Journal of
Data Mining in Genomics & Proteomics

ISSN: 2153-0602

Research Article Open Access

A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS

Xiaoli Jiao[1#], Xin Zheng[1#], Liang Ma[3#], Geetha Kutty[3], Emile Gogineni[3], Qiang Sun[2], Brad T. Sherman[1], Xiaojun Hu[1], Kristine Jones[4], Castle Raley[4], Bao Tran[4], David J. Munroe[4], Robert Stephens[2], Dun Liang[1], Tomozumi Imamichi[1], Joseph A. Kovacs[3], Richard A. Lempicki[1] and Da Wei Huang[1]*

[1]*Laboratory of Immunopathogenesis and Bioinformatics, SAIC-Frederick, Inc., Frederick National Laboratory, MD 21702, USA*
[2]*Advanced Biomedical Computing Center, SAIC-Frederick, Inc., Frederick National Laboratory, MD 21702, USA*
[3]*Critical Care Medicine Department, National Institutes of Health, Bethesda, MD, 20892, USA*
[4]*Advanced Technology Program, SAIC-Frederick, Inc., Frederick National Laboratory, MD 21702, USA*
[#]*These authors equally contributed to the study*

Abstract

PacBio RS, a newly emerging third-generation DNA sequencing platform, is based on a real-time, single-molecule, nano-nitch sequencing technology that can generate very long reads (up to 20-kb) in contrast to the shorter reads produced by the first and second generation sequencing technologies. As a new platform, it is important to assess the sequencing error rate, as well as the quality control (QC) parameters associated with the PacBio sequence data. In this study, a mixture of 10 prior known, closely related DNA amplicons were sequenced using the PacBio RS sequencing platform. After aligning Circular Consensus Sequence (CCS) reads derived from the above sequencing experiment to the known reference sequences, we found that the median error rate was 2.5% without read QC, and improved to 1.3% with an SVM based multi-parameter QC method. In addition, a *De Novo* assembly was used as a downstream application to evaluate the effects of different QC approaches. This benchmark study indicates that even though CCS reads are post error-corrected it is still necessary to perform appropriate QC on CCS reads in order to produce successful downstream bioinformatics analytical results.

Keywords: PacBio; CCS read; quality control (QC); pass number; quality value (QV); SVM regression; assembly

Introduction

The PacBio RS platform, a newly emerging third-generation DNA sequencer produced by Pacific Biosciences, Inc., is based on a real-time, single-molecule, nano-nitch technology [1-3]. Besides several advantages over earlier generation sequencers, such as no PCR-amplification, single molecule sequencing, and shorter turn-around time, the most distinct feature of PacBio is the very long reads that are produced ranging up to ~10 kb for raw reads and ~2.5 kb for the error corrected, Circular Consensus Sequence reads (see definition in next paragraph) [4]. In contrast, the earlier generation sequencers typically generate much shorter reads with median lengths of ~100-200 bp for Illumina and ~500 bp for 454 [1,3,5-8].The longer reads produced by the PacBio platform is a key progression in the high-throughput sequencing field, which is expected to benefit many aspects of genomic projects in the near future. For example, assembling a genome with highly repetitive DNA, closing gaps in genome assemblies, phasing analysis of DNA polymorphisms, discovering rare isoforms of a highly conserved gene family, and identification of rare gene alternative splicing, which all remain challenging tasks using the shorter reads derived from earlier generation sequencers, would benefit from this approach [9-11].

Although PacBio's longer reads provide new power to researchers, careful error and Quality Control (QC) of the reads is essential to effectively use such power. Regardless of the ~15% error rate reported for the raw sub reads of PacBio [1,10], one of the standard outputs from the platform is the Circular Consensus Sequence(CCS) read (the throughput is ~10-20 k per SMRT cell), which is an error-corrected consensus read derived from the multiple alignment consensus of sub reads belonging to the same single-molecule circular sequencing [1-3,5].The Pass number is a unique feature of the PacBio platform when forming CCS reads [4]. It represents how many rounds the same single-molecule is sequenced in a hairpin structure during the PacBio circular sequencing procedure [1,2,4]. Since the CCS reads are post error-corrected, users often optimistically treat them as high quality reads without a upstream QC step before the downstream assembly or other bioinformatics analysis. In fact, because PacBio RS is a new platform, the CCS read QC related questions of the individual CCS reads remains largely unanswered even though some studies have addressed the accuracy and sequencing bias in a global view of raw reads [12,13]. Is the accuracy of all CCS reads good enough for downstream analysis? How much can the accuracy of the CCS reads be improved by applying appropriate QC filters? What is the impact of CCS read QC on the downstream assembly analysis? To answer these questions, we prepared a complex DNA mixture sample with 10 closely related and known DNA amplicons, which serves as a standard benchmark dataset. After sequencing the mixed DNA sample with the PacBio RS platform, we assessed the accuracy of the CCS reads from different angles in order to answer the above questions. This study can help analysts understand the general characteristics of CCS read accuracy as well as the gain and tradeoff of QC filters, in order to appropriately QC CCS reads for different study purposes. In addition, the benchmark dataset and QC-matrix query script in this study are freely available (http://david.abcc.

***Corresponding author:** Da Wei Huang, Laboratory of Immunopathogenesis and Bioinformatics, National Institutes of Health, Bethesda, MD, 20892, USA, E-mail: huangdawei@mail.nih.gov

Received May 20, 2013; **Accepted** July 25, 2013; **Published** July 31, 2013

Citation: Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, et al. (2013) A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. J Data Mining Genomics Proteomics 4: 136. doi:10.4172/2153-0602.1000136

ncifcrf.gov/manuscripts/pacbio_qc) for researchers to facilitate their own QC work.

Materials and Methods

Construction of a pacbio benchmark dataset

Ten plasmid clones, containing 10 different known *Pneumocystis jirovecii* Major Surface Glycoprotein (MSG) isoforms (sharing 80-90% similarity, ~3 kb length) [14], were mixed together in an even amount of 1 µl each, as PCR templates. A PCR reaction with a pair of primers in the MSG conservative regions amplified a ~1.5 kb mixture product of the 10 MSG isoforms in plasmids. The mixed PCR products (for sequences see http://david.abcc.ncifcrf.gov/manuscripts/pacbio_qc) provide a good control dataset to test the sequencing capability of the PacBio RS platform.

1 µg of the above mixed PCR products were used to construct one PacBio DNA library using the PacBio standard 2 kb template prep protocol. In addition, the 608-bp DNA fragment consisting of a randomly generated sequence, provided in the PacBio library kit, was spiked into the true DNA sample as a technical control in the PacBio sequencing procedure. Thereafter, the samples were sequenced on the PacBio RS platform on a single SMRT Cell (part number 001-350-385). C2 Polymerase (part number 001-672-551) was used for the sequencing reaction and ninety-minute movie windows were used for signal detection. After raw sequence data was generated, the base calling and CCS read generation was done using version 1.3.0 of PacBio's instrument control and SMRT Analysis software (http://www.pacificbiosciences.com). 10712 CCS reads, including 9812 study CCS reads and 900 PacBio spike-in control CCS reads, were obtained from the above procedure in FASTA and FASTQ files. Moreover, all PacBio sequencing related data was archived in a H5 formatted file for custom queries later.

Bioinformatics analysis of the pacbio ccs read accuracy and associated qc parameters

To evaluate the accuracy of CCS reads, each of the 9812 study CCS reads in the FASTA file were compared to the 10 prior known MSG isoform sequences. The comparisons were performed using the NCBI standalone BLAST program with default parameters. For a given CCS read, the hit with the best BLAST bit score was selected, and then an adjusted BLAST identity percentage (ABIP) based on the BLAST result (ABIP%=matches/[(matches+mismatches+deletions+insertions+(non-aligned bases at the two ends of the CCS read)]) was calculated as the final assessment value to represent the global true accuracy of the CCS read. When an ABIP value for a given CCS read is less than 95%, the CCS read is classified as low quality.

An in-house Perl script (http://david.abcc.ncifcrf.gov/manuscripts/pacbio_qc) was developed to query the H5 archive file in order to create a summary read-QC-parameter report. The script outputs common QC values (read length, overall mean QV), as well as QC parameters (i.e. pass #, read Quality Score, read deletion mean QV, read insertion mean QV, read substitution mean QV, read minimum mean QV) that are not routinely reported in the default output of the SMRT Analysis software. These values were summarized in a read-QC-parameter report file for this study, containing 9812 CCS read names and the associated QC values. To filter and assess the CCS reads with associated QC values, the QC-matrix file was analyzed using MS Excel and Partek Genomic Suite 6.6.

The assembly pipeline is not in the scope of the discussion in this paper. In brief, all of the input CCS reads went through multiple steps using the software packages Uclust, Muscle and Sequencher.

The SVM-Based QC strategy

The multiple QC parameters were used in a Support Vector Machine (SVM) regression model [15] for training using the 900 CCS reads generated from the PacBio spike-in positive control. (http://david.abcc.ncifcrf.gov/manuscripts/pacbio_qc). One of the advantages of the SVM is that it can avoid the difficulties of using linear functions in the high dimensional feature space by implicit mapping via kernels. For the dataset in this study, we chose a radial basis function (RBF) kernel with the default settings [15] to build the model with the training set of PacBio spike-in CCS reads in Matlab. The SVM regression model was built on two input variables, which are mean Quality Value and CCS pass number. We then used it to predict the accuracy of each of the 9812 non spike-in CCS reads in the same dataset. As a result, each of the CCS reads was assigned a predicted accuracy value in the range of 0 to 100 (percentage value) which was merged into the read-QC-parameter report file.

Results

The median accuracy of total CCS reads is 97.5%

In principle, each of the 9812 CCS reads should belong to one of the 10 MSG isoform reference sequences. If some are not, there must be sequencing errors derived during the course of sample prep, library prep and the PacBio sequencing procedure. For a given CCS read, the adjusted BLAST identity percentage (ABIP, see method section for formula) value to the closest MSG sequence was used to represent the global accuracy of the CCS read. The median accuracy (ABIP) value of the 9812 CCS reads is 97.5% in this study (Table 1). Importantly, the 2.5% error rate may not only come from PacBio sequencing procedure, but also from the sample prep (PCR error) and library prep procedures. Assuming a threshold of 95% accuracy of CCS reads is what assembly programs can tolerate, then ~20.5% of CCS reads are below the threshold and thus of low-quality (Figure 1 and Table 1). After examining the details of the BLAST alignments, we found that sequencing errors frequently occur in the 3' and 5' ends while the central region usually represents good accuracy, which suggests that the PacBio algorithm should more precisely process the two ends. Further evidence on this point is that we performed a static 50-bp trimming on the two ends of all CCS reads, which subsequently improves the CCS read accuracy and assembly quality (Table 1). In addition, adaptor sequences and chimeric sequences are also observed in some cases. Overall, while a majority (7791) of the 9812 total CCS reads is high-

QC method	None	50-bp trimmed at both ends	QV-Based		spike-in trained SVR	
# of CCS reads selected	all 9812	all 9812	top 3000	top 5000	top 3000	top 5000
90% percentile of read accuracy	99.44%	99.48%	99.62%	99.56%	99.62%	99.56%
50% percentile of read accuracy	97.48%	97.63%	99.12%	98.61%	99.12%	98.67%
10% percentile of read accuracy	92.98%	93.06%	98.44%	94.56%	98.54%	95.09%
De Novo Assembly: # of Contigs	13 (3 FP*)	10* (0 FP)	11 (1 FP)	12 (2 FP)	10 (0 FP)	10 (0 FP)

Note*: final assembled length is 100 bp shorter.
*FP denotes False Positive

Table 1: A Comparison of read accuracy (ABIP) improvements across three quality control (QC) strategies and their impacts on the De Novo assembly results.

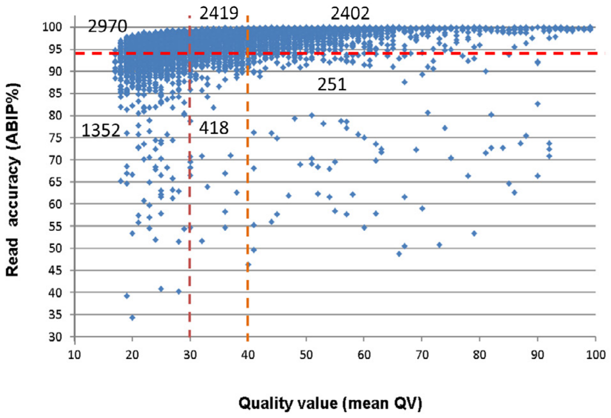


Figure 1: Read accuracy (ABIP) versus quality value (mean QV). The mean QV is correlated with the CCS read accuracy, particularly, at the range of QV-40 and above. It suggests that QV can be a useful QC parameter to remove low-quality CCS reads. However, the plot also shows that the majority of CCS reads, including the low-quality CCS reads, are below QV-40. Thus, a QV-40 cutoff might have a higher tradeoff by removing a large amount of high-quality CCS reads. A QV-30 cutoff may be more balanced.

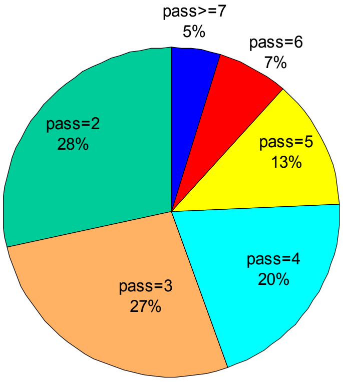


Figure 2: The distribution of pass number for 9812CCS MSG reads.

quality, users should be aware that ~20% (2021) low-quality CCS reads may need to be filtered (Figure 1). A QC step is necessary to remove the problematic CCS reads, otherwise, downstream assembly or other bioinformatics analysis might be adversely affected [16,17], such as resulting in false positive SNPs, isoforms, or assembly.

The Phred-like quality value can effectively filter low-quality reads, but with a tradeoff

PacBio's software package reports a Phred-like Quality Value (QV) for each CCS read, wherein the median QV of the 9812 CCS reads is 31. The actual CCS read accuracy (ABIP) and QV have positive correlation, particularly for CCS reads with a QV of 40 or greater (Figure 1). Therefore, QV, as expected, can be used to filter the problem CCS reads. QV cutoffs of 30 or 40 can improve the median accuracy values of CCS reads to 98.5% and 99.0% respectively, while ~66.9% and ~77.7% of the problem CCS reads (ABIP<95%) can be filtered out. However, the tradeoff is that ~38.1% and ~69.2% of the high-quality CCS reads, respectively, are also filtered out (Figure 1). Thus, users should be particularly careful of the high tradeoff with QV cutoffs of 40 or above, even though it can remove more low quality CCS reads. QV-40 is a very sensitive threshold because a majority (>73%) of the total CCS reads are below QV-40. Overall, QV-30 seems to be a more balanced cutoff line, in general, considering the tradeoff [16]. However, there is ~33.1% of the problem CCS reads remaining with a cutoff of QV-30. The question is whether other QC parameters can be used to improve this situation?

A support vector machine (svm)-based integrative qc strategy

Pass number is a unique feature of the PacBio platform. It represents how many rounds the same single-molecule is sequenced in a hairpin structure during the PacBio circular sequencing procedure [1,2]. Logically, one would presume that a higher number of passes can produce more multiple alignment information resulting in better-quality CCS reads (Figures 2 and 3) however, this relationship is not linear. As the pass number gets higher, the increase of the read quality value slows down. Actually, the pass number and the read quality value (mean QV) are correlated with a coefficient squared value of 0.74.

There are many different ways to use multiple QC parameters to filter the high dimensional data. To simultaneously use the pass number and other QC parameters in one single QC step, we built a SVM [15] regression model with CCS reads derived from PacBio spike-in sequences as the training set to predict MSG CCS read accuracy. Since the PacBio standard spike-in control DNA and the true sample are sequenced at the same time, the quality of the control CCS reads can largely represent the entire course of PacBio sequencing procedure, i.e., the efficiency of library prep, sample loading, polymerase activity, ligase activity, accuracy of fluorescent signal detection, threshold of bioinformatics, etc. In this sense, it is a good and universal source as the training dataset when a perfectly matched control to the true sample is not available. However, this universal control may also have potential biases or differences from the true sample, such as differing DNA length from that of the true sample and thus, we did not use length as an input feature in the SVM.

The advantage of the SVM model is that it allows us to simultaneously use the multiple QC parameters in a non-linear space. We used CCS pass number and read mean QV as input variables and ABIP as the dependent variable associated with PacBio spike-in sequences to train the SVM model. The trained SVM model is then used to predict the accuracy for each of the CCS reads in the dataset. As a result, each CCS

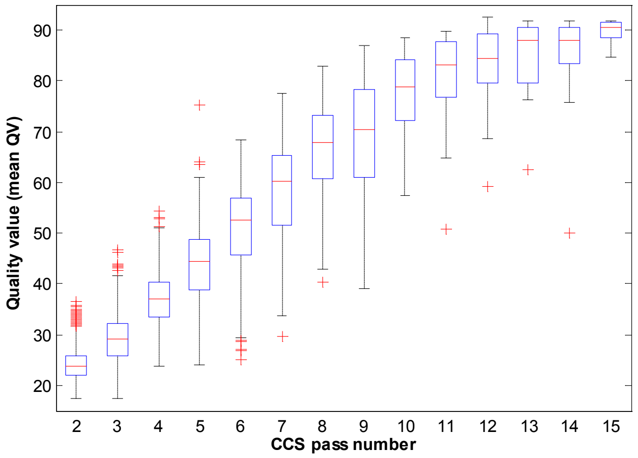
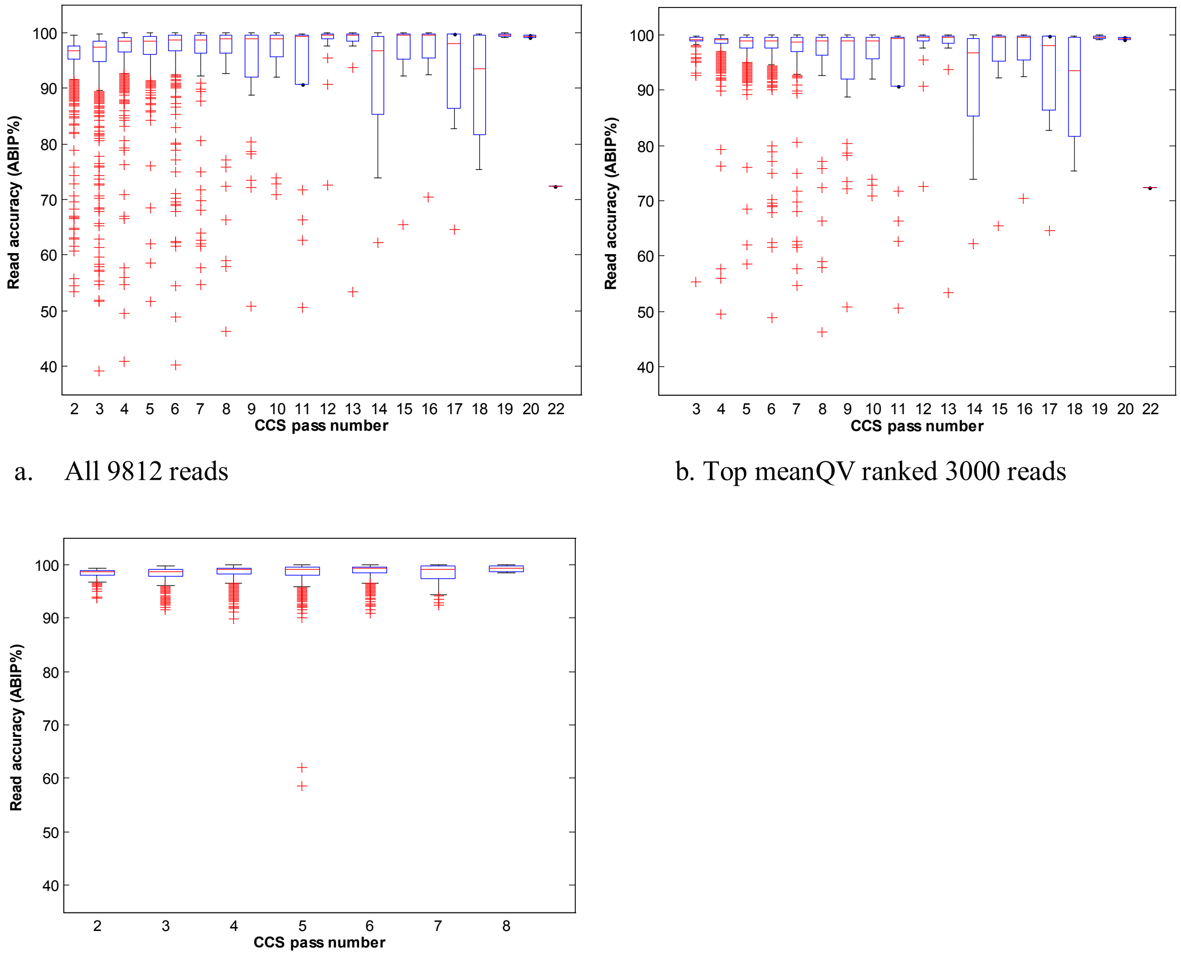


Figure 3: Read quality value (mean QV) vs CCS pass number. As the pass number increases the read quality value in general increases, however, the correlation is not linear, when the pass number getting higher, the increase of the read quality value slows down. (Note: the box plots for pass number greater than 15 were not shown in the figure due to insufficient data points).



a. All 9812 reads

b. Top meanQV ranked 3000 reads

c. Top SVM ranked 3000 reads

Figure 4: Box plots of CCS read accuracy (ABIP) for different pass numbers. **Figure 4a)** shows the box plots for all 9812 reads without doing QC, most of the outliers denoted by red crosses are low-quality reads; **Figure 4b)** shows the box plots for the top 3000 reads ranked by mean QV. The low-quality reads for pass numbers lower than 7 have significantly been removed but none of those with pass numbers greater than 7 have been removed, meanwhile, no reads with a pass number of 2 have been selected and most good reads for low pass number have also been screened out with the high mean QV threshold. **Figure 4c)** Shows the box plots for the top 3000 reads ranked by the predicted accuracy value by SVM. Obviously, most of the low-quality reads have been cleaned and all of the reads selected are those with pass numbers less than 9. The two figures **4b)** and **4c)** illustrate the different effects of the two QC methods due to different ranking mechanisms.

read is assigned a SVM predicted accuracy score from 0 to 100 (100 is the best score meaning 100% match).These scores were used to select better CCS reads for QC purposes.

At this point, there are three QC scenarios: trim 50bp at both ends of the read, QV-based QC, and SVM-based QC. To assess the effectiveness of the QC methods, we compare their 90%, 50% (median) and 10% percentile values of ABIP as shown in Table 1. The comparison shows that SVM-based QC, which is a multiple-parameter QC, is more effective in improving low end CCS read accuracy in the 10% percentile range (Table 1). Figure 4b and Figure 4c illustrate the different QC effects for the QV-based method and SVM-based method compared to the box plots of read accuracy versus CCS pass number for all 9812 CCS reads shown in Figure 4a.

The impact of different qc strategies on the downstream *de novo* assembly

The ultimate goal of the analysis is the *De Novo* assembly of the CCS reads to generate a number of contigs which should correspond to the 10 known given MSG sequences. We use the *de novo* assembly results to evaluate the effectiveness of each of the QC strategies. Five CCS read datasets derived from three different QC strategies, 9812 reads with 50 bp trimmed from both ends, the top 3000 and 5000 CCS reads selected by QV-based QC and the top 3000 and 5000 CCS reads selected by SVM-based QC scores, in addition to the original data set of 9812 MSG reads (without QC), were respectively run through an in-house de novo assembly pipeline with the exact same program settings. Since the assembly analysis is not in the scope of this QC-centric paper, we only focus on the assembly results without discussing the details of the assembly pipeline. The assembly shows that without QC the original 9812 reads resulted in 13 contigs, among which 10 matched with the 10 given MSG sequences and 3 are false positives (FP) (Table 1). The one with 50 bp trimmed at both ends for each of the 9812 reads resulted in 10 contigs which were 100 bp shorter than each of the matched 10 MSG sequences. The top 3000 and 5000 QV-selected reads gave 1 and

2 false positives respectively while the top 3000 and 5000 SVM-selected reads generated 10 contigs with 0 FP, exactly corresponding to the 10 MSG reference sequences with high accuracy (~99%). The comparison of assembly results indicates that the SVM-based QC, integrating multiple-parameters, is useful for more accurate assembly results.

Conclusion

The PacBio benchmark study in this paper demonstrates that PacBio targeted amplicon sequencing yields ~20% of total CCS reads in low-quality. The percentage of low quality CCS reads may be underestimated by many optimistic users. Thus, it is very necessary to apply appropriate QC filters to remove low-quality CCS reads even though all CCS reads are post error-corrected. This study indicates that inefficient or no QC could result in some false positive contigs after assembly. A combination of multiple QC parameters can be more powerful than a single measure alone in order to effectively remove low-quality CCS reads. Users should balance tradeoffs by applying an appropriate QC stringency depending on their needs of different downstream analysis.

Acknowledgement

This work was supported by the National Cancer Institute, National Institutes of Health (Contract HHSN261200800001E). This research was also supported in part by the National Institute of Allergy and Infectious Disease. The contents do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

References

1. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science 323: 133-138.
2. Otto TD (2011) Real-time sequencing. Nat Rev Microbiol 9: 633.
3. Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11: 759-769.
4. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 38: e159.
5. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.
6. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, et al. (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27: 1013-1023.
7. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
8. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Anal Chem 83: 4327-4341.
9. Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, et al. (2012) A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 30: 701-707.
10. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30: 693-700.
11. Zhang X, Davenport KW, Gu W, Daligault HE, Munk AC, et al. (2012) Improving genome assemblies by sequencing PCR products with PacBio. Biotechniques 53: 61-62.
12. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, et al. (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics 13: 375.
13. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14: R51.
14. Kutty G, Maldarelli F, Achaz G, Kovacs JA (2008) Variation in the major surface glycoprotein genes in Pneumocystis jirovecii. J Infect Dis 198: 741-749.
15. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:1-27.
16. Reumers J, De Rijk P, Zhao H, Liekens A, Smeets D, et al. (2011) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol 30: 61-68.
17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

This article was originally published in a special issue, **Bioinformatics for Highthroughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA