Longer Reads and More Robust Assemblies with the KB Basecaller / P135

James Labrenz,[1] Curtis Gehman[1], Primo Baybayan,[1] Lisa L. Cook[3], Quynh Doan[1], Amy Egan[2], George Fry[1], Richard A. Gibbs[2], Christie L. Kovar[2], Lora R. Lewis[2], Elaine R. Mardis[3], Stephanie M. Moore[2], Donna M. Muzny[2], Anjali Pradhan[1], Stephanie Schneider[1], Graham B.I. Scott[2], Jon Sorenson[1], David L. Steffen[2], Donna M. Villasana[2], David Wheeler[2], and Shiaw-Pyng Yang[3] — [1]Applied Biosystems, Foster City, CA, [2]Baylor College of Medicine – Human Genome Sequencing Center, Houston, TX, [3]Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO

ABSTRACT

One critical aspect of genome sequencing is the useful read length, or number of high quality bases, produced from each sample. Longer read lengths contribute to more robust assemblies and to higher genomic coverage using fewer contributing reads, thus making the sequencing process more efficient and cost-effective. We present results demonstrating that the median useful read length of the KB basecaller exceeds that of other popular basecalling methods by about 100 bases. The test data consist of over 20,000 genomic BAC samples sequenced on Applied Biosystems 3730 and 3730*xl* DNA Analyzers, with the majority of reads generated by the production lines of the Baylor College of Medicine and Washington University genome sequencing centers. Our evaluation of basecalling and quality value accuracy on these reads uses alignments to the finished consensus sequences. For both the KB and *phred* basecalling algorithms, we compute Q20 scores, accuracy-based read lengths, predicted read lengths or *clear ranges*, and actual clear range error rates. Comparative statistics on these metrics demonstrate that the KB basecaller provides a substantial increase in read length over *phred*.

INTRODUCTION

The KB Basecaller was developed at Applied Biosystems to provide a complete and integrated basecalling solution, with quality value predictions on each base call that are statistically valid according to the standard relation $Q = -10 \times \log_{10}(P_E)$, where P_E denotes the probability that a basecall is in error [1]. The common approach, used at most genome centers, is to utilize the ABI basecaller in conjunction with *phred*. In this scenario, the ABI basecaller converts "raw" color data to *processed* traces, and *phred* uses the processed traces to re-call bases and assign quality values.

The KB Basecaller offers several advantages over the ABI-*phred* approach, both in terms of functionality and ease of workflow. These include

- Support for run modules and chemistries available on the Applied Biosystems 3730/*xl* DNA Analyzers and the ABI PRISM® 310, 3100-Avant and 3100 Genetic Analyzers
- Full and on-going quality value calibration support for these AB platforms
- Integrated options for heterozygous base calling
- Calibrated heterozygous quality values
- Fully integrated primary analysis (basecalling and quality values)
- Support of .SCF and PHD.1 files through Sequencing Analysis Software

In this poster we compare the actual performance of the KB v1.0 algorithm to the ABI[2] -*phred*[3] hybrid approach using metrics that characterize (1) basecalling accuracy, (2) length of read, (3) quality value accuracy, and (4) the predictive power, or discrimination ability, of the quality values. Finally, we show initial results comparing *phrap*[4] assemblies using the two basecalling algorithms.

MATERIALS AND METHODS

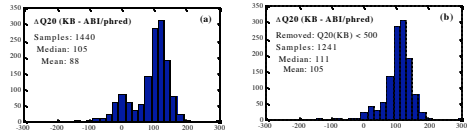
See the attached data sheet for an overview of the data sets. All samples from genome centers (BCM, WashU and JGI) were sequenced on the Applied Biosystems 3730*xl* DNA Analyzer using BigDye® Terminator v3 or v3.1 chemistry. Samples from AB were sequenced on 3730*xl* systems, using BigDye® Terminator v3, v3.1, v1 and v1.1 chemistries.

All reads have been aligned to an annotated reference sequence that is derived from known vector and consensus, and then analyzed for basecalling error. Descriptions of the metrics used to evaluate the basecallers are provided in the following sections.

RESULTS

Q20 Scores: The Q20 score for a read is defined as the number of basecalls that were assigned a quality value of 20 or greater by the basecaller.

Figure 1. Delta Q20 Scores



ΔQ20 is a per sample difference of the Q20 score, (KB - ABI/*phred*). Figure 1(a) shows a histogram of ΔQ20 for an arbitrary subset of the data from one of the genome centers. The histogram is bi-modal, indicating that there are two populations of sample types in the data—one for which the KB Basecaller improves over phred by over 100 bases, and one for which there is, on average, little or no improvement.

The quality of samples from production sequencing can be highly variable, and we expect that the lower mode represents poorer quality data. To verify this, we place a cut on the Q20 score of the KB analysis, keeping only those samples with Q20 ≥ 500. The resulting histogram of ΔQ20 indicates that the KB basecaller provides a median increase in Q20 of over 100 bases in the higher-quality reads.

Veracity Clear Range: The VCR is defined as the range of bases between 5′ and 3′ trim points that are computed based on local basecalling accuracy. We trim using a sliding window that tolerates no more than 3/100 errors at either end.

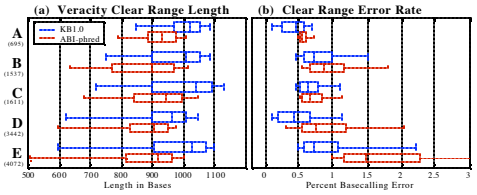


Figure 2. Veracity Clear Range Distributions, Long Read Module

Figure 2 shows distributions of (a) veracity clear range length, and (b) corresponding error rate within the clear range, for both algorithms on the data sets sequenced using the Long Read module. The latter metric provides a check that the identified clear range has a low overall rate of error. The colored bars show the median and middle 50% of distributions; the whiskers denote the 10th and 90th percentiles. The median clear range length of the KB v1.0 algorithm consistently exceeds that of *phred* by approximately 100 bases and the clear range error rates are generally lower.

Q20 Clear Range: The Q20CR is defined as the range of bases between 5′ and 3′ trim points that are computed based on a median quality value threshold of 20, using a sliding window of 30 bases.

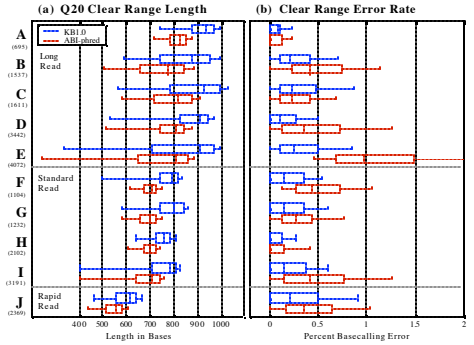


Figure 3. Q20 Clear Range Distributions

Figure 3 shows (a) predicted Q20 clear range lengths, and (b) observed basecalling error rates within the clear range, as measured by alignment to the reference. These data indicate that the basecalls and quality value predictions from the KB basecaller can increase useable length of read in most genomic samples by over 10% and provide a concurrent reduction in CR error.

Quality Value Accuracy: Statistical accuracy of the quality values is measured by comparing the observed quality values (based on observed error rates) for all base calls in the data set, binned according to predicted quality value.

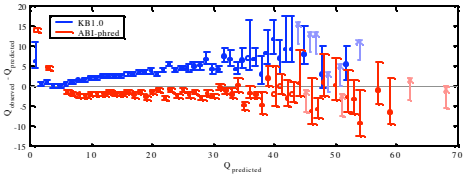


Figure 4. Quality Value Accuracy, Genome Center Data Only

Figure 4 shows a comparison of quality value accuracy on combined data from the genome centers (which were not used in KB v1.0 quality value calibration). We note that *phred* tends to slightly over-predict quality, while KB v1.0 tends to under-predict. We believe the accuracy deviations of KB v1.0 were caused by annotation errors in the training set that have since been corrected. The KB v1.1 calibration will use a "scrubbed" training set of over 20M bases.

Assembly Statistics: We have compared *phrap*[4] assemblies of the two BAC data sets sequenced at the HGSC, Baylor College of Medicine. Each experiment at a given assembly size consists of three random selections of input reads from the data set, with the same subsets used for each basecaller.

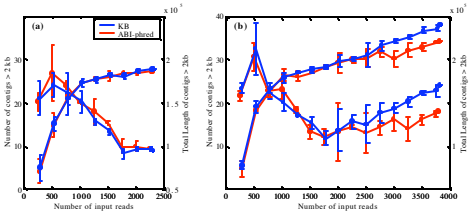


Figure 5. Comparison of *phrap* Assemblies

Figures 5 (a) and (b) each show two metrics used to characterize the state of the assemblies: the number of contigs > 2 kb (left axes, lines with decreasing trends), and the total length of contigs > 2 kb (right axes, lines with increasing trends). These results demonstrate that reads from the KB basecaller are fully compatible with *phrap* and produce assemblies that are at least as good. While there appears to be a trend in the second BAC at the larger assembly sizes (> 3000 reads) for the KB Basecaller to produce a longer total contig length, the significance of this result is not yet clear and requires further investigation. From these initial experiments, we have not seen a clear indication that the improvement in Q20 clear range lengths from the KB Basecaller substantially affects a *phrap* assembly at the early stages.

CONCLUSIONS

Using over 20,000 BAC reads, the majority collected from three major genome sequencing centers, we have shown that the KB Basecaller consistently produces Q20 clear range lengths that exceed those of the standard ABI-*phred* approach by more than 10% (over 100 bases using the Long Read module of the AB 3730*xl* DNA Analyzer), with a substantial reduction in the clear range error rates.

Experiments designed to determine how these basecalling improvements affect shotgun assemblies are in the early stages and are at this point inconclusive. While the improvement using *phrap* appears to be marginal, we note that one key feature of this program is its use of the lower-quality portion of the reads in forming pair-wise alignments. We anticipate that similar comparisons using a whole-genome shotgun assembler will show marked improvement in the KB assemblies, since reads must be trimmed at a Q20 threshold. Verification of this hypothesis is work in progress.

REFERENCES

1. B. Ewing and P. Green, *Genome Research*, **8**:186-194, 1998.
2. ABI Basecaller version 1.5.1.
3. *phred* software version 0.020425.c.
4. *phrap* software, http://www.phrap.org.

ACKNOWLEDGEMENTS

Applied Biosystems gratefully acknowledges the ongoing contributions of sample files and consensus sequences from our customer test sites for basecaller development. In particular, we thank members from GSC Washington University, St. Louis and those from HGSC at Baylor College of Medicine. In addition, we would like to thank JGI for their data contribution.