Frequently Asked Questions

KB Basecaller Software v1.4

August 2007

SUBJECT: KB Basecaller Frequently Asked Questions

This user bulletin contains the following topics:

Applied Biosystems

Executive Summary

Applied Biosystems introduces the KB™ Basecaller Software designed to reduce manual data review time, elongate the read length of high-quality bases in sequences and thereby substantially reducing sequencing costs. This algorithm is designed to accurately extract more bases out of the sequencing data generated on current instrument and chemistry platforms provided by Applied Biosystems. KB Basecaller Software v1.4 supports all BigDye® Terminator chemistries and run modules available on the ABI PRISM® 310, 3100-Avant, 3100 and Applied Biosystems 3130/*xl* Genetic Analyzers and the Applied Biosystems 3730/*xl* DNA Analyzers.

Integration for Auto-Analysis

Products integrated with KB Basecaller Software v1.4

- Sequencing Analysis Software v5.3
- SeqScape® Software v2.6
- Variant Reporter™ Software v1.0

Products not integrated with KB Basecaller Software v1.4 for auto-analysis

- MicroSeq® ID Software v1.0 and v2.0
- 3100, 3100-Avant & 3730/*xl* Data Collection Software v2.0 and v3.0
- Sequencing Analysis Software versions prior to v5.3
- SeqScape® Software versions prior to v2.6

During the installation of Sequencing Analysis Software v5.3 and SeqScape® Software v2.6, KB™ Basecaller v1.4 will also be installed into your Data Collection Software v3.0 on the same computer. After you analyze your data, you can see the basecaller version in the Annotation view and the quality value bars in the Electropherogram view.

Extensive testing has been conducted on more than 50,000 of sequencing samples generated by Applied Biosystems and Applied Biosystems customers. Test results show that this new algorithm offers many advantages, including longer accurate read length.

Details of the test and validation process are published in a poster titled *Longer Reads and More Robust Assemblies with the KB Basecaller*. (See "Conference Posters and References" on page 16 for more information.)

IMPORTANT! Applied Biosystems strongly recommends using the KB Basecaller.

Key Benefits of Using the KB Basecaller

There are several benefits of using the KB Basecaller. These are listed below, with detailed information on each item below the list.

- Increased length of read
- Provides per-base quality value predictions using equation 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
- Provides per-sample quality value that facilitates determining the quality of each read
- Optional detection of PCR stop
- Optional assignment of Ns
- Optional generation of .phd.1 files

Increased length of read

The KB Basecaller uses advanced algorithms to accurately extract more bases from the 3' and 5' ends of the sequence. Our tests on genomic BAC samples indicate a measurable improvement of roughly 100 bases in length-of-read as compared to the same data analyzed by the ABI Basecaller and Phred software (v0.020425.c). The tests were performed on a data set generated by Applied Biosystems and several customer sites using 3730*xl* instruments. 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.

Provides per-base quality value predictions using equation standardized by Phred software

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 pipelines requiring Phred software or output.

Quality value calibration was performed using a controlled 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 there will not be a need for you to manually set start points for each sample.

Optional detection of mixed-base with quality values

The KB™ Basecaller provides the option to 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 associated with 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 extracted from low-signal regions or in data partially contaminated by secondary sequence or by other sources of "chemistry noise."

Basecalling errors caused by anomalous chemistry and/or instrument signals (e.g., dye blobs, fluorescent spikes) are substantially reduced. These artifacts are often found in otherwise high-quality "clear-range" data, resulting in the loss of high quality bases downstream from the noise region. Our tests indicate that KB Basecaller can better distinguish between target DNA peaks and the most common artifacts, thus allowing the basecaller to better "read through" the noise.

Analysis of 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. It is possible to basecall products with less than 100 bases; however these types of sample files were not tested.

Detection of failed samples

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

Provide the option to trim data using per-base quality value

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

Provide per-sample quality value (QV) that facilitates determining quality of reads

Software with the KB Basecaller integrated uses the QV provided by the KB Basecaller to trim and also determine a sample score. The sample score is the average QV in the clear range, or in the entire read when no clear range is determined. This single number is a useful metric to determine 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

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

Optional assignment of Ns

By default, the KB Basecaller will not generate Ns; however, you may choose an option to reassign Ns to bases with QV below a user-specified threshold for both pure and mixed base positions.

Optional generation of .Phd.1 files

.phd.1 files can be generated via auto-analysis or in analysis software. The .phd.1 files may be used for further analysis by down-stream software such as Phrap software.

Future Support of ABI and KB Basecaller

Applied Biosystems will continue to provide technical support for the ABI Basecaller; however, further development and defect fixes will only be done 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 will be removed from the software wherever there is duplicate support in the KB Basecaller.

What's New in KB Basecaller Software v1.4?

- Improvements over all earlier versions of the KB Basecaller (v1.0, v1.1, v1.1.1, v1.1.2, v1.2 and v1.3).
- The .scf files generated using the KB Basecaller will contain quality values.
- Content of "comment" block in .phd.1 output files conforms better to standards established by Phred.
- Calibration for TargetSeq Run module.
- Improved accuracy in detection of mixed bases, though sensitivity is marginally diminished.
- Decreased susceptibility to basecalling errors when modifications are made to the electrophoresis voltage.
- Decreased likelihood of calling one mixed base where two pure bases should be called when the two peaks are unusually close.
- Reduced false-positive mixed calls related to residual mobility shifts near 5' end.
- New quality metric output including PUP score.

  Note: In the comment block, the lines labelled TRIM and TRACE_PEAK_AREA_RATIO will always contain the following default values:

  - TRIM: -1 -1 -1.000000e+000
  - TRACE_PEAK_AREA_RATIO: -1.000000e+000

Comparison of the ABI and KB Basecallers

Table 1 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 & mixed calls • Provides quality values • Generates .phd.1 and .scf files • Provides a sample score
What are the resulting basecalls?	One option available: • Mixed bases are assigned as Ns. Further processing (either manually or via additional software) is required to assign IUB codes to the Ns or pure bases	Four options available: • Assigns ACG or T and Q value to each peak • Assigns ACGT and Q value to each peak, any peak with Q value below a defined threshold will be reassigned an N • Assigns ACG T or a mixed base and Q value to each peak • Assigns ACG T or a mixed base and Q value to each peak, any peak with Q value below a defined threshold will be reassigned an N
How are failed samples handled? (no signals, chemistry failure)	Will attempt to call all bases, so sample will result with many Ns	Will assign 5 Ns to the entire sample to signal that the sample has failed analysis Analysis report will flag these files
Baseline in processed data	Appears smoother	Appears less smooth. (See the FAQ on page 9.)
Steps to process data	Use ABI Basecaller to call bases on Windows OS	Use KB Basecaller to call bases and estimate QVs on Windows OS
Data and future support	ABI PRISM® 310, 3100-Avant, 3100, Applied Biosystems 3130/*xl* & 3730/*xl* instruments. Development has ceased	ABI PRISM 310, 3100-Avant, 3100, Applied Biosystems 3130/*xl* & 3730/*xl* instruments. On-going development

Differences Between the ABI and KB Basecallers

Table 2 Differences between the ABI and KB Basecallers

Question	Answer
Can the KB™ Basecaller be used to 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. It may be possible to basecall products with less than 100 bases; however such sample files have not been tested. Samples significantly shorter than 100 bases may not contain enough signal information needed by the basecaller to process the sample file.
Why does the baseline look less smooth when the data is analyzed with the KB Basecaller?	Processed signals or traces provided by the ABI Basecaller will appear smoother than those provided by the KB Basecaller because each algorithm uses distinct code that processes the signals somewhat differently. With the ABI Basecaller, only AGCT and Ns are assigned 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 a more aggressive baseline estimate to present a cleaner baseline in the processed signals. The KB Basecaller can determine pure and mixed bases and therefore there is no need for second stage processing, which allows less aggressive baseline subtraction. The processed traces will 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 simply look at the baseline.
What is the signal to noise value found with data analyzed with the KB Basecaller?	KB Basecaller calculates this information and presents the data in the Annotation view and analysis report. The ABI Basecaller will calculate only the signal intensity. The signal to noise ratio is more informative of data quality than the signal intensity value alone. Both properties are important in determining quality.

Table 2 Differences between the ABI and KB Basecallers *(continued)*

Question	Answer
What are the two scaling options available with the KB™ Basecaller?	With the KB Basecaller, you have two display options for scaling data: • 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 (e.g., 1000). The profile of the processed traces will be 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 (e.g., 1000). The profile of the processed traces will be flat on an intermediate scale (> about 40 bases). You should decide which option is better suited to your particular circumstances. The sequence and QVs called by the KB Basecaller are *independent* of the selected scaling option. Options for scaling data are not provided with the ABI Basecaller. The ABI Basecaller employs a scaling method closer to the “True profile” option than the “Flat profile” option
Will I get more “good” sample files using the KB Basecaller?	Our tests show that medium and high quality data will yield more usable bases (i.e. longer read length) when analyzed by the KB Basecaller as compared to results produced by the ABI Basecaller. For extremely poor quality data, the KB Basecaller will not provide more bases but instead will fail these samples – i.e. no signal, extremely low signals or extremely noisy signals. By calling a string of “NNNNN” for the failed samples (instead a sequence all containing low QVs), the KB Basecaller is signaling that the sample is unusable.
Can the KB Basecaller analyze data generated on ABI PRISM® 373, 377 or 3700 instruments?	No, the KB Basecaller is not calibrated for this task. It is calibrated to basecall and estimate the basecall quality for BigDye® Terminator chemistries on ABI PRISM® 310, 3100-Avant, 3100, and 3130/*xl* Genetic Analyzers and 3730/*xl* DNA Analyzers. Applied Bioystems ceased to support the 373, 377 or 3700 instruments and data analysis.
How can I tell which basecaller was used to analyze each sample file?	The Annotation view for each sample file and the print header contains the basecaller name and version number. When displaying samples files, files analyzed by the KB Basecaller will have QV value bars displayed above the electropherogram.
Are there any known incompatibilities when a sample file is analyzed with the KB™ Basecaller?	To our knowledge there are no known incompatibility issues when a sample file (.ab1) is analyzed with the KB Basecaller and used in third party software.

Processing Data with Phred Software and .phd.1 Files FAQs

Table 3 Processing Data with Phred Software and.phd1 Files FAQs

Question	Answer
Can I analyze sample files with the KB™ Basecaller and then reprocess it with Phred software?	In principle, yes, but this is not recommended. The resulting quality values from Phred software will not be 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, since Phred replaces (and ignores) the initial called sequence, re-processing KB-analyzed samples with Phred will, on average, degrade the accuracy of the analysis in terms of actual sequence error. Analysis improvements provided by KB Basecaller outlined above will be essentially lost. **Note:** Our studies indicate that running Phred software on sample files processed by the KB Basecaller significantly *degrades* the quality of the results. Analysis with KB Basecaller can output .phd.1 files, which are interchangeable with any pipeline that currently depends 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 330/*xl* Data Collection Software v3.0 • Sequencing Analysis Software v5.2 and higher • SeqScape® Software v2.5 and higher • MicroSeq® ID Software v1.0 and higher • Variant Reporter™ Software v1.0

Quality Values FAQs

Table 4 Questions and Answers about quality values

Question	Answer
How should I use quality values to review data?	When analyzing data with pure bases, we recommend that you set Low QV = <15, Medium QV= 15 to 19 and High QV= 20+ (default). When reviewing data with pure bases, use the quality values to briefly review bases with high QV >20. Pay close attention to bases with medium QVs as you may need to make edits. Quickly review low QV bases, though most likely you will discard these bases from further analysis. When reviewing mixed bases, your 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 10. In all cases, keep in mind that, by definition, the predicted probability of error for a particular basecall is equal to $10^{-q/10}$.
What are the differences between quality values of mixed bases and pure bases?	The definition of quality values is the same for pure and mixed bases. In both cases the probability of error for the associated basecall is $10^{-q/10}$. The distribution of quality values assigned to mixed bases, however, will differ dramatically from that for pure bases. Typically, high quality pure bases will be assigned QVs of 20 or higher. Good mixed bases, on the other hand, may be assigned quality values as low as 10. The reason that a high quality mixed base may receive such low QVs is that the probability of error with more complex signals is higher. Do not discard mixed bases with QV between 10 and 20. It is a good idea to review them. For mixed bases, quality values greater than 30 are rare.
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 4 Questions and Answers about quality values *(continued)*

<table>
<tr><th>Question</th><th>Answer</th></tr>
<tr><td>Is there a table mapping each quality value and the corresponding probability of error?</td><td>Below is a table mapping each quality value to the corresponding probability of error. For a more extensive table, look in the Help menu or the Sequencing Analysis Software or 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>0.0000000013%</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 provide the option to display or hide quality value bars in displays and printouts. You can customize the color and range for low, medium and high quality values. For QV ≤ 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 that defined for a QV of 50. To see the quality value for a particular base, hover 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 alter the basecall the quality value is no longer applicable to the new base, therefore it will be displayed as a gray bar.

Also when you choose to reassign Ns to bases below a certain QV, the QV bar is not applicable to the N basecall, therefore it will be displayed as a gray bar.</td></tr>
</table>

Table 4 Questions and Answers about quality values *(continued)*

Question	Answer
Are quality value bars printed for the Electropherogram or Sequence views?	You may choose to show or hide QV bars when printing the Electropherogram or Sequence view of the sample file. QV bars may not be printed if you choose to print more than 7 panels per page, due to space limitations. The actual quality value numbers cannot be printed.
Which Applied Biosystems software can display the quality values?	Sequencing Analysis Software v5.X, SeqScape Software v2.X, MicroSeq® ID Software v1.X, v2.X and Variant Reporter™ Software v1.X can all display quality values. Sequencing Analysis Software v3.X and SeqScape Software v1.X can open and display the sample files with quality values, however the QVs will not be displayed.
Will I be able to view quality values provided by KB™ Basecaller with other software?	Quality value graphic views are customized for software provided by Applied Biosystems. The design allows for additional functionality such as clear range trimming and more streamlined editing.

Miscellaneous FAQs

Table 5 Questions about Ns, spacing values and providing feedback

Question	Answer
When will I see Ns in samples analyzed by the KB™ Basecaller Software?	When using the KB Basecaller, you will see the sequence “NNNNN” when the sample failed 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 with QV bars, you may also see N’s and gray QV bars when you choose to reassign Ns to all bases before the user-specified QV threshold. This option is present to allow those who analyze data with the KB Basecaller but share data with others who do not have software than can display quality values. This allows you to take advantage of the longer read length and more accurate basecalling provided by the 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 will appear as in red in the sample manager and the Annotation view will display “-12.00”.

Table 5 Questions about Ns, spacing values and providing feedback *(continued)*

Question	Answer
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 particular instrument/polymer/chemistry/run condition used to generate the sample file. This number will appear in red in the sample manager and the Annotation view will display –1 times this value.
How do I provide feedback to the KB Basecaller product team?	Please send detailed information that illustrates your feedback to your local Applied Biosystems applications support representative. You may also send email to US technical support at **GAlab@appliedbiosystems.com**. Whenever possible, please include sample files and detailed instructions (including analysis settings) on how to reproduce your observation.

Conference Posters and References

Posters

- 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 Services Literature**. Search with the keyword *KB*.

References

1. B. Ewing and P. Green, Genome Research, 8:186-194, 1998.
2. **http://www.genome.washington.edu/UWGC/protocols/#DataAnalysisTools**

Worldwide Sales and Support

Applied Biosystems vast distribution and service network, composed of highly trained support and applications personnel, reaches 150 countries on six continents. For sales office locations and technical support, please call our local office or refer to our Web site at **www.appliedbiosystems.com**.

Applera is committed to providing the world's leading technology and information for life scientists. Applera Corporation consists of the Applied Biosystems and Celera businesses.

Headquarters
850 Lincoln Centre Drive
Foster City, CA 94404 USA
Phone: +1 650.638.5800
Toll Free (In North America): +1 800.345.5224
Fax: +1 650.638.5884

08/2007

Applied Biosystems

Part Number 4362968 Rev. B