Roberts *et al. Genome Biology* 2013, **14**:405
http://genomebiology.com/2013/14/7/405

Genome Biology

CORRESPONDENCE

The advantages of SMRT sequencing

Richard J Roberts[1]*, Mauricio O Carneiro[2] and Michael C Schatz[3]

Abstract

Of the current next-generation sequencing technologies, SMRT sequencing is sometimes overlooked. However, attributes such as long reads, modified base detection and high accuracy make SMRT a useful technology and an ideal approach to the complete sequencing of small genomes.

Pacific Biosciences' single molecule, real-time sequencing technology, SMRT, is one of several next-generation sequencing technologies that are currently in use. In the past, it has been somewhat overlooked because of its lower throughput compared with methods such as Illumina and Ion Torrent, and because of persistent rumors that it is inaccurate. Here, we seek to dispel these misconceptions and show that SMRT is indeed a highly accurate method with many advantages when used to sequence small genomes, including the possibility of facile closure of bacterial genomes without additional experimentation. We also highlight its value in being able to detect modified bases in DNA.

Extending read lengths

So-called next-generation technologies for sequencing DNA are penetrating every aspect of biology thanks to the immense amount of information that is encoded within nucleic acid sequences. However, today's next-generation sequencing technologies, such as Illumina, 454 and Ion Torrent, have several significant limitations, especially short read lengths and amplification biases, that restrict our ability to fully sequence genomes. Unfortunately, with the rise of next-generation sequencing, even less emphasis is being placed on trying to understand at the biological and biochemical levels just what functions newly discovered genes have and how those functions allow an organism to work, which is surely why we are sequencing DNA in the first place. Now a new technology, SMRT sequencing from Pacific Biosciences [1], has been developed that not only produces considerably longer and highly accurate DNA sequences from individual unamplified molecules, but can also show where methylated bases occur [2] (and thereby provide functional information about the DNA methyltransferases encoded by the genome).

SMRT sequencing is a sequencing-by-synthesis technology based on real-time imaging of fluorescently tagged nucleotides as they are synthesized along individual DNA template molecules. Because the technology uses a DNA polymerase to drive the reaction, and because it images single molecules, there is no degradation of signal over time. Instead, the sequencing reaction ends when the template and polymerase dissociate. As a result, instead of the uniform read length seen with other technologies, the read lengths have an approximately log-normal distribution with a long tail. The average read length from the current PacBio RS instrument is about 3,000 bp, but some reads may be 20,000 bp or longer. This is roughly 30 to 200 times longer than the read length from a next-generation sequencing instrument, and more than a four-fold improvement since the original release of the instrument two years ago. It is notable that the recently announced PacBio RS II platform claims to have a further four-fold improvement, with twice the mean read length and twice the throughput of the current machine.

Applications of SMRT sequencing

The SMRT approach to sequencing has several advantages. First, consider the impact of the longer reads, especially for *de novo* assemblies of novel genomes. While typical next-generation sequencing can provide abundant coverage of a genome, the short read lengths and amplification biases of those technologies can lead to fragmented assemblies whenever a complex repeat or poorly amplified region is encountered. As a result, GC-rich and GC-poor regions, which tend to be poorly amplified, are particularly susceptible to poor quality sequencing. Resolving fragmented assemblies requires additional costly bench work and further sequencing. By also including the longer reads of SMRT sequencing runs, the read set will span many more repeats and missing bases, thereby closing many of the gaps automatically and

*Correspondence: roberts@neb.com
[1]New England Biolabs, 240 County Road, Ipswich, MA 01938, USA
Full list of author information is available at the end of the article

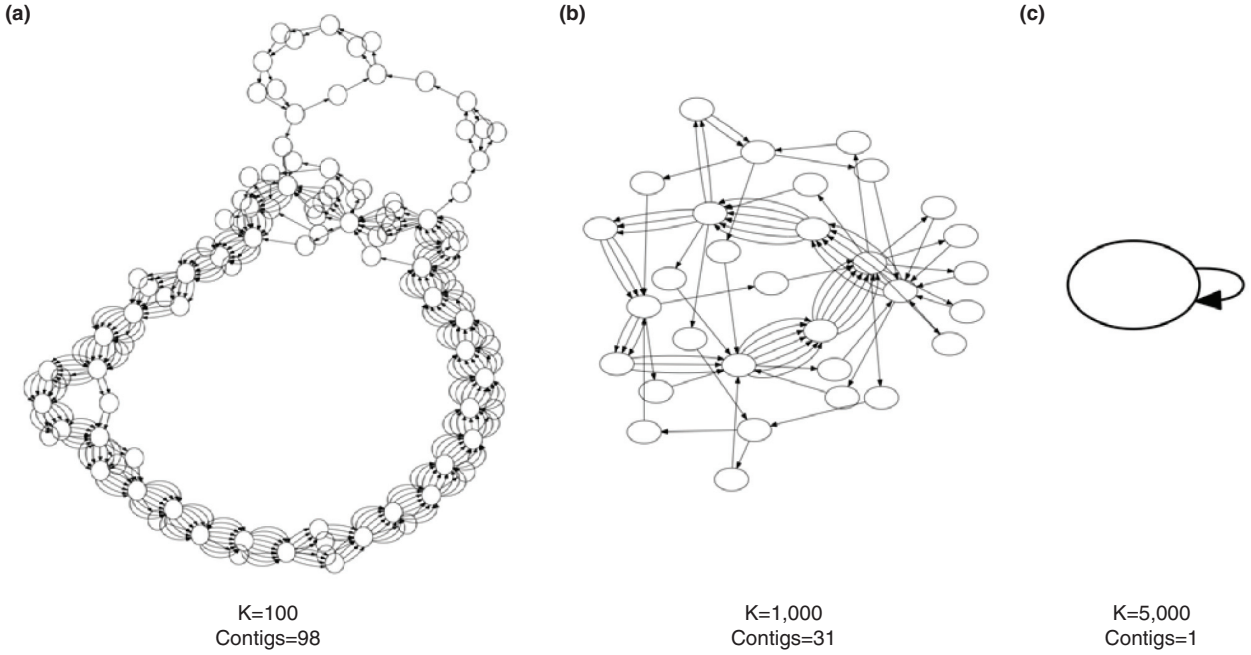

Figure 1. Idealized assembly graphs [18] of the 5.2 megabase-pair *B. anthracis* Ames Ancestor main chromosome using (a) 100 bp, (b) 1,000 bp and (c) 5,000 bp reads. The graphs encode the compressed de Bruijn graph derived from infinite coverage error-free reads, effectively representing the repeats in the genome and the upper bound of what could be achieved in a real assembly. Increasing the read length decreases the number of contigs because the longer reads will span more of the repeats. Note the assembly with 5,000 bp reads has a self-edge because the chromosome is circular.

simplifying, or even eliminating, the finishing time (Figure 1). It is becoming routine for bacterial genomes to be completely assembled using this approach [3,4], and we expect this practice will translate to larger genomes in the near future. A complete genome is far more useful than the poor quality draft sequences that litter GenBank because it provides a complete blueprint for the organism; the genes encoded therein represent the full biological potential of that organism. With only draft assemblies available, one is always left with the nagging feeling that some crucial gene is missing - perhaps the one in which you are most interested! The long read lengths also have more power to reveal complex structural variations present in DNA samples, such as pinpointing precisely where copy number variations have occurred relative to the reference sequence [5]. They are also extremely powerful for resolving complex RNA splicing patterns from cDNA libraries, since a single long read may contain the entire transcript end-to-end, thus eliminating the need to infer the isoforms [6].

Second, consider DNA methyltransferases. These can exist as solitary entities or as parts of restriction-modification systems. In both cases, they methylate relatively short sequence motifs that can easily be recognized from SMRT sequencing data because of the change in DNA polymerase kinetics, as it moves along the template molecule, that result from the presence of epigenetic modifications. The altered kinetics cause a change in the timing of when the fluorescent colors are observed, thus enabling direct detection of epigenetic modifications, which can ordinarily only be inferred, and bypassing the usual necessity of enrichment or chemical conversion. Often, thanks to bioinformatics, the gene responsible for any given modification can be matched to the sequence motif in which the modification lies [7,8]. When it cannot, then simply cloning the gene into a plasmid, which is subsequently grown in a non-modifying host and re-sequenced, can provide the match [9]. Moreover, SMRT sequencing has also been able to identify RNA base modifications through the same approach as DNA base modifications, but using an RNA transcriptase in place of the DNA polymerase [10]. In fact, SMRT sequencing represents an important step toward uncovering the biology that happens between DNA and proteins, including not only the study of mRNA sequences but also the regulation of translation [11,12]. Thus, functional information emerges directly from the SMRT sequencing approach.

Third, we must consider the persistent rumor that SMRT sequencing is much less accurate than other next-generation sequencing platforms, which has now been demonstrated to be untrue in several ways. First, a direct

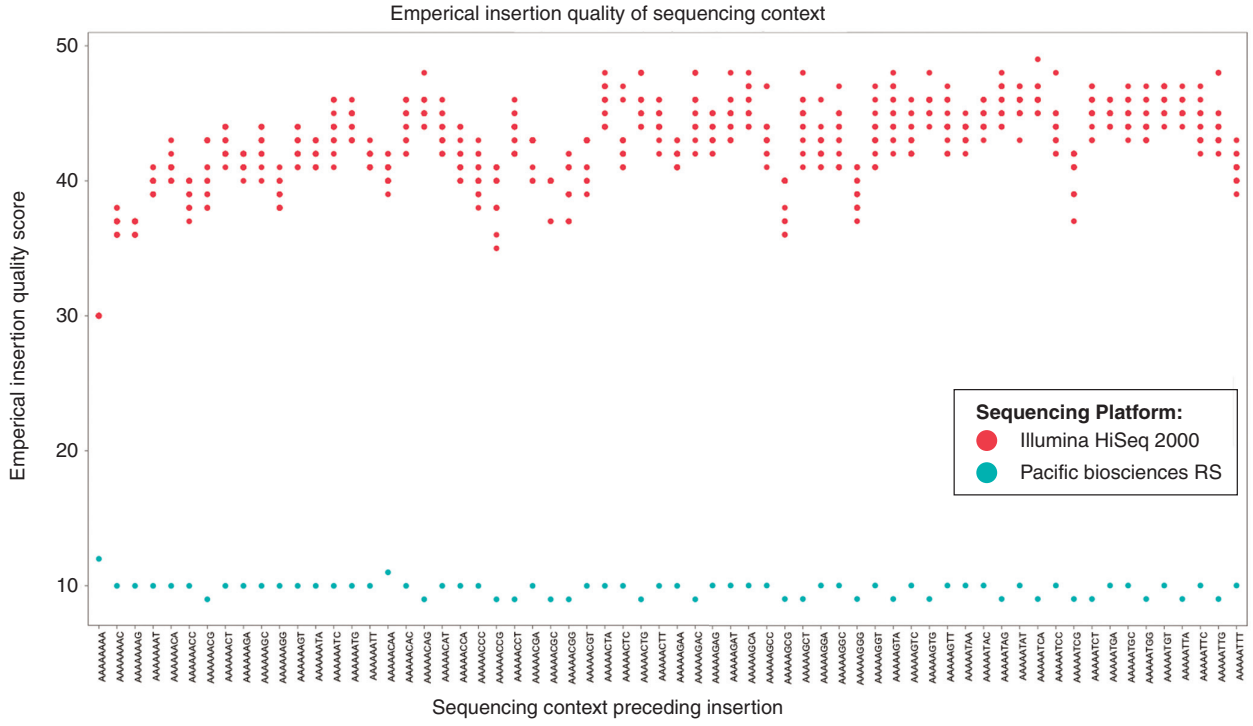

Figure 2. A sequencing context breakdown of the empirical insertion error rate of the two platforms on NA12878 whole genome data. In this figure we show all contexts of size 8 that start with AAAAA. The empirical insertion quality score (*y*-axis) is PHRED scaled. Despite the higher error rate (approximately Q12) of the PacBio RS instrument, the error is independent of the sequencing context. Other platforms are known to have different error rates for different sequencing contexts. Illumina's HiSeq platform, shown here, has a lower error rate (approximately Q45 across eight independent runs), but contexts such as AAAAAAAA and AAAAACAG have extremely different error rates (Q30 versus Q55). This context-specific error rate creates bias that is not easily clarified by greater sequencing depth. Empirical insertion error rates were measured using the Genome Analysis Toolkit (GATK) - Base Quality Score Recalibration tool.

comparison of several approaches to determining genetic polymorphisms has shown that SMRT sequencing has comparable performance to other sequencing technologies [13]. Second, the accuracy of assembling a complete genome using SMRT sequencing in combination with other technologies has proved to be as reliable and accurate as more traditional approaches [3,6,14]. Moreover Chin *et al*. [15] showed that an assembly using only long SMRT sequencing reads achieves comparable or even higher performance than other platforms (99.999% accuracy in three organisms with known reference sequences), including 11 corrections to the Sanger reference of these genomes. Koren *et al*. [6] showed that most microbial genomes could be assembled into a single contig per chromosome with this approach; it is by far the least expensive option for doing so.

Debunking the error myth

The power of SMRT sequencing data lies both in its long read lengths and in the random nature of the error process (Figure 2). It is true that individual reads contain a higher number of errors: approximately 11% to 14% or Q12 to Q15, compared with Q30 to Q35 from Illumina and other technologies. However, given sufficient depth (8x or more, say), SMRT sequencing provides a highly accurate statistically averaged consensus perspective of the genome, as it is highly unlikely that the same error will be randomly observed multiple times. Notoriously, other platforms have been found to suffer from systematic errors that need to be resolved by complementary methods before the final sequence is produced [16].

Another approach that benefits from the stochastic nature of the SMRT error profile is the use of circular consensus reads, where a sequencing read produces multiple observations of the same base in order to generate high-accuracy consensus sequence from single molecules [17]. This strategy trades read length for accuracy, which can be effective in some cases (targeted re-sequencing, small genomes) but is not necessary if one can achieve some redundancy in the sequencing data (8x is recommended). With this redundancy, it is preferable to benefit from the improved mapping of longer inserts than opt for circular consensus reads, because the longer reads will be able to span more repeats and high accuracy will still be achieved from their consensus.

Conclusions

The considerations above make a strong case for combining the more traditional, sequence-dense data from other technologies with at least moderate coverage of SMRT data so that genomes can be improved, their methylation patterns obtained, and the functional activity of their methyltransferase genes deduced. We would especially urge all groups currently sequencing bacterial genomes to adopt this policy. That said, SMRT sequencing has also substantially improved eukaryotic genome assemblies, and we expect it to become more widely applied in this context over time, in light of the greater read lengths and throughput of the PacBio RS II instrument.

Perhaps it would even be worth redoing many genomes so that existing shotgun dataset-based assemblies could be closed and their complete methylomes obtained. The resultant assembled (epi)genomes would be inherently more valuable: the usefulness of a closed genome with associated functional annotation of its methyltransferase genes is far greater than the uncertainties left with a shotgun data set. Whereas we currently know much about the importance of epigenetic phenomena for higher eukaryotes, very little is known about the epigenetics of bacteria and the lower eukaryotes. SMRT sequencing opens a new window that may have a dramatic effect on our understanding of this biology.

Abbreviations

bp, base pair.

Competing interests

None of the authors have competing financial interests, but RJR and MCS have collaborated extensively with scientists from Pacific Biosciences leading to several publications cited in the text.

Authors' contributions

All three authors contributed to the writing of this article.

Acknowledgements

MOC acknowledges Ryan Poplin for kindly sharing data on error rates. RJR acknowledges support from New England Biolabs and NIH (4R44GM105125). MCS acknowledges support from NIH (R01-HG006677).

Author details

[1]New England Biolabs, 240 County Road, Ipswich, MA 01938, USA. [2]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. [3]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11743, USA.

Published: 3 July 2013

References

1. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, *et al.*: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323:**133-138.
2. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing.** *Nat Methods* 2010, **7:**461-465.
3. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, Berlin AM, Montmayeur A, Shea TP, Walker BJ, Young SK, Russ C, Nusbaum C, Maccallum, I, Jaffe DB: **Finished bacterial genomes from shotgun sequence data.** *Genome Res* 2012, **22:**2270-2277.
4. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey DS, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing** [http://arxiv.org/abs/1304.3752]
5. Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE, Danilova TV, Kudrna D, Magalhaes JV, Piñeros MA, Schatz MC, Wing RA, Kochian LV: **Aluminum tolerance in maize is associated with higher MATE1 gene copy number.** *Proc Nat Acad Sci U S A* 2013, **110:**5241-5246.
6. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM: **Hybrid error correction and *de novo* assembly of single-molecule sequencing reads.** *Nat Biotechnol* 2012, **30:**693-700.
7. Murray, IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, Fomenkov A, Turner SW, Korlach J, Roberts RJ: **The methylomes of six bacteria.** *Nucleic Acids Res* 2012, **40:**11450-11462.
8. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado O, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE: **Genome-wide detection of methyladenine residues in an HUS-linked pathogen.** *Nat Biotechnol* 2012, **30:**1232-1239.
9. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J: **Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing.** *Nucleic Acids Res* 2012, **40:**e29.
10. Vilfan ID, Tsai Y, Clark TA, Wegener J, Dai Q, Yi C, Pan T, Turner SW, Korlach J: **Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription.** *J Nanobiotechnol* 2013, **11:**8.
11. Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD: **Real-time tRNA transit on single translating ribosomes at codon resolution.** *Nature* 2010, **464:**1012-1017.
12. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE: **The birth of the Epitranscriptome: deciphering the function of RNA modifications.** *Genome Biol* 2012, **13:**175.
13. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, Depristo MA: **Pacific biosciences sequencing technology for genotyping and variation discovery in human data.** *BMC Genomics* 2012, **13:**375.
14. Bashir A, Klammer AA, Robins WP, Chin CS, Webster D, Paxinos E, Hsu D, Ashby M, Wang S, Peluso P, Sebra R, Sorenson J, Bullard J, Yen J, Valdovino M, Mollova E, Luong K, Lin S, LaMay B, Joshi A, Rowe L, Frace M, Tarr CL, Turnsek M, Davis BM, Kasarskis A, Mekalanos JJ, Waldor MK, Schadt EE: **A hybrid approach for the automated finishing of bacterial genomes.** *Nat Biotechnol* 2012, **30:**701-707.
15. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.** *Nat Methods* 2013, **10:**563-569.
16. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly, M: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.
17. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW: **A flexible and efficient template format for circular consensus sequencing and SNP detection.** *Nucleic Acids Res* 2010, **38:**e159.
18. Kingsford C, Schatz MC, Pop M: **Assembly complexity of prokaryotic genomes using short reads.** *BMC Bioinformatics* 2010, **11:**21.

doi:10.1186/gb-2013-14-7-405
Cite this article as: Roberts RJ, *et al.*: The advantages of SMRT sequencing. *Genome Biology* 2013, **14:**405.