User Bulletin

KB™ Basecaller Software v1.4.1

November 2009

SUBJECT: KB™ Basecaller Software v1.4.1 User Bulletin

This user bulletin covers:

applied biosystems

Executive Summary

Applied Biosystems KB™ Basecaller Software v1.4.1 reduces manual data review time and elongates the read length of high-quality bases in sequences. This algorithm accurately extracts more bases out of the sequencing data generated on Life Technologies Corporation instrument and chemistry platforms. KB Basecaller Software v1.4.1 supports all BigDye® Terminator chemistries and run modules available on the Life Technologies Corporation:

- 310 Genetic Analyzer
- 3100/3100-*Avant* Genetic Analyzers
- 3130/3130*xl* Genetic Analyzers
- 3730/3730*xl* DNA Analyzers
- 3500 Dx and 3500 Dx/3500xL Dx Genetic Analyzers
- 3500 and 3500/3500xL Genetic Analyzers.

Software integration

KB Basecaller Software v1.4.1 is integrated with:

- Sequencing Analysis Software v5.4
- SeqScape® Software v2.7
- Variant Reporter™ Software v1.1
- 3130/3130*xl* and 3730/3730*xl* Data Collection Software v3.1
- 3500/3500xL and 3500 Dx/3500xL Dx Data Collection Software v1.0.
- MicroSEQ® ID Analysis Software v2.2

KB Basecaller Software v1.4.1 is *not* integrated with:

- MicroSeq® ID software versions 2.1 and older
- Any versions of Data Collection software for the 310 and 3100/3100-*Avant*
- 3130/3130*xl* and 3730/3730*xl* Data Collection Software versions before v3.1
- Sequencing Analysis Software before v5.4
- SeqScape® Software versions before v2.7
- Variant Reporter™ Software versions before v1.1.

During the co-installation of Sequencing Analysis Software v5.4 and SeqScape® Software v2.7 with Data Collection software v3.1, KB Basecaller Software v1.4.1 is installed into your Data Collection Software v3.1 on the same computer.

Testing on more than 50,000 sequencing samples shows that version 1.4.1 of the algorithm offers many advantages, including longer, accurate read lengths.

Details of the test and validation process are in the poster *Longer Reads and More Robust Assemblies with the* KB Basecaller.

IMPORTANT! Life Technologies Corporation strongly recommends using the KB Basecaller.

Benefits of using the KB™ Basecaller

Some benefits of using the KB Basecaller include:

- Increased length of read
- Per-base quality value predictions using an equation that is standardized by Phred software
- Optional detection of mixed-base with quality values
- Analysis of short PCR products
- Accurate start point detection
- Increased accuracy in regions of low signal-to-noise or anomalous signal artifacts
- Detection of failed samples
- Trimming of data using per-base quality value
- Per-sample quality value that helps to determine the quality of each read
- Optional detection of PCR stop
- Optional assignment of Ns
- Optional generation of .phd.1 files

Increased length of read

KB Basecaller accurately extracts more bases than ABI Basecaller from the 3′ and 5′ ends of a sequence. Tests on genomic BAC samples, performed on data generated using 3730/3730*xl* instruments, indicate an improvement of approximately 100 bases in length-of-read as compared to the same data analyzed by the ABI Basecaller and Phred software (v0.020425.c). The gain in read length varies depending on the run module used to collect the data. The accuracy of start point estimation and the first 50 bases of called sequence is substantially increased. Typically, ~10 more correct calls on average are identified at the 5′ end as compared to the ABI Basecaller.

Per-base quality value predictions

The KB Basecaller assigns quality values to every basecall. The quality prediction algorithm is calibrated to return Q values that conform to the industry-standard relation established by the Phred software. The KB Basecaller and its output are, therefore, interchangeable in processes requiring Phred software for output.

Quality value calibration was performed using a set of correct-sequence annotated sample files, representative of production sequencing data generated on capillary electrophoresis platforms. Over 23 million basecalls were used to calibrate KB Basecaller and over 12 million distinct basecalls were used to test the calibration.

Accuracy in start point detection

Improved start point detection contributes to better mobility shift corrections and greater basecalling accuracy in the first 50 bases. Because the KB Basecaller detects the start point accurately, you do not need to manually set start points for each sample.

Optional detection of mixed-base with quality values

The KB Basecaller can detect mixed base positions and assign IUB codes and quality values to those positions. Quality values are assigned to mixed basecalls using an algorithm similar to that for pure bases.

The definition conforms to the Phred relation. Quality values for mixed bases are inherently lower than those of pure bases due to the higher error risk of interpreting more complex signals. Note that when using the ABI Basecaller or ABI Basecaller and Phred software, a separate analysis stage is required to determine mixed bases.

Increased accuracy in regions of low signal-to-noise or anomalous signal artifacts

The KB Basecaller increases the accuracy of sequence reads from low-signal regions or from data that are partially contaminated by a secondary sequence or by other sources of "chemistry noise."

Basecalling errors caused by anomalous chemistry and/or instrument signals such as dye blobs and fluorescent spikes are substantially reduced. These artifacts often occur in otherwise high-quality "clear-range" data. They result in the loss of high-quality bases that are downstream from the noise region. Tests indicate that KB Basecaller distinguishes between target DNA peaks and the most common artifacts better than ABI Basecaller.

Analysis of short PCR products

The KB Basecaller has been tested for accuracy in basecalling and quality value estimates on PCR products as short as 100 bases. Although KB Basecaller may be able to basecall products with less than 100 bases, these types of sample files were not tested.

Detection of failed samples

The KB Basecaller indicates the gross sample quality of each analysis as "Success without warnings," "Success with warnings," or "Failure due to poor data quality." A common failure mode is no signal–insufficient detection of DNA peaks. For failed samples, the KB Basecaller uses "NNNNN" as the sequence, indicating that the sample quality is very low and may need to be omitted from further analysis. Failed samples are flagged in reports in the analysis software. Note that this behavior is different from the ABI Basecaller, which *always* tries to call bases, resulting in sequences of many Ns.

Option to trim data using per-base quality value

You can use software with KB Basecaller to automatically determine the clear range region by trimming the ends using the per-base quality values. The parameters used for trimming are similar to those in other tools used by the genome community.

Per-sample quality value (QV) evaluates quality of reads

Software with the KB Basecaller uses the QV from the KB Basecaller to trim and determine a sample score. The sample score is the average QV in the clear range, or, if no clear range is determined, in the entire read. This single number value is a measure of the quality of the data. The sample score appears in reports generated by Sequencing Analysis Software, SeqScape® Software, Sequence Scanner Software, Variant Reporter™ Software, and/or MicroSeq® ID Software.

Optional detection of PCR stop

You can set the KB Basecaller to end basecalling at a PCR stop. Note that samples with enzymatic failure may have signal properties similar to those in PCR stop conditions. The KB Basecaller may not be able to distinguish between these two conditions.

Optional assignment of Ns

By default, the KB Basecaller does not generate Ns. However, you may choose to reassign Ns to bases with QVs below a user-specified threshold for both pure and mixed base positions.

Optional generation of .phd.1 files

.phd.1 files can be generated by autoanalysis or in analysis software.You can use the .phd.1 files for further analysis by downstream software such as Phred software.

Future support of ABI and KB™ Basecaller

Life Technologies Corporation will continue to provide technical support for the ABI Basecaller. However, further development and defect fixes will occur only on the KB Basecaller. If you encounter a defect in the ABI Basecaller, please use the KB Basecaller instead. In future releases, ABI Basecaller support files are removed from the software wherever they duplicate support in the KB Basecaller.

Features in KB™ Basecaller Software v1.4.1

- A basecalling algorithm that supports Applied Biosystems 3100/3100-*Avant*, 3130/3130*xl*, 3730/3730*xl*, 3500/3500xL, and 3500 Dx/3500xL Dx Genetic Analyzers
- Improvements over all earlier versions of KB Basecaller (v1.0, v1.1, v1.1.1, v1.1.2, v1.2, v1.3, and v1.4)

Note: Basecalling results with KB Basecaller Software v1.4.1 may differ slightly from results obtained with previous versions of KB Basecaller.

Comparison of the ABI and KB™ Basecallers

Question	ABI Basecaller	KB™ Basecaller
What does the software do?	• Processes raw traces • Provides processed traces • Provides AGCTN calls	• Processes raw traces • Provides processed traces • Provides pure bases only *or* • Provides pure and mixed calls • Provides quality values • Generates .phd.1 and .scf files • Provides a sample score
What are the resulting basecalls?	One option available: Only mixed bases are assigned as Ns. Further processing (either manual or using additional software) is required to assign IUB codes to the Ns or pure bases.	Four options are available. The software can assign an: • ACGT and Q value to each peak. • ACGT and Q value to each peak. Any peak with a Q value below a defined threshold is reassigned an N. • ACGT or a mixed base and a Q value to each peak. • ACGT or a mixed base and a Q value to each peak. Any peak with a Q value below a defined threshold is reassigned an N.
How are failed samples handled (for example, no signals, chemistry failure)?	Attempts to call all bases so a sample results with many Ns.	Assigns five Ns to the entire sample to indicate that the sample failed analysis. The analysis report flags these files.
Baseline in processed data	Appears smoother than in KB Basecaller.	Appears less smooth than in ABI KB Basecaller.
What are the steps to process data?	Calls bases on Windows OS.	Calls bases and estimates QVs on Windows OS.
Data and future support	Supports the 310, 3100, 3100-*Avant*, Applied Biosystems 3130/3130*xl* and 3730/3730*xl* instruments. Further development has stopped.	Applied Biosystems 3100/3100-*Avant*, 3130/3130*xl,* 3730/3730*xl,* 3500/3500xL, and 3500 Dx/3500xL Dx Genetic Analyzers. Development is ongoing.

Differences between the ABI and KB™ Basecallers

Table 1 Differences between the ABI and KB™ Basecallers

Question	Answer	
	ABI Basecaller	KB™ Basecaller
Can the KB Basecaller basecall short PCR products?		The KB Basecaller has been tested for accuracy in basecalling and quality value estimation on PCR products as short as 100 bases. Although it may be possible to basecall products with less than 100 bases, such sample files have not been tested. Samples shorter than 100 bases may not contain enough signal information to basecall the sample file.
Why is the baseline less smooth when the data are analyzed with the KB Basecaller?	Processed signals or traces from the ABI Basecaller appear smoother than those from the KB Basecaller because each software application uses an algorithm that processes the signals differently. The ABI Basecaller assigns only AGCT and Ns to each peak. Therefore, you must manually search for mixed bases or use a secondary software to complete the task. To facilitate this secondary process, the ABI Basecaller subtracts an aggressive baseline estimate to show a cleaner baseline in the processed signals.	The KB Basecaller can determine pure and mixed bases. Therefore, second-stage processing, which allows less aggressive baseline subtraction, is not needed. The processed traces have a higher baseline. If you have mixed bases, turn on the mixed-base detection option and allow KB Basecaller to call mixed bases. Use the mixed base calls and the associated QVs to review mixed bases – do not look only at the baseline.
What is the signal to noise value found with data analyzed with the KB Basecaller?	The signal-to-noise value is the average of the signal intensity of the A, C, G, or T base divided by the average of the noise for that base. The ABI Basecaller calculates only the signal intensity. The signal-to-noise value is more indicative of data quality than the signal intensity value alone. Both properties are important in determining quality.	KB Basecaller calculates the information and presents the data in the Annotation view and analysis report.

Question	Answer	
	ABI Basecaller	KB™ Basecaller
What scaling options are available with the KB Basecaller?	The ABI Basecaller uses a scaling method closer to the "True profile" option than the "Flat profile" option.	The KB Basecaller can display scaled data in two ways: • True profile scaling With this method, the processed traces are scaled uniformly so that the average height of peaks in the region of strongest signal is about equal to a fixed value (for example, 1000). The profile of the processed traces is very similar to that of the raw traces. • Flat profile scaling The processed traces are scaled semi-locally so that the average height of peaks in any region is about equal to a fixed value (for example, 1000). The profile of the processed traces is flat on an intermediate scale (> about 40 bases). You must decide which option is better suited to your circumstances. The sequence and QVs called by the KB Basecaller are independent of the selected scaling option.
Does the KB Basecaller produce more usable sample files than the ABI Basecaller?		Tests show that medium- and high-quality data result in more usable bases (longer read length) when analyzed by the KB Basecaller than by the ABI Basecaller. For very poor-quality data (samples with no, low, or noisy signal), the KB Basecaller does not provide more bases but instead fails the samples. By calling a string of "NNNNN" for the failed samples (instead of a sequence containing low QVs), the KB Basecaller indicates that the sample is unusable.
Can the KB Basecaller analyze data generated on the ABI PRISM® 373, 377, or 3700 instruments?		No, the KB Basecaller is calibrated to basecall and estimate the basecall quality for BigDye® Terminator chemistries on ABI PRISM® 310, 3100, 3100-*Avant*, and 3130/3130*xl* Genetic Analyzers, 3730/3730*xl* DNA Analyzers, and 3500/3500xL and 3500/3500 Dx/3500xL Dx Genetic Analyzers. Life Technologies Corporation has stopped support for the 373, 377, and 3700 instruments and data analysis.

Question	Answer	
	ABI Basecaller	KB™ Basecaller
How can I determine which basecaller was used to analyze each sample file?		The Annotation view for each sample file and for the print header displays the basecaller name and version number. When displaying samples files, files analyzed by the KB Basecaller have QV value bars displayed above the electropherogram.
Are there any known incompatibilities when a sample file is analyzed with the KB Basecaller?		Life Technologies Corporation does not know of any incompatibility issues when a sample file (.ab1) is analyzed with the KB Basecaller and used in third-party software.

FAQs: Processing data with Phred software and .phd.1 Files

Question	Answer
Can I analyze sample files with the KB™ Basecaller and then reprocess them with Phred software?	In principle, yes, but this is not recommended. The resulting quality values from Phred software are not calibrated—i.e., it is possible that Phred will over or under-predict quality in certain circumstances because it has not been trained on the type of processed electropherogram produced by the KB Basecaller. (Phred has been trained using the ABI Basecaller to produce the processed traces.) In addition, Phred replaces (and ignores) the initial called sequence. Reprocessing KB-analyzed samples with Phred, on average, degrades the accuracy of the analysis in terms of actual sequence error. Analysis improvements in KB Basecaller outlined above are lost. Studies by Life Technologies Corporation indicate that running Phred software on sample files processed by the KB Basecaller degrades the quality of the results. Analysis with KB Basecaller can generate .phd.1 files, which are interchangeable with any processes that currently depend on Phred.
Which Applied Biosystems software generates .phd.1 files?	The following software products have KB Basecaller (version varies for each software) integrated and can generate .phd.1 files: • ABI PRISM® 3100-*Avant* Data Collection Software v2.0 • ABI PRISM® 3100 Data Collection Software v2.0 • Applied Biosystems 3130/*xl* and 3730/*xl* Data Collection Software v3.0 and later • Sequencing Analysis Software v5.2 and later • SeqScape® Software v2.5 and later • MicroSeq® ID Software v1.0 and later • Variant Reporter™ Software v1.0 and later

FAQs: Quality values

Table 2 Quality values questions and answers

Question	Answer
How do I use quality values to review data?	When analyzing data with pure bases, Life Technologies Corporation recommends that you use the following settings: **Pure bases** – Low QV = <15, Medium QV= 15 to 19, High QV= 20+ (default) When reviewing data with pure bases, use the QVs to briefly review bases with high QV(>20). Pay close attention to bases with medium QVs because you may need to make edits. Quickly review low-QV bases, although you will likely discard these bases from further analysis. Mixed base quality values will be lower than pure bases. For mixed bases, you may want to review and accept basecalls with quality values as low as 5. **Mixed bases** – Low QV = <5, Medium QV = 5 to 10 (investigate to determine the best range for your application) In all cases, keep in mind that, by definition, the predicted probability of error for a particular basecall is $10^{-q/10}$.
What are the differences between quality values of mixed bases and pure bases?	Pure bases and mixed bases have the same probability of error for the associated basecall (10-q/10). Note the following: • High-quality pure bases typically have QVs of 20 or higher. • The distribution of quality values for mixed bases differs dramatically from that of pure bases. • For mixed bases, quality values greater than 30 are rare. • Good mixed bases may be assigned quality values as low as 5, because the probability of error with mixed bases is higher. Review mixed bases with QVs between 5 and 20.
Can I trim my data using quality values?	Yes. When using Data Collection, you can set trimming using QVs in the analysis protocols. When using Sequencing Analysis Software, SeqScape® Software, MicroSeq® ID Software or Variant Reporter™ Software, you can set trimming using QVs in the Analysis settings.

<table>
<tr><th>Question</th><th>Answer</th></tr>
<tr><td>Is there a table that shows each quality value and its corresponding probability of error?</td><td>The following table shows each quality value and its corresponding probability of error. For a more extensive table, look in the Help menu or the Sequencing Analysis Software or the SeqScape® Software user guides.

<table>
<tr><th>QV</th><th>Pe</th><th>QV</th><th>Pe</th></tr>
<tr><td>1</td><td>79.0%</td><td>35</td><td>0.032%</td></tr>
<tr><td>5</td><td>32/0%</td><td>40</td><td>0.010%</td></tr>
<tr><td>10</td><td>10.0%</td><td>41</td><td>0.0079%</td></tr>
<tr><td>15</td><td>3.2%</td><td>45</td><td>0.0032%</td></tr>
<tr><td>20</td><td>1.0%</td><td>50</td><td>0.0010%</td></tr>
<tr><td>21</td><td>0.79%</td><td>60</td><td>0.00010%</td></tr>
<tr><td>25</td><td>0.32%</td><td>99</td><td>0.00000000013%</td></tr>
<tr><td>30</td><td>0.10%</td><td></td><td>–</td></tr>
</table>
</td></tr>
<tr><td>Where can I see quality value bars and numbers?</td><td>Sequencing Analysis Software, SeqScape® Software, MicroSeq® ID Software, and Variant Reporter™ Software allow you to display or hide quality value (QV) bars in displays and printouts. You can customize the color and range for low, medium, and high quality values. For QVs ≤ 50, the length of a bar is proportional to the corresponding quality value. Quality values above 50 will have the same color and QV bar length as those defined for a QV of 50. To see the quality value for a particular base, place the computer mouse over the QV bar.

In SeqScape® Software, MicroSeq® ID Software, and Variant Reporter™ Software, the per-base quality values also appear in the reports corresponding to bases identified as mutations.</td></tr>
<tr><td>Why are the quality value bars displayed in gray?</td><td>A quality value is assigned to a specific basecall. When you change a basecall, the quality value does not apply to the new base, and therefore, it is displayed as a gray bar.

Also when you reassign Ns to bases below a certain QV, the QV bar does not apply to the N basecall, and therefore it is displayed as a gray bar.</td></tr>
<tr><td>Are quality value bars printed for the Electropherogram or Sequence views?</td><td>You can show or hide the QV bars when printing the Electropherogram and Sequence views of the sample file. QV bars are not printed if you print more than seven panels per page (due to space limitations). The quality value numbers cannot be printed.</td></tr>
<tr><td>Which Life Technologies Corporation software can display the quality values?</td><td>Sequencing Analysis Software v5.X, SeqScape Software v2.X, MicroSeq® ID Software v1.X, v2.X, and Variant Reporter™ Software v1.X can display quality values.</td></tr>
<tr><td>Can I view quality values from KB™ Basecaller with other software?</td><td>Quality value graphics from KB Basecaller are customized for processing by other Life Technologies Corporation software. The KB Basecaller allows other Life Technologies Corporation software to perform additional functions, such as clear range trimming and more streamlined editing.</td></tr>
</table>

Miscellaneous FAQs

Some frequently asked questions regarding Ns, spacing values, and providing feedback are shown below.

Question	Answer
When do Ns appear in samples analyzed by the KB™ Basecaller Software?	When using the KB Basecaller, the sequence “NNNNN” appears in the sample file when the sample fails analysis. Omit this file from further analysis. The Analysis Report in Sequencing Analysis Software will also flag these files. In addition to pure and mixed bases shown with QV bars, N’s and gray QV bars are also shown when you reassign Ns to all bases before the user-specified QV threshold. This allows you to view the longer read length and more accurate basecalling of KB Basecaller while still viewing data with software that does not display QVs.
Why does the spacing value sometimes appear in red?	When the ABI Basecaller fails to determine a spacing value for a sample file, it uses a default value of 12.00 for all run conditions. This number appears as in red in the Sample Manager, and the Annotation view displays “–12.00”.
Why does the spacing value sometimes have a negative value?	When the KB Basecaller fails to determine a spacing value for a sample file, it uses a default value specific to the instrument/polymer/chemistry/run condition used to generate the sample file. This value appears in red in the Sample Manager. The Annotation view displays –1 times this value.
How can I provide feedback to the KB Basecaller product team?	Email information to your local Applied Biosystems applications support representative. You can also email U.S. technical support at **GAlab@appliedbiosystems.com**. If applicable, please include sample files and details (including analysis settings) on how to reproduce your observation.

Conference posters and reference

Posters

- ABRF 2007 – *Improved Accuracy for Mutation and SNP Detection: Variant Reporter™ Software*, Ming Li et. al.
- ESHG 2007 – Direct Sequencing Quality Control
- AGBT 2004 – Longer Reads with the KB Basecaller
- ABRF 2004 – Integrated Sequencing Analysis Solutions using the KB™ Basecaller from Applied Biosystems

These posters and other literature can be found at:

http://www.appliedbiosystems.com.

Click **Support**, then **Products and Technical Literature**. Search with the keyword *KB*.

Reference

B. Ewing and P. Green, Genome Research, 8:186-194, 199.

This page is intentionally left blank.

Part Number 4401733 Rev. B 11/2009

Stock Number 117UB36-02

applied biosystems™

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm