCAGAAAAGGCAATTGTCTGGATAAT-
GCTGTGGTGGAGTGTTTCTTTGGAACCTTAAAGTCGGAGTGTTTTTATCTTGAT-
GAGTTCAGTAATATAAGCGAACTGAAGGATGCTGTTACGGAATATATTGAATACTACAACAGCAGAA-
GAATTAGCCTGAAATTAAAAGGTCTGACTCCAATTGAATATCGGAATCAGACCTATATGCCTCGTGTT-
TAACTGTCCAACTTTTTGGGGTCAGTACAAACTTTGATTTATAGTCAGGTGGGGCTTTTCTCTGTCTGC-
CTTTCGGTGAATACCTGAGACAAACAGTCTCAAGCACCCGTGGCTATTCTAGCT-
TAATAAGTTTGTTTCTTCTCCTTGATATAATCCTAAAAAAATCTCATAAAATTAATATATGAGATAATCTT-
TATTCAGCAGAAGATTATTAAAGGTTGCTGTATTATTTAGCGATAAAAAAAGCCTGCCAGATGGCAGGC-
TATTTAATAACGGCGTTATTATTGCAACAGCGAAATATCCGCAACGCGCAGGAACAGTTCGCGCAGTTTC-
CTCAGCATGGTCAGACGGTGGATACGCACTCTTTTGTCATCACCATGACCATCACTTTATCGAAGAAAG-
CATCAACGGGTCACGCAGCTAGCAGTTCGACAGCGCA~~TCTGGTACGACCTCGTAAGTCGGCTCGCCT-~~
~~GTCGTGACCACCTCATGCAGTTATTCTCGCCTCAGGGTAGCATACCGGTCCACCTGTGATGCAAGTAAT-~~
~~GTTGTCCGCCGCACCCATAGAGAGTATCTTATTT~~

Figure 8A. *Sequence Text results after analyzing using Call Threshold of 0.0. Bases that would be trimmed using quality-based criterion as described in the text are shown in red and strikethrough.*

Conclusion

Many applications involving the use of automated DNA sequencing data benefit from the inclusion of the estimated accuracy of each called base. Processes involving multiple sequence fragments (such as sequence assemblies and alignments) often profit most from quality assessments that discriminate best between bases with the lowest error probabilities. Quality Values produced by Beckman Coulter's CEQ™ Sequence Analysis software provide such an assessment. Beckman Coulter has included Call Scores in the CEQ Sequence Analysis software in the hope that single-sequence applications (such as sequence editing and EST searching) will take advantage of its increased discriminating power among lower-quality bases.

16	16	16	16	12	14	16	31	47	47	47	47	47	47	47	47	39	39	39	39	21	18	21	18	39	31	39	39	31	39
39	39	31	30	30	39	39	47	39	31	39	31	31	39	31	31	31	31	30	30	31	47	39	31	30	39	47	47	47	47
47	47	47	47	47	47	47	47	39	39	31	39	39	39	31	39	47	47	39	47	47	31	39	39	39	47	47	47	47	39
39	47	39	39	39	47	47	47	47	47	47	47	47	47	39	31	47	47	39	39	47	39	47	39	39	39	39	39	39	47
47	39	47	47	47	47	47	47	47	47	47	39	47	47	47	39	47	47	47	47	47	47	47	47	47	47	47	47	47	47
47	39	22	22	38	38	47	47	47	47	47	39	47	47	47	47	47	47	47	47	47	47	39	47	47	47	47	39	31	39
47	47	47	47	47	39	31	47	39	47	31	17	30	47	39	31	39	39	47	39	39	47	47	47	47	47	47	47	47	47
39	39	39	47	47	47	47	47	47	47	47	39	31	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	39	39
39	47	47	47	47	47	47	47	47	47	47	39	47	39	47	39	39	47	31	39	36	31	31	31	31	47	47	36	47	47
47	36	36	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47
47	47	47	47	47	47	47	47	47	47	36	47	47	47	47	47	47	47	47	47	47	39	39	26	19	26	36	36	47	47
47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47	47
47	47	47	47	36	47	47	47	47	47	47	36	47	47	31	36	47	47	47	47	47	47	47	47	47	47	47	47	47	47
47	47	33	36	47	47	47	47	47	47	47	47	47	47	47	47	47	36	47	47	47	47	47	47	36	31	31	31	31	31
47	47	47	47	47	47	47	47	47	47	47	47	36	47	47	47	47	36	47	47	47	47	47	47	47	47	47	47	47	31
36	47	47	36	31	36	47	38	38	31	17	31	38	38	31	47	47	47	47	47	47	31	47	47	21	15	18	31	38	38
38	38	47	47	47	47	47	36	47	47	47	47	47	47	47	47	47	47	47	36	36	31	31	31	17	18	15	15	31	38
31	38	38	38	38	38	38	38	31	22	38	26	17	14	10	9	7	10	15	18	31	38	22	18	14	11	14	11	5	12
10	10	9	5	12	5	12	12	9	12	18	18	31	17	17	10	6	15	12	10	11	20	38	22	38	38	31	26	21	18
8	6	10	8	10	20	17	17	16	12	20	22	18	20	17	17	12	10	12	5	12	18	15	16	12	10	11	12	7	7
12	12	17	12	31	21	14	7	10	15	10	10	16	10	10	15	15	10	11	14	17	22	26	17	17	6	10	16	12	15
9	10	16	16	6	6	8	9	9	10	12	9	4	8	8	5	15	12	10	8	5	12	15	12	15	15	15	10	4	9
10	8	4	7	9	10	10	11	11	10	8	5	9	9	12	12	15	11	4	5	5	4	10	9	8	12	12	10	7	5
5	6	5	4	9	4	9	7	7	9	7	5	5	6	4	8	10	6	8	9	5	7	9	9	11	11	12	5	15	4
8	3	8	9	10	4	4	10	4	4	15	15	10	8	8	12	9	9	5	11	4	8	5	8	10	4	10	9	12	9
9	10	7	9	8	10	5	11	10	11	8	8	10	8	8	6	8	6	6	5	3	5	4	5	5	8	4	10	9	4
9	11	8	3	6	4	8	8	9	5	8	9	9	7	7	9	3	7	5	7	6	6	6	3	3	5	3	8	4	5
3	7	6	5	3	6	9	9	4	8	10	5	10	8	8	6	5	7	6	6	5	6	3	8	4	7	6	5	6	8
8	4	6	8	5	3	7	10	5	5	5	5	5	6	6	3	5	6	3	5	3	6	6	9	4	6	9	7	4	7
7	3	5	5	3	7	3	5	6	9	8	9	8	8	8	4	9	6	5	5	6	6	5	5	5	5	7	5	8	5
5																													

Figure 8B. *Quality Values for Sequences in 8A, 8C.*

GTGCCAAGCTTAATCCTCACGAGCATCCTGTTCTGCACTCTGACCAGGGATGGCAGTATCGTATGA-
GAAGATATCAAAATATCCTTAAAGAACATGGTATTAAACAAAGCATGTCCAGAAAAGGCAATTGTCTG-
GATAATGCTGTGGTGGAGTGTTTCTTTGGAACCTTAAAGTCGGAGTGTTTTTATCTTGAT-
GAGTTCAGTAATATAAGCGAACTGAAGGATGCTGTTACGGAATATATTGAATACTACAACAGCAGAA-
GAATTAGCCTGAAATTAAAAGGTCTGACTCCAATTGAATATCGGAATCAGACCTATATGCCTCGTGTT-
TAACTGTCCAACTTTTTGGGGTCAGTACAAACTTTGATTTATAGTCAGGTGGGGCTTTTCTCTGTCT-
GCCTTTCGGTGAATACCTGAGACAAACAGTCTCAAGCACCCGTGGCTATTCTAGCT-
TAATAAGTTTGTTTCTTCTCCTTGATATAATCCTAAAAAAATCTCATAAAATTAATATATGAGATAATCTT-
TATTCAGCAGAAGATTATTAAAGGTTGCTGTATTATTTAGCGATAAAAAAAGCCTGCCAGATG-
GCAGGCTATNTAATAACGGCGTTATTATTGCAACAGCGAAATATCCGCAACGCGCAGGAACAGT-
NCGCGCAGNTTCN~~TCAGCATGGTCAGACGGTNGATACGCANTCNTTTGTCATCACCATGANCAT-
CACTTTANCGANGAAAGCATCAACGGNTCACGCAGCNAGCAGNNCGACAGCNCATNNGGTAC-
GACCTCGNAAGNCNGCTCGCNTNTCGNNNCNANNNCATNCAGNNATTNTCGC-
NNNNGGGTNNNATANNNNTCCACCTNTNNNNCNNNTANNGTTGNNNNCNGCANNCATNGN-
NAGTATCTTNTTT~~

Figure 8C. *Sequence Text results after analyzing using Call Threshold of 0.6. Bases that would be trimmed using the criterion of 3 Ns in a 25-base window are shown in red and strikethrough.*

```
Read 1:    G  T  G  T  A  C  A  A  A  C  C  A  G  G  G  T  A  C
Read 2:    G  T  G  T  A  C  A  A  A  C  T  A  G  G  G  T  A  C
Consensus: G  T  G  T  A  C  A  A  A  C  R  A  G  G  G  T  A  C
```

Figure 9A. *Consensus sequence determined without the use of Quality Values.*

```
Read 1:    G  T  G  T  A  C  A  A  A  C  C  A  G  G  G  T  A  C
Read 2:    G  T  G  T  A  C  A  A  A  C  T  A  G  G  G  T  A  C
Consensus: G  T  G  T  A  C  A  A  A  C  C  A  G  G  G  T  A  C
           28 33 47 47 47 47 55 65 65 43 32 37 45 55 65 65 65 65
```

Figure 9B. *Consensus sequence determined using Quality Values. Differences in the consensi appear in* ***bold*** *type.*

References

1. Dear, S. and R. Staden. A standard file format for data from DNA sequencing instruments. *DNA Sequence 3,* 107-110 (1992).
2. Lawrence, C. B. and V. V. Solovyev. Assignment of position-specific error probability to primary DNA sequence data. *Nucleic Acids Research 22,* 1272-1280 (1994).
3. Berno, A. J. A graph theoretic approach to the analysis of DNA sequencing data. Genome Research. 6: 80-91 (1996).
4. Ewing, B. and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research 8,* 186-194 (1998).
5. Richterich, P. Estimation of Errors in "Raw" DNA Sequences: A Validation Study. *Genome Research 8,* 251-259 (1998).
6. Dang, V., P. Martz, and J. Vaks. "CEQ DNA Sequencer: Optimal Correlation of Peak Attributes to Errors Associated with Basecalls." June, 2002.
7. Ewing, B., L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research 8,* 175-185 (1998).
8. Sequencher , version 4.0. Gene Codes Corporation, Ann Arbor MI.

SCIEX
www.sciex.com/ce

View SCIEX products at www.sciex.com
Find your local office at www.sciex.com/offices

AB SCIEX Headquarters
500 Old Connecticut Path | Framingham, MA 01701 USA
Phone 508-383-7700
www.absciex.com