RESEARCH ARTICLE

REVISED Investigation of chimeric reads using the MinION [version 2; referees: 2 approved]

Ruby White, Christophe Pellefigues, Franca Ronchese, Olivier Lamiable, David Eccles

Malaghan Institute of Medical Research, Wellington, 6242, New Zealand

v2 **First published:** 05 May 2017, **6**:631 (doi: 10.12688/f1000research.11547.1)
Latest published: 16 Aug 2017, **6**:631 (doi: 10.12688/f1000research.11547.2)

Abstract

Following a nanopore sequencing run of PCR products of three amplicons less than 1kb, an abundance of reads failed quality control due to template/complement mismatch. A BLAST search demonstrated that some of the failed reads mapped to two different genes -- an unexpected observation, given that PCR was carried out separately for each amplicon. A further investigation was carried out specifically to search for chimeric reads, using separate barcodes for each amplicon and trying two different ligation methods prior to sample loading. Despite the separation of ligation products, chimeric reads formed from different amplicons were still observed in the base-called sequence. The long-read nature of nanopore sequencing presents an effective tool for the discovery and filtering of chimeric reads. We have found that at least 1.7% of reads prepared using the Nanopore LSK002 2D Ligation Kit include post-amplification chimeric elements. This finding has potential implications for other amplicon sequencing technologies, as the process is unlikely to be specific to the sample preparation used for nanopore sequencing.

This article is included in the Nanopore Analysis gateway.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	Invited Referees 2
REVISED version 2 published 16 Aug 2017		✓ report
version 1 published 05 May 2017	✓ report	? report

1 **Keith E. Robison**, Warp Drive Bio, USA

2 **Winston Timp**, Johns Hopkins University, USA

Discuss this article

Comments (0)

Corresponding author: David Eccles (bioinformatics@gringene.org)

Author roles: White R: Investigation, Methodology, Writing – Review & Editing; **Pellefigues C**: Investigation, Writing – Review & Editing; **Ronchese F**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Lamiable O**: Conceptualization, Data Curation, Investigation, Methodology, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Eccles D**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: The R9.4 flow cell and sequencing kit (SQK-LSK208) used for this experiment were provided free of charge by ONT as replacements for a purchased kit and flow cell where the phenomena of chimeric reads was initially discovered. ONT provided advice regarding the sample preparation protocols, including the suggestion of a slow overnight ligation step.

How to cite this article: White R, Pellefigues C, Ronchese F *et al.* **Investigation of chimeric reads using the MinION [version 2; referees: 2 approved]** *F1000Research* 2017, **6**:631 (doi: 10.12688/f1000research.11547.2)

Grant information: This work was funded in full by Health Research Council of NZ Independent Research Organization (IRO) funding to the Malaghan Institute of Medical Research (grant number HRC14/1003).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 05 May 2017, **6**:631 (doi: 10.12688/f1000research.11547.1)

REVISED **Amendments from Version 1**

We have added additional sub-diagrams in Figure 4–Figure 6 to indicate likely structure. We have revised the text to improve grammar and consistency. Also, we added additional discussion about hairpin detection with nanopore basecallers.

See referee reports

Introduction

High-throughput DNA sequencing is a rapidly evolving field with new methods and applications introduced almost weekly[1]. One of the most recent sequencing technologies available on the market is the MinION sequencing device from Oxford Nanopore Technologies (ONT)[2]. A brief overview of MinION sequencing technology is discussed in our previous study on mitochondrial genome assembly[3].

Instead of exploiting base-pairing as in the sequencing-by-synthesis approach used by Illumina and others, nanopore sequencing uses an electronic sensor to detect DNA via a change in electric current (reviewed in 4). The MinION's flow cell is comprised of 2048 wells containing a membrane perforated by nanopores. Ligated with a molecular motor, a single stranded DNA molecule passes through the pore, altering the recorded current. After the electronic sequencing is carried out, a software base-calling algorithm transforms the current trace into a modelled DNA sequence. The advantages of the MinION are rapid library preparation, portability[5,6], long molecule sequencing[7], and sequencing of non-model modifications of the DNA strand[8]. Recent improvements in the chemistry of the MinION have overcome the majority of issues associated with low yield and high error rates that have limited the range of its application. The MinION sequencing device has now been successfully used to sequence genomes of a wide range of sizes, from bacterial and viral genomes[9,10], amplicon sequencing such as bacterial 16S rRNA sequencing[11], and more recently a human genome[12]. The MinION has also been used for cDNA sequencing[13], for detecting DNA methylation patterns without chemical treatment[8,14], and for direct RNA sequencing with detection of modified 16S rRNA nucleotides[15].

Using R9.4 flow cells we have evaluated the MinION technology for the amplicon sequencing of highly similar genes. Since we have an interest in the interferon response during helminth infection[16], we sequenced the type I interferon (IFN) family. Type I IFNs are a family of intronless antiviral response genes comprised, in mice, of 14 highly homologous *Ifna* members, as well as the genes *Ifnb*, *Ifnk* and *Ifne*[17]. In humans, sequence similarity across the 14 members of the *Ifna* genes is 70–80%, with a further 35% sequence similarity between *Ifna* and *Ifnb*. Type I IFN has both an important role in innate antiviral immunity and in mounting adaptive T helper cell responses[16,18]. Building on previous observations, we aimed to identify which type I IFN member(s) were responsible for driving the type I IFN signalling in our infection model.

Due to the high homology between the *Ifna* family genes, accurately detecting quantitative expression of the different gene members by Sanger sequencing or next generation sequencing is difficult. We instead employed nanopore sequencing, which allowed us to acquire full-length reads from each individual sequence that were amplified by the PCR reaction. We aimed to determine the relative quantities of the various *Ifna* family and *Ifnb* transcripts, in helminth-treated mouse ear tissue using the MinION; therefore enabling both the differentiation between the various *Ifna* genes, and the potential to perform quantitative analysis.

Methods

Nippostrongylus brasiliensis was originally sourced from Lindsey Dent of the University of Adelaide, South Australia and has been maintained for 22 years by serial passage at the Malaghan Institute. Female Lewis rats were bred and used for maintenance of the *N. brasiliensis* life cycle when 4 months of age (and weight over 150g), as outlined in Camberis *et al.*[19].

Two 8-week-old C57BL6/J male mice (Jackson Laboratories, approx 23g), housed and bred at the MIMR under specific pathogen free conditions respecting the local and New Zealand ethic guidelines, were chosen for the investigation. 300 dead infective *N. brasiliensis* L3 larvae were injected intradermally in each ear of one mouse in 30uL PBS after anaesthesia with an intraperitoneal injection of 200uL ketamine/xylazine. The other mouse was similarly euthanised and injected intradermally in each ear with 30uL PBS. The mice were euthanised in a CO_2 chamber 3h post injection and ears (approx 27–30mg in weight) were immediately harvested and conserved in RNALater at 4C for <1h. RNA extraction of each whole ear was done in 1mL of Trizol following the products' guidelines (ThermoFisher). cDNA was synthesised using the High Capacity RNA-to-cDNA kit (Applied Biosystems), according to the manufacturer's instructions. Only the cDNA from the *N. brasiliensis*-treated mouse was used for this investigation.

Ifna, *Ifnb*, and *Actb* amplicons were generated using specific primers: *IfnaF* (ATGGCTAGRCTCTGTGCTTTCCT) and *IfnaR* (AGGGCTCTCCAGAYTTCTGCTCTG)[20]; *IfnbF* (CTGGCTTCCATCATGAACAA) and *IfnbR* (GCAACCACCACTCATTCTGA); and *ActbF* (AGGGAAATCGTGCGTGACAT) and *ActbR* (ACGCAGCTCAGTAACAGTCC). PCR amplification was performed using Phusion High-Fidelity PCR Kit (Thermo Scientific), see Figure 1. PCR products were cleaned using QIAquick PCR Purification Kit (QIA-GEN) and verified by gel electrophoresis.

Ifna cDNA were amplified by PCR using primers designed across a highly-conserved region of all *Ifna* coding sequences, which resulted in a mixed PCR product containing all 14 *Ifna* genes. cDNAs of *Ifnb* and *Actb* were amplified separately and used as quantification controls. Altogether, the three pooled amplicons were loaded into a flow cell and sequenced. Among the reads that we obtained, we noticed long chimeric reads comprising of two or more sequences from different amplicons. We decided to further examine this phenomenon.

Ethics approval for maintenance of the *N. brasiliensis* life cycle is overseen and approved by the Victoria University of Wellington Animal Ethics Committee. C57BL/6J mice were originally obtained

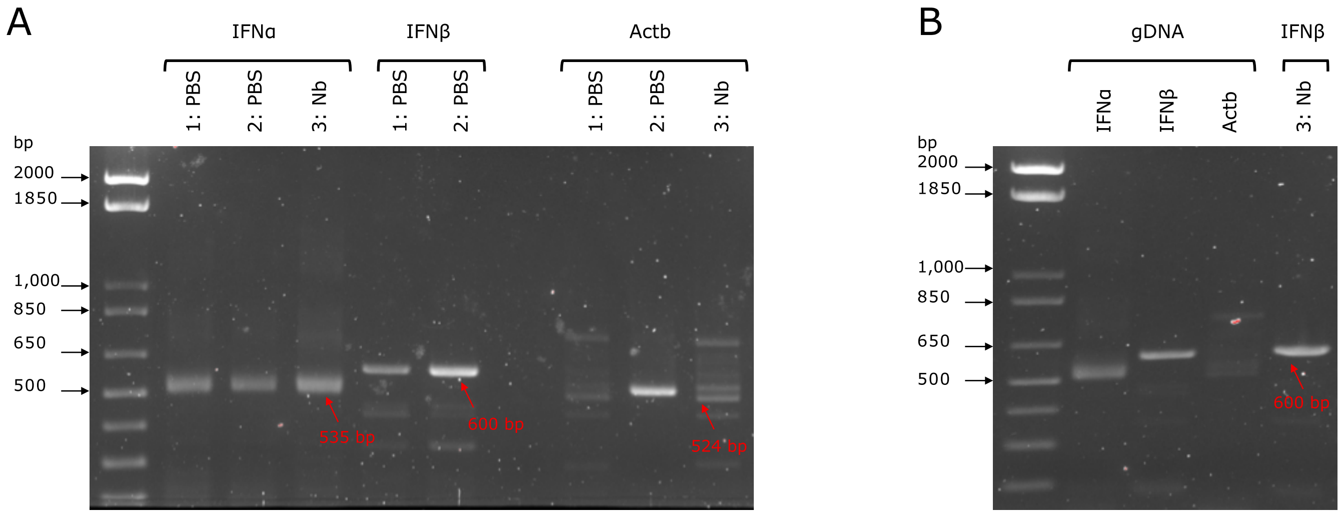

Figure 1. Gel-electrophoresis image of PCR products amplified for this investigation. (**A**) Amplicons were observed for Ifna and Actb from both PBS treated (1&2, not sequenced in this investigation) and *Nb*-treated (3) samples at the expected sizes of 535 bp, and 524 bp respectively. The Ifnb gene from the *Nb*-treated sample (3) failed to amplify during this first attempt. (**B**) A repeat amplification of Ifnb from *Nb*-treated sample was carried out, producing a single band of approximately 600bp. This was run alongside amplicons of Ifna, Ifnb and Actb from genomic DNA; however genomic amplicons were not used for subsequent MinION sequencing.

from The Jackson Laboratory, Bar Harbour, Maine, USA, and maintained at the Biomedical Researc Unit of the Malaghan Institute of Medical Research by brother X sister mating. Breeding pairs were refreshed regularly to maintain the genetic integrity of the strain. Mice were maintained in specific pathogen-free conditions and all mouse experiments were approved by the Victoria University Animal Ethics Committee (permit number 23907) and carried out according to institutional guidelines.

Library preparation

The ONT Native Barcoding Kit (EXP-NBD002) and 2D Ligation Sequencing Kit (SQK-LSK208) were used to prepare the samples for sequencing, as per the manufacturer's protocol. Briefly, purified PCR amplicon products were blunt-ended, ligated with barcode sequences, pooled in approximately equimolar amounts, then ligated with flow cell adapters and a hairpin linker. In order to explore the effect of ligation method on the degree of chimerism, two different adapter/hairpin ligation reactions were carried out: one using the standard quick (10-minute) ligation, and the other using an overnight ligation at 4° Celsius. No additional adapter-free controls were used; it has been our prior experience that sequencing does not proceed in a callable fashion unless adapter sequences are present. The barcoding scheme used in the library preparation is shown in Figure 2. Samples were quantified after barcoding for overnight ligation (2.14 *ng/µl*, 2.54 *ng/µl* and 2.56 *ng/µl* for *Ifna*, *Ifnb*, and *Actb* respectively) and for quick ligation (2.13 *ng/µl*, 2.68 *ng/µl* and 2.45 *ng/µl* for *Ifna*, *Ifnb*, and *Actb* respectively). These samples were normalised and pooled together to give 26.6ng each in 33.1*µl* distilled water for ligation. After adapter ligation, the quick ligation method showed no detectable nucleic acid, as seen using a fluorescence quantitation with the Quantus fluorometer (Promega), while the overnight ligation quantified at 0.239ng/*µl*. We decided to pool the samples together anyway, and were pleasantly surprised to discover a substantial proportion of reads from quick-barcoded sequences.

Base-calling

Reads were initially base-called during the sequencing runs in January 2017 using Metrichor 2D basecalling, from MinKNOW v1.3.25. An initial analysis of called reads demonstrated substantial disagreement between base-calls and the raw signal (e.g. hairpin adapter sequences matching multiple times when the signal showed only one present), so reads were recalled as in March 2017 using Albacore v0.7.5.

Results and discussion

During the initial MinION sequencing run to investigate the expression of *Ifna*-family members in mice (comparing with *Ifnb* and *Actb* transcripts), we encountered issues with 2D base-calling through the Metrichor web service, which seemed to be due to failed alignment of component 1D strands. A BLAST search on some of the longest base-called 1D reads led to a discovery that some reads had multiple mappings to our target *Ifna*-family members. Further exploration of the data demonstrated a situation in which both *Ifna* and *Actb* sequences were present in the same read (see Figure 3). This was an unexpected result; we had carried out separate PCR reactions for each transcript, so were not expecting reads to appear that mapped to different transcripts. Our conclusion was that chimeric ligation of input DNA was occurring at some stage during the sample preparation process, but all we were able to determine at the time was that this chimerism was happening some time after the PCR, but before the sequencing. The present experiment was designed in light of these prior results to more easily quantify the degree of ligation that was happening.

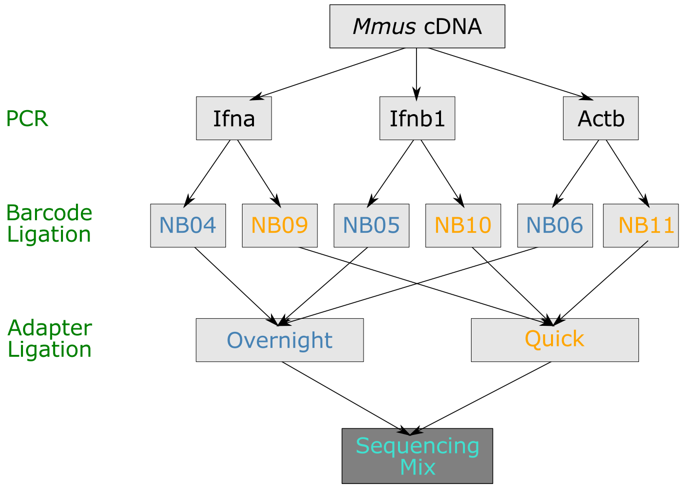

Figure 2. Sample preparation workflow demonstrating the steps used to aid in the identification of the stage at which chimeric reads were formed. Mouse cDNA was extracted and separately amplified for three different amplicons. The amplified product was then separated and barcoded based on the intended ligation process. Barcoded products were pooled and ligated to adapters via the overnight or the quick ligation method, then finally pooled together for sequencing.

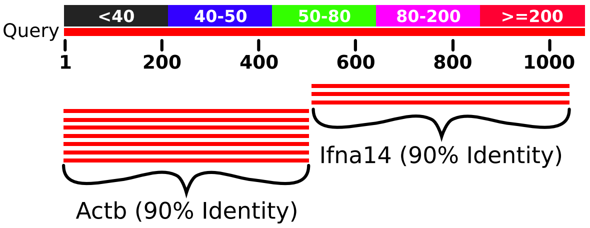

Figure 3. A chimeric read that was discovered during the preliminary investigation of interferon expression. This read mapped to both beta-actin and interferon alpha, suggesting that a ligation of sequence had occurred, either during sample preparation or *in-silico*.

Read counts

Despite using a 2D ligation chemistry in the sample preparation, and selecting out hairpin-containing reads using streptavidin beads, the majority of reads could not be called as an aligned 2D sequence: of 329,591 sequenced reads, 299,124 were base-called by Albacore, and 1005 (0.3%) of these base-called reads had an aligned 2D sequence (see Supplementary File 1). Any called reads that were not called as 2D were processed further as 1D sequence, i.e. the remaining 298,119 (99.7%) of called reads.

Discussions with ONT staff, in particular Forrest Brennen, during the London Calling conference in 2017 provided insight into what had caused the failure in 2D base-calls. Oxford Nanopore Technologies introduced a chemistry upgrade for their 2D ligation sequencing kits that produced a different, and more obvious, hairpin signal with three peaks rather than two. This modified hairpin signal was the one that the Metrichor and Albacore base-callers were looking for in January 2017 and March 2017 respectively. However, the 2D barcoding kit that we used still had the old hairpin adapter included, and this meant that the base-callers ignored the hairpin region and attempted to call the entire sequence as a 1D read. Oxford Nanopore Technologies subsequently updated their Albacore base-caller to correct this error for 2D barcoded reads, but due to discontinuing the 2D chemistry in preference to the faster and more accurate $1D^2$ chemistry, the 2D base-caller is no longer developed or included in Albacore. We were able to obtain from ONT the latest, and only, Albacore version that included this fix (version 1.2.4), and recalling reads showed substantial improvement in detecting 2D sequence: 40.8% of reads were called as 2D reads, which was much closer to the 48.6% of reads that we found with a detectable hairpin adapter in the 1D base-called sequence.

Read mapping

Called 1D reads were mapped to *Actb*, *Ifnb1*, an *Ifna* consensus sequence, additional interferon sequences, the ONT control strand sequence, and known ONT adapter sequences (see Supplementary File 2) using LAST v833[21]. A total of 261,183 reads (87.6% of called 1D reads) were discovered that mapped to at least one known amplicon and/or barcode sequence.

Categorisation of chimeric reads

Using a process of elimination, a total of 4563 reads (1.7% of amplicon or barcode-mappable 1D reads) were discovered with base-called sequences that were definitively chimeric (see Supplementary File 5). These reads mapped at least once to either one of the three amplicon sequences, or at least once to one of the six barcode sequences. These were broken into four categories (with some overlap) based on the observed combinations of barcode and amplicon sequences (see Figure 4):

1. Repeated identical amplicons aligned in the same direction
2. At least two distinct amplicons
3. At least two distinct barcode
4. Disagreement between barcode and amplicon

A more complete count of different categories of chimerism (for those observed at least five times) can be found in Table 1. The highest proportion of chimeric reads were associated with repeated identical amplicons, with 3441 reads seen (75% of all definitively chimeric reads). This suggests that an amplicon sequencing procedure will be particularly susceptible to read chimerism, as the same sequence will appear in increased abundance compared to an untargeted sequencing approach. One potential mechanism for this is that the identical sequences encourage the formation of complex base-pairing structures (e.g. quadruplexes) that bring the ends of similar sequences closer to each other. The low-temperature overnight ligation resulted in a much higher proportion of repeated amplicons than the quick ligation; in this case it appears that the quick ligation was better at reducing the occurrence of chimeric reads, despite prior expectations.

Of the definitively chimeric reads, 2869 included at least one overnight barcode (1.8% of 159,188 amplicon-mapped reads with an overnight barcode), and 1203 included at least one quick barcode (2.6% of 45,850 amplicon-mapped reads with a quick barcode). While it appears that the use of overnight ligation has helped somewhat to reduce chimeric reads, a substantial proportion of chimeric reads still remain.

If a cassette of adjacent *Ifna* genes were transcribed together, it is possible that this cassette could be amplified together as a single sequence. These sequences would appear to be chimeric (and fall into the “Repeated amplicons” category), but would not have any intermediate barcodes. The count similarities for repeated *Ifna*, *Ifnb1* and *Actb* genes suggest that this cassette amplification is not happening at any significant level.

Categorisation of non-chimeric reads

After elimination of definitively chimeric reads, 256,620 reads remained that appeared to map uniquely to single sequences (see Figure 5). A small proportion of these sequences (14,223; or

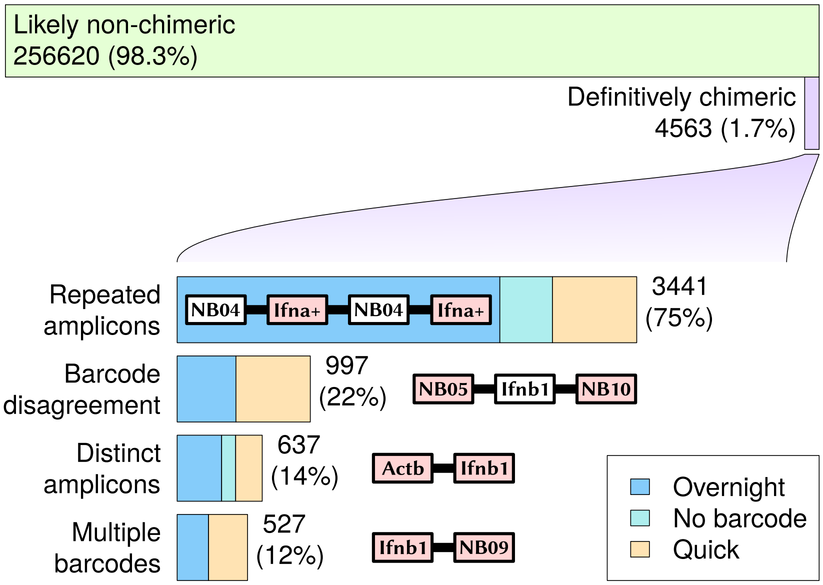

Figure 4. Definitively chimeric reads mapped during the sequencing run. Chimeric read categories are not disjointed: different categories may intersect with each other. Reads that mapped to repeated identical, but reverse-complemented sequences, are not included in these chimeric results, as it was not possible to distinguish at the base sequence level between such a duplicated sequence fragment and a 2D read with hairpin.

Table 1. Chimeric read counts split into categories depending on the number of amplicons and barcodes seen. Only categories with a count of 5 or more are displayed.

Amplicon count			Barcodes seen			Read count
Ifna	Ifnb1	Actb	overnight	quick	disagreement	
0	0	2	○			876
2	0	0	○			803
0	2	0	○			704
2	0	0		○		378
1	0	0		○	○	246
0	0	2		○		224
2	0	0				201
1	0	1	○		○	140
0	2	0				125
1	0	1		○	○	108
1	1	0	○		○	99
0	1	1	○		○	79
0	0	2				67
0	1	0		○	○	64
1	1	0		○	○	48
0	0	1	○		○	42
1	1	0				41
0	0	1		○	○	37
1	0	1				36
0	1	0	○		○	34
0	1	1		○	○	31
1	0	0	○		○	27
0	1	1				25
0	0	0		○		19
3	0	0	○			15
0	2	0		○		15
2	0	0		○	○	7
0	0	3	○			7
0	3	0	○			6
0	0	0	○			5

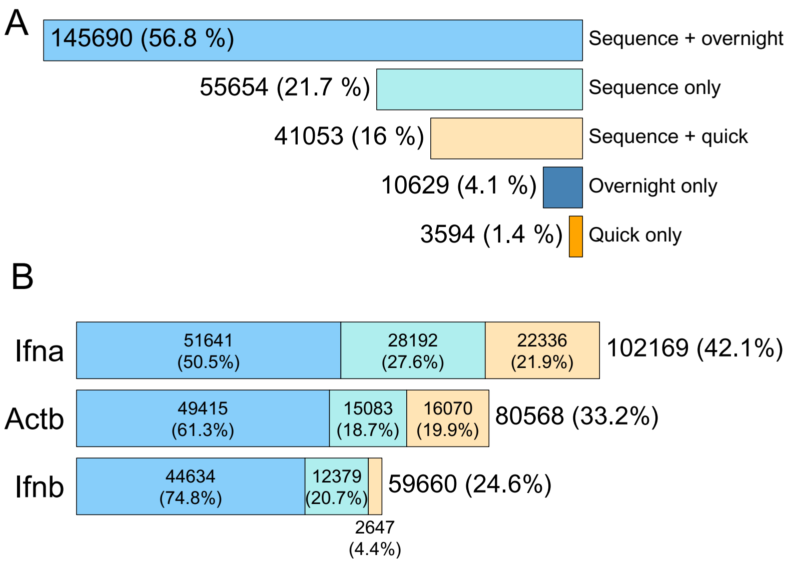

Figure 5. Amplicons mapped from basecalled non-chimeric reads. (**A**) Amplicon counts split by barcode type. (**B**) Sequence only, quick barcode, and overnight barcode counts for amplicon-mapped sequences.

5.5%) had detectable barcode sequences, but did not map to any amplicons (i.e. mappable to an overnight or quick barcode sequence only). It is expected that these unmapped barcoded sequences were unamplified mouse cDNA sequences.

A difference in read counts was observed between overnight-barcoded sequences and quick-barcoded sequences (77.8% overnight, 22.2% quick), which was consistent with the difference in input amount observed during sample preparation. An attempt was made during sample preparation to add in the three different amplicon preparations in equimolar quantities, which was more successful for the *Actb* preparation (33.6%) than it was for the *Ifna* and *Ifnb* preparations (42.7% and 23.7%, respectively).

An additional categorisation of *Ifna* family members (see Supplementary File 3) was attempted, but is not presented here as it detracts from the main chimeric read investigation. Intermediate results and a processing script from this categorisation are available in verbose form as Supplementary File 4.

Read signal confirmation of chimerism

A few of the reads were investigated at the raw signal level to make sure that the electrical trace was in agreement with the base-called signal. A demonstrative signal trace for a non-chimeric 2D read comprising of a single barcode-adapted amplicon is shown in Figure 6. Read traces typically began with a high-current (but relatively uniform) open pore state, followed by an intermediate stall signal (also fairly uniform), after which the highly variable sequence trace begins. Hairpin adapters could be easily identified in the raw signal as a bridge structure a little over halfway through a 2D sequence.

A number of situations were observed in the base-called sequence where ligation during sample prep seems to have occurred, and in some cases this ligation resulted in multiple hairpin adapters being ligated in the same sequence. One such occurrence of this is seen in Figure 7, where two barcoded overnight sequences from two different amplicons (*Ifnb1* and *Ifna2*) were joined together. Because two amplicons were concatenated, this ligation must have happened

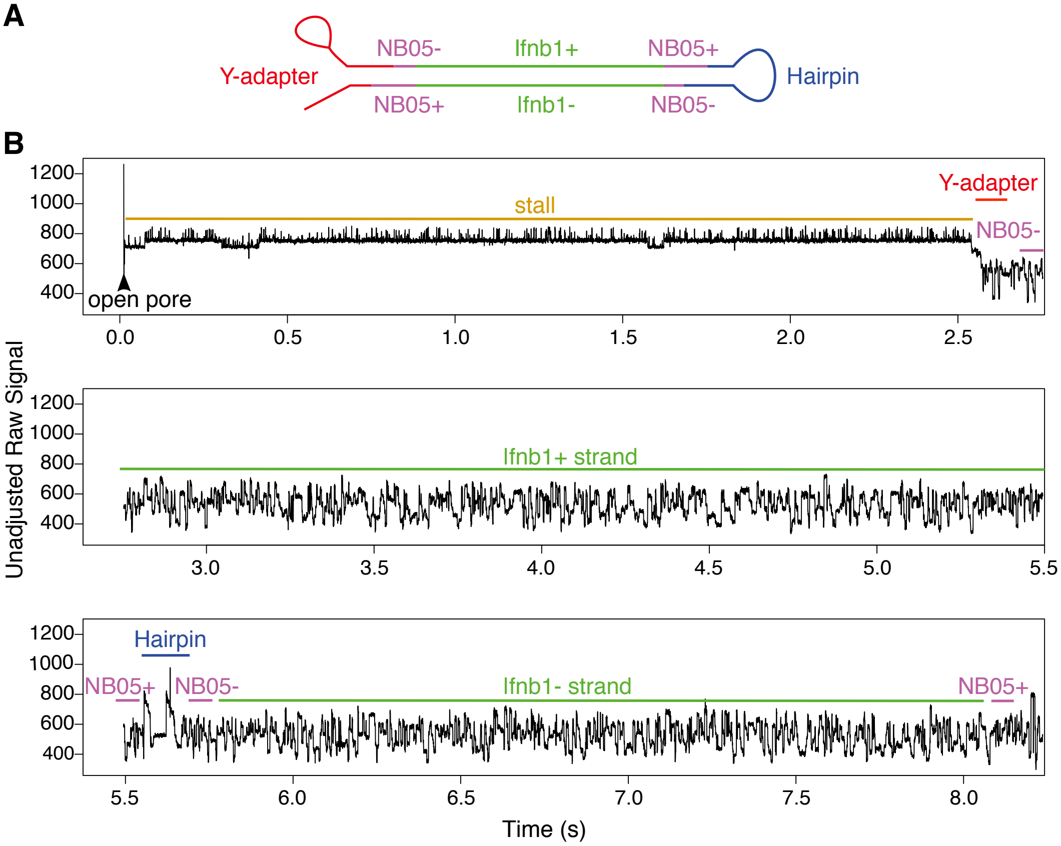

Figure 6. Demonstrative raw signal for a non-chimeric read (from an *Ifnb1* amplicon). The recorded signal for this read starts with a very long period of 7s in the open pore state, followed by a short stall of 0.3s, then a coding *Ifnb1* sequence that took 2.5s to transition through the pore, then an NB05-flanked non-coding *Ifnb1* sequence that took 2s seconds to transition through the pore. *Note: These figures have been annotated with approximate region boundaries based on the order of hits to the base-called sequence.*

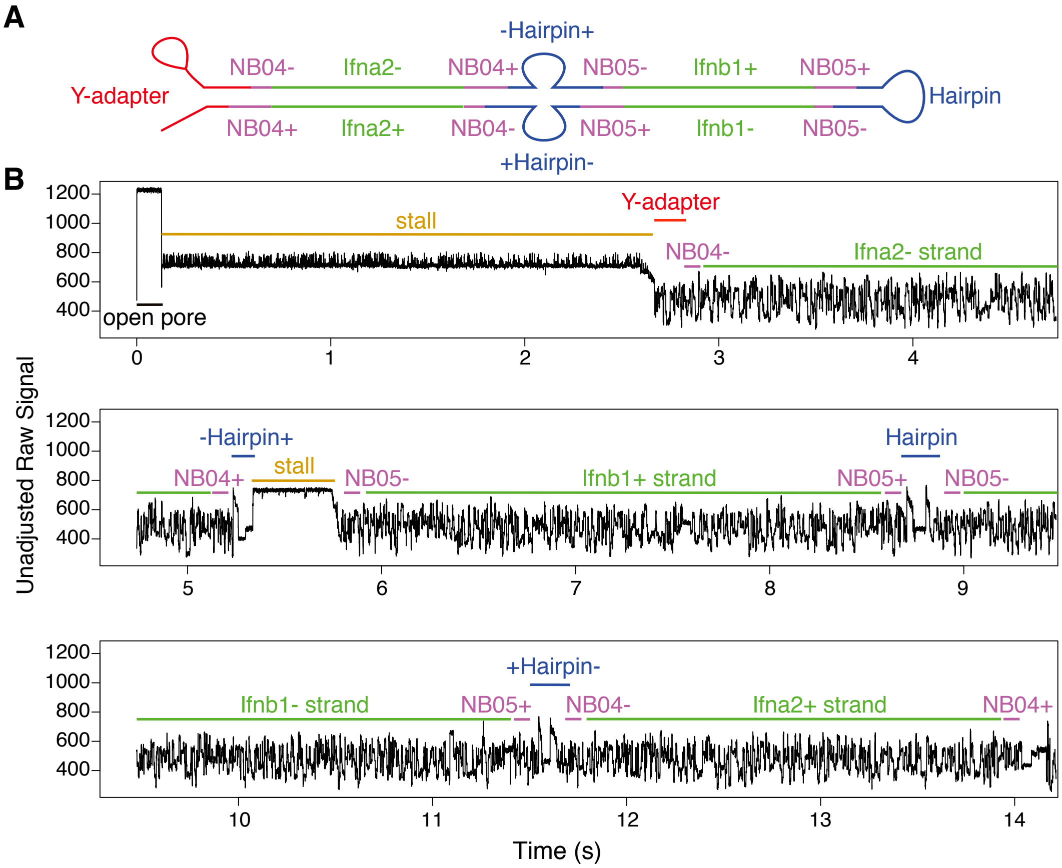

Figure 7. Demonstrative raw signal for a chimeric read (containing a strand dissociation event and two separate hairpin events within the same base-called sequence). The recorded signal begins with a very short open pore state (0.1s), followed by a long stall (2.5s), then an NB04-flanked *Ifna2* non-coding sequence with a transition time of 2.5s. At this stage there appears to be the beginning of a hairpin sequence that is finished by a pore stall. This was followed by a coding *Ifnb1* sequence with a transition time of 2s, then a hairpin, then an NB05-flanked non-coding *Ifnb1* sequence (2.5s), and finally an NB04-flanked coding *Ifna2* sequence (2.5s). Barcodes detected from this read (NB04/NB05) suggest that the chimeric sequence was likely formed during overnight ligation.

after the barcoding step of sample preparation (i.e. during adapter ligation).

This finding has potential implications for other sequencing technologies, as the ligation process used for sample preparation is unlikely to be specific for nanopore sequencing. The formation of chimeric reads during sample preparation may be one explanation for the index switching phenomenon seen in Illumina-sequenced reads (e.g. see 22–24), and presents a substantial problem for dual-indexed reads where identical indexes are used for different samples. Where dual-indexed reads are not used, ligation of reads with the same index may still be problematic depending on the particular sequencing application.

In-silico chimerism (1D^2)

There were 8 instances where both an overnight and a quick barcode were observed in the base-called sequence. In all such cases, there appears to have been a very short pore-protein dissociation between the sequencing of the two sequences (i.e. these were chimeric reads formed from *in-silico* ligation). The dissociation was only noticeable after inspecting the raw signal: a very short blip in the signal that matched the open pore current (e.g. see Figure 8).

It is likely to be the case that similar situations involving fast pore reloading are present in other reads, but not easily detectable from the called sequence because other barcode/amplicon combinations fit the expected base calling pattern. Considering that this situation can happen with non-identical sequences, software that is able to flag the presence of dissociation and/or stall events that are not at the start of the raw signal would be useful, as these features suggest that the base call is not likely to be a correct single sequence.

The release of ONT's R9.5 flow cells and 1D^2 base-calling exploits this phenomenon of fast sequence loading into pores in order to produce high-accuracy reads derived from a combined template/complement base-call (i.e. replacing the current hairpin-based 2D call). This replaces the 2D sample preparation process that we used for this investigation (see 25).

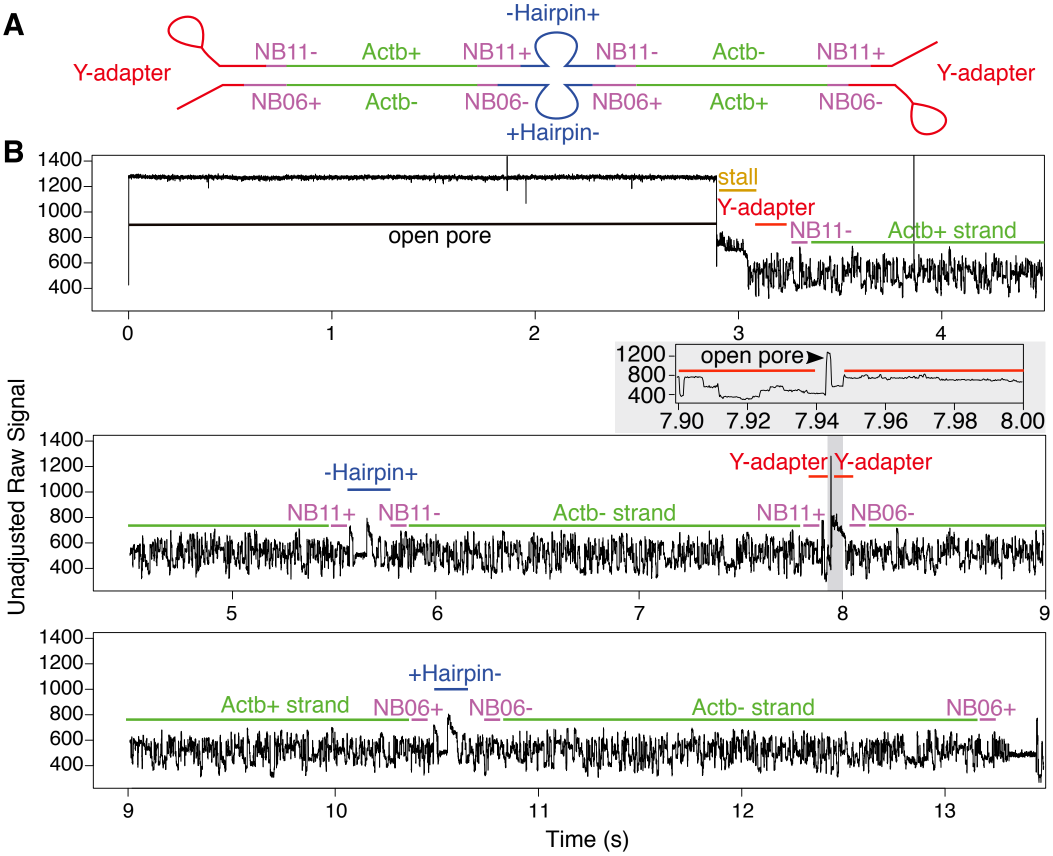

Figure 8. Demonstrative raw signal for a read (base-called as chimeric) that appeared to be from two different ligation preparations. The recorded signal begins with a long open pore period (2.9s), and a short stall (0.1s), followed by NB11-flanked coding and non-coding *Actb* sequences (transition time of 2.5s for each). There is a very short open-pore blip at around 8s, followed by a short stall (0.1s), then NB06-flanked coding and non-coding *Actb* sequences (transition time of 2.5s for each).

Conclusions

It is apparent from our investigation that chimeric reads can exist in the output of sequencing runs, and we recommend that researchers consider this possibility when interpreting their own results. As a result, it is advisable to include easily-detectable adapters when sequencing DNA. These adapters, particularly if present at both ends of a sequence, will help substantially in the identification (and if necessary, filtering) of concatenated sequences that are not native to the sample.

Although a non-negligible 1.7% of reads were found to have post-amplification chimeric elements, careful quality control of reads after long-read sequencing should be able to identify and exclude the majority of chimeric reads that are produced during a sequencing run.

Data availability

Raw read signal and basecalled reads have been uploaded to ENA under accession number PRJEB20601. Additional supplementary scripts used for FASTQ file filtering, mapping, and raw signal investigation are available as part of David Eccles' bioinformatics script repository (doi, 10.5281/zenodo.556966)[26]. The following scripts from that repository were used for intermediate discovery and result generation:

maf_bcsplit.pl Converting MAF format to machine-readable CSV with forward-oriented location information

pos_aggregate.pl Merging adjacent MAF matches to the same target sequence in the same orientation

fastx-fetch.pl Retrieving sequences from a FASTQ/FASTA file given a a list of identifiers (possibly as a text file)

fastx-length.pl Generating length information and aggregate statistics for a FASTQ/FASTA file

length_plot.r Generating "digital electrophoresis" image and read density plots given a file containing length information

porejuicer.py Extracting raw data and called FASTQ files from FAST5 files

A rough shell command script (including additional dead-end attempts at discovery & analysis) is provided for reproduction and/or extension of these findings to other investigations (see Supplementary File 6).

Author contributions

RW: Sample preparation and QC; CP: Mouse injections, RNA extraction; FR: Project oversight; OL: Sample preparation, project design and oversight; DE: DNA sequencing and bioinformatics analysis. All authors contributed towards the preparation of the manuscript.

Competing interests

The R9.4 flow cell and sequencing kit (SQK-LSK208) used for this experiment were provided free of charge by ONT as replacements for a purchased kit and flow cell where the phenomena of chimeric reads was initially discovered. ONT provided advice regarding the sample preparation protocols, including the suggestion of a slow overnight ligation step.

Grant information

This work was funded in full by Health Research Council of NZ Independent Research Organization (IRO) funding to the Malaghan Institute of Medical Research (grant number HRC14/1003).

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

Acknowledgements

We would like to thank the Nanopore Community for providing help and insightful discussion for this investigation.

Supplementary material

Supplementary File 1: Base calling summary from Albacore v0.7.5.

Click here to access the data.

Supplementary File 2: Reference sequences used for the initial amplicon mapping.

Click here to access the data.

Supplementary File 3: Reference sequences used for *Ifna* paralog mapping.

Click here to access the data.

Supplementary File 4: R script and intermediate data files used for *Ifna*-family gene counting.

Click here to access the data.

Supplementary File 5: R script and intermediate data files used for chimeric read filtering.

Click here to access the data.

Supplementary File 6: Shell/process script for reproducing the data analysis.

Click here to access the data.

References

1. Levy SE, Myers RM: **Advancements in Next-Generation Sequencing.** *Annu Rev Genomics Hum Genet.* 2016; **17**(1): 95–115.
PubMed Abstract | Publisher Full Text

2. Mikheyev AS, Tin MM: **A first look at the Oxford Nanopore MinION sequencer.** *Mol Ecol Resour.* 2014; **14**(6): 1097–1102.
PubMed Abstract | Publisher Full Text

3. Chandler J, Camberis M, Bouchery T, *et al.*: **Annotated mitochondrial genome with nanopore r9 signal for *nippostrongylus brasiliensis* [version 1; referees: 1 approved, 2 approved with reservations].** *F1000Res.* 2017; **6**: 56.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Reuter JA, Spacek DV, Snyder MP: **High-throughput sequencing technologies.** *Mol Cell.* 2015; **58**(4): 586–597.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Walter MC, Zwirglmaier K, Vette P, *et al.*: **MinION as part of a biomedical rapidly deployable laboratory.** *J Biotechnol.* 2016; **250**: 16–22.
PubMed Abstract | Publisher Full Text

6. Castro-Wallace SL, Chiu CY, John KK, *et al.*: **Nanopore dna sequencing and genome assembly on the international space station.** *bioRxiv.* 2016.
Publisher Full Text

7. Urbanc JM, Bliss J, Lawrence CE, *et al.*: **Sequencing ultra-long dna molecules with the oxford nanopore minion.** *bioRxiv.* 2015.
Publisher Full Text

8. Simpson JT, Workman RE, Zuzarte PC, *et al.*: **Detecting DNA cytosine methylation using nanopore sequencing.** *Nat Methods.* 2017; **14**(4): 407–410.
PubMed Abstract | Publisher Full Text

9. Deschamps S, Mudge J, Cameron C, *et al.*: **Characterization, correction and *de novo* assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*.** *Sci Rep.* 2016; **6**(1): 28625.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Quick J, Grubaugh ND, Pullan ST, *et al.*: **Multiplex pcr method for minion and illumina sequencing of zika and other virus genomes directly from clinical samples.** *bioRxiv.* 2017.
Publisher Full Text

11. Benítez-Páez A, Portune KJ, Sanz Y: **Species-level resolution of 16s rrna gene**

amplicons sequenced through the minion™ portable nanopore sequencer. *Gigascience.* 2016; **5**(1): 4.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Jain M, Koren S, Quick J, *et al.*: **Nanopore sequencing and assembly of a human genome with ultra-long reads.** *bioRxiv.* 2017.
Publisher Full Text

13. Hargreaves AD, Mulley JF: **Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing.** *PeerJ.* 2015; **3**: e1441.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Rand AC, Jain M, Eizenga JM, *et al.*: **Mapping DNA methylation with high-throughput nanopore sequencing.** *Nat Methods.* 2017; **14**(4): 411–413.
PubMed Abstract | Publisher Full Text

15. Smith AM, Jain M, Mulroney L, *et al.*: **Reading canonical and modified nucleotides in 16s ribosomal rna using nanopore direct rna sequencing.** *bioRxiv.* 2017.
Publisher Full Text

16. Connor LM, Tang SC, Cognard E, *et al.*: **Th2 responses are primed by skin dendritic cells with distinct transcriptional profiles.** *J Exp Med.* 2017; **214**(1): 125–142.
PubMed Abstract | Publisher Full Text | Free Full Text

17. van Pesch V, Lanaya H, Renauld JC, *et al.*: **Characterization of the murine alpha interferon gene family.** *J Virol.* 2004; **78**(15): 8219–8228.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Brinkmann V, Geiger T, Alkan S, *et al.*: **Interferon alpha increases the frequency of interferon gamma-producing human CD4+ T cells.** *J Exp Med.* 1993; **178**(5): 1655–1663.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Camberis M, Le Gros G, Urban J Jr: **Animal model of *Nippostrongylus brasiliensis* and *Heligmosomoides polygyrus.*** *Curr Protoc Immunol.* 2003; **Chapter 19**: Unit 19.12.
PubMed Abstract | Publisher Full Text

20. Démoulins T, Baron ML, Kettaf N, *et al.*: **Poly (I:C) induced immune response in lymphoid tissues involves three sequential waves of type I IFN expression.** *Virology.* 2009; **386**(2): 225–236.
PubMed Abstract | Publisher Full Text

21. Frith MC, Hamada M, Horton P: **Parameters for accurate genome alignment.** *BMC Bioinformatics.* 2010; **11**(1): 80.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Sinha R, Stanley G, Gulati GS, *et al.*: **Index switching causes "spreading-of-signal" among multiplexed samples in illumina hiseq 4000 dna sequencing.** *bioRxiv.* 2017.
Publisher Full Text

23. Hadfield J: **Index mis-assignment between samples on hiseq 4000 and x-ten.** Core-Genomics Blog, 2016.
Reference Source

24. Bushnell B: **Introducing crossblock, a bbtool for removing cross-contamination.** Seqanswers discussion thread, 2017.
Reference Source

25. Brown C: **Gridion x5 - the sequel.** Technology presentation, 2017.
Reference Source

26. Eccles D (gringer): **gringer/bioinfscripts: Chimeric read update.** 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓

Version 2

Referee Report 13 September 2017

doi:10.5256/f1000research.13356.r25051

✓ **Winston Timp**

Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

The question regarding the low number of 2D sequences was addressed by the authors as the effect of different versions of basecalling algorithm and sequencing chemistry. The authors further explained the differences in ligation conditions, concluding that the short ligation is better at preventing chimeric reads than long ligation. Based on these observations I am satisfied with the report.

Competing Interests: Patents licensed to Oxford Nanopore Technologies

Referee Expertise: Biophysics, sequencing, epigenetics

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.

Version 1

Referee Report 19 June 2017

doi:10.5256/f1000research.12473.r23252

? **Winston Timp**

Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

In this manuscript, the authors discuss the detection and potential sources of chimeric reads from minION (nanopore) sequencing. Though the manuscript has some interesting analyses and ideas, I have some problems with the results that should be addressed, specifically points 1, 3, 4, 5 and 7:

1. Though 2D libraries were prepared - only *0.3%* of the reads were actually 2D, the vast majority were 1D. This is in my view quite surprising - though with 2D libraries I have seen plenty of 1D reads mixed in, this level of 2D/1D for a 2D prep suggests something strange upstream of sequencing is occurring. However, the authors decline to address it, “The reasons behind this basecall failure were not investigated”. I think this must be addressed more carefully to understand the results.

2. I feel the low % of 2D reads is important because it may play into the source of the chimeras - if the 2D calling is failing due to heterogenous DNA strands - i.e. hybridization of an IFN to an Actb for example, then end polishing and adapting would lead to a 2D read where the two strands don’t match, hence called as 1D.

3. The authors suggest that amplicon sequencing is more susceptible to chimeras because “the same sequencing will appear in increased abundance” - I’m not clear on why that makes chimeras more frequent, just that it makes them more likely to be easily detectable.

4. The authors discuss “multiple hairpin adapters being ligated in the same sequence”. I don’t understand how the authors think this is possible? There are only two free ends of DNA, and if there are hairpins on both, the DNA will not be able to enter the pore. Instead I suggest it could be the proposed “in silico” chimerism the authors later discuss.

5. PCR Chimeras are not unknown in the literature - having been described, for example, in Sanger and 454 here (PMID: 21212162, 20833233). The authors’ assumption that the chimerism is occurring downstream of PCR needs to be demonstrated - Figure 3 suggests that the length of the chimeric is not outside the range of either Illumina or Sanger sequencing, so could be easily validated with these technologies.

6. But - given the multiple ligation steps of this protocol, it seems likely that the dA-tailing failing some fraction of the time could results in blunt-end ligation and chimeric reads.

7. How does enrichment look comparing the overnight to quick ligation for the different categories of chimerism detailed? The only results given are overall chimerism.

8. The authors only tried overnight ligation/quick ligation for the last ligation step, but not for the barcode ligation step. I also wonder if a PCR-barcode may have given better results - the multiple ligations may have led to a higher rate of chimeras, as the end-polishing likely had some fraction of blunt-ended amplicons.

9. Another possible point was the the authors may have not added enough (relative) adaptors - the relative high concentration of template allowed self-ligations to be more frequent. Adaptor dimers are probably easier to eliminate in this case than chimeras.

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?
Partly

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

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?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: I have two patents licensed to Oxford Nanopore.

Referee Expertise: Biophysics, Sequencing, Epigenetics

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, however I have significant reservations, as outlined above.

Author Response 20 Jun 2017

David Eccles, Malaghan Institute of Medical Research, New Zealand

Thanks for you comments. We will be incorporating your suggested changes into a revised manuscript. To make sure we've got the right idea about your suggestions, here are my initial impressions:

1. I understand what happened now after talking to Forrest Brennen at the London Calling conference this year. 2D reads were not called because the older hairpin was used. Even though we used a newer 2D kit, the barcode kit had the old hairpin sequence in it (which wasn't detected by the Albacore caller used at the time). This mis-call has either been fixed now, or will be fixed soon -- I'll re-call the reads with the most recent Albacore and if no improvement will talk to ONT about the correct software tweaks to fix it. In any case, plenty of hairpin sequences were detected in the linear/1D base calls.

2. We initially thought that the failure of 2D was due to chimeric reads, because it seemed to be occurring at the 2D alignment step. However, reads that were very obviously not chimeric were still failing the 2D calling. There is a chance that a properly called chimeric read will fail the 2D alignment step, but I expect it will still have a called template + complement sequence. I don't recall seeing many situations where the chimerism had happened on a single strand; it was mostly double-stranded fragments that had joined together.

3. My theory on why amplicon sequencing is more susceptible to chimeras is that it encourages the formation of base pairing structures (e.g. quadruplexes) that bring the ends of similar sequences closer to each other. I don't know how deep we should go into this; it's a hypothesis about why they could be in higher abundance for our specific experiment, but we haven't tested whether or not amplicon sequencing runs have a higher rate of chimeric reads.

4. Multiple hairpin adapters do make sense; David Stoddart (ONT sample prep guru) was with me when I was doing some "napkin drawings" of the structure that was formed by a 3-hairpin sequence, and helped correct a bit of the structure. We should add that into the paper (I think he kept the drawing, but I can make another one).

5. PCR chimeras may exist in our results, but appear to be of low abundance according to the electrophoresis plot, and I've tried to analyse the results in such a way that PCR chimeras would be excluded. Our experimental design was such that barcoding (and mixing of

separate amplicons) happened after the PCR was done, so the results should be at worst an underestimate of chimerism, and PCR chimeras should be observable in the data.

6. Yes; dA-tailing failure seems to be a likely explanation. Due to the physics of chemical reactions, there are going to be some that don't work. However, it is a little bit curious that the overnight ligation produced more chimeric reads. If the chimerism were due solely to dA-tailing failure, then increased abundance for overnight ligation doesn't make sense.

7. Different categories of chimerism are outlined in Table 1, and the four categories are broken down into overnight/quick in figure 5. I've included a full text description of each ***read*** in the supplementary information, but we felt that it was overly complex to include all those details in a graph.

8. The overnight/quick difference was unexpected (particularly in the direction that it happened). While ONT have discontinued their 2D hairpin kit, we would still be able to carry out a subsequent investigation in the future to look at overnight vs quick ligation during barcoding.

9. We tried to add the recommended amount of adapters to the samples, and the called results suggest that adapter dimers were minimal. Those situations are very easy to pull out at the analysis stage, because there are no amplicon sequences between adapters.

Competing Interests: No competing interests were disclosed.

Referee Report 17 May 2017

doi:10.5256/f1000research.12473.r22553

✓ **Keith E. Robison**
Warp Drive Bio, Cambridge, MA, USA

The authors have provided a useful report on technical aspects of the emerging Oxford Nanopore MinION DNA sequencing technology. As noted in the paper, the library preparation methods scrutinized here are commonly found in multiple advanced DNA sequencing technologies, and so lessons learned in this work are likely applicable elsewhere.

In the methods section, the authors report injected "dead infectious" worms, but not how the worms were killed.

It would be greatly preferable for all of the input DNA amounts to library preparations to be given in both mass and fmol. While it is common to report masses of DNA, the ligations really are dependent on the availability of DNA ends.

The point Figure 8 is trying to convey would be greatly enhanced by adding a zoom of the region around 8s in the plot in which the temporal sequence barcodeNB11-open pore--stall-barcode NB06 is seen. Zooms of other transitions should be considered as well.

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.

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.