Oxford Nanopore Sequencing and *de novo* Assembly of a Eukaryotic Genome

Sara Goodwin*, James Gurtowski*, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz†, W. Richard McCombie†

*These authors contributed equally to this work

†Co-corresponding authors

Cold Spring Harbor Laboratory, 1 Bungtown Rd. Cold Spring Harbor, NY 11724

Monitoring the progress of DNA through a pore has been postulated as a method for sequencing DNA for several decades[1,2]. Recently, a nanopore instrument, the Oxford Nanopore MinION, has become available[3]. Here we describe our sequencing of the *S. cerevisiae* genome. We describe software developed to make use of these data as existing packages were incapable of assembling long reads at such high error rate (~35% error). With these methods we were able to error correct and assemble the nanopore reads *de novo*, producing an assembly that is contiguous and accurate: with a contig N50 length of 479kb, and has greater than 99% consensus identity when compared to the reference. The assembly with the long nanopore reads was able to correctly assemble gene cassettes, rRNAs, transposable elements, and other genomic features that were almost entirely absent in an assembly using Illumina sequencing alone (with a contig N50 of only 59,927bp).

Most current DNA sequencing methods are based on either chemical cleavage of DNA molecules[4] or synthesis of new DNA strands.[5] In the more common synthesis based methods, base analogues of one form or another are incorporated into a nascent DNA strand that is labeled either on the primer from which it originates or on the newly incorporated bases. This is the basis of the sequencing method used for most current sequencers and their earlier predecessors [6]. Alternatively, it is been speculated that individual DNA molecules could be sequenced by monitoring their progress through various types of pores, originally envisioned as being pores derived from bacteriophage particles[7]. The advantages of this approach include potentially very long and unbiased sequence reads as no amplification nor chemical reactions are necessary for sequencing. In principle a pore-based instrument could be very inexpensive, sequencing would occur as fast as the molecules can pass through the pores, and the sequencing read lengths would be limited only by the lengths of the molecules in the sample. The combination of inexpensive and high throughput long read sequencing technology could be revolutionary to genomics as no alternative technology available today can resolve the most complex regions of large genomes at low cost.

Recently we began testing a sequencing device using nanopore sequencing technology from Oxford Nanopore Technologies through their early access program. This device, the MinION, is a nanopore based device in which pores are embedded in a membrane placed over an electrical detection grid. As DNA molecules pass through the pores they create measureable alterations in the ionic current. The fluctuations are sequence dependent and thus can be used by a base-calling algorithm to infer the sequence of nucleotides in each molecule[8,9]. As part of the library preparation protocol, a hairpin adapter is ligated to one end of a double stranded DNA sample

while a "motor" protein is bound to the other to unwind the DNA and control the rate of nucleotides passing through the pore[10]. Under ideal conditions the leading template strand passes through the pore followed by the hairpin adapter and then the complement strand. In such a run where both strands are sequenced, a consensus sequence of the molecule can be produced; these consensus reads are termed "2D reads" and have been shown to be of higher accuracy than reads from only a single pass of the molecule ("1D reads").

We chose to sequence the yeast genome so that we could carefully measure the accuracy and other data characteristics of the device on a tractable and well-understood genome. Our initial flow cells had somewhat low reliability and throughput, but improved substantially over time (Supplemental Figure S1, Supplemental Note 8). This is due to a combination of improvements in chemistry, protocols, instrument software, and shipping conditions. Some runs have produced upwards of 490 Mb of sequencing data per flow cell over a 48 hour period. All together, we generated a total of 120x coverage of the genome with an average read length of 5473bp but with a long tail extending to a maximum read length of 146,992bp (Supplemental Note 3)

Alignment of the reads to the reference genome using BLAST gave us a deeper analysis of the per base error rate. Of the 267,768 reads produced by our 30 sequencing runs, 64,849 reads (24%) aligned to the reference yeast genome, and 31,013 (11%) aligned to the known spike-in sequence used for calibrating the instruments. The remaining 65% of reads did not show significant similarity to the W303 genome or spike-in sequence, presumably because of insufficient read quality (Supplemental note 4). Supplemental Figure S5A shows that the mean identity to the reference of "1D" reads was 64% while "2D" reads produced many reads

exceeding 75% identity. The overall alignment identities of both 1D and 2D reads are summarized in Figure 1A that compares both read length and percent identity.

Figure 1a) Oxford Nanopore Read lengths and Accuracy

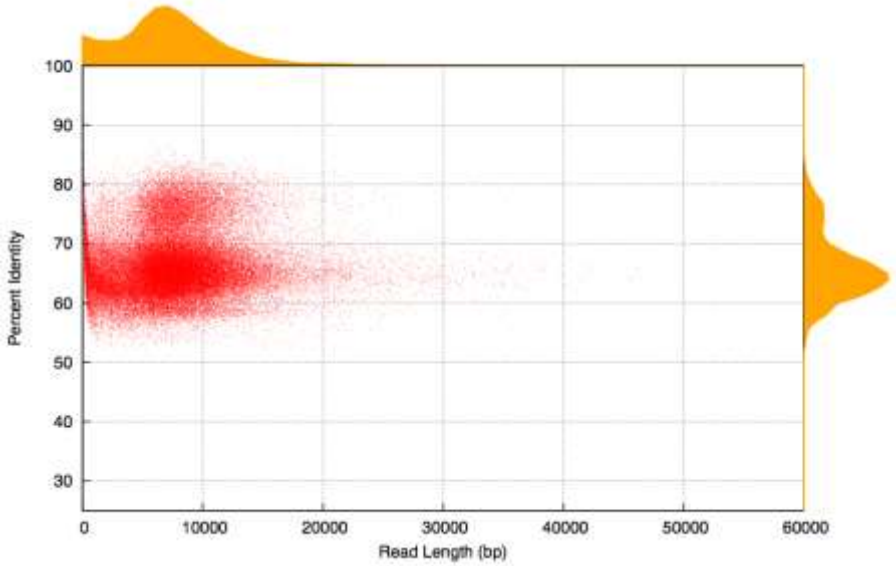

Figure 1b) Heatmap of Oxford Nanopore Read Lengths and Accuracy

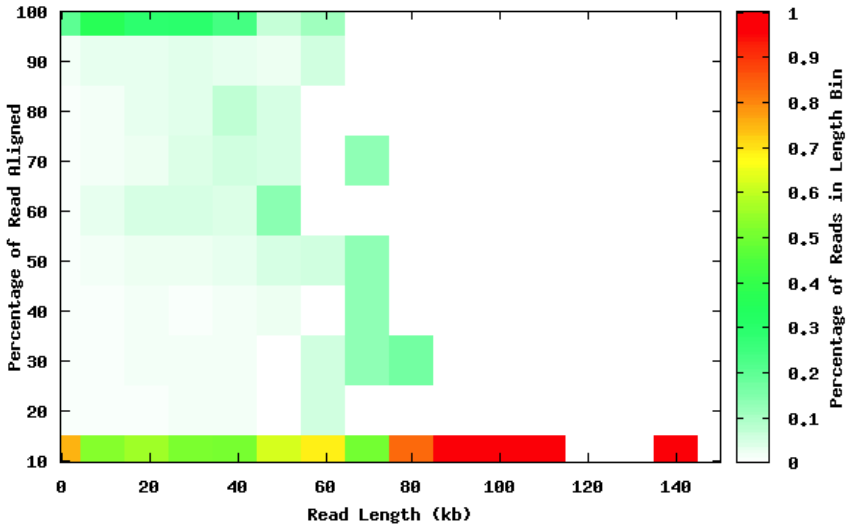

Figure 1. A) Scatter plot of read length versus accuracy with marginal histograms summarizing the raw ONT alignments. B) Heatmap of read length versus percent of read aligned to the W303 genome. Each cell represents a summary of how reads of different lengths align. Each color represents the percent of reads in a given read length bin. Maximal alignment efficiency is observed between 10 and about 30 kb, while fragments longer than 70kb are virtually unalignable.

Overall read quality is further summarized by Figure 1B showing a heatmap of the lengths of the alignments relative to the full length of the reads. On the lower end (below 50kbp), a substantial number (up to 50%) of the reads do not align to the reference in any capacity. However, those that can be aligned have matches that span nearly their entire length. As reads get longer (>50kb) portions of the reads can be aligned which suggests that reads are composed of both high and low quality segments. However, this local variability in quality does not seem to be position specific as can be seen in Supplemental Figure S5B that shows that per base error rate is consistent across the length of a read on average. The very longest reads tend not to be alignable at all, suggesting that the longest reads may be extremely low quality or include other artifacts of the sequencing process.

Of the reads that align, the overall coverage distribution approximated a Poisson distribution centered at 35x coverage, although some over-dispersion was observed that was better modeled by a Negative Binomial distribution (Supplemental Figure S5C). To examine some of the sources of the over-dispersion we also examined the coverage as a function of the GC composition of the genome. Between 20% and 60% GC content, the coverage was fairly

uniform, while at high and low GC content the coverage is more variable partially explaining some of the regions of the genome lacking raw read coverage (Supplemental Figure S5D).

As a demonstration of the utility of the Oxford Nanopore device, we developed a novel algorithm called Nanocorr to error correct the reads for *de novo* genome assembly or other purposes. Nanocorr uses a hybrid strategy for error correction, using high quality MiSeq short reads to error correct the long highly erroneous nanopore reads. It follows the design of hybrid error correction pipelines for PacBio long read sequencing[11], although in our testing none of the available algorithms were capable of utilizing the nanopore reads and therefore required an entirely new algorithm. Briefly, it uses the sequence aligner BLAST to align MiSeq reads to the nanopore reads, and then uses a dynamic programming algorithm based on the longest-increasing-subsequence (LIS) problem to select the optimal set of short read alignments that span each long read. The consensus reads are then calculated using a finite state machine of the most commonly observed sequence transitions using the open source algorithm *pbdagcon* (Figure 3A). Overall, this process increased the percent identity from an average of 65% for uncorrected reads to >97% (Figure 2B, Supplemental Figure S6A) allowing for 40% of the total reads to be aligned. The error corrected long reads can be used for any purpose, including de novo genome assembly, structural variation analysis, or even isoform resolution when applied to cDNA sequencing.

Figure 2a) Nanocorr workflow

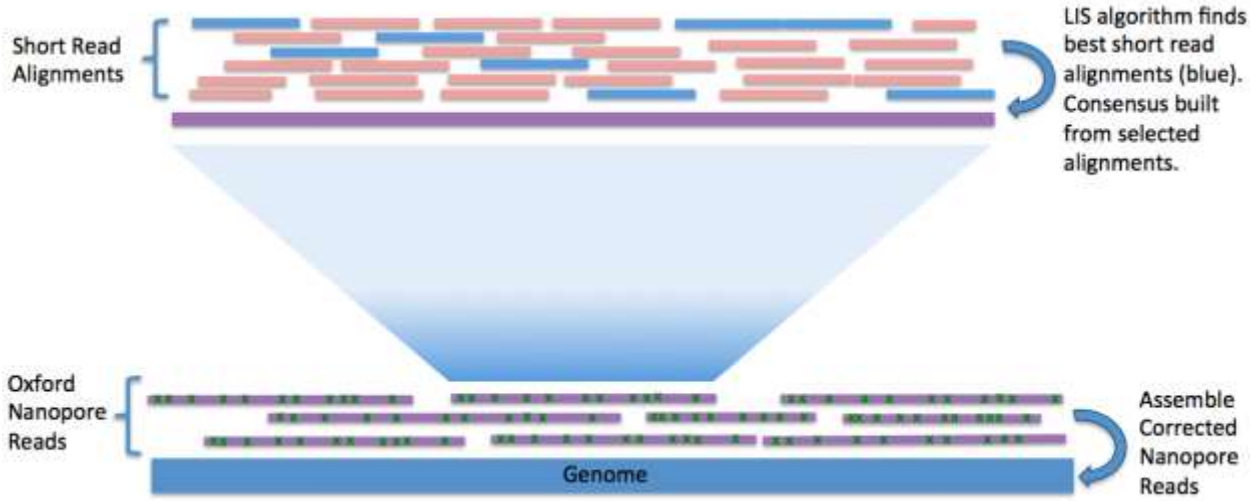

Figure 2b) Post-Nanocorr correction read length and accuracy

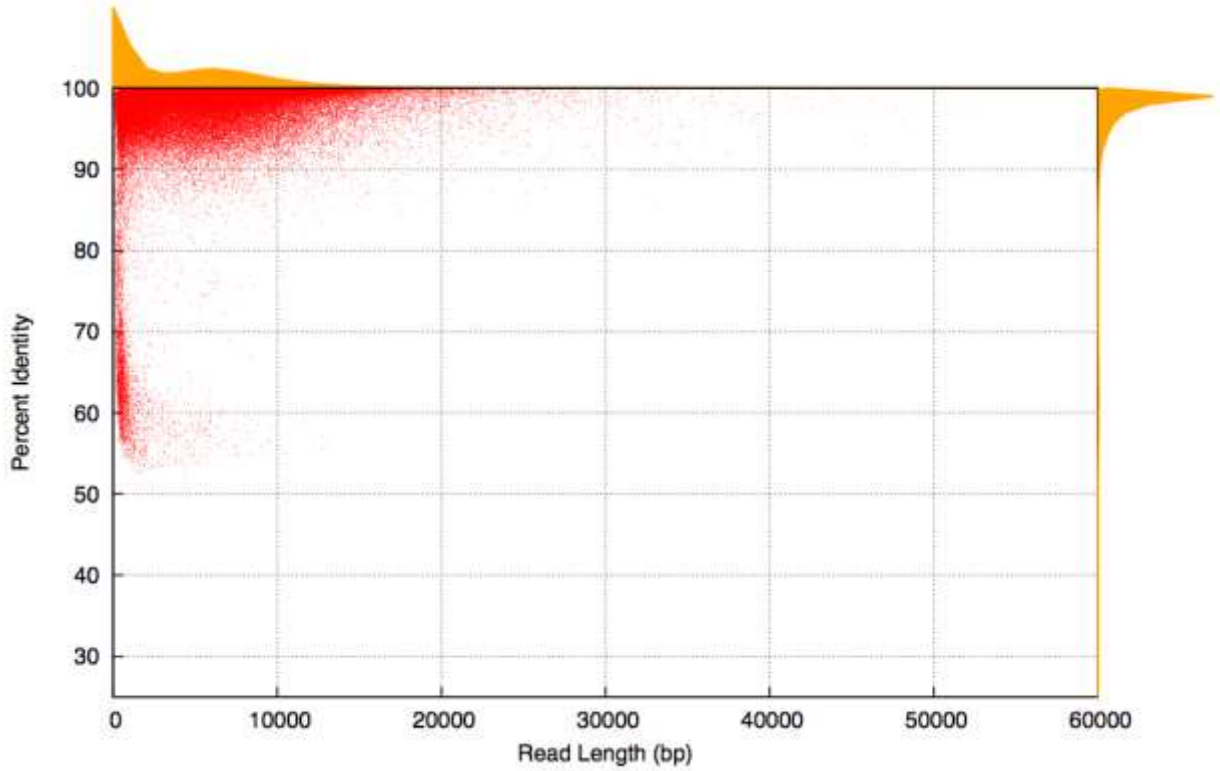

Figure 2 A) Nanocorr workflow. Short high identity reads are aligned to raw ONT reads. The best overlapping set is determined by the LIS algorithm and a consensus sequence of these alignments is built using pbdagcon. Error corrected reads can then be assembled using a long read assembler. B) Scatterplot with marginal histograms summarizing the percent identity of reads after correction for W303. Average identity before correction is ~65%, average post-correction identity of > 97%.

To demonstrate the utility of the long reads, we assembled the error corrected reads *de novo* using the Celera Assembler, which constructs a string graph from the overlaps between the reads to assemble reads up to 500kbp long. This created an assembly with 235 non-redundant contigs with a contig N50 size of 479kbp. Upon alignment to the reference sequence we found that over 99% of the reference genome aligned to our assembly and the per-base accuracy of our assembly was more than 99.78%. Supplemental Figure S6B shows the dot plot Oxford Nanopore-based assembly to the reference genome, and highlights the highly contiguous nature of the assembly with only a few contigs needed to span each chromosome. This assembly has substantially better resolution of the genome compared to an assembly of the MiSeq reads on their own which produced an assembly with a contig N50 size of only 59kbp (the Nanopore-based assembly is more than 7 times more contiguous).

Aligning the assemblies against the reference yeast genome allowed us to evaluate how well the two assemblies represented the various classes of annotated genomic features. As can be seen in Figure 3, while both the Illumina-only and nanopore-based assemblies could correctly assemble short genomic features, the nanopore-based assembly was able to substantially outperform the Illumina-only assembly of long genomic features averaging ~1393bp or longer. In particular, rRNAs, gene cassettes, telomers, and transposable elements were substantially better represented

in the nanopore assembly, and nearly completely absent from in Illumina-only assembly. Only the very longest repeats in the genome, such as the 20kbp telomeric repeats, remain unresolved in the Oxford Nanopore assembly and become fragmented in both assemblies. The MiSeq assembly slightly outperforms for "binding site" features, although these are binding sites within the telomeric repeats that were not well assembled by either technology.

Figure 3) Genomic features assembly by Oxford Nanopore and Illumina sequencing

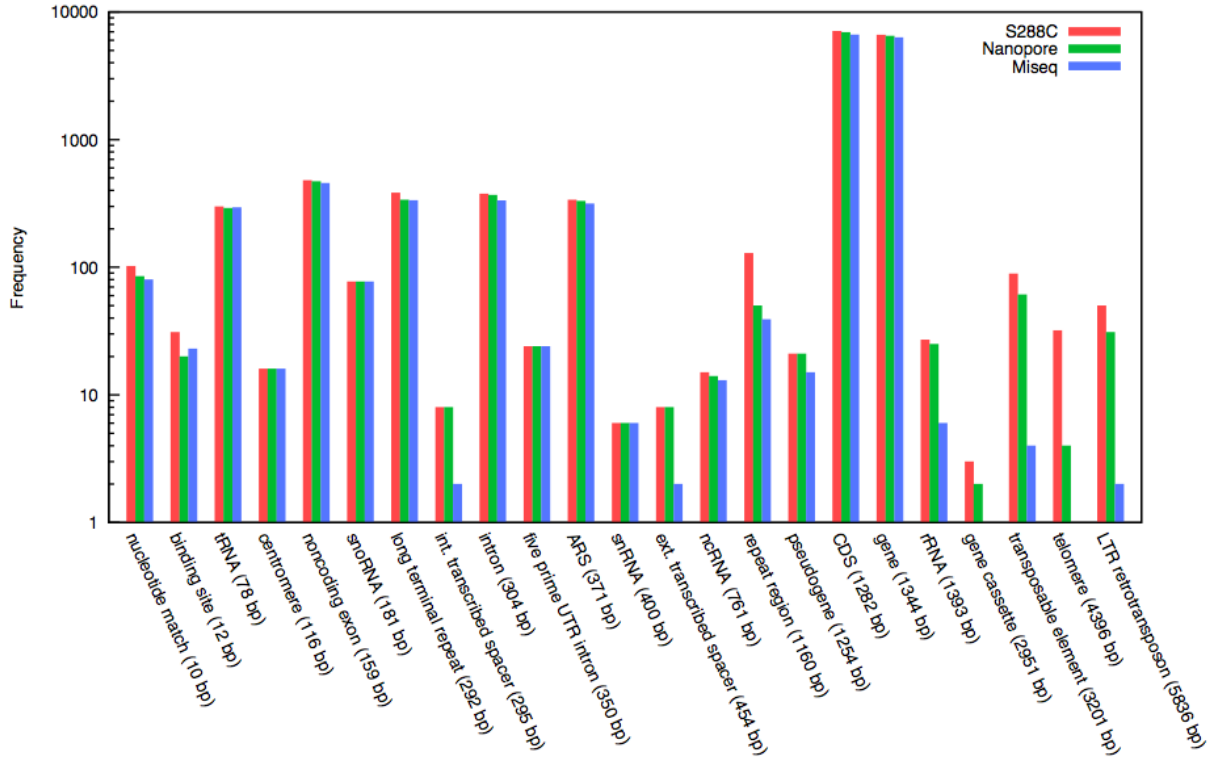

Figure 3. Quantification of different annotated genomic features assembled completely by the Nanopore and Illumina/MiSeq only assembly relative to the complete S288C reference annotation. The Nanopore-based assembly produces an assembly with many more of the longer features assembled compared to the Illumina/MiSeq-only assembly.

In order to validate the utility of this workflow, we also error corrected and de novo assembled the Oxford Nanopore reads generated by Nicholas Loman et al.[12] of *E. coli* K12 using the same approach (Supplemental Note 7). In this experiment, a total of 145x Oxford Nanopore read coverage of the genome was error corrected with the Nanocorr pipeline using 30x Illumina MiSeq coverage to improve the average identity to over 99%. The final assembly was essentially perfect: one single 4.6Mbp chromosome length contig with >99.99% identity. In contrast, the Illumina-only assembly produced an assembly with hundreds of contigs and a contig N50 size of only 176kbp.

The results of this study indicate the Oxford Nanopore sequence data currently have substantial errors (~25% to 40% error) and a high proportion of reads that completely fail to align (~65%). This is likely due to the challenges of the signal processing the ionic current measurements[13] as well as the challenges inherent in any type of single molecule sequencing. Oxford Nanopore has indicated that the pores are more than a single base in height so that the ionic signal measurements are not of individual nucleotides but of approximately 5 nucleotides at a time. Consequently the base calling must individually recognize at least 4^5=1024 possible states of ionic current for each possible *5-mer*. We also observed the potential for some bias in the signal processing and basecaller, particularly for homopolymers. Despite the limitations of this early phase device, there has been notable improvement over the course of this program, and well performing flowcells of the current iteration (R7 at the time this publication was written) can generate upwards of 400Mb on a single run and we remain optimistic to see yields improve with future generations of the technology.

While short read sequencers in general have lower error rates and, to date, have become the standard approach of genomics, the short reads are not sufficient to generate long continuous assemblies of complex genomes. To this day the reference human genome remains incomplete as do the reference genomes for most higher species, especially larger plant genomes. Instead, long reads such as those generated by the Pacific Bioscience System and now the Oxford Nanopore MinION, are necessary to span repetitive elements and other complex sequences to generate high quality, highly contiguous assemblies. Improving the contiguity of a genome assembly enables more detailed study of its biological content and function in every aspect. Genes will more often be correctly assembled along with their flanking sequences, enabling deeper study of regulatory elements. Longer reads will also resolve more repetitive sequences as well, especially transposable elements, high copy genes, segmental duplications, and centromeric/telomeric repeats that are difficult to assemble with short reads. Finally, high-quality assemblies are also essential to study high-level genome structures such as the evolution and synteny of entire chromosomes across species. Even in genome resequencing, short reads can be problematic, with some (perhaps many) structural variants unresolved, obscuring the true gene content of a member of a species or obscuring clinically relevant structural variants in an affected individual[14].

Modern genome assemblers are not equipped to natively handle reads with error rates above a few percent. Consequently, before the Oxford Nanopore reads can be used for *de novo* assembly they must first be error corrected. These general strategies are helpful for other single molecule, long read sequences such as that from Pacific Biosciences, although existing algorithms were not capable of resolving the Oxford Nanopore errors[11,15] . We successfully developed a new hybrid

error correction approach that can improve the average per base identity of the Oxford Nanopore reads from 65% to greater than 97% and generates nearly perfect or extremely high quality assemblies given sufficient coverage and read lengths. Using the error corrected data, we were able to fully reconstruct an entire microbial genome and produce an extremely high quality assembly of yeast that had many important genomic features that were almost entirely lost in the Illumina-only assembly. This work has demonstrated how single molecule, long read data generated by the Oxford MinION can be successfully used to create highly contiguous genome assemblies, paving the way for essentially any lab to create perfect or high quality reference sequences for their microbial or small eukaryotic projects using an inexpensive, handheld long read sequencer.

Data Access

All data and software used in the study are available open source at:

http://schatzlab.cshl.edu/data/nanocorr/

Acknowledgments

This project was supported in part by National Science Foundation award DBI-1350041 and National Institutes of Health award R01-HG006677 to MCS. We would like to express our thanks to Oxford Nanopore for affording us the opportunity to participate in the MinION early Access program (MAP). In particular we would like to thank Clive Brown, James Brayer and all

the members of the technical support staff for their support and assistance during this research. Finally, we would like to thank all the members of the MAP community for their on-going insight and dedication into the novel device.

Author contribution

SG Performed data analysis, library preparation, managed flow cells and is the MAP lead. JG Performed data analysis, developed Nanocorr and library preparation. SE performed library preparation, PD performed library preparation. MCS assisted in data analysis and in the overall design of the project. WRM developed the overall design of the study and assisted with library preparation. SG, JG, MCS and WRM wrote the manuscript. All authors reviewed and approve the final manuscript.

W.R.M. has participated in Illumina sponsored meetings over the past four years and received travel reimbursement and an honorarium for presenting at these events. Illumina had no role in decisions relating to the study/work to be published, data collection and analysis of data and the decision to publish. W.R.M. has participated in Pacific Biosciences sponsored meetings over the past three years and received travel reimbursement for presenting at these events. W.R.M. is a founder and shared holder of Orion Genomics, which focuses on plant genomics and cancer genetics.

References

1 Kasianowicz, J. J., et al. Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A.* **93**, 13770-13773 (1996).

2 Venkatesan, B.M, and Bashir, R. Nanopore sensors for nucleic acid analysis. *Nature Nanotech.* **6**, 615-624 (2011)

3 Eisenstein, M. Oxford Nanopore announcement sets sequencing sector abuzz. *Nature biotech.* **30**, 295-296 (2012): 295-296

4 Maxam, A. M. and Gilbert W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A.* **74**, 560-564. (1977)

5 Sanger, F., Nicklen, S. and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* **74**, 5463-5467 (1977)

6 Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387-402 (2008)

7 Sanger, F., et al. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J mo.l bio.* **143**, 161-178 (1980)

8 Stoddart, D., et al. Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci U S A.* **106**, 7702-7707 (2009)

9 Clarke, J, et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotech.* **4**, 265-270 (2009)

10 Clive Brown "Oxford Nanopore" http://www.globalengage.co.uk/pgcasia/Brown.pdf (2014)

11 Koren, S., et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotech.* **30**, 693-700 (2012)

12 Quick J., Quinlan, A., and Loman N. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *bioRxiv* doi: http://dx.doi.org/10.1101/009613 (2014)

13 Schreiber, J., et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands *Proc Natl Acad Sci U S A.* **110**, 18910-18915 (2013)

14 Chaisson, Mark JP, et al. "Resolving the complexity of the human genome using single-molecule sequencing." Nature (2014). doi:10.1038/nature13907.

15 Lee, H., et al. Error correction and assembly complexity of single molecule sequencing reads." *bioRxiv* doi: http://dx.doi.org/10.1101/006395 (2014)