Check for updates

RESEARCH ARTICLE

MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry [version 1; referees: 1 approved, 2 approved with reservations]

Miten Jain[1*], John R. Tyson[2*], Matthew Loose[3*], Camilla L.C. Ip[4,5*], David A. Eccles[6], Justin O'Grady[7], Sunir Malla[3], Richard M. Leggett[8], Ola Wallerman[9], Hans J. Jansen[10], Vadim Zalunin[11], Ewan Birney[11*], Bonnie L. Brown[12*], Terrance P. Snutch[2*], Hugh E. Olsen[1*], MinION Analysis and Reference Consortium

[1]University of California at Santa Cruz, Santa Cruz, CA, USA
[2]Michael Smith Laboratories and Djavad Mowfaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada
[3]School of Life Sciences, University of Nottingham, Nottingham, UK
[4]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[5]Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK
[6]Malaghan Institute of Medical Research, Wellington, New Zealand
[7]Norwich Medical School, University of East Anglia, Norwich, UK
[8]Earlham Institute, Norwich Research Park, Norwich, UK
[9]Science for Life Laboratory, IGP, Uppsala University, Uppsala, Sweden
[10]ZF-screens B.V., Leiden, Netherlands
[11]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK
[12]Virginia Commonwealth University, Richmond, VA, USA

[*] Equal contributors

v1 **First published:** 31 May 2017, **6**:760 (doi: 10.12688/f1000research.11354.1)
Latest published: 31 May 2017, **6**:760 (doi: 10.12688/f1000research.11354.1)

Abstract

Background: Long-read sequencing is rapidly evolving and reshaping the suite of opportunities for genomic analysis. For the MinION in particular, as both the platform and chemistry develop, the user community requires reference data to set performance expectations and maximally exploit third-generation sequencing. We performed an analysis of MinION data derived from whole genome sequencing of *Escherichia coli* K-12 using the R9.0 chemistry, comparing the results with the older R7.3 chemistry.
Methods: We computed the error-rate estimates for insertions, deletions, and mismatches in MinION reads.
Results: Run-time characteristics of the flow cell and run scripts for R9.0 were similar to those observed for R7.3 chemistry, but with an 8-fold increase in bases per second (from 30 bps in R7.3 and SQK-MAP005 library preparation, to 250 bps in R9.0) processed by individual nanopores, and less drop-off in yield over time. The 2-dimensional (“2D”) N50 read length was unchanged from the prior chemistry. Using the proportion of alignable reads as a measure of base-call accuracy, 99.9% of “pass” template reads from 1-dimensional (“1D”)

Open Peer Review

Referee Status: ? ✓ ?

	Invited Referees		
	1	2	3
version 1 published 31 May 2017	? report	✓ report	? report

1 **Wigard P. Kloosterman**, University Medical Center Utrecht, Netherlands
Jose Espejo Valle Inclan, University Medical Center Utrecht, Netherlands
Mircea Cretu Stancu, University Medical Center Utrecht, Netherlands

experiments were mappable and ~97% from 2D experiments. The median identity of reads was ~89% for 1D and ~94% for 2D experiments. The total error rate (miscall + insertion + deletion) decreased for 2D “pass” reads from 9.1% in R7.3 to 7.5% in R9.0 and for template “pass” reads from 26.7% in R7.3 to 14.5% in R9.0.

Conclusions: These Phase 2 MinION experiments serve as a baseline by providing estimates for read quality, throughput, and mappability. The datasets further enable the development of bioinformatic tools tailored to the new R9.0 chemistry and the design of novel biological applications for this technology.

Abbreviations: K: thousand, Kb: kilobase (one thousand base pairs), M: million, Mb: megabase (one million base pairs), Gb: gigabase (one billion base pairs).

This article is included in the Nanopore Analysis gateway.

2 **Martin C. Frith**, National Institute of Advanced Industrial Science and Technology (AIST), Japan
The University of Tokyo, Japan
National Institute of Advanced Industrial Science and Technology (AIST), Japan

3 **Titus Brown**, University of California, Davis, USA
Lisa Cohen, University of California, USA

Discuss this article

Comments (1)

Corresponding authors: Miten Jain (miten@soe.ucsc.edu), John R. Tyson (jtyson@msl.ubc.ca), Matthew Loose (matt.loose@nottingham.ac.uk), Camilla L.C. Ip (camilla.ip@well.ox.ac.uk), Ewan Birney (birney@ebi.ac.uk), Bonnie L. Brown (blbrown@vcu.edu), Terrance P. Snutch (snutch@msl.ubc.ca), Hugh E. Olsen (heolsen@soe.ucsc.edu)

Competing interests: All flow cells and library preparation kits were provided by ONT free of charge. Ewan Birney is a paid consultant of ONT. MJ, HEO, JT, ML, CI, HJ, JOG and BB have accepted reimbursement for conference travel expenses from ONT. VZ was funded for his work on this project from Oxford Nanopore through an agreement with EMBL.

How to cite this article: Jain M, Tyson JR, Loose M *et al.* **MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2017, **6**:760 (doi: 10.12688/f1000research.11354.1)

Grant information: The following grants supported the research of the following authors: NHGRI, USA award numbers HG006321 and HG007827 (MJ and HEO, UCSC), Canadian Institutes of Health Research #10677 (JT and TS, UBC), Brain Canada Multi-Investigator Research Initiative Grant with matching support from Genome British Columbia, the Michael Smith Foundation for Health Research and the Koerner Foundation (JT and TS, UBC), BBSRC grant BB/M020061/1 (ML and SM, Nottingham), Wellcome Trust grant 090532/Z/09/Z (CI, WTCHG), UK Antimicrobial Resistance Cross Council Initiative supported by the seven research councils (MR/N013956/1) and Rosetrees Trust grant A749 (JOG, UEA), BBSRC grant BB/J010375/1 (RML, Earlham), E o R Börjessons foundation (OW, Uppsala), and National Science Foundation DEB-1355059 (BB, VCU).

First published: 31 May 2017, **6**:760 (doi: 10.12688/f1000research.11354.1)

Introduction

The Oxford Nanopore Technologies (ONT) MinION Access Programme (MAP) released the MinION™ nanopore sequencer to early access users in June 2014. The MinION Analysis and Reference Consortium (MARC) was formed by a subset of MAP participants to perform independent evaluation of the platform, share standard protocols, collaboratively produce reference data for the nanopore community, and to address biological questions. The Phase 1 MARC analysis of October 2015[1] was an evaluation of the library preparation chemistry version SQK–MAP005, R7.3 flow cell chemistry, and a base-calling algorithm derived from a Markov model (HMM) using a 5-mer model. The R9.0 chemistry and protocol, (https://www.youtube.com/watch?v=nizGyutn6v4) was made available to users in June 2016 (https://londoncallingconf.co.uk/lc/2016-plenary#168687629). This substantial upgrade to the platform included the CsgG membrane protein for the pore and a recurrent neural network (RNN) for base-calling. In part, ONT claimed these changes made substantial improvements to data yields and quality, to the extent that 1-dimensional (“1D”) reads, without a hairpin, could be used for analyses in many use-cases.

Before embarking on further analyses, MARC performed “bridging experiments” to evaluate the effect of the R9.0 changes on data yield, quality, and accuracy. To capture variability and reproducibility among experiments using R9.0 chemistry, two labs concurrently sequenced *Escherichia coli* strain K-12 substrain MG1655, the same strain used for MARC Phase 1[1]. Sequencing was performed using both the 2-dimensional (“2D”) “ligation” kit and the newer 1D “rapid” kit. The analyses performed included characterizing throughput, read quality, and accuracy. This work also marks the release of MinION Phase 2 data for both sequencing modes with the R9.0 chemistry. Although the newer R9.4 flow cell chemistry has become available to the community since the Phase 2 experiments were performed in late July and early August 2016, ONT have stated that R9.4 flow cell chemistry has similar base-calling characteristics compared to R9.0, as it uses the same pore and base-calling strategy. Thus, this data release and analysis is of interest as it describes the major changes introduced with the R9 chemistry. It is a resource to aid further developments in nanopore informatics as well as the development of biological applications using the MinION.

Materials and methods

Two laboratories each performed a 1D and a 2D experiment using the protocol described in MARC Phase 1[1] to obtain total genomic DNA from freshly grown cells (Supplementary File 1) and slightly modified protocols for 1D “rapid” and 2D “ligation” library preparation and sequencing.

Cell culture and DNA extraction of the *E. coli* K-12 target sample

E. coli cells were cultured and DNA was extracted using the protocol described in MARC Phase 1 (Supplementary File 1).

2D sequencing library preparation

Sequencing libraries were prepared according to the ONT recommended 2D protocol (SQK-NSK007 kits), which included addition of the lambda control sample, with the following changes:

(i) genomic DNA was sheared to ~10 kb; and

(ii) both labs performed a 0.4x AMPureXP cleanup post-FFPE treatment.

1D sequencing library preparation

Sequencing libraries were prepared according to the ONT recommended 1D protocol (SQK-RAD001 kits, referred to as 1D “rapid” sequencing) with the following changes:

(i) a 0.4x AMPureXP cleanup was performed prior to 1D library preparation;

(ii) an unsheared input DNA sample of 400 ng was used for the library;

(iii) 0.4 µl Blunt/TA Ligase was added; and

(iv) a 10 min incubation was used in the final step.

Note that this protocol does not include addition of the lambda control sample DNA.

Sequencer configuration and sequencing run conditions

All sequencing runs used MinKNOW (version 1.0.3) and Metrichor Desktop Agent. The experiments are henceforth referred to as P2-Lab6-R1-2D, P2-Lab7-R1-2D, P2-Lab6-R1-1D and P2-Lab7-R1-1D following a “phase-lab-replicate-kit” format. All flow cells used for sequencing underwent the standard MinION Platform QC for analysis of overall quality and number of functional pores. This was followed by the recommended priming step, after which the prepared library was loaded onto an R9.0 flow cell. Final library volume for the 1D runs was 11.2 µl, which was loaded once with running buffer at the start of the experiment. A 500 µl flush with running buffer alone was performed at 24 hrs on the P2-Lab6-R1-1D run. The final volume of 2D libraries was 25 µl, of which 12 µl was loaded with running buffer at the start of the sequencing run followed by addition of another 12 µl library aliquot 16 hours into the run. All sequencing runs were performed on MinION Mk1b devices using the standard MinKNOW 48-hour sequencing protocol (NC_48Hr_Sequencing_Run_FLO-MIN104).

Base-calling and data formats

The sequencing data for 1D MinION runs were base-called using the Metrichor 1D Base-calling RNN for the SQK-RAD001 (v1.107) workflow. This workflow classified base-called sequence data into “pass” and “fail” categories based on the mean Phred-scaled quality score for that read. The threshold for a read to be categorized as “pass” was a Q-value of 6. The sequencing data for 2D MinION runs were base-called using the Metrichor 2D Base-calling RNN for the SQK-NSK007 (v1.107) workflow. Similarly, this workflow classified reads into “pass” and “fail” with a Q-value threshold of 9 required for pass reads.

European Nucleotide Archive data pre-processing pipeline

As in Phase 1, the base-called FAST5 files and meta-data were collated on a server at the European Nucleotide Archive (ENA). These data were then processed using several tools. The base-calls in FASTQ format were extracted using poretools (version 0.5.1)[2] and then aligned against the *E. coli* K-12 reference genome (NCBI RefSeq, accession NC_000913.1) using BWA-MEM (version 0.7.12-41044), parameter "-x ont2d"[3] and LAST (version 460)[4], parameters "-s 2 -T 0 -Q 0 -a 1" as recommended by 5. Both alignments were then improved with marginAlign (version 0.1)[6], and were statistics computed using marginStats[6].

Data analyses

The R9.0 data were characterized by collating statistics for a typical run from MARC Phase 1 (P1b-Lab2-R2, hereafter referred to as P1b-Lab2-R2-2D for consistency with the Phase 2 experiment naming convention) and the four Phase 2 experiments. In keeping with the MARC Phase 1 analyses[1], we computed alignments and error-rate measurements using BWA-MEM and LAST, followed by re-alignment using marginAlign[6]. Real-time evaluation of the runs was performed by minoTour[7] (more information available from: http://minotour.github.io/minoTour), run locally at the two experimental laboratories. The "pass" and "fail" reads from each experiment were evaluated with NanoOK (version 0.95)[8] using bwa alignments. Additional metrics and analyses were performed with bespoke Python and R scripts, (available at https://github.com/camilla-ip/marcp2)[9].

Results

Experimental conditions

The MARC Phase 2 experiments were performed by two laboratories (Supplementary File 1) between 27 July and 2 August 2016 (Table 1). The total number of functional g1 pores prior to sequencing on R9.0 flow cells was ~94%, an improvement from ~88% for R7.3 (Table 1). The operating ASIC (chip) temperature on the R9.0 flow cell ranged from 30 to 34°C, and the temperature regulation of the flow cell heat sink was a uniform 34°C across all flow cells (Table 1). All experiments ran for at least 40 hours of the 48 hour run script. However, experiment P2-Lab6-R1-2D crashed when the controlling computer's hard-drive reached capacity; it was restarted ~42 hours after the initial experiment start time using modified recipe scripts, but produced few further reads. Experiment P2-Lab7-R1-2D was terminated after ~44 hours. Experiment P2-Lab7-R1-1D was restarted twice between 24 and 32 hours and terminated at 41.5 hours (Table 1).

Data format and experimental constants

One challenge of MinION data analysis is referencing the proper data format after major upgrades, such as the switch from an HMM to an RNN base-caller. The new or superseded fields in the resulting table after introduction of R9 chemistry are shown in Supplementary File 2[9].

Base yield and read lengths

The read count, base yield, and read lengths of the 2D and 1D R9.0 experiments compared to a typical R7.3 experiment (Table 2 and Table 3, and Figure 1) were inferred from NanoOK reports (Supplementary File 3) and bespoke scripts[9]. There was considerable variability between the quantity of data produced by the two 2D experiments and the two 1D experiments, but overall, the R9.0 chemistry showed an increase in data yield and read length when compared with a typical Phase 1 R7.3 experiment.

Improvements in base yield and read length were observed for the 2D R9.0 experiments compared with a typical R7.3 experiment (Table 2 and Table 3). The 2D R9.0 experiments sequenced

Table 1. Experimental conditions. P1 refers to a typical R7.3 run from MARC Phase 1[1]. P2 refers to the MARC Phase 2 R9.0 data presented in this study. NA: not available.

	P1b-Lab2-R2-2D	P2-Lab6-R1-2D	P2-Lab7-R1-2D	P2-Lab6-R1-1D	P2-Lab7-R1-1D
Library & base-call type	2D	2D	2D	1D	1D
Flow cell version	R7.3	R9	R9	R9	R9
MinION device	Initial version	Mk1b	Mk1b	Mk1b	Mk1b
Experiment start date	2015-07-25	2016-07-27	2016-07-27	2016-08-02	2016-07-29
Active g1 pores (% of 512)	87.9	94	NA	94	NA
Active g2 pores (% of 512)	60.7	77	NA	69	NA
Mean ASIC temperature (°C)	24.4	30.5	33.7	31.9	33.8
Mean heat-sink temp (°C)	37.1	34.0	34.0	34.0	34.0
Experiment run time (h)	48.0	41.5	44.0	48.0	48.0
Experimental notes	Full run	Hard drive filled up at ~41.5 h	Terminated early ~44h as no more data generated	Full run; no lambda control sample	Two restarts between 24 and 32 h, no lambda control sample

Table 2. Read counts and base yields. ("-") indicates not applicable.

	P1b-Lab2-R1-2D	P2-Lab6-R1-2D	P2-Lab7-R1-2D	P2-Lab6-R1-1D	P2-Lab7-R1-1D
Read count (K)					
files - total	48.5 (100%)	126.8 (100%)	216.5 (100%)	96.2 (100%)	57.4 (100%)
pass	14.8 (30.5%)	41.7 (32.9%)	80.6 (37.2%)	56.9 (56.1%)	35.3 (60.5%)
fail	33.7 (69.5%)	85.1 (67.1%)	135.9 (62.8%)	39.3 (40.9%)	22.1 (38.5%)
template - total	48.0 (99.0%)	126.7 (99.9%)	216.5 (100%)	96.2 (100%)	57.4 (100%)
pass	14.8 (30.8%)	41.7 (32.9%)	80.6 (37.2%)	56.9 (59.1%)	35.3 (60.5%)
fail	33.2 (69.2%)	85.0 (67.1%)	135.8 (62.7%)	39.3 (40.9%)	22.1 (38.5%)
comp - total	34.2 (70.5%)	89.2 (70.3%)	146.6 (67.7%)	-	-
pass	14.8 (43.3%)	41.7 (46.7%)	80.6 (55.0%)		
fail	19.4 (56.7%)	47.4 (53.1%)	66.0 (45.0%)		
2D - total	21.4 (44.1%)	63.6 (50.2%)	111.4 (51.5%)	-	-
pass	14.8 (69.2%)	41.7 (65.6%)	80.6 (72.4%)		
fail	6.6 (30.8%)	21.9 (34.4%)	30.8 (27.6%)		
Base yield (Mb)					
template - total	242.4 (100%)	790.5 (100%)	1268.8 (100%)	829.7 (100%)	410.6 (100%)
pass	92.1 (38.0%)	330.5 (41.8%)	546.3 (43.1%)	526.3 (63.4%)	276.3 (67.3%)
fail	150.3 (62.0%)	460.0 (58.2%)	722.5 (56.9%)	303.3 (36.6%)	134.3 (32.7%)
comp - total	180.3 (74.4%)	437.6 (55.6%)	681.1 (53.7%)	-	-
pass	87.7 (48.7%)	286.2 (65.4%)	481.3 (70.7%)		
fail	92.5 (51.3%)	151.4 (34.6%)	199.8 (29.3%)		
2D - total	128.2 (52.9%)	414.9 (52.5%)	665.6 (52.5%)	-	-
pass	94.1 (73.4%)	320.8 (77.3%)	533.8 (80.2%)		
fail	34.2 (26.6%)	94.1 (22.7%)	131.8 (19.8%)		

Table 3. Read lengths. ("-") indicates not applicable.

	P1b-Lab2-R1-2D		P2-Lab6-R1-2D		P2-Lab7-R1-2D		P2-Lab6-R1-1D		P2-Lab7-R1-1D	
	total	pass	total	pass	total	pass	total	pass	total	pass
Mean length (Kb)										
template	5.0	6.2	6.2	7.9	5.9	6.8	8.6	9.2	7.2	7.8
complement	5.3	5.9	4.9	6.9	4.6	6.0	-	-	-	-
2D	6.0	6.4	6.5	7.7	6.0	6.6	-	-	-	-
**Longest read (Kb)										
template	244.5	36.1	119.0	50.3	110.8	34.6	478.8	141.2	353.1	151.2
complement	50.2	33.9	403.7	47.0	253.0	34.9	-	-	-	-
2D	35.2	35.2	50.9	50.9	40.1	35.8	-	-	-	-
Longest aligned read (> 75% length aligned) (Kb)										
template	35.6	34.6	56.4	50.3	42.7	33.3	141.2	141.2	151.2	151.2
complement	33.9	33.9	47.0	47.0	32.2	32.2	-	-	-	-
2D	35.2	35.2	50.9	50.9	33.6	33.6	-	-	-	-
N50 length (Kb)										
template	6.9	7.5	9.4	10.0	7.6	7.9	15.7	16.2	13.1	13.6
complement	6.8	7.1	8.0	8.8	6.5	7.0	-	-	-	-
2D	7.4	7.6	9.1	9.8	7.3	7.8	-	-	-	-

** : Longest read here is pre-alignment.

127–217 K molecules (compared with ~49 K molecules for the typical Phase 1 R7.3 experiment). Of these, ~50% resulted in 2D reads (an improvement from ~44% for the typical R7.3 experiment) and a total of 64–111 K 2D pass reads (compared with 21 K for the typical R7.3 experiment). The proportion of "pass" reads with a Q-value threshold of 9 was 66% to 72%, about the same as that observed for the typical R7.3 experiment, with a base quality threshold of 9.0. Average read lengths of "pass" 2D base-calls were higher at 6.6–7.7 Kb (compared with 6.4 Kb for the typical R7.3 experiment), and for "all" 2D base-calls at 6.0–6.5 Kb (compared with 6.0 Kb for the typical R7.3 experiment). The longest 2D reads observed in R9.0 (50.9 Kb, Table 3) were comparable to those observed in R7.3 experiments (59.7 Kb)[1]. However, the longest 2D aligned read observed increased to 50.9 Kb (from 35.2 Kb in the typical R7.3 experiment) (Table 3). The increase in N50 read length to 7.3–9.1 Kb for all 2D reads in the R9.0 experiments (compared with 7.4 Kb for the typical R7.3 experiment) and 7.8–9.8 Kb for "pass" R9.0 reads (compared with 7.6 Kb for the R7.3 experiment) indicates, as for the 1D data, an overall increase in the proportion of longer 2D base-called reads.

The 1D R9.0 experiments sequenced 57–96 K molecules (compared with 49 K for the typical Phase 1 R7.3 experiment), resulting in a total template base yield of 410–830 Mb (compared with 242 Mb for the typical R7.3 experiment), of which ~60% were higher-quality "pass" reads with a Q-value threshold of 6.0 (compared with ~31% for the typical Phase 1 experiment classified with 2D base quality threshold of 9.0) (Table 2). Read lengths also improved, with the mean template length for "pass" reads increasing to 7.2–8.6 Kb (from 5.0 for R7.3) and increasing to 7.8–9.4 for "fail" reads (from 6.2 for the R7.3 experiment). The longest mappable template read observed across all of the R9 runs was 151.2 kb and the artefactually long reads, detectable by a discrepancy between the longest read lengths and the longest mappable read lengths, were comparably rare (Table 3). Read length N50 increased to 13.1–15.7 Kb for "pass" reads (compared with 6.9 Kb for the typical R7.3) and 13.6–16.2 Kb (compared with 7.5 Kb for the typical R7.3), indicating that more of the base-calls were contained in longer reads.

We observed that the speed and convenience of the 1D "rapid" library protocol came at a cost. The distribution of template "pass" read lengths was skewed toward shorter reads peaking closer to 1 Kb rather than the ~6.5 Kb obtained through the 2D "ligation" library protocol. However, one benefit was that a greater proportion of longer reads was also produced (Figure 1). The addition of the lambda control sample in the 2D library protocol resulted in a variable ratio of "target" to "control" sample reads, evident in the relative sizes of the bimodal read length distributions for the 2D library experiments (Figure 1).

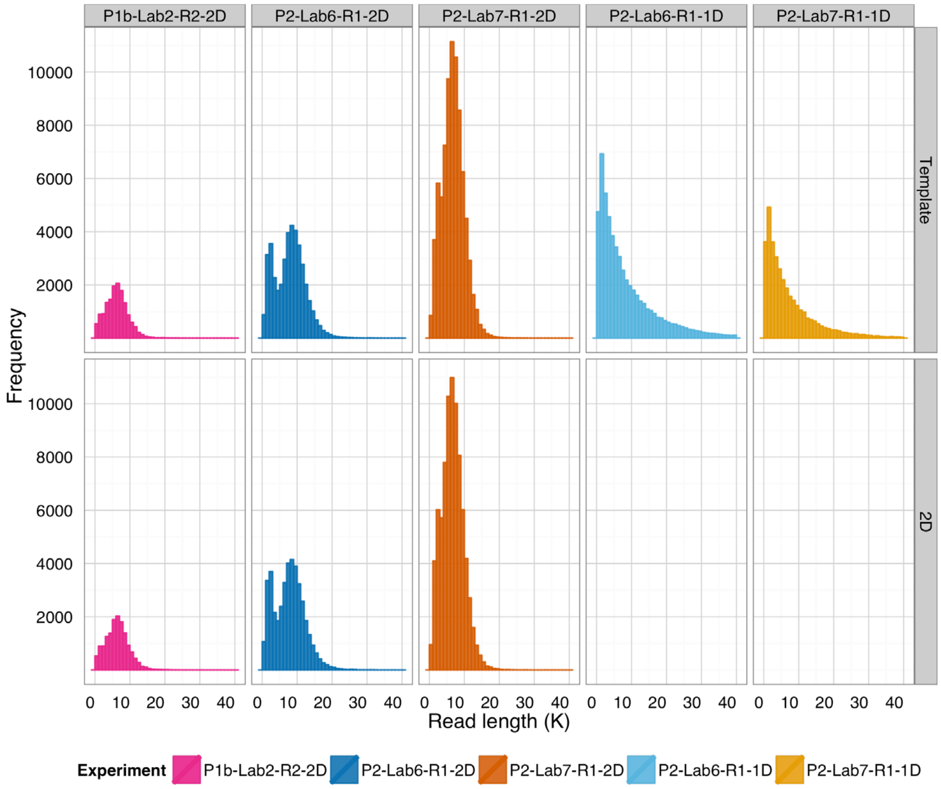

Figure 1. Read length distribution for template and 2D "pass" reads. The distribution of template ("1D") read lengths for experiments based on 1D "rapid" libraries (P2-Lab6-R1-1D and P2-Lab7-R1-1D) was skewed toward shorter read lengths due to enzymatic, rather than mechanical, DNA fragmentation. The long tails of the distributions were truncated at 40,000 bases for clarity.

Alignment identity and accuracy

The proportion of alignable reads is a measure of the accuracy of the base-calls. For template reads from both 1D and 2D experiments, 99.9% of "pass" reads were alignable from both 1D and 2D experiments, and 60% and 83% for "fail" reads from 1D and 2D experiments, respectively (Table 4).

The median identity of reads from 1D and 2D experiments (Table 4) was similar to that observed for the R7.3 chemistry in MARC Phase 1. The median identity for 1D template reads was ~88% and ~76%, for "pass" and "fail", respectively (compared with 78% and 75% for the typical R7.3 experiment). For the 2D experiments, the read identity was ~89% and ~85%, for "pass" and "fail", respectively (compared with ~92% and ~82%, respectively, for the typical R7.3 experiment).

Another metric of overall error, the longest perfectly aligned subsequence, showed improvement associated with the R9.0 chemistry. The longest perfectly aligned subsequences in the R9.0 1D runs were 235 and 273 bases (compared with 87 in the typical R7.3 experiment), and in the 2D runs were 713 and 750 bases (compared with 333 bases in the typical R7.3 experiment).

Table 4. Per-read accuracy metrics for target E. coli sample. ("-") indicates the metrics were not applicable for that experiment. NA: not available.

	P1b-Lab2-R1-2D		P2-Lab6-R1-2D		P2-Lab7-R1-2D		P2-Lab6-R1-1D		P2-Lab7-R1-1D	
	pass	fail	pass	fail	pass	fail	pass	fail	pass	fail
Identity %										
template	77.9	74.6	89.4	85.0	89.7	85.6	88.0	76.0	88.8	75.6
complement	79.7	76.8	88.5	83.1	88.9	83.4	-	-	NA	-
2D	92.4	82.2	93.5	69.9	94.0	69.1	-	-	NA	-
Reads mapped %										
template	96.3	47.5	95.9	72.7	98.0	77.2	99.9	62.6	99.9	57.5
complement	96.1	59.8	95.6	61.5	97.8	63.0	-	-	-	-
2D	96.7	84.0	95.9	83.7	98.0	82.4	-	-	-	-
Longest perfectly aligned subsequence										
template	87	85	275	248	274	271	235	235	273	273
complement	75	75	228	228	203	203	-	-	-	-
2D	333	333	713	713	750	750	-	-	-	-
Total error %										
template	26.7	32.8	14.5	19.2	14.0	18.5	15.3	30.5	14.5	31.3
complement	27.9	32.4	17.4	23.0	16.7	22.6	-	-	-	-
2D	9.1	19.7	7.8	25.4	7.2	25.4	-	-	-	-
Miscall %										
template	10.3	12.2	6.1	8.0	5.9	8.0	6.4	15.1	5.9	15.7
complement	9.9	10.6	6.5	8.6	6.4	8.7	-	-	-	-
2D	1.9	5.2	2.1	7.2	2.0	7.2	-	-	-	-
Insertion %										
template	6.5	7.6	2.7	3.8	2.6	3.6	3.2	5.8	3.0	5.9
complement	6.3	7.6	3.0	4.5	2.9	4.4	-	-	-	-
2D	3.1	6.5	2.0	7.9	1.9	8.1	-	-	-	-
Deletion %										
template	9.9	13.0	5.8	7.4	5.5	6.9	5.7	9.7	5.6	9.8
complement	11.8	14.2	7.9	9.9	7.4	9.5	-	-	-	-
2D	4.1	8.0	3.7	10.3	3.3	10.2	-	-	-	-

Miscall, insertion, and deletion rates

The total error of "pass" reads in the 1D sequencing experiments reduced from 26.7% in R7.3 to 15.0% in R9.0 (miscalls 6.2%, insertions 3.1%, deletions 5.7%) (Table 4). Little change was observed for the "fail" template reads, between the 32.8% observed for a typical R7.3 experiment and the 31.1% for the R9.0 experiments (miscalls 15.4%, insertions 5.9%, deletions 9.8%) (Table 4).

Total error of the 2D reads was reduced from 9.1% in R7.3 to 7.3% in R9.0 for "pass" reads, whereas the total error increased for "fail" reads from 19.7% in R7.3 to 25.4% in R9.0 (Table 4).

Sequencing performance over time

In the MARC Phase 1 analysis of R7.3 chemistry experiments, the quantity and quality of data produced during an experiment varied as material passed from one side of the membrane to the other. This was punctuated by periodic changes in voltage every 4 hours, and a switch to the group 2 pores at 24 hours[1]. To enable a direct comparison between the performance of the R7.3 and R9.0 chemistry, key metrics were plotted for 15 minute windows over the course of the 48 hour experiment for the typical R7.3 experiment (P1b-Lab2-R2-2D) and the four R9.0 experiments on the same scale (Figure 2). The mean of each time window was computed from "pass" reads

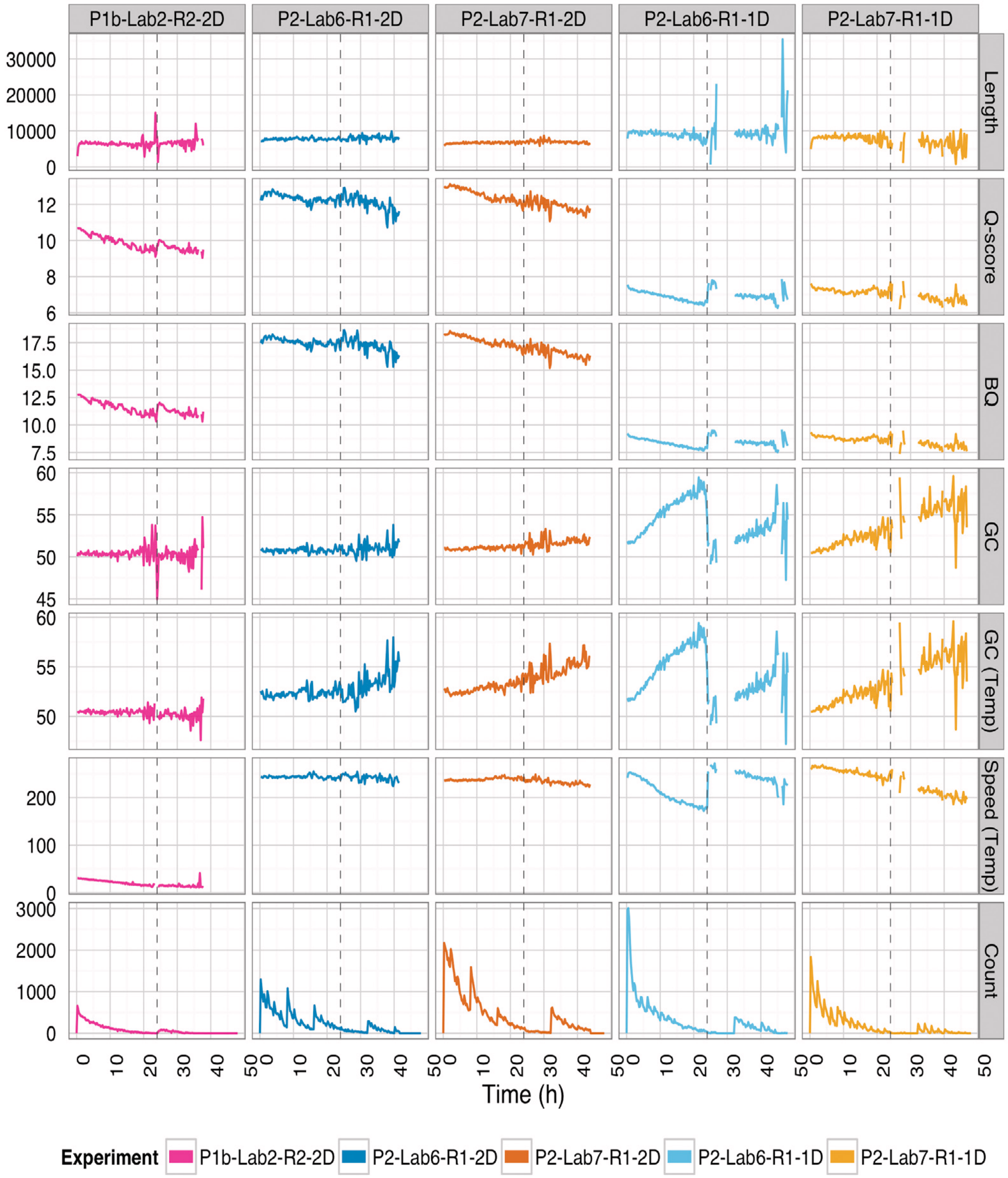

Figure 2. Sequencing performance over time. The mean read length (kb), Q-score, base quality (BQ), and GC%, speed (bases per second), and throughput (count) for each experiment, computed from "pass" reads that mapped to the *E. coli* reference, were plotted for 15 minute intervals. The values for template reads ("1D") are plotted for the 1D libraries (P2-Lab6-R1-1D and P2-Lab7-R1-1D) whereas the values for 2D reads were plotted for the 2D libraries (P1b-Lab2-R2-2D, P2-Lab6-R1-2D, and P2-Lab7-R1-2D).

that mapped to the *E. coli* reference genome, to remove irregularities due to poor quality reads. The metrics computed from template base-called reads were plotted for the 1D library experiments, and those from 2D base-called reads for the 2D library experiments.

The plots show some irregularities due to lower throughput before the pore group switch at 24 hours, towards the end of the runs, during run script restarts (in P2-Lab7-R1-1D and P2-Lab6-R1-2D), and at the early termination (P2-Lab7-R1-2D). However, in general, the read lengths and GC% varied around a constant value over time and the Q-score and base quality dropped at a similar rate (Figure 2) This was despite sequencing speed increasing (measured in bases per second) from about 30 bps to 250 bps. Differences in the Lab6 and Lab7 1D "rapid" run plots around the 24hr point can be attributed to flushing of the flowcell with 500µl of fresh running buffer in the case of Lab6. This appears to be of benefit for speed and quality, but would require further investigation on a chemistry no longer in use. This procedure may be worth bearing in mind going forward, however, for possible beneficial effects with newer chemistries.

We noticed an increase in the GC content of the template reads from the 1D "rapid" library experiments and to a lesser extent for the 2D reads from the 2D experiments (Figure 2). These plots should have shown stochastic variation throughout the run around the mean GC of 50.8% for the *E. coli* sample. We considered a number of possible factors that could account for this artefact including: (i) low data density; (ii) an over-representation of poorer-quality "fail" reads; (iii) an over-representation of unmappable reads; or (iv) high-GC repetitive motifs. We found a negative correlation for the R9.0 1D data between %GC and average QV scores and also a decrease in base qualities over time. This was particularly pronounced for 1D "fail" reads (Q 3–10), but persisted even for 2D reads, likely due to 1D consensus follow through. The current report is for the initial R9.0 chemistry, and the GC-bias seems to be less pronounced with the improved version of the R9 pore (R9.4 data not shown).

Discussion

The MARC Phase 2 experiments were performed with the MinION Mk1b device to provide an independent evaluation of the performance, data yield, and data quality of the R9.0 chemistry and scripts. By comparing the data from four R9.0 experiments on the same *E. coli* isolate sequenced with R7.3 chemistry in MARC Phase 1[1], we have established new benchmarks for data from the 1D "rapid" and 2D "ligation" protocols and kits available in late July 2016. (Table 1).

We have verified that the MinION Mk1b device reliably maintains the R9.0 flow cell at an appropriate temperature (Table 1). The R9.0 flow cells improve overall data yield through provision of a higher proportion of available functional pores during an experiment, with 94% functional group 1 pores observed in this study (Table 1). With higher yields comes an increased chance of experiment failure as the file system accepting the data is likely to reach capacity during a run (Table 1). This suggests scripts should be deployed routinely to move the data from the file system during the sequencing run. The FAST5 data format continues to evolve and improve (Supplementary File 2) to store more comprehensive metadata in a more logical internal structure, and is now beginning to be documented on the MAP Community Forum (available via https://nanoporetech.com).

In the 12 months between the MARC Phase 1 and Phase 2 experiments (Table 1), we observed that for 2D base-calls, the distribution of read lengths remained the same (Figure 1, Table 3). The yield of higher-quality "pass" base-calls increased from ~100 Mb to ~450 Mb per flow cell (Table 2), and the total error of the "pass" base-calls reduced from 9.1% to 7.5% (Table 4). The read length and GC% over the course of the experiment remained uniform (Figure 2). The initial mean Q-scores increased from ~11 to over 12. The initial mean base qualities increased from ~12.5 to over 17.5, and both decreased gradually over the course of an experiment as observed previously (Figure 2). Finally, the proportion of mappable reads remained comparable, between 96 and 98% (Table 4) despite the sequencing speed increasing from 50 to 250 bases per second (Supplementary File 2). The yield improvements are a result of higher speeds and proportion of available pores, and the increase in data quality is attributed to the newer RNN basecaller.

The new 1D "rapid" library protocol, which sequences a single DNA strand, has the potential to query twice as many molecules during the lifetime of a flow cell. We found that this technique is a viable alternative to 2D library chemistry for use-cases where rapid scanning of the population of library molecules is important. The higher total error of 15.3% for "pass" template base-calls, compared with 7.5% for "pass" 2D base-calls (Table 4), is an acceptable trade off.

We confirm that the yield and quality of MinION data continues to improve. The data released in this study provide a benchmark to compare the newer R9.4 chemistry to and can be used to develop bioinformatic tools tailored to the newer chemistry. The updated reports of achievable data yield and quality, along with the characteristics of data production during the lifetime of a flow cell, will enable the design of new biological applications for this third-generation sequencing technology. Although a newer R9.4 chemistry has recently become available, ONT has emphasized that R9 platforms that use the CsgG nanopore will be backward compatible. This study provides the first comprehensive description of data from R9.0 flow cells and RNN base-calling software. We anticipate that it will serve as a framework for evaluating changes resulting from subsequent R9-based chemistries.

Data and software availability

All data presented in this study are available via ENA with accession PRJEB18053.

Archived source code as at the time of publication: http://dx.doi.org/10.5281/zenodo.582311[10]

License: CC BY 4.0

Author contributions

MJ and JT coordinated the study. The MARC group collectively designed the study. ML, SM, and JT performed the experiments. VZ, RL, ML, MJ, RL and CI ran data pre-processing steps. MJ, CI and JT analysed the data. MJ and BB drafted the manuscript. All authors participated in discussions relating to the generation and analysis of the data and edited and approved the final manuscript for submission.

Competing interests

All flow cells and library preparation kits were provided by ONT free of charge. Ewan Birney is a paid consultant of ONT. MJ, HEO, JT, ML, CI, HJ, JOG and BB have accepted reimbursement for conference travel expenses from ONT. VZ was funded for his work on this project from Oxford Nanopore through an agreement with EMBL.

Grant information

The following grants supported the research of the following authors: NHGRI, USA award numbers HG006321 and HG007827 (MJ and HEO, UCSC), Canadian Institutes of Health Research #10677 (JT and TS, UBC), Brain Canada Multi-Investigator Research Initiative Grant with matching support from Genome British Columbia, the Michael Smith Foundation for Health Research and the Koerner Foundation (JT and TS, UBC), BBSRC grant BB/M020061/1 (ML and SM, Nottingham), Wellcome Trust grant 090532/Z/09/Z (CI, WTCHG), UK Antimicrobial Resistance Cross Council Initiative supported by the seven research councils (MR/N013956/1) and Rosetrees Trust grant A749 (JOG, UEA), BBSRC grant BB/J010375/1 (RML, Earlham), E o R Börjessons foundation (OW, Uppsala), and National Science Foundation DEB-1355059 (BB, VCU).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Rosemary Dokos and her colleagues at ONT for promptly responding to questions and ONT for providing the flow cells used in these experiments free of charge.

Supplementary materials

Supplementary File 1. Laboratories. List of laboratories that generated data for this study.

Click here to access the data.

Supplementary File 2. Experimental constants. Table of metadata fields and values shared across experiments.

Click here to access the data.

Supplementary File 3. NanoOK experiment reports. NanoOK PDF reports for three sets of reads (pass only, fail only, and both pass and fail) for each experiment.

Click here to access the data.

References

1. Ip CL, Loose M, Tyson JR, *et al.*: **MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved]** *F1000Res.* 2015; **4**: 1075.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Loman NJ, Quinlan AR: **Poretools: a toolkit for analyzing nanopore sequence data.** *Bioinformatics.* 2014; **30**(23): 3399–3401.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** *arXiv [q-bio.GN].* 2013.
Reference Source

4. Kiełbasa SM, Wan R, Sato K, *et al.*: **Adaptive seeds tame genomic sequence comparison.** *Genome Res.* 2011; **21**(3): 487–93.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Quick J, Quinlan AR, Loman NJ: **A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer.** *Gigascience.* 2014; **3**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Jain M, Fiddes IT, Miga KH, *et al.*: **Improved data analysis for the MinION nanopore sequencer.** *Nat Methods.* 2015; **12**(4): 351–6.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Loose M: **minoTour - a platform for real-time analysis and management of Oxford Nanopore minION reads.** 2014.
Publisher Full Text

8. Leggett RM, Heavens D, Caccamo M, *et al.*: **NanoOK: multi-reference alignment analysis of nanopore sequencing data, quality and error profiles.** *Bioinformatics.* 2016; **32**(1): 142–144.
PubMed Abstract | Publisher Full Text | Free Full Text

9. **MARC Phase 2 analysis documentation and scripts.**
Reference Source

10. Jain M, Tyson JR, Loose M, *et al.*: **Code used in analysis titled "MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry".** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ? ✓ ?

Version 1

Referee Report 28 July 2017

doi:10.5256/f1000research.12257.r23369

?

Titus Brown [1], **Lisa Cohen** [2]

[1] Department of Population Health and Reproduction, University of California, Davis, Davis, CA, USA

[2] Department of Population Health and Reproduction, University of California, Davis, CA, USA

This is a followup data release and analysis by the MinION Analysis and Reference Consortium (MARC) to the Phase 1 release by Ip et al. in 2015 (https://f1000research.com/articles/4-1075/v1), who compared the consistency, rate, volume and quality of E. coli K-12 data produced by 5 labs from 2 R7.3 flow cell runs each. In this Phase 2 data release and analysis, the MARC characterized the throughput, read quality, and accuracy of one run each of 1D and 2D library preps of R9.0 chemistry.

Aside from a fundamental issue with this study that it would have been nice to have data from more than 2 labs, and more than 1 flowcell run of each 1D and 2D (only 4 total flowcells were used in this study), there were several really nice features:

Great idea to compare!
"Sequencing was performed using both the 2-dimensional (“2D”) “ligation” kit and the newer 1D “rapid” kit."

Excellent, thank you for providing these data!
"It is a resource to aid further developments in nanopore informatics as well as the development of biological applications using the MinION."

Cool!
"overall, the R9.0 chemistry showed an increase in data yield and read length when
compared with a typical Phase 1 R7.3 experiment"

A few criticisms and questions -

1. Would have liked more robust comparison and discussion on differences between 2D and 1D sequencing since the consensus in the community seems to be now that 1D sequencing libraries are fine (nobody uses 2D anymore). "The higher total error of 15.3% for “pass” template base-calls, compared with 7.5% for “pass” 2D base-calls (Table 4), is an acceptable trade off."

2. ENA accessions could be more clear, such as Table S10 from MARC Phase 1 paper (Ip et al. 2015).
   * These data are clearly generated by experts (some of whom are long-term experts and paid consultants supported by ONT), with available pore numbers and sequencing yields representing

best case scenarios. While perhaps beyond the scope of this benchmark, it would be nice to see similar data comparisons by novice labs trying to figure this technology out.

3. Why are 1D and 2D library preparation modifications made in this study not part of standard ONT protocols? What was reasoning behind making these changes? One of the hardest parts of figuring out ONT is troubleshooting the little modifications like the ones mentioned in this study. Modifications indicated in the manuscript: genomic DNA was sheared to ~10 kb and 0.4x AMPureXP cleanup treatment. And 1D: 0.4x AMPureXP cleanup prior to prep, unsheared DNA input of 400ng, 0.4ul blunt/TA ligase; 10 min incubation used in final step.

4. This might be obvious, but I'm not sure: why was the lambda control DNA not included in the 1D runs?

5. Why does the % of active pores decrease from g1 to g2? It is difficult to compare the percentage of active pores between flowcells since, as the manuscript states, the computer from Lab7 crashed in the middle of the experiment and these numbers were not available. And there were only 4 flowcells used in this study. What are some of the reasons why the number of active pores fluctuates between flow cells?

6. Were these spot on flowcells? This feature was added recently sometime after R9.0 was released, and I am curious what effect this had on sequencing yield.

7. Perhaps include data analyses software versions?
   "we computed alignments and error-rate measurements using BWA-MEM and LAST, followed by re-alignment using marginAlign. Real-time evaluation of the runs was performed by minoTour (more information available from: http://minotour.github.io/minoTour), run locally at the two experimental laboratories. The “pass” and “fail” reads from each experiment were evaluated with NanoOK (version 0.95) using bwa alignments. Additional metrics and analyses were performed with bespoke Python and R scripts, (available at https://github.com/camilla-ip/marcp2).

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Referee Report 27 July 2017

doi:10.5256/f1000research.12257.r24143

✓ **Martin C. Frith** [1,2,3]

[1] Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

[2] Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

[3] Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

This study tests version R9.0 of the MinION nanopore sequencer, and describes its error rate, read lengths, throughput, and other characteristics. The article's timing is unfortunate, because I believe R9.0 had already been superceded at the time of publication, but this does not affect the study's soundness. As far as I can tell, this study is basically sound, but there are some careless mistakes:

- This text does not match Table 2: maybe pass -> total? "64–111 K 2D pass reads (compared with 21 K for the typical R7.3 experiment)".
- Is this really an "increase"? "increase in N50 read length to 7.3–9.1 Kb for all 2D reads in the R9.0 experiments (compared with 7.4 Kb". The abstract says: "The 2-dimensional (“2D”) N50 read length was unchanged".
- This text does not match Table 3, maybe pass/fail -> total/pass? "mean template length for “pass” reads increasing to 7.2–8.6 Kb (from 5.0 for R7.3) and increasing to 7.8–9.4 for “fail” reads (from 6.2":
- This text also does not match Table 3: "Read length N50 increased to 13.1–15.7 Kb for “pass” reads (compared with 6.9 Kb for the typical R7.3) and 13.6–16.2 Kb"
- This is true only for 1D reads: "For template reads from both 1D and 2D experiments, 99.9% of “pass” reads were alignable from both 1D and 2D".
- 88% is not similar to 78%: "The median identity of reads from 1D and 2D experiments (Table 4) was similar... The median identity for 1D template reads was ~88% and ~76%, for “pass” and “fail”, respectively (compared with 78% and 75%".
- This is not comparing like with like (R9.0 template versus R7.3 2D): "For the 2D experiments, the read identity was ~89% and ~85%, for “pass” and “fail”, respectively (compared with ~92% and ~82%, respectively, for the typical R7.3 experiment)."

A few things should be clarified:

- Why do "Identity %" and "Total error %" not sum to 100?

- Fig 2:
  - what is the difference between Q-score and BQ?
  - what is "(Temp)"?
  - what is the difference between "GC" and "GC (Temp)"?
  - what is "throughput": count of what per what?

- What is "1D consensus follow through"?

- What does "bridging experiment" mean?

- What is a "run script"?

Other minor comments:

- The title should be shortened to something like "Analysis of MinION R9.0 chemistry". The rest is not scientifically meaningful, and might be perceived as "appeal to authority".

- The abstract "methods" section is incorrectly brief.

- The abstract "conclusions" section should probably not say "new" R9.0 chemistry.

- Is this really "higher"? "higher at... 6.0–6.5 Kb (compared with 6.0 Kb".

- Page 4: "were statistics computed" -> "statistics were computed".

- The LAST usage is likely suboptimal (though I guess it matters little here). The currently-recommended usage has been here since 2016-11-22: https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md

Data availability:

I found it excessively hard to obtain the data. The PRJEB18053 link leads to three "component projects", each of which has numerous files. Which file is which dataset? This should be better organized, or at least described. The best I could do was to mouse-over the links: a name like "Nott_R9_run2_1D_pass_f74a133aa1ac903384a928a51051582db2cc412b_0.fastq" gives me a clue, but a name like "ERR2025969.fastq" is hopeless.

For example, what is the difference between these files?
Nott_R9_run2_1D.pass.1D.fastq
Nott_R9_run2_1D_pass_f74a133aa1ac903384a928a51051582db2cc412b_0.fastq

Suggestions for future studies of this type:

- Characterize the substitution errors further, e.g. is A->G more frequent?

- Are the insertions and deletions long-and-rare, or short-and-numerous?

- Are the base quality scores accurate/useful? (And what do they even mean when indels are the main error?)
- Characterize context-dependence of errors, e.g. homopolymers, CCXGG context (http://www.biorxiv.org/content/early/2017/06/29/157040).
- Can rearrangement errors be characterized? Long reads are promising for finding rearrangements (e.g. inversions, translocations), but do artifactual rearrangements occur?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Computational biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 28 June 2017

doi:10.5256/f1000research.12257.r23414

? **Wigard P. Kloosterman** , **Jose Espejo Valle Inclan** , **Mircea Cretu Stancu**
Department of Genetics, University Medical Center Utrecht, Utrecht, Netherlands

In this work the Minion Analysis and Reference Consortium describes the analysis of Oxford Nanopore sequencing data generated from *E.coli* using the R9.0 sequencing chemistry. The R9.0 data characteristics were benchmarked against previous R7.3 data. The work is of much interest to (new) users of the Oxford Nanopore sequencing technology, as it provides a realistic overview of how the technology has developed over the course of 2016 and what can be expected in terms of data throughput and quality.

I have the following remarks:

Major point

- Oxford Nanopore sequencing technology has developed rapidly over the last year. Yet, the data presented in the paper are derived from older R9.0 and R7.3 chemistries. The value of the data analysis and comparisons would much increase if one or two runs of the more recent R9.4 data will be added as an extra column to each of the plots/tables. R9.4 chemistry is mentioned several times in the manuscript, but unfortunately no data are shown.

Other points

- Materials & methods:
    - The authors mention that DNA extraction procedures are described in MARC Phase 1, but no reference is given. Instead, the authors do refer to Supplementary File 1, but this file only contains a list of affiliations. I would suggest that the authors provide the appropriate reference here and/or refer to a Supplementary file that describes the DNA extraction procedures (or add this in the methods).
    - 1D/2D library preparation: The authors list some modifications with respect to ONT protocols. For the more ignorant reader, the authors could spell out why these modifications were added.
    - Sentence page 4: “Both alignments were then improved with marginAlign.” What does ‘improved’ mean in this case? If specific marginAlign settings were used, then these should be listed as well.

- Results:
    - Base yield and read lengths:
        - There appear some inconsistencies regarding claims about read lengths for R7.3 vs R9.0. The abstract states that read length N50 was not different, while page 4 states that “R9.0 chemistry showed an increase in [...] read length” compare to R7.3. Are differences in yield and read lengths statistically significant between R7.3 and R9.0?
        - Related to this: the authors mention that median read length is longer for R9.0 compared to R7.3 (6.6kb - 7.7kb and 6.4kb), yet they mention that the maximum read length is comparable (50.9kb and 59.7 kb). Why are median values regarded as different, while maximum values are regarded as comparable?

    - Base quality:
        - “The proportion of “pass” reads with a Q-value threshold of 9 was 66% to 72%, about the same as that observed for the typical R7.3 experiment, with a base quality threshold of 9.0.” Is Q-value equivalent to base quality value here? Or do the authors mean Q-value instead of base quality? A similar statement is made later in this paragraph.

    - Alignment identity and accuracy:
        - From the Table, it appears that R9.0 “fail” reads are worse than R7 failed reads. Could the authors comment why this is the case?
        - Page 7, second column: There appears to be a mistake in the given read identities for 2D R9.0 experiments (89% and 85% given, while these numbers appear for R9.0 template reads; should be ~94% and ~70%).
        - The authors make a point about read quality and mention that the longest subsequence that perfectly aligns increases around 2-4 times for 1D runs, going from R7.3 to R9. Does this mean that the errors are less randomly distributed in R9 data,

given that the median percent identity does not change substantially? The authors could improve this analysis, by evaluating the randomness of the error distribution within reads, or across the genome and how this relates to genome sequence context.

- Performance over time (Figure 2):
  - Final sentence of results: “The current report is for the initial R9.0 chemistry, and the GC-bias seems to be less pronounced with the improved version of the R9 pore (R9.4 data not shown).” It would be better if the authors draw a clear conclusion whether this bias is present or not, and include data to support this.
  - Page 9: the authors mention that GC content differs for different run and read types. It would be good if the authors quantify these differences and provide the numbers in the text.
  - Figure 2: What does count mean here? Read counts or event counts?
  - Figure 2 legend: Read length is referred to as ‘kb’, but probably the authors mean ‘b’ (looking at the y-axis of the length plot).
  - Page 8: “the quantity and quality of data produced during an experiment varied as material passed from one side of the membrane to the other.” What does this mean exactly? This could be replaced by a more precise statement.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Genomics, bioinformatics

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Discuss this Article

Version 1

Reader Comment 07 Jun 2017

Wouter De Coster, VIB Department of Molecular Genetics Antwerp, Belgium

Hi,

I noticed the link to the scripts (https://github.com/camilla-ip/marcp2) is no longer up to date.

Cheers,
Wouter

Competing Interests: No competing interests were disclosed.