UCL

Rotation report

Investigating the feasibility of nanopore sequencing in human prion diseases

Author Francois Kroll

Supervisors Simon Mead/Emmanuelle Vire

Leonard Wolfson
Experimental Neurology Centre

Abstract

MinION is a portable nanopore sequencing device. Reads generated by this technology have no set maximum length, which has the potential to greatly simplify sequencing of complex regions of the human genome. We performed sequencing of 1 kb *PRNP* coding region amplicons with the MinION. We were able to specifically detect the codon 129 single-nucleotide polymorphism at the correct zygosity and filter false positive variant calls using criteria such as allele frequency and strand bias. While studies in cattle have identified two deletions in the *PRNP* promoter and intron which increase susceptibility to bovine spongiform encephalopathy, human *PRNP* non-coding regions have been little studied to date, and all known pathogenic variants remain located within the 762-bp coding region. We attempted the first sequencing of the *PRNP* gene and promoter in a single 17.5 kb read. However, amplification of the region proved difficult, potentially because of a secondary structure or high GC content. While the existence of novel variants in the *PRNP* non-coding region remains unknown, the successful detection of the codon 129 variant with this pocket-size device suggests the feasibility of on-site same-day genetic diagnosis for individuals at risk of prion diseases.

Background

Nanopore long-read sequencing

Using nanopores to sequence polynucleotide molecules was first suggested in 1996[1]. In 2014, Oxford Nanopore, a spin-off from University of Oxford, was the first company to release on the market an accessible nanopore sequencing device called MinION[2].

MinION functions by recording the electrical current circulating through pores of 1 nm in diameter. As the DNA or RNA molecule is translocated through the pore by a motor protein, each nucleotide causes disruptions in current in a specific pattern. These can be detected by a basecalling algorithm which translates the electrical signal into a nucleotide sequence[3,4] (**Figure 1**).

One of the main advantages of nanopore sequencing is the generation of long reads. As opposed to the vast majority of next-generation sequencing technologies, there is no theoretical limit to the read length, and reads over 2 Mb have recently been recorded[5].

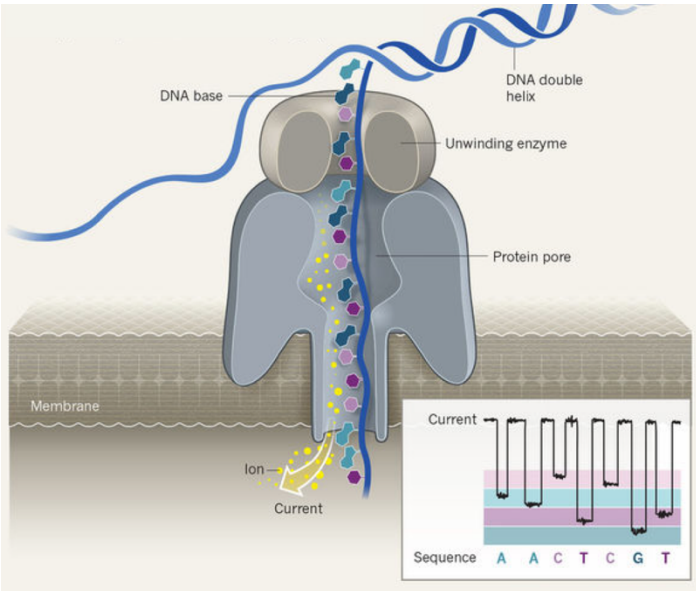


Figure 1 | Nanopore sequencing. During sequencing, a motor protein (unwinding enzyme) translocates the DNA molecule through the pore. Nucleotides cause disruptions in current which can be detected and translated to a sequence by a basecalling algorithm (image adapted from Perkel, 2017).

Illumina technologies, for instance, generate reads of a maximum length of 300 bp[6]. The user thus heavily relies on algorithms for *de novo* assembly or alignment to a reference sequence. However, these often show limitations in complex regions of the genome. Long-reads, as they span larger regions or whole of the targeted sequence, often allow for much more reliable downstream analyses.

Specifically, in neurological disorders, MinION has been used to sequence *CACNA1C*, a calcium channel gene involved in schizophrenia. It contains at least 50 exons forming more than 40 predicted isoforms, which makes its sequencing especially difficult with short reads. Nanopore sequencing of *CACNA1C* cDNA transcripts revealed 38 novel exons and additional predicted isoforms not previously identified[7]. Another example is the *GBA* gene. Mutations in this gene cause Gaucher disease and can be a risk factor for Parkinson's disease. Its sequencing using short-read technologies is complicated due to the neighbouring and virtually identical *GBAP1* pseudogene. As recombination events occur between these two regions, it is especially important that *GBA* is specifically sequenced. This is a challenge with short reads, with up to 35% of Illumina-generated *GBA* reads misaligning to *GBAP*[8]. On the other hand, only 1.1% of nanopore-generated *GBA* reads misalign, which ease the detection of known variants at the correct zygosity[9].

The main drawback to nanopore sequencing is accuracy. The most pervasive error is numerous incorrect calls of short (1-3 bp) indels, which is mostly problematic for whole-genome assemblies[10]. Another typical error is wrong basecalling in around 3-5% of the reads. The main

way around this is sequencing at high coverage, which can be straightforward with the MinION thanks to high sequencing speed over 500 bp per second.

PRNP

The prion gene *PRNP* is a 15.3 kb sequence on chromosome 20. It is composed of two exons flanking a single large intron. The coding region is a 762-bp sequence within the second exon[11].

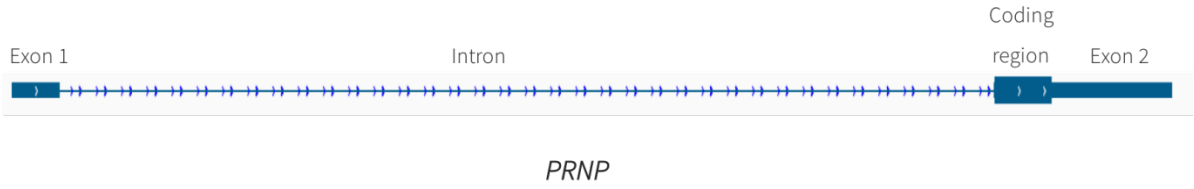


Figure 2 | Structure of the *PRNP* gene on chromosome 20.

All 69 known potentially-pathogenic *PRNP* variants are located within the 762-bp coding region[12]. As such, genetic testing at the Institute of Prion Diseases is typically performed by amplification of this sequence. A common single-nucleotide polymorphism (SNP) is at codon 129, where either Methionine or Valine can be encoded. It represents the strongest association between a genotype and any disease known to date, with all cases of variant Creutzfeldt-Jakob disease (CJD) genotyped being Methionine homozygous at PRNP amino acid 129[13].

Studies from cattle have identified novel variants in *PRNP* non-coding regions. Specifically, two 23 and 12-bp deletions, respectively in the promoter region and the intron, have been shown to significantly increase the animal's risk to develop bovine spongiform encephalopathy upon consumption of infected feedstuffs[14–16].

The non-coding sequences of human *PRNP* have been little studied to date, including in cases of CJD. Here, we set out to investigate the practicality of nanopore technology to sequence full length *PRNP* from CJD patients' samples.

Materials and methods

PCR

Human genomic DNA samples used in these experiments were from blood collected and processed by the Institute of Prion Diseases.

PRNP coding region amplicons

PCR amplification of the *PRNP* coding region was performed with primers routinely used for genetic diagnosis at the Institute of Prion Diseases. Forward primer sequence was 5'-

CTATGCACTCATTCATTATGC-3'. Reverse primer sequence was 5'-GTTTTCCAGTGCCCATCAGTG-3'. Each PCR well contained 12.5 µL 2x MegaMix Royal (Microzone, #2MMR), 10.5 µL H_2O, 1 µL primers (12.5 µM each), 1 µL genomic DNA. Initial denaturation was 5 min at 95˚C, followed by 35 cycles of: 95˚C for 30 sec (denaturation); 58˚C for 40 sec (annealing); 72˚C for 1 min (extension). A total of ten 25 µL reactions were performed to reach a yield of 900 ng which is recommended by Nanopore Technologies. The PCR product was cleaned and concentrated using Zymo DNA Clean & Concentrator-5 kit and eluted in 50 µL TE buffer. Quantification of the PCR product mass was performed with a Qubit (dsDNA Broad Range assay). Size, quality and concentration of the fragments were checked using TapeStation 2200 (D1000 tape).

PRNP gene and promoter

Four pairs of forward/reverse primers were designed using online tools Primer3, PrimerBlast and IDT PrimerQuest.

#	Forward primers	#	Reverse primers
1	5'-TCCAAGAAATCCCAGGCCAT-3'	**2**	5'-GGAGAAAATGCAAAAGCCGC-3'
3	5'-CAAGAAATCCCAGGCCATTTG-3'	**4**	5'-CTTGAGAAGGGAGGACAGAAAG-3'
5	5'-GAGCCCTTTCCTGATGGGAG-3'	**6**	5'-GTGGGTGACTGGGAAGTGAG-3'
7	5'-TTGAGCCCTTTCCTGATGGG-3'	**8**	5'-CTCCTGCTGCAATAGTGCCT-3'

The 16 pairwise combinations of a forward and a reverse primer were tested with various PCR programmes and four different DNA polymerases: KAPA HiFi, KAPA LongRange, Phusion High-Fidelity and Roche Expand. The results reported here were produced with the Roche Expand DNA polymerase and primer pairs 3/2 and 5/6 with the following PCR programme: 2 min at 94˚C (initial denaturation), then 10 cycles of: 10 sec at 94˚C (denaturation), 30 sec at 62˚C (annealing), 14 min at 68˚C (elongation); followed by 25 cycles where the elongation time is increased by 20 sec every cycle, finished by a final elongation of 7 min at 68˚C. Each PCR well contained 37.5 µL H_2O, 5 µL buffer 3 (provided with Roche Expand), 2.5 µL dNTPs (10 mM each), 1.5 µL forward primer (10 µM stock), 1.5 µL reverse primer (10 µM stock), 1.25 µL genomic DNA template (393 ng/µL) and 0.75 µL Roche Expand enzyme mix. A total of twenty 50 µL reactions were performed to reach a sufficient yield as measured by Qubit dsDNA Broad Range. PCR product was cleaned with the Zymo Genomic DNA Clean and Concentrator-10 kit and eluted in 50 µL TE buffer. PCR products' lengths were checked on TapeStation 2200 with the genomic DNA tape.

Library preparation

Library preparation prior to sequencing was performed according to Nanopore Technologies' 1D amplicon by ligation protocol (version *ADE_9003_v108_revU_18Oct2016*). All reagents were provided by the SQK-LSK108 kit, except: NEBNext Ultra II End Repair/dA-Tailing buffer and enzyme mix (#E7546), NEB Blunt TA/Ligase Master Mix (#M0367), and Agencourt AMPure XP beads (#A63880).

The *PRNP* coding region amplicons were diluted after end-repair/dA-tailing to bring only 0.2 pmol of DNA fragments to the ligation reaction.

With the *PRNP* gene and promoter amplicons, the only amendment was longer incubation times of the end-repair/dA-tailing reaction (30 min at 20˚C, 30 min at 65˚C).

Sequencing

PRNP coding region

23 ng of final library were loaded onto the MinION flow cell. Sequencing was performed for 53 minutes and followed live on the MinKNOW software.

PRNP gene and promoter (primer pair 3/2)

331 ng final library were loaded on the MinION flow cell and sequenced for 5 hours.

PRNP gene and promoter (primer pair 5/6)

6.9 µg final library were loaded on the MinION flow cell and sequenced for 8 hours.

Data analysis

Basecalling was performed on the fast5 files generated by MinKNOW using Albacore v2.3.1[17]. Alignment to the hg38 human reference genome was performed with NGMLR v0.2.6[18] and variants were called with the variant calling algorithm of Nanopolish v0.9.1[19]. Samtools v1.7[20] was used for conversions from *bam* to *sam* alignment files, sorting and indexing. The typical command line workflow was:

```
read_fast5_basecaller.py -i fast5 -t 8 -s /path_to_fast5/ -f FLO-MIN106 -k SQK-LSK108 -o
fastq -q 0 -n 0 --disable_filtering

ngmlr -t 8 -r hg38.fa.gz -q reads.fastq -o alignment.sam -x ont

samtools view alignment.sam -o alignment.bam

samtools sort alignment.bam > alignment_sort.bam

samtools index alignment_sort.bam
```

Final sorted and indexed *bam* alignments were visualised in IGV v2.4.10[21].

Genomic positions are reported for the hg38 assembly.

Results

Known *PRNP* variants can be specifically detected by nanopore sequencing

As a proof of principle that *PRNP* variants could correctly be identified using the MinION, we aimed to reproduce the genetic diagnosis of a codon 129 heterozygous individual previously sequenced with Sanger sequencing. After amplification of the coding region and library preparation, we sequenced the amplicons for an hour using the MinION. Electrical current traces were basecalled using the Oxford Nanopore-designed algorithm Albacore. After alignment to the reference genome with NGMLR, an algorithm designed specifically for nanopore-generated long reads, we obtained around 800 times coverage of the targeted region.

We ran on these data the variant calling algorithm Nanopolish. It called four potential variants: two SNPs (chr20:4,699,605; chr20:4,699,996) and two one-nucleotide deletions (chr20:4,699,802; chr20:4,699,840).

As this sample was previously sequenced on the Sanger sequencer, we only expected a single true positive variant, namely heterozygous Methionine/Valine at codon 129 (chr20:4,699,605). We thus expected the other three called variants to be false positives. These could be filter by a combination of four criteria: indels, low quality scores, inconsistent allele frequencies and strand biases (see Discussion). The allele frequency (50% A/47% G) is consistent with a heterozygous call, which is also confirmed by Nanopolish output.

As a conclusion, codon 129 *PRNP* variant can be identified at the right zygosity using nanopore sequencing. Some false positives were called by the variant calling algorithm, but non-biased filtering methods allow for their detection.

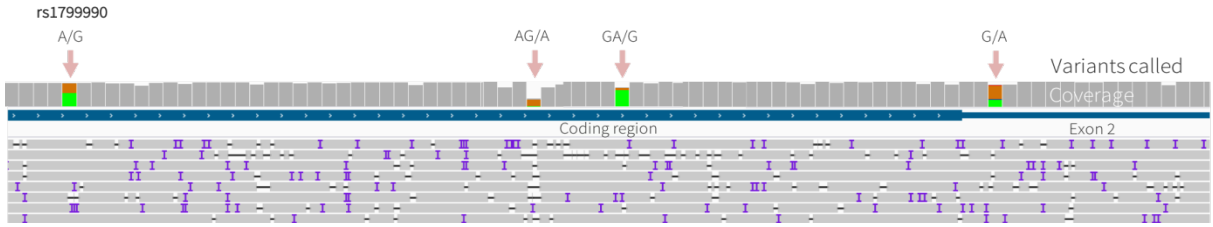


Figure 3 | Alignment of nanopore-sequenced *PRNP* coding region reads to the reference genome. Variants called by Nanopolish are indicated by pink arrows and written as reference/alternative allele. The coverage track is a boxplot showing nucleotide calls for each potential variant position (red is T, brown is G, blue is C, green is A). The first variant to the left corresponds to the position of the codon 129 SNP (rs1799990). In the reads, purple bars and white spaces represent short insertions and deletions, respectively.

Amplification of a 17.5 kb genomic region spanning the *PRNP* gene and promoter

As the non-protein coding regions of *PRNP* remain little studied to date, we aimed to amplify and sequence the *PRNP* gene and promoter in a single read. We designed four forward and reverse primers spanning the whole region for a 17.5 kb product. We tested four different long-range DNA polymerases and more than 10 PCR programmes with the 16 possible pairs but were unable to sequence a specific *PRNP* product of this length. Interestingly, the DNA polymerase seems able to produce products of at least 17.5 kb or even longer, albeit non-specific (**Supplementary figure 1**). Sequencing of the PCR product also revealed that one pair of primers (5/6) successfully binds to the correct genomic sites, generating a high number of short reads at both ends of the targeted region (**Figure 4B**).

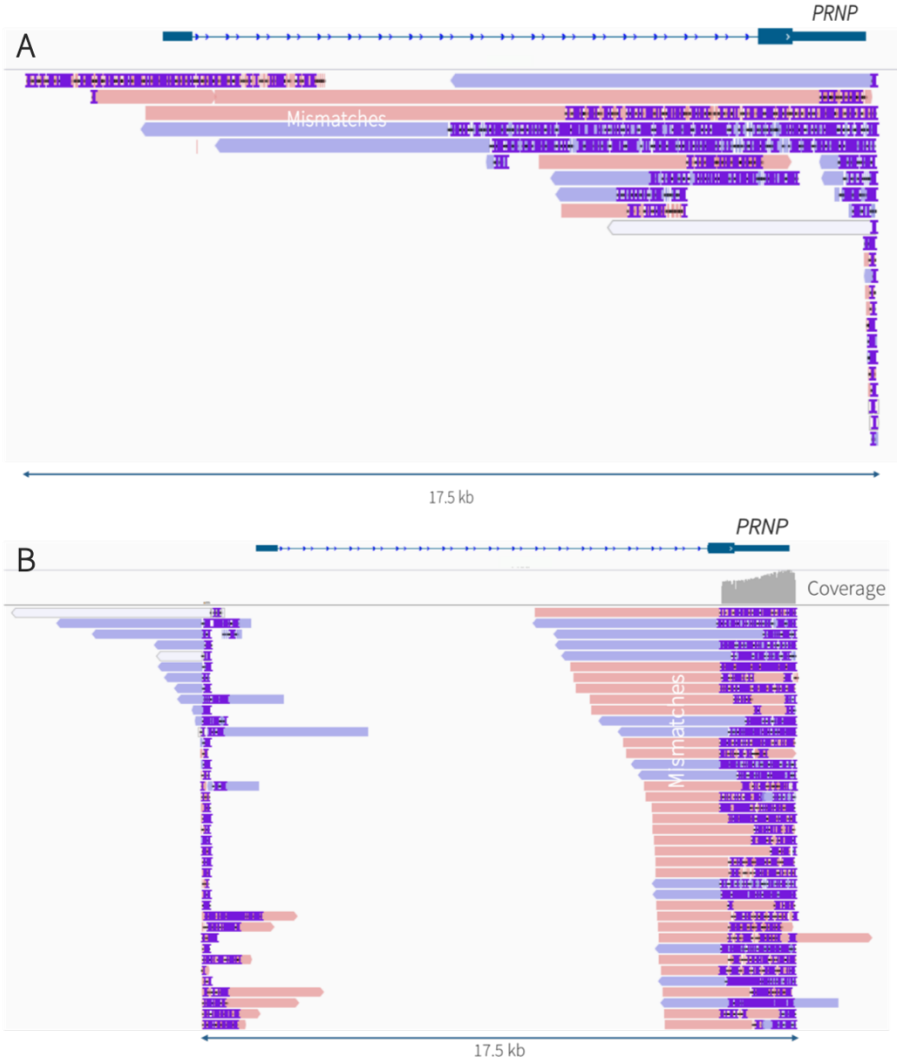


Figure 4 | Alignments of potential *PRNP* gene and promoter reads. **(A)** Reads obtained with the 3/2 primer pair. Parts of the reads clear of short insertions (purple bars) and deletions (white spaces) correspond to mismatches. All the reads aligning to *PRNP* are shown, while a total of ~ 430,000 reads were generated during this run. **(B)** Reads obtained with the 5/6 primer pair. Not all reads are shown.

Discussion

This study pioneers the use of the nanopore-sequencing device MinION in human prion diseases. We successfully sequenced *PRNP* coding region amplicons routinely used for genetic testing and were able to detect the codon 129 variant. However, three false positive variants were also called by the Nanopolish algorithm. We show here an unbiased method to identify these false positives. The strategy is to use a combination of four criteria. The first obvious one is to consider all short indels variant calls as dubious as this is the commonest mistake with nanopore sequencing. This may be too stringent though, and true short indels variants can probably be correctly detected by nanopore sequencing. Another criteria is to look at the quality scores computed by Nanopolish[19]. A particularity of this algorithm is to perform another alignment of the electrical current trace to the basecalled nucleotide sequence. This allows the variant calling to consider how unambiguous the electrical trace is at the position of each potential variant. Low quality scores represent a more ambiguous electrical trace at this position, and the corresponding variant should thus be less trusted. This was the case for both indels (respectively 3210.9 and 2329.6 for both SNPs vs. respectively 35.9 and 639.9 for both 1-bp deletions). Furthermore, assuming the absence of genetic mosaicism at these positions, allele frequencies can help decipher which variant can be trusted or not. In the case of a homozygous variant, around 100% (± 5% to account for MinION basecalling accuracy) of the reads should call for the alternative allele, while if the variant is heterozygous, around 50% of the reads should call for the alternative allele and the other 50% for the reference allele. Frequencies significantly differing from these are suspicious and should be considered as potential false positives. This was the case for all three variants except codon 129, which had frequencies consistent with a heterozygous call (50% reference/47% alternative, **Figure 3**). Finally, a powerful criterion which to our knowledge is not mentioned in literature for nanopore sequencing is a phenomenon called strand bias. In brief, evidence of strand bias exists if there is a bias towards a specific nucleotide call depending on which DNA strand is sequenced. When reads were coloured in IGV based on their strand, all variants except codon 129 showed evidence of such strand bias (**Figure 5**).

Given the fact that a long read is more likely to span a region containing more than one variant, nanopore sequencing can also make haplotype phasing much more reliable compared to short-reads technologies. This is probably the strongest strategy to identify true positive variants and their zygosity. This could not be performed on our data as only one true positive variant is present (codon 129 SNP, chr20:4,699,605).

With this proof of principle that *PRNP* amplicons could be sequenced and variant-called using MinION, we attempted the first amplification and sequencing of *PRNP* gene and promoter in a single 17.5 kb read. We were however unable to obtain a specific *PRNP* product of this length. By sequencing one of these PCR products using MinION, we identified a pair of primers that seemed to successfully anneal to the correct targeted sites, producing a high number of short reads at both ends of the *PRNP* region (**Figure 4B**). These short reads all seem to have similar lengths and efforts were made to avoid shearing during library preparation. This leads us to propose that the difficulties encountered when amplifying this 17.5 kb genomic region encompassing human *PRNP* may arise from the genomic context. This could be caused by the presence of a secondary structure[22], high GC content or a "closed" chromatin state. Interestingly, *PRNP* is also known to be a difficult locus for recombination studies[23].

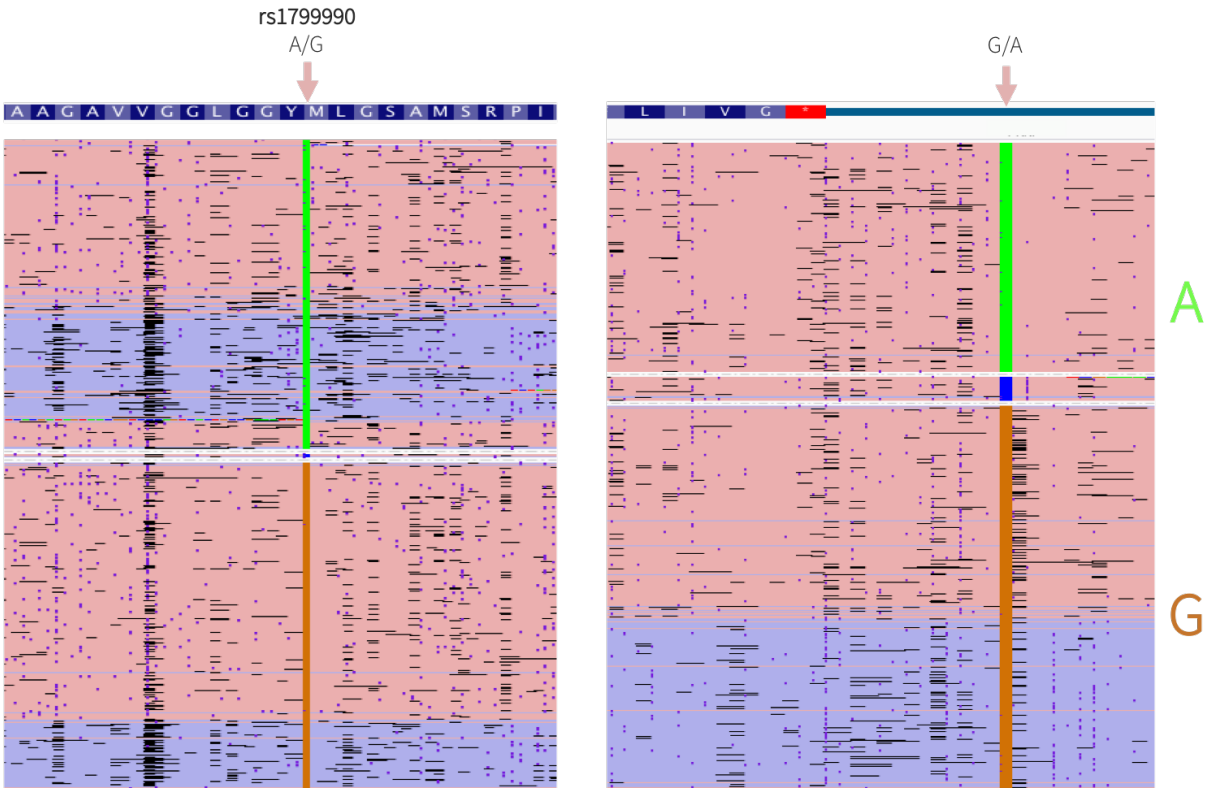


Figure 5 | Strand bias as a false positive variant call filtering criterion. Reads are coloured in IGV based on their strand. Codon 129 SNP (rs1799990, left) is not showing any evidence of strand bias as both blue and pink reads are calling for the reference and alternative alleles. The second SNP called by Nanopolish (right) showed evidence of strand bias as none of the blue reads is calling for the alternative allele.

Future directions

With its size smaller than a smartphone, MinION is the first sequencing device which has been used outside the laboratory in remote environments such as rainforest[24], the Arctic[25] or the International Space Station[26]. In more clinical settings, it was shown to be a promising tool for rapid diagnosis of aneuploidy in prenatal samples[27] and same-day genomic/epigenomic diagnosis of brain tumours[28]. By showing the successful sequencing and detection of the codon 129 *PRNP* variant in a blood genomic DNA sample, our first experiment suggests the feasibility of same-day on-site genetic diagnosis for individuals at risk of prion diseases. This is likely to be even eased in the future with new Nanopore Technologies products soon to be released such as the Flongle[29] which should provide cheap single-use flow cell and the VolTRAX[30] which aim to automate PCR and library preparation in a small programmable device.

Understanding of prion diseases has benefited from the study of kuru, a rare disease found in cannibalistic tribes of Papua New Guinea[31]. From 1997 to 2001, blood samples from infected individuals were collected on-field and shipped to the Institute of Prion Diseases for sequencing. The transport of these samples was a particularly complicated manoeuvre, starting from the dangerous regions of Papua New Guinea and across several airport customs. If a similar study was to be performed again, MinION would be an ideal tool to sequence samples directly on-field. Such use of the device was shown feasible during the Ebola outbreak when it was used to monitor the virus' evolution in Guinea. All instruments, reagents and consumables could be packed in a single standard aircraft baggage[32].

Whether *PRNP* non-coding regions include previously unidentified pathogenic variants remains an open question. One strategy here could be to use two different pair of primers to generate two overlapping PCR products of ~ 8 kb each. If this is successful, the subsequent step would be to barcode, pool and sequence on MinION sets of 12 genomic DNA samples from patients with prion disease.

References

1. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 13770–3 (1996).

2. Oxford Nanopore Technologies. Oxford Nanopore: Company History. (2018).

3. Oxford Nanopore Technologies. How it works. (2018). Available at: https://nanoporetech.com/how-it-works. (Accessed: 31st July 2018)

4. Perkel, J. TechBlog: The nanopore toolbox. *Naturejobs* (2017). Available at: http://blogs.nature.com/naturejobs/2017/10/16/techblog-the-nanopore-toolbox/.

5. Payne, A., Holmes, N., Rakyan, V. & Loose, M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. *bioRxiv* 1–28 (2018). doi:https://doi.org/10.1101/312256

6. illumina. lllumina sequencing platforms. (2018).

7. Clark, M. *et al.* Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain. *bioRxiv* (2018). doi:https://doi.org/10.1101/260562

8. Zampieri, S., Cattarossi, S., Bembi, B. & Dardis, A. GBA Analysis in Next-Generation Era. *J. Mol. Diagnostics* **19,** 733–741 (2017).

9. Leija-Salazar, M. *et al.* Detection of GBA missense mutations and other variants using the Oxford Nanopore MinION. *bioRxiv* (2018). doi:https://doi.org/10.1101/288068

10. Biomickwatson. How accurate is the nanopore-only assembly of GM12878? (2018). Available at: http://www.opiniomics.org/how-accurate-is-the-nanopore-only-assembly-of-gm12878/. (Accessed: 31st July 2018)

11. UCSC. Genome Browser. Available at: https://genome.ucsc.edu/. (Accessed: 15th June 2017)

12. Vallabh Minikel, E. List of reportedly pathogenic PRNP variants. (2015). Available at: http://www.cureffi.org/2015/01/13/list-of-reportedly-pathogenic-prnp-variants/. (Accessed: 1st August 2018)

13. Mead, S. *et al.* Genetic risk factors for variant Creutzfeldt–Jakob disease: a genome-wide association study. *Lancet Neurol.* **8,** 57–66 (2009).

14. Hills, D. *et al.* Complete genomic sequence of the bovine prion gene (PRNP) and polymorphism in its promoter region. *Anim. Genet.* **32,** 231–232 (2001).

15. Juling, K., Schwarzenbacher, H., Williams, J. L. & Fries, R. A major genetic component of BSE susceptibility. *BMC Biol.* **4,** 33 (2006).

16. Sander, P. *et al.* Analysis of sequence variability of the bovine prion protein gene (PRNP) in German cattle breeds. *Neurogenetics* **5,** 19–25 (2004).

17. Oxford Nanopore Technologies. Software Downloads. (2018).

18. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15,** 461–468 (2018).

19. Simpson, J. Nanopolish. (2018).

20. Samtools. (2018).

21. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

22. Nelms, B. L. & Labosky, P. A. A predicted hairpin cluster correlates with barriers to PCR, sequencing and possibly BAC recombineering. *Sci. Rep.* **1,** 106 (2011).

23. Kaczmarczyk, L., Mende, Y., Zevnik, B. & Jackson, W. S. Manipulating the Prion Protein Gene Sequence and Expression Levels with CRISPR/Cas9. *PLoS One* **11,** e0154604 (2016).

24. Pomerantz, A. *et al.* Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *Gigascience* **7,** (2018).

25. Goordial, J. *et al.* In Situ Field Sequencing and Life Detection in Remote (79°26′N) Canadian High Arctic Permafrost Ice Wedge Microbial Communities. *Front. Microbiol.* **8,** 2594 (2017).

26. Castro-Wallace, S. L. *et al.* Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Sci. Rep.* **7,** 18022 (2017).

27. Wei, S. & Williams, Z. Rapid Short-Read Sequencing and Aneuploidy Detection Using MinION Nanopore Technology. *Genetics* **202,** 37–44 (2016).

28. Euskirchen, P. *et al.* Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol.* **134,** 691–703 (2017).

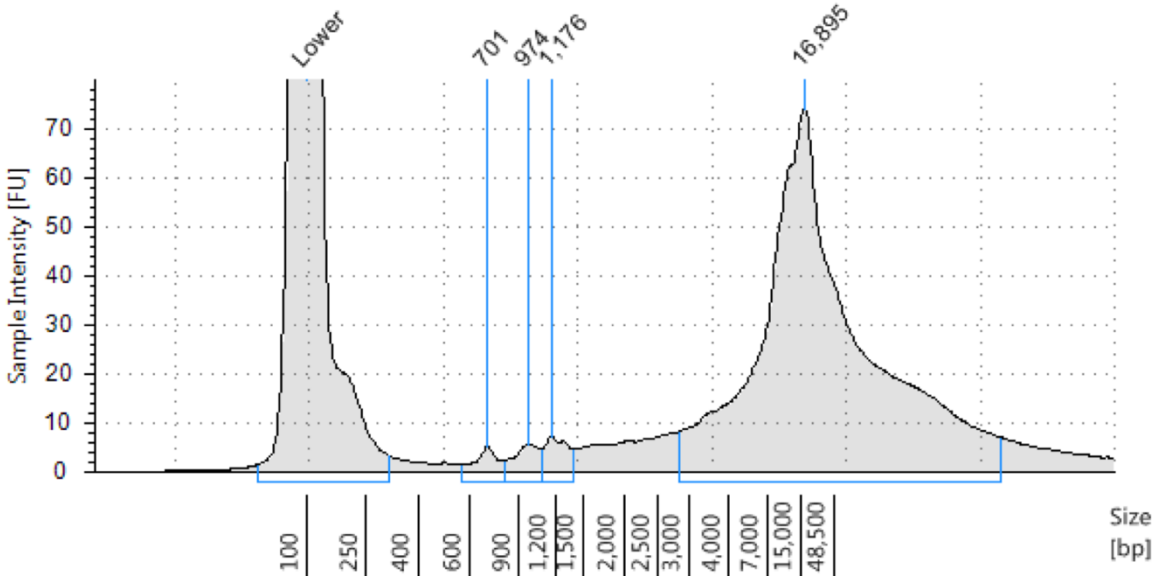
29. Oxford Nanopore Technologies. Flongle. (2018). Available at: https://nanoporetech.com/products/flongle. (Accessed: 2nd August 2018)

30. Oxford Nanopore Technologies. About VolTRAX. (2018).

31. Mead, S. *et al.* Genetic susceptibility, evolution and the kuru epidemic. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **363,** 3741–6 (2008).

32. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).

Supplementary data



Supplementary figure 1 | Tapestation run of the PCR product obtained with primer pair 3/2.