SequelQC: Analyzing PacBio Sequel Raw Sequence Quality

David E. Hufnagel [1,*], Matthew B. Hufford [1] and Arun S. Seetharam [2,*]

April 2019

[1]Department of Ecology, Evolution and Organismal Biology and [2]Genome Informatics Facility, Iowa State University, Ames, IA, 50011, USA.

[*]To whom correspondence should be addressed.

Abstract

Summary: PacBio sequencing is an incredibly valuable third-generation DNA sequencing method due to very long read lengths, ability to detect methylated bases, and its real-time sequencing methodology. Yet hitherto no tractable program exists for analyzing the quality of PacBio Sequel raw sequence data. Here we present *SequelQC*, a bash tool that quickly processes PacBio Sequel raw sequence data from multiple SMRTcells producing multiple statistics and publication-quality plots describing the quality of the data including N50, read length and count statistics, PSR, and ZOR.

Availability and implementation: *SequelQC* is implemented in bash, R, and Python and is freely available at https://github.com/ISUgenomics/SequelQC

Contact: davehuf@iastate.edu or arnstrm@iastate.edu

Supplementary information: Supplementary data are available at BioRxiv online.

1 Introduction

The third generation of sequencing is here and making tremendous impact on the sequencing field. These long-read sequencing platforms are undergoing active development, pushing the boundaries in terms of total output, read length, sequencing time, cost reduction and read accuracy (Schadt *et al.*, 2010; Goodwin *et al.*, 2016). One example is the PacBio Sequel platform, the most widely used long-read sequencing platform, which uses Single Molecule Real Time (SMRT) sequencing technology (Rhoads and Au, 2015; Goodwin *et al.*, 2016). Unlike second generation approaches, PacBio sequencing can provide much longer length reads, in much less time, with greatly reduced content bias, the ability to distinguish between methylated and unmethylated bases, and with almost as much accuracy (Schadt *et al.*, 2010; Ross *et al.*, 2013; Rhoads and Au, 2015; Goodwin *et al.*, 2016). Due to improvements in data formats and the technology itself, previous base quality programs for PacBio RSII data are no longer valid for benchmarking

PacBio Sequel data. Currently, the only program that provides quality assessment for the Sequel raw sequence data is the instrumentation software itself, SMRT Link, a linux-only, computationally intensive webtool where the user must upload their data files one at a time. Furthermore, SMRT Link can only be installed by root users, requiring the installation of 23 external programs to run, and generates non-downloadable plots after setting up a web server (https://www.pacb.com/support/software-downloads/). The development of a fast, easy-to-install-and-use, third-party program to assess raw sequence quality is therefore crucial for the genomic community. Here we present SequelQC, an efficient and user-friendly program that calculates multiple standardized statistics and creates publication-quality plots describing the quality of raw PacBio Sequel data.

2 Implementation and Usage

In order to create tables and plots summarizing the quality of PacBio Sequel data, *SequelQC* uses standard libraries and packages within bash, R and Python. PacBio sequence files include both subreads files containing reads of interest and scraps files with additional reads generated during the sequencing process. *SequelQC* can be run with or without scraps files. With scraps files *SequelQC* takes longer to run, but also produces more plots and analysis in standard plots using additional information concerning continuous long reads (CLRs). The main script is written in bash, which calls samtools (Li *et al.*, 2009) and awk to convert BAM files to SAM format and extracts only needed information. Then Python is used to make all necessary calculations, producing intermediate data files that are passed to R. Python 2 or 3 can be used, and the version is determined automatically. By default these intermediate data files will be deleted at the end of the program's operation, but they will be retained if the user selects the appropriate parameters. At this point, reads are organized into up to four read groups: 1) subreads, 2) longest subreads, 3) CLRs, and 4) subedCLRs (CLRs containing subreads). When scraps are included, the default is to use all four read groups, but the user can request only two groups if preferred: 1) subreads and 2) subedCLRs. Alternatively, if scraps are not included the two read groups are 1) subreads and 2) longest subreads. Final plots and tables are produced in R, including a table of summary statistics, which can be viewed easily in Microsoft Excel, as well as several publication-quality PDF plots. A subset of these plots are included as Supplemental Figures to this manuscript.

The summary statistics table includes information for all chosen read groups for each SMRTcell. Statistics include number of reads, total bases, mean and median read length, N50, L50, PSR, and ZOR. PSR is the polymerase-to-subread ratio and is calculated as follows: total bases from the longest subreads per CLR divided by the total bases from subreads. ZOR is the ZMW (zero-mode waveguide) occupancy ratio and is calculated as follows: the number of CLRs with subreads divided by the number of subreads. While the summary statistics table is always produced, the user can request more or fewer plots based on their needs. The full suite of plots with scraps files includes barplots of A) N50s, B) L50s, C) total bases, and D) read length; frequency plots of E) subreads per subedCLR, and F) adapters per CLR; boxplots of G) subread lengths and H) subedCLR lengths; and I) ZOR and J) PSR plots (see Supplemental Figures). The user can also request an intermediate (A,C,G,H,I, & J) or basic (A & C) suite of plots. Without

scraps files, the full suite of plots is A,B,C,D,G,I, & J, the intermediate collection is A,C,G,I, & J, and the basic set is A & C (see Supplemental Figures). With or without scraps the intermediate selection of plots is default.

SequelQC works on any operating system and command-line operation is quite simple; In its most basic form with scraps files:

```
bash SequelQC.sh -u subFiles.txt -c scrFiles.txt
```

or without scraps files:

```
bash SequelQC.sh -u subFiles.txt
```

Where subFiles.txt and scrFiles.txt are files containing the names of subreads and scraps BAM files, respectively, with one name per line. While '-u' is the only required parameter, the user can choose from other parameters including the number of threads to use for running samtools, whether to keep intermediate files, and how many read groups and plots are desired. While it can be run on a single SMRTcell, *SequelQC* is designed to run across multiple SMRTcells simultaneously. The summary statistics table as well as all plots, except the read length histograms and frequency plots, present the data from all SMRTcells together.

Some users may want to modify the provided R script to change plot format or to make entirely new plots. This can be done by copying the R script, making the desired modifications, and modifying the name of the R script in SequelQC.sh to reflect the modified file.

3 Validation

SequelQC was tested on Condo, a High-Performance Computing Cluster, at Iowa State University, running the Red Hat Linux operating system. Varying number of CPU's (1-16) and SMRTCells (1-8) from the PacBio Human Genome were used for bench-marking. The run-time for SequelQC for all SMRTcells was approximately 30 mins when using two or more CPUs (see Supplemental Figures). This did not vary with increasing number of CPUs. This may be because only samtools uses multiple CPU's and all runs used 128Gb of RAM. The PacBio Human Genome can be found at https://www.pacb.com/blog/puerto-rican-genome/

4 Conclusion

SequelQC is an easy to install and use program that calculates key statistics and generates publication-quality plots for raw PacBio Sequel data. SequelQC provides all standard metrics for overall sequence quality including N50, read length and count statistics, PSR, and ZOR. SequelQC can evaluate eight SMRTcells from the PacBio human genome dataset in about 30min on our HPC system. Other than the proprietary PacBio SMRTlink program, which requires the user to set up a web server to install, there is currently no program available to compute these statistics. We therefore conclude that SequelQC is the only reasonable choice for most users of PacBio Sequel sequencing data.

Acknowledgements

We would like to thank Maggie Woodhouse for proving comments on this manuscript, as well as Jim Coyle, director of HPC at Iowa State University and Andrew Severin at the Genome Informatics Facility for supporting the development of this tool.

Funding

This work was supported by the National Science Foundation Plant Genome Research Program [IOS-1744001].

References

Goodwin, S. *et al.* (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**(6), 333–351.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Rhoads, A. and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*, **13**(5), 278–289.

Ross, M. G. *et al.* (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**(5), R51.

Schadt, E. E. *et al.* (2010). A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**(R2), R227–240.