Received Date : 30-Jun-2014

Revised Date : 21-Aug-2014

Accepted Date : 25-Aug-2014

Article type : Opinion

A first look at the Oxford Nanopore MinION sequencer

Running head: Reviewing Oxford Nanopore's MinION

Alexander S. Mikheyev, Mandy M.Y. Tin

Ecology and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa, Japan 904-0495

Keywords: genomics, next-generation sequencing, nanopore sequencing, bioinformatics

Corresponding Author: Alexander S. Mikheyev

sasha@homologo.us

Abstract

Oxford Nanopore's third-generation, single-molecule sequencing platform promises to decrease costs for reagents and instrumentation. After a two-year hiatus following the initial announcement, the first devices have been released as part of an early access program. We explore the performance of this platform by re-sequencing the lambda phage genome, and amplicons from a snake venom gland transcriptome. Although the handheld MinION sequencer can generate more than 150 megabases of raw data in one run, at most a quarter of the resulting reads map to the reference, with less than average 10% identity. Much of the sequence consists of insertion/deletion errors, or is seemingly without similarity to the template. Using the lambda phage data as an example, although the reads are long, averaging 5kb, at best 890±1,932 bases per mapped read could be matched to the reference without soft clipping. In the course of a 36 hour run on the MinION it was possible to re-sequence the 48kb lambda phage reference at 16x coverage. Currently substantially larger projects would not be feasible using the MinION. Without increases in accuracy, which would be required for applications such as genome scaffolding and phasing, the current utility of the MinION appears limited. Library preparation requires access to a molecular lab, and is of similar complexity and cost to that of other next-generation sequencing platforms. The MinION is an exciting step in a new direction for single-molecule sequencing, though it will require dramatic decreases in error rates before it lives up to its promise.

Introduction

We live in a period of amazing technological change. In the past 10 years, an explosion of sequencing technologies has challenged dominance of venerable, decades-old Sanger sequencing. These include massively parallel, second-generation sequencing technologies, such as the widely used pyrosequencing (Roche 454), sequencing by synthesis (Illumina), ion semiconductor sequencing (Ion Torrent), and sequencing by ligation (SOLiD) platforms (Margulies *et al.* 2005; Bentley *et al.* 2008; McKernan *et al.* 2009; Rothberg *et al.* 2011). Before second-generation sequencing technologies peaked, single-molecule third generation sequencing platforms started to appear, such as those developed by Helicos and Pacific Biosciences (Pushkarev *et al.* 2009; Korlach *et al.* 2010). Some technologies came and went in a matter of years, such as those by Helicos and Roche, being replaced by their more rapidly evolving competitors. Others, like Ion Torrent and SOLiD, have struggled to gain market share from Illumina, which capitalized on an early edge with more rapid improvements in sequencing chemistry, and more extensive bioinformatic tools. Use of all next-generation sequencing platforms requires substantial investments – in the sequencers themselves, which run in the hundreds of thousands of dollars, and in sequencing reagents, which cost in the thousands (Glenn 2011; Quail *et al.* 2012). In addition, developing analytical pipelines involves considerable expenditure of human capital. Therefore, choosing the right platforms has become a critical decision for individual researchers, and even for sequencing centers, as they pool their resources to afford emerging technologies.

Each of the new technologies comes with significant limitations. Most of them, such as Illumina, SOLiD and Ion Torrent generate short reads, rarely exceeding several hundred bases. Pacific Biosciences real time single-molecule sequencing offers reads on the scale of kilobases, but at the cost of significantly increased error rates (Carneiro *et al.* 2012). Illumina's sequencing platforms currently attain the best balance between read lengths, error rates, and cost (Loman *et al.* 2012). While the overall cost per base has been dropping at a rate exceeding Moore's law, the rate of decrease has recently slowed, with recent advances providing steady but incremental improvements (Wetterstrand, 2014).

The promise of longer reads, and at a lower cost, has fueled excitement about new sequencing applications, unattainable by current technologies. Although short-read platforms are the current standard for re-sequencing applications, short reads can produce mis-mappings and mis-alignments, making regions heterozygous and repetitive regions of the genome inaccessible (McKenna *et al.* 2010; Li 2011). Similarly, short reads alone cannot resolve genomic structural variation, or haplotypic structure, which are important for many applications from population genetics to disease mapping (Kitzman *et al.* 2011). The relatively high cost of sequencing has impeded the progress of personal genomics, as few can afford it at the current prices (Desai & Jere 2012). Finally, the extreme expense of the sequencers, which may cost more than a million dollars (Glenn 2011), prevents most labs and even many universities, from having one on site. In general, the relatively high cost of next-generation sequencing can only be borne by relatively well-funded laboratories, and even they often have to spend time waiting in queues before their runs are scheduled.

Despite their limitations, advances in next-generation sequencing have been so rapid, that they have ceased to amaze. New single molecule methods, particularly nanopore sequencing, promise to significantly expand the realm of possibilities by greatly decreasing costs (Stoddart *et al.* 2009). Two years ago, a startup called

Oxford Nanopore announced the launch of a new third-generation single-molecule sequencing platform. By moving single-pore sequencing technologies to market, Oxford Nanopore promised read lengths orders of magnitude longer than existing technologies, together with low per-base costs, and a tiny futuristic USB-powered sequencer. The hand-held $1,000 sequencer, with a simple library preparation process, promised to democratize sequencing, making it affordable to a larger community, and perhaps even to citizen-scientists. After announcing their disruptive technology, and generating a viral amount of interest on the Internet, the company went silent for two years (Figure 1).

After the long hiatus, Oxford Nanopore finally released its portable MinION sequencer for beta testing by a broader community of users as part of the MinION Access Program (MAP). This report reviews the MinION's performance, and compares it to existing technologies. A comparison between the earliest available versions of a technology and that of its more mature competitors may seem premature and unfair. We also caution readers that the results we report here are among the very earliest available for the MinION, and its performance will likely improve. For instance, base calling for the MinION sequencers is carried out using a hidden Markov model (Timp *et al.* 2012), and the massive amount of data obtained during the MAP should allow the model to be substantially improved. However, it is also important for the scientific community to receive timely objective peer-reviewed evaluations of the new technologies, to allow potential users to determine whether the level of performance is appropriate for their applications, or whether they should await a major upgrade. We can also use the history of other platforms as a guide to see whether any shortcomings of the new technology will be easy to overcome. Here we overview early results, and discuss their utility for existing applications in molecular ecology.

Library preparation and sequencing on the MinION

Detailed methods for these analyses are available as an electronic supplement; source code and data are likewise available (see Data Accessibility). We used two of the library preparation kits provided by Oxford Nanopore, genomic DNA, and amplicon sequencing kits. The former was used to sequence the lambda phage genome (Sanger *et al.* 1982), while the latter was used to sequence cDNA from the venom gland of the pitviper, *Protobothrops flavoviridis*, which has been extensively characterized by Illumina sequencing and mass spectrometry (Aird *et al.* 2013). The runs were performed on separate flow cells.

Library preparation is similar to that for other next-generation applications, requiring DNA shearing, end repair, adaptor ligation and size selection. Once these steps are complete, the library must be conditioned, and then it is ready to load on the sequencer. Library preparation takes about half a day, and is of comparable complexity and cost to library preparation for other platforms. Finally DNA is 'conditioned' by the addition of a motor protein, a step that requires a 30-min incubation that can be extended to overnight, for a greater number of 2D calls. The libraries are then mixed with buffer and a 'fuel mix', and loaded directly into the sequencer. The run parameters are configured using Oxford Nanopore's MinKNOW software. As the sequencer runs, base calling takes place in real time using Oxford Nanopore's Metrichor cloud service. For optimal yield, the MinKNOW must be re-filled with additional library every four hours during a 48 hour run, which is the lifetime of a flow cell. To comply with the refill schedule, we worked in shifts, except for an eight-hour break at night when there were no refills. Data can be analyzed using either 1D or 2D workflows. The ingenious 2D workflow uses a hairpin adaptor, which links the top and bottom strands of double-stranded DNA into one strand. The base caller recognizes the hairpin sequence and aligns both strands of the template molecule, with the goal of improving sequencing accuracy.

Read lengths, yield and accuracy

We acquired sequence data from Enterobacteria phage lambda genomic DNA for 36 hours. This period of data acquisition is close to the life of one flow cell, which fully ceased to generate data by about 48 hours. For all libraries, reads were mapped to the reference using BLASR, a mapper developed for long, relatively inaccurate reads of the Pacific Biosciences sequencer (Chaisson & Tesler 2012), and also using BLASTN (Altschul *et al.* 1990) (Table 1).

Although the raw yield of the MinION is impressive, both in terms of read lengths and data output, the accuracy was extremely low. For the lambda phage, about 10% of the reads actually mapped to the reference using BLASR; their identities were 2.2% and 8.9% for the 1D and 2D workflows, respectively. This means that much less than 1% of all the generated sequence faithfully matches the reference. BLASTN produced similar results, although more short sequences were detected, resulting in a higher percentage of mapped reads, but not leading to greater overall coverage (Table 1). Many of the mapped reads had only short segments of similarity, although there were a number of 7-10kb reads with matches along much of their length (Figure 3A). However, the mean identity was low for each mapped read, with fewer than one in ten sites being correctly sequenced (Table 1). The major sources of error in MinION data were indels, particularly insertions that introduce spurious data (Figure 3B). Most of these errors do not appear systematic, and the consensus sequence can be called in most cases using 16x coverage, although even deeper coverage would be necessary for an error-free haploid sequence. The high rates of indels may be due to thermodynamic noise causing uneven movement of the DNA strand through the pore, resulting in regions without signal, or with spurious signal. The reliance on a HMM may also make this method sensitive to reaction conditions, which may be different from those used in the model's training.

Results were far worse for the snake cDNA library amplicon sequencing run, which generated only 1,429 1D reads (1.0 mb), in the course of 24 hours (and just 16 2D reads). Reads were mapped to a reference assembly of the *P. flavoviridis* transcriptome (Aird *et al.* 2013), but few 1D reads aligned (10 by BLASR and 21 by BLASTN). In an attempt to possibly correct errors in the MinION data, we also mapped GAII reads from Aird *et al.* (2013) to the MinION reads. This was not a successful strategy; most of the mapped Illumina reads were localized to relatively short regions within MinION reads that had homology to the template, with most of the read having no coverage. Despite aggressive alignment settings, permitting a divergence of 12% between template and read, overall alignment rate was only 1.4%. Poor mapping performance by the MinION, was not due to incorrect transcriptome assembly, as the most abundant proteins were previously validated by mass spectrometry (Aird *et al.* 2013). Only 5 of the 22 BLAST hits from MinION reads to NCBI's nr database actually matched snake sequences, though not necessarily the correct species, and one read matched a synthetic spike-in added to the total RNA (See Supplementary Methods). Curiously, the GC content of the MinION reads was close to that of the lambda reference even for the snake libraries, despite the fact that the transcript assembly is highly AT biased (Appendix S1). This suggests that the base caller model may be over-trained on lambda genomic sequence.

Utility and possible applications

Sequencing by synthesis currently dominates molecular ecology (Narum *et al.* 2013). One of the great potential advantages of having a portable low-cost sequencer, such as the MinION, and of simple library prep, is that lab work could move closer to the field. Many applications would benefit from shortening the time from specimen collection to first sequence data. For instance, mobile labs could conduct DNA associated with disease outbreaks in the field, or teams deployed at a remote field station could use molecular barcoding to track endangered species. Provided that there is an Internet connection plus access to basic lab facilities and refrigeration, these applications might actually work with the MinION. Given an anonymous sample, the MinION data could be matched to a given reference, and may work for some diagnostic applications. With the use of a spike-in control, the amount of the target sequence in the sample might even be quantifiable, although given that the number of reads from the MinION is relatively low, it will not have sufficient dynamic range to resolve anything but course-grained quantitative differences.

Templates that are difficult to amplify, such those high in GC content or containing inverted repeats, pose problems for sequencing by synthesis platforms (Nakamura *et al.* 2011). Because it does not rely on PCR, Oxford Nanopore's technology may be useful for such samples. Direct sequencing could also benefit a range of molecular ecology applications, such as sequencing of gut contents without PCR-introduced biases, which would be useful for trophic ecology (Andrew *et al.* 2013). Currently, the low number of reads and relatively high DNA input requirements (1 μg), offset any benefits of eliminating PCR. If these parameters can be improved, the MinION might become a useful tool for direct sequencing of field-collected samples.

Because of its extraordinarily high error rates, the current iteration of Oxford Nanopore technology is close to useless for genotyping applications, which account for much of genomics research. Together with other kinds of data, such as more accurate Illumina reads, it may assist genome assembly through scaffolding, similar to the way Pacific Biosciences reads are used now (Quail *et al.* 2012). However, our attempt to do this has not been successful, because large regions of a MinION read have no homology to the template. Although there is a need for a low-cost scaffolding solution, the yield and accuracy of the MinION will need to increase dramatically for this to work for large genomes, where scaffolding is most needed. Scaffolding might even work without an overall decrease in error rate, but reads would need to contain multiple regions homologous to the template to be present, allowing different contigs to be joined. Because of their length, Oxford Nanopore reads may also be useful in phasing Illumina data. However, a series of niche applications in symbiosis with another platform may not make this technology worthwhile, or even commercially viable.

Conclusions and outlook

The current iteration of the MinION is not ready for routine use. A 36-hr run was largely able to reconstruct the sequence of 48kb-lambda phage by mapping (Figure 3). By contrast, it would take more than 100 MinION-days (and 50 or more flow cells) to sequence the 4.6mb genome of its host bacterium (*Escherichia coli*) (Blattner *et al.* 1997) to the same depth. By comparison, existing desktop platforms, such as the 454 GS Junior and the Illumina MiSeq, can accomplish this in two runs, and about 1% of a run, respectively. The snake cDNA data were useless for any practical purpose. Poor performance does not automatically damn an early version of a technology, as all next-generation sequencing platforms started with major shortcomings, typically short reads. Some, most notably Illumina, have managed to overcome them; others like Roche 454 and Helicos, have not, and have gone bankrupt, or have been shut down. So, how much better does the MinION need to be, and is it likely to get there?

We can use the evolution of other platforms as a guide. Yields have increased across the board, in the case of Illumina technology, by three orders of magnitude, from a gigabase to a terabase. By contrast, platform error rates have not improved comparably. In addition to improvements in sequencing chemistry, which gradually enhance performance, platforms like Pacific Biosciences mitigate errors by re-sequencing the same DNA fragment multiple times (Travers *et al.* 2010). Such strategies depend on the topology of the template and will be much more difficult with a pore-based technology, although Oxford Nanopore increases its accuracy four-fold with 2D sequencing (Table 1). It seems likely that Oxford Nanopore can greatly increase the yield by increasing the number of pores used for sequencing. However, since less than 1% of the raw base calls were identical to the reference, an improvement of at least two orders of magnitude is in order for the technology to find a wide range of uses. If the error rates will be lowered so drastically, such an improvement would be unprecedented in the history of next-generation sequencing. Then again, with a fundamentally new technology, anything seems possible.

Acknowledgements. We would like to thank Oxford Nanopore for including us in the MinION Access Program. We are grateful to Miguel Gray-Lopez for assistance with bioinformatics, and to Steven D. Aird for providing the *P. flavoviridis* RNA and for comments on the manuscript. We thank B. Faircloth and two other reviewers for comments on an earlier draft. This work was funded by the Okinawa Institute of Science and Technology.

Data accessibility. Bioinformatic scripts used in the analysis can be found on GitHub (https://github.com/mikheyev/MinION-review). Data to be used with these scripts are available on DataDryad (doi:10.5061/dryad.5p0c3)

References

Aird SD, Watanabe Y, Villar-Briones A *et al.* (2013) Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). *BMC Genomics*, **14**, 790.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.

Andrew RL, Bernatchez L, Bonin A *et al.* (2013) A road map for molecular ecology. *Molecular Ecology*, **22**, 2605–2626.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Blattner FR, Plunkett G, Bloch CA *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453–1462.

Carneiro MO, Russ C, Ross MG *et al.* (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, **13**, 375.

Chaisson MJ, Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

Desai AN, Jere A (2012) Next-generation sequencing: ready for the clinics? *Clinical genetics*, **81**, 503–510.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Kitzman JO, Mackenzie AP, Adey A *et al.* (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnology*, **29**, 59–63.

Korlach J, Bjornson KP, Chaudhuri BP *et al.* (2010) Real-time DNA sequencing from single polymerase molecules. *Methods in enzymology*, **472**, 431–455.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Loman NJ, Misra RV, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

McKernan KJ, Peckham HE, Costa GL *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**, 1527–1541.

Nakamura K, Oshima T, Morimoto T *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, **39**, e90.

Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–2847.

Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, **27**, 847–850.

Quail MA, Smith M, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Rothberg JM, Hinz W, Rearick TM *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology*, **162**, 729–773.

Stoddart D, Heron AJ, Mikhailova E *et al.* (2009) Single-Nucleotide Discrimination in Immobilized DNA Oligonucleotides with a Biological Nanopore. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **106**, 7702–7707.

Timp W, Comer J, Aksimentiev A (2012) DNA base-calling from a nanopore using a Viterbi algorithm. *Biophysical journal*, **102**, L37–9.

Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, **38**, e159.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed June 26, 2014.

Table 1. Sequencing and mapping statistics for one MinION run of the lambda genome. “Aligned bases” include all those without soft clipping, but not necessarily with the correct sequence. “Mapped read identity” refers to the percentage of bases in an aligned read that exactly match the reference. Overall, considering that the vast majority of the reads did not map, much less than one percent of the overall sequence generated by the MinION is identical to the reference. We do not present quality score data, since they are derived by solving the HMM and are back calibrated to the data; they do not correspond to quality scores produced by other platforms. BLASR and BLASTN produced similar results overall, although the total number of reads was higher with BLASTN, which tended to find shorter matches. Means for aligned bases per read are give ± standard deviation.

	Workflow	
	1D	**2D**
Reads	29,458	11,094
Total sequenced bases	155,370,698	55,854,289
Reads mapped by BLASTN	7,997 (27%)	2,746 (25%)
Reads mapped by BLASR	3,472 (12%)	909 (8%)
BLASR aligned bases	648,756	808,539
BLASTN aligned bases	687,148	772,443
BLASR aligned bases/read	187±243	890±1932
BLASTN aligned bases/read	23±73	70±423
BLASR coverage (fold)	13.5	16.8
BLASTN coverage (fold)	14.3	16.1
BLASR mapped read identity (%)	2.2%	8.9%
BLASTN mapped read identity (%)	0.4%	1.1%

Figure 1. Buzz surrounding Oxford Nanopore sequencing technology as measured by Google Trends based on search volume. The plot shows a continuing increase in interest in illumina's sequencing technology, and rapid decline in the popularity of Roche's. Oxford Nanopore's announcement of its sequencing technology and the MinION went viral, when it was unveiled in early 2012, but interest gradually decreased as the actual platform failed to materialize for the next two years. The search was conducted by searching for platform names as listed in the legend, plus the word "sequencing."

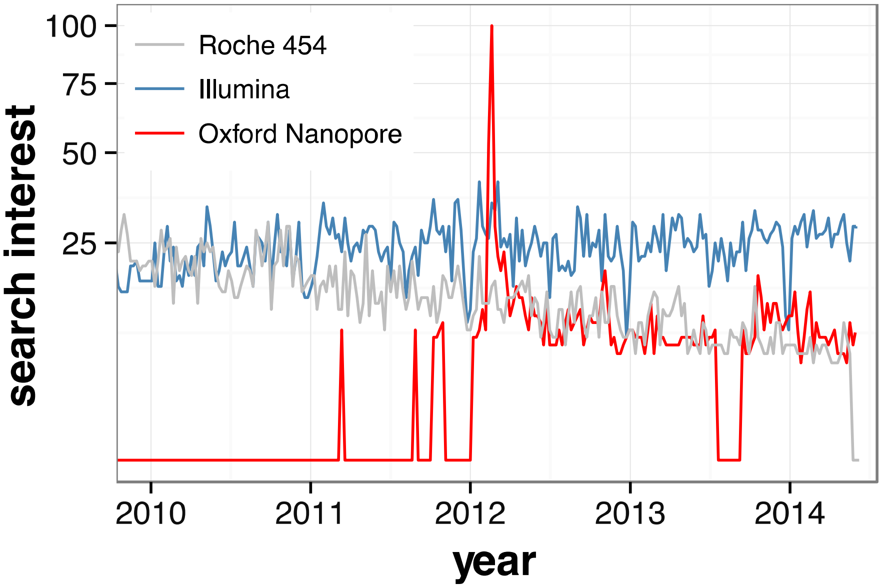


Figure 2. Read length distribution of raw data visualized as a stacked histogram. A black trace overlays the DNA fragment size distribution of the input DNA, as measured on an Agilent Bioanalyzer. Read lengths were comparable for both workflows, although the 2D workflow produced only about one-third as many reads. Sequenced read lengths were about 2kb shorter than those in the input library. Although data acquisition software indicated that 7kb reads were the most common class of reads during data acquisition, they were under-represented in the final analysis.

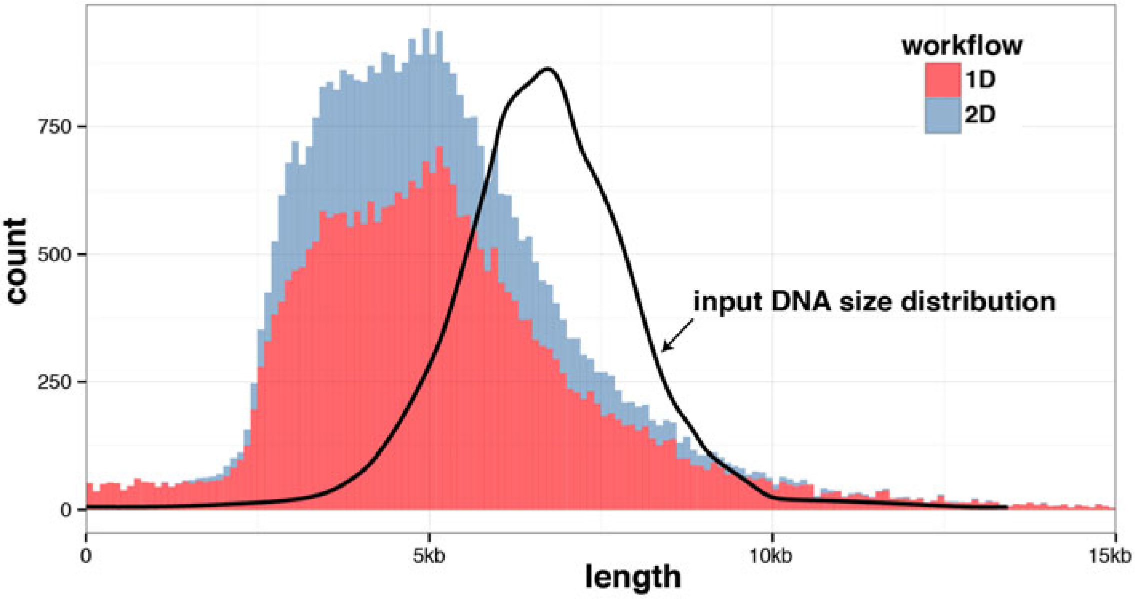
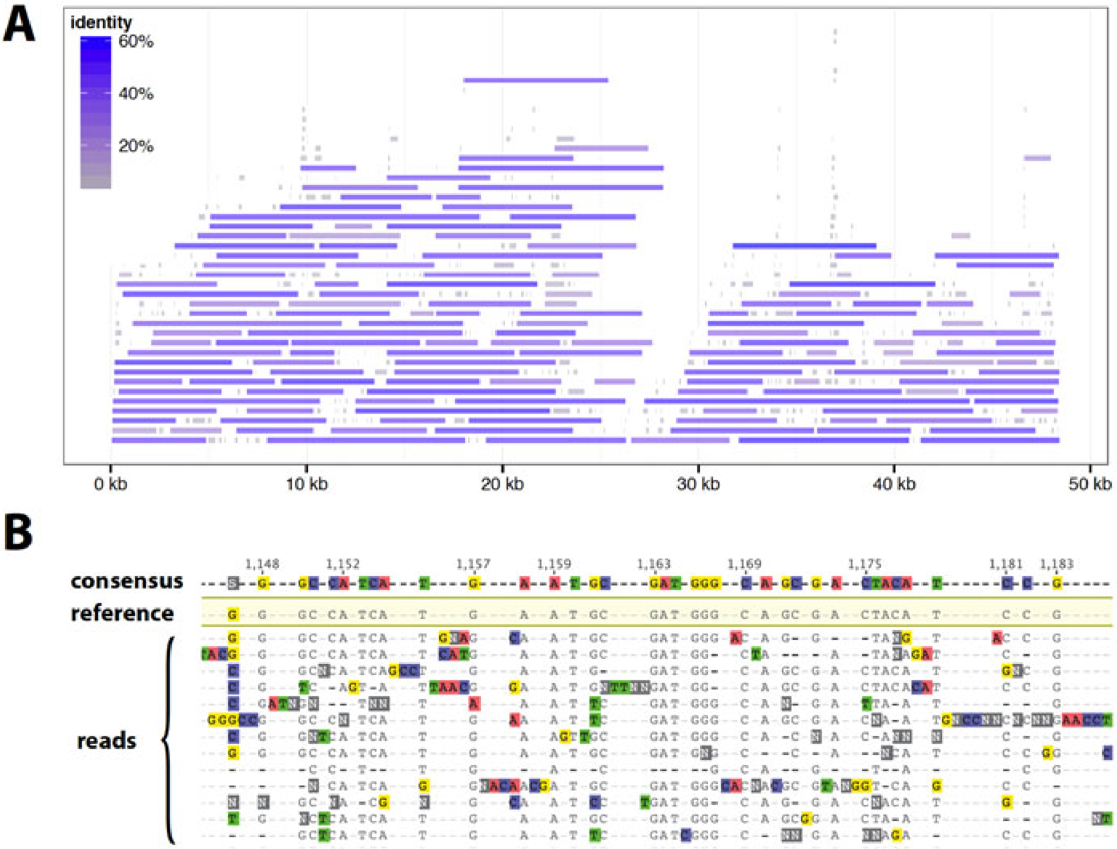


Figure 3. Alignment of 909 MinION reads processed by the 2D workflow to the lambda phage reference genome. (A) Positions of reads are color-coded by their identity to the reference. There were relatively few long reads with more than 20% identity; however some of the mapped reads reached nearly 10kb in length. The longest read from the 2D workflow was 21kb, far longer than the mean fragment size. (B) Representative alignment over a short genomic region, highlighting errors made by the MinION. Most errors resulted from insertions and deletions, resulting in low overall sequence identity. Insertions, in particular, are associated with spurious sequence that does not match the template. For insertions within BLAST-aligned reads, all the nucleotides (A,C,G,T and N) were approximately equally distributed, but with a slight excess of Gs and Cs (55%), relative to As and Ts. It was possible to call a consensus sequence at most, but not all of the sites.



An evaluation of Oxford Nanopore’s MinION platform

```
library(ggbio)
library(GenomicRanges)
library(GenomicAlignments)
library(biovizBase)
library(seqinr)
```

Summary statistics for aligned lambda reads

```
param <- ScanBamParam(what=c("cigar"),tag=("NM"))
bam1 <- readGAlignments("../data/lambda/align/1d.bam", param=param)
bam2 <- readGAlignments("../data/lambda/align/2d.bam", param=param)
cigar.matrix1 <- cigarOpTable(bam1@cigar)
values(bam1)$indels<-rowSums(cigar.matrix1[,2:3])/rowSums(cigar.matrix1[,1:3])
values(bam1)$identity <- (cigar.matrix1[,1]-values(bam1)$NM)/qwidth(bam1)
sum(cigar.matrix1[,1]) #mapped bases
```

```
## [1] 648756
```

```
mean(cigar.matrix1[,1]) #average mapped read length
```

```
## [1] 186.9
```

```
sum(cigar.matrix1[,5]) #soft clipped bases
```

```
## [1] 18766371
```

```
mean((cigar.matrix1[,1]-values(bam1)$NM)/qwidth(bam1)) # mean identity
```

```
## [1] 0.02291
```

```
mean(rowSums(cigar.matrix1[,2:3])/rowSums(cigar.matrix1[,1:3])) # mean indels
```

```
## [1] 0.1997
```

```
cigar.matrix2 <- cigarOpTable(bam2@cigar)
values(bam2)$indels<-rowSums(cigar.matrix2[,2:3])/rowSums(cigar.matrix2[,1:3])
values(bam2)$identity<-(cigar.matrix2[,1]-values(bam2)$NM)/qwidth(bam2)
sum(cigar.matrix2[,1]) #mapped bases
```

```
## [1] 808539
```

```
mean(cigar.matrix2[,1]) #average mapped read length
```

```
## [1] 889.5
```

```
sum(cigar.matrix2[,5]) #soft clipped bases
```

```
## [1] 4151416
```

```
mean(cigar.matrix2[,5]/sum(cigar.matrix2[,5])) #soft clipped %
```

```
## [1] 0.0011
```

```
mean(rowSums(cigar.matrix2[,2:3])/rowSums(cigar.matrix2[,1:3])) # mean indels
```

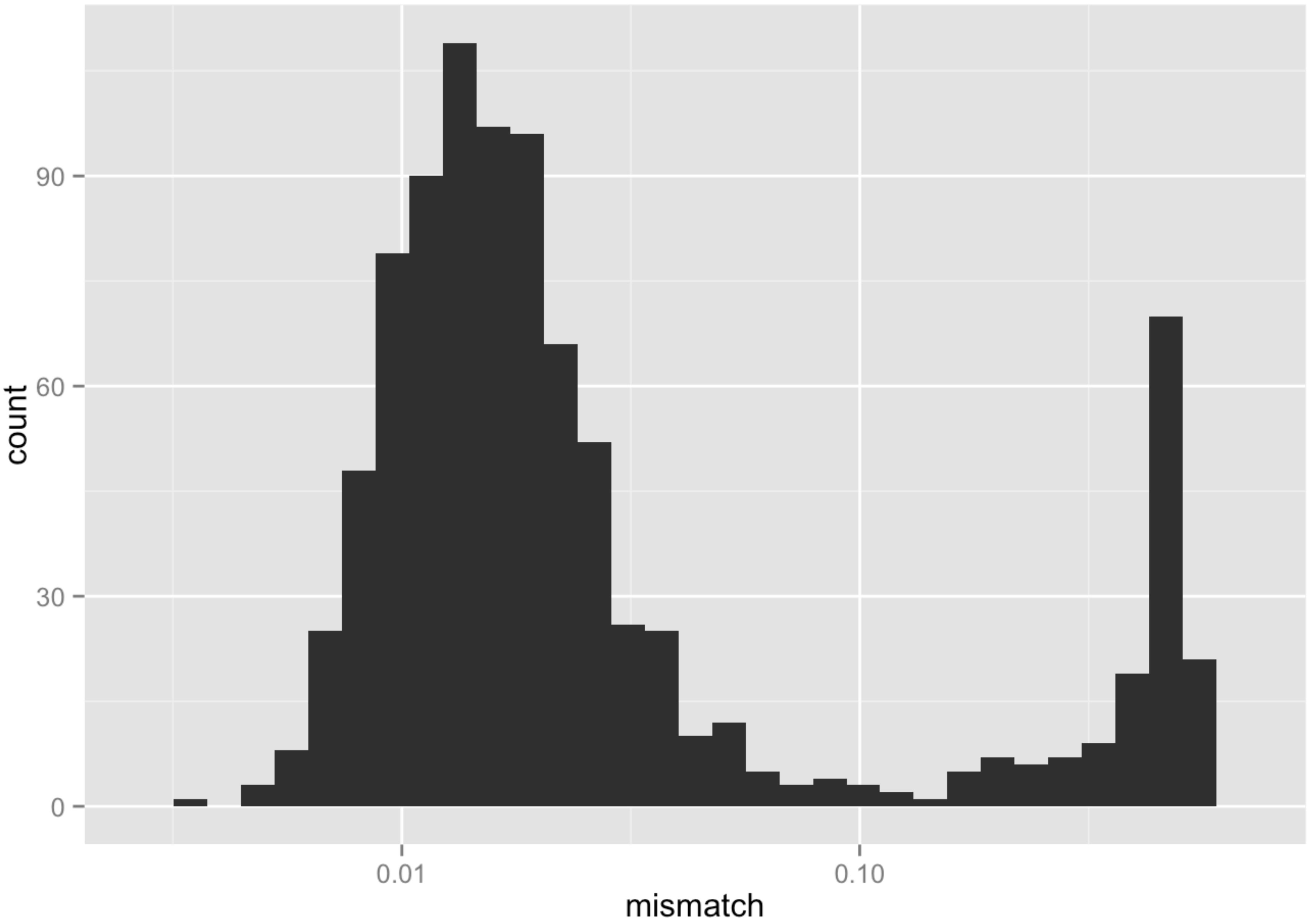
```
## [1] 0.1886
```

```
mean((cigar.matrix2[,1]-values(bam2)$NM)/qwidth(bam2)) #mean identity
```

```
## [1] 0.08926
```

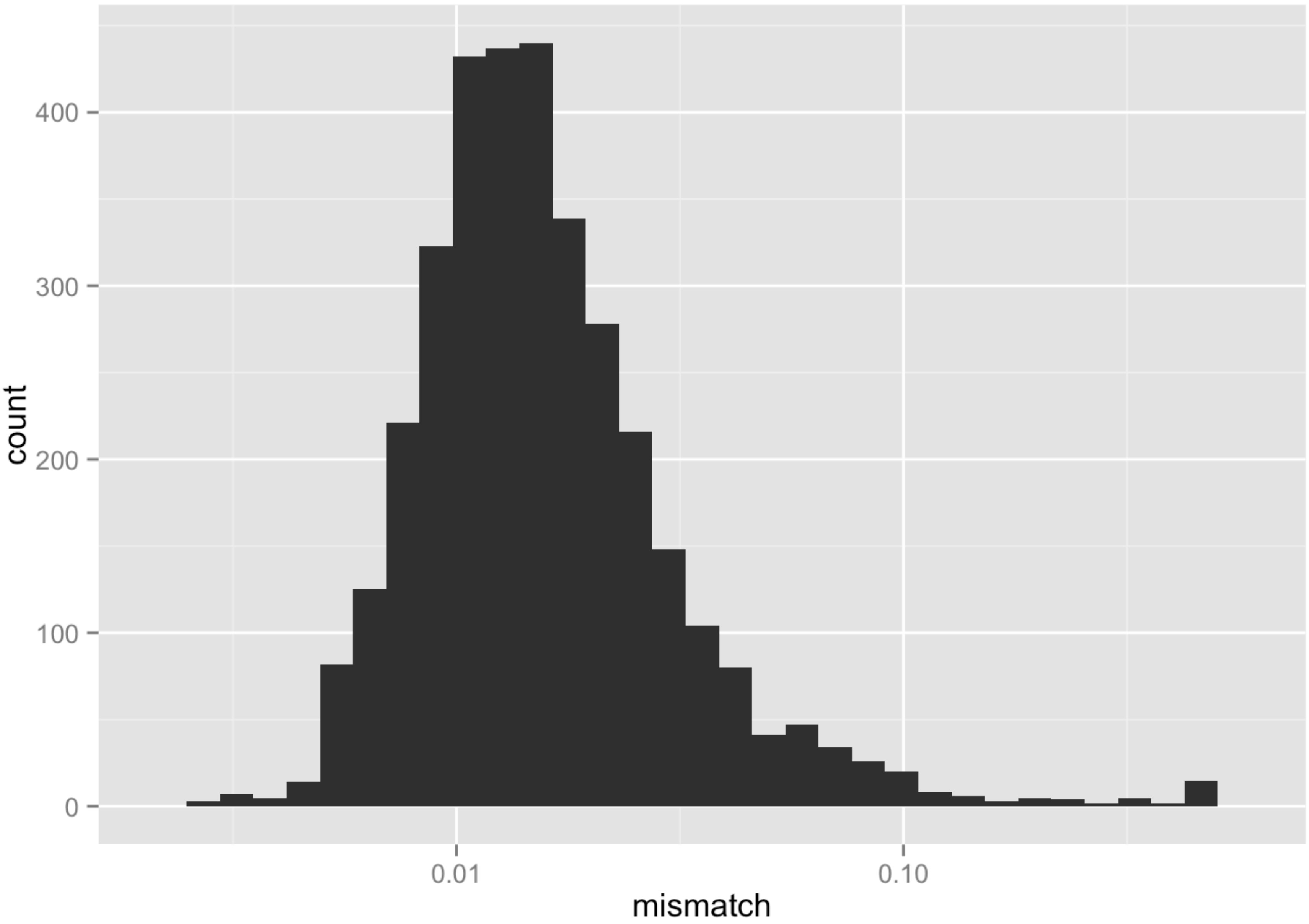
```
ggplot(data.frame(mismatch=(cigar.matrix2[,1]-values(bam2)$NM)/rowSums(cigar.matrix2)))+geo
m_histogram(aes(x=mismatch)) +scale_x_log10() #mean mismatches for aligned sections of read
s
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



```
ggplot(data.frame(mismatch=(cigar.matrix1[,1]-values(bam1)$NM)/rowSums(cigar.matrix1)))+geom_histogram(aes(x=mismatch)) +scale_x_log10() #mean mismatches for aligned sections of reads
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

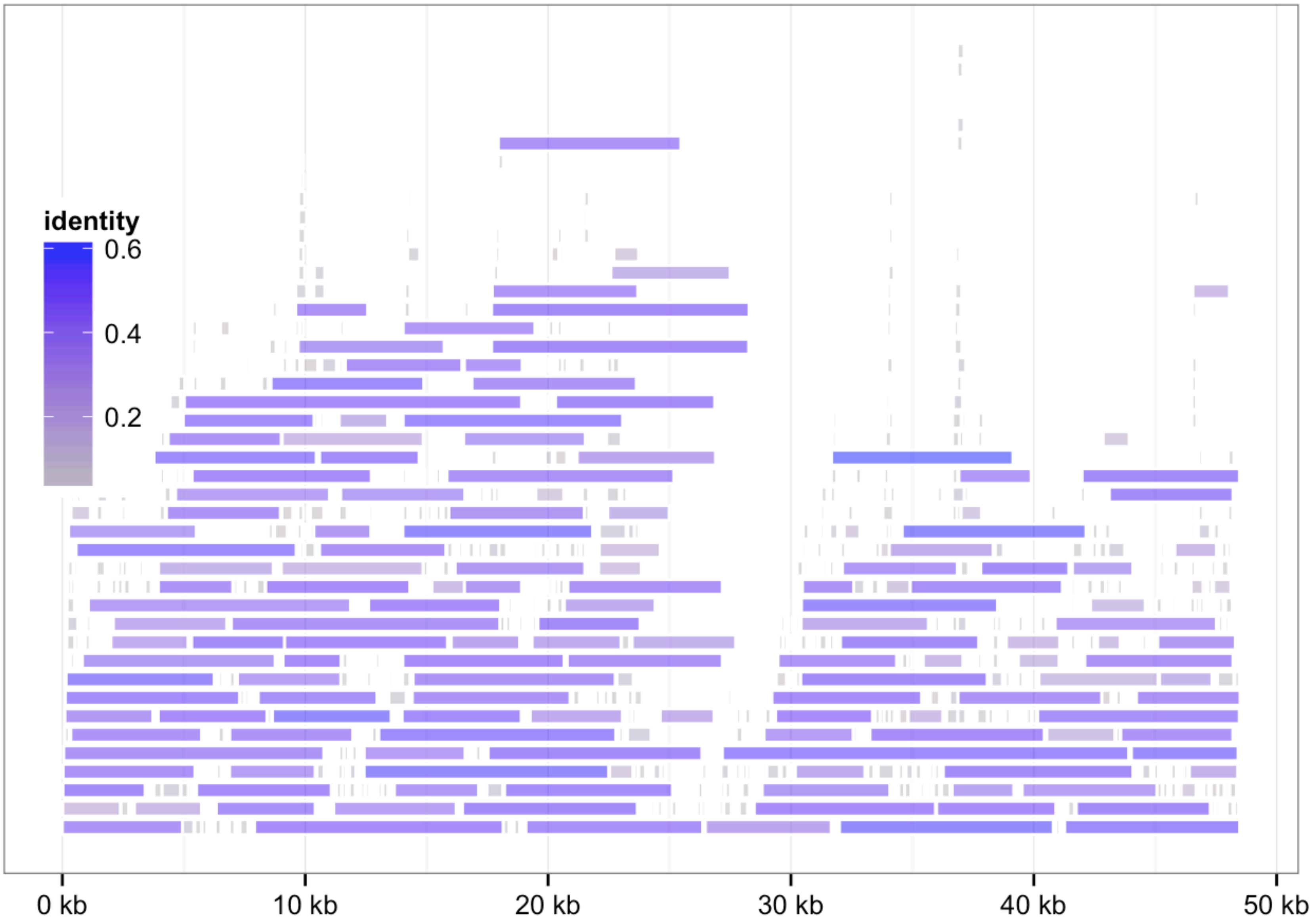


```
bam1 <-as(bam1, "GRanges")
bam2 <-as(bam2, "GRanges")
p1 <- autoplot(bam1,geom="rect",aes(color=factor(1),alpha=0))+theme_bw()+scale_color_manual
(values=c("white"))+guides(colour=FALSE,alpha=FALSE)
```

```
## Warning: the condition has length > 1 and only the first element will be
## used
```

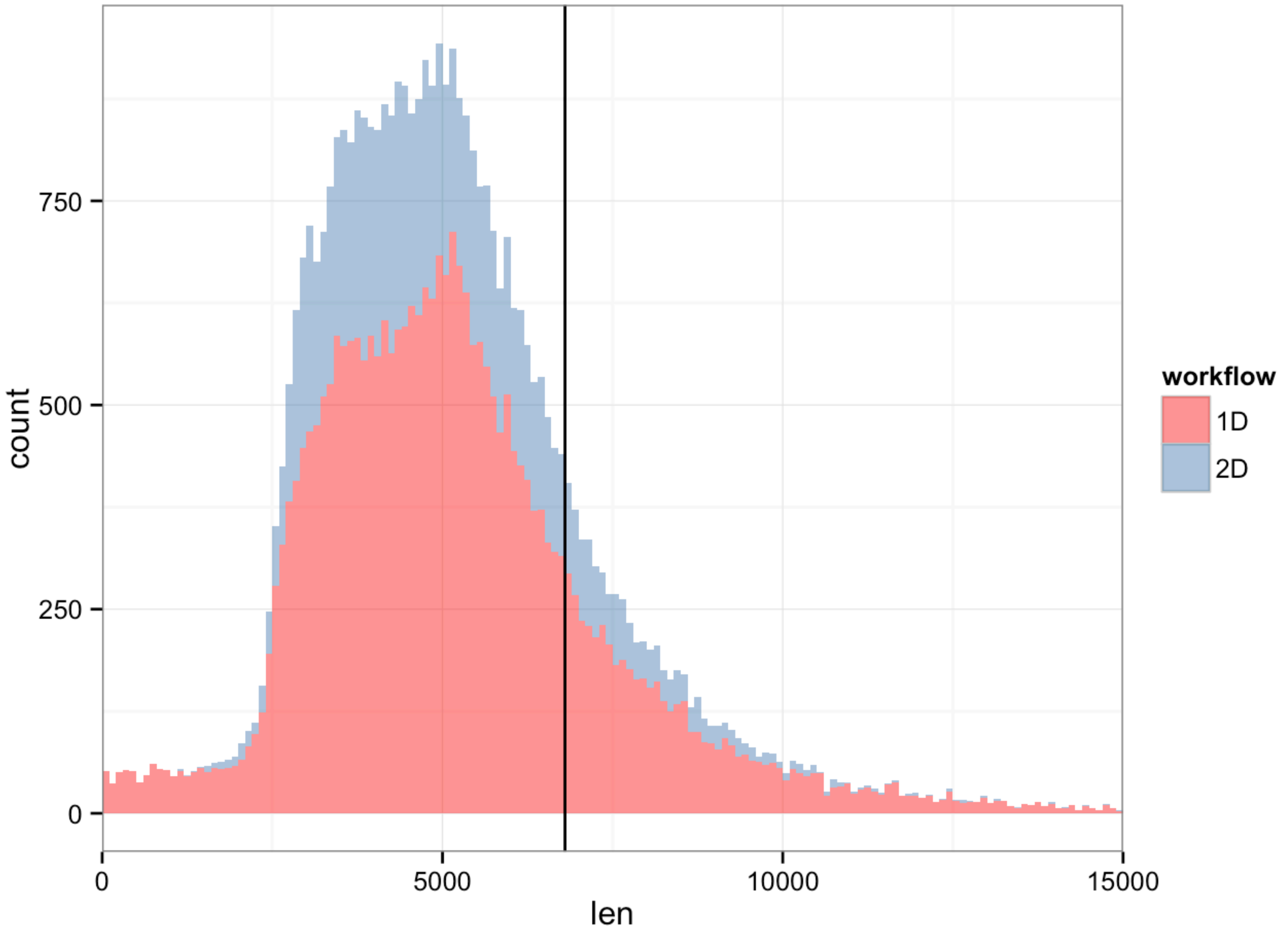
```
autoplot(bam2,geom="rect",aes(color=factor(1),alpha=0,fill=identity))+theme_bw()+scale_colo
r_manual(values=c("white"))+guides(colour=FALSE,alpha=FALSE)+scale_fill_gradient(low="grey"
, high="blue")+theme(legend.justification=c(0,0), legend.position=c(0,.4))
```

```
## Warning: the condition has length > 1 and only the first element will be
## used
```



Read length distribution in 2D data from lambda phage

```
len <- read.csv("../data/processed/lengths.csv",header=T)
ggplot(len,aes(x=len,fill=workflow))+geom_histogram(binwidth=100,alpha=.5)+xlim(0,15000)+theme_bw()+scale_fill_manual(values=c("red","steelblue"))+geom_vline(xintercept=6800)
```

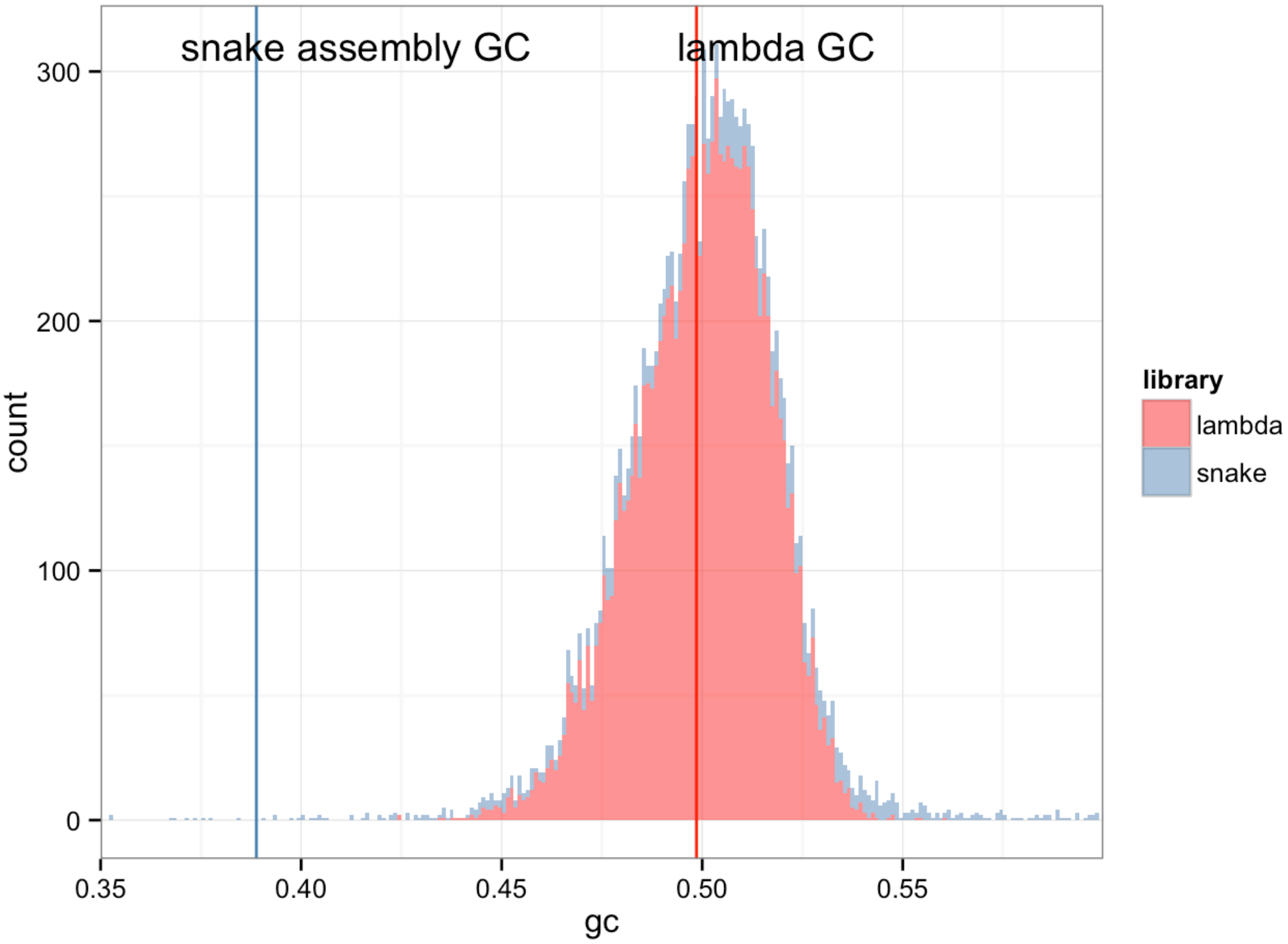


```
#ggsave("../plots/readlength.pdf")
```

Examine GC content of the reads and compare it to the reference

```
lambda_ref <- read.fasta(file = "../ref/NC_001416.fa")
lambda <- read.fasta(file = "../data/lambda/2d.fasta")
snake_ref <- read.fasta(file = "../ref/protobothrops_ref.fasta")
snake_ref_gc <- mean(sapply(snake_ref,GC))
snake <- read.fasta(file = "../data/snake/1d.fasta")
dat <- data.frame(gc = c(sapply(snake,GC),sapply(lambda,GC)),library = c(rep("snake",length(snake)),rep("lambda",length(lambda))))
ggplot(dat,aes(x=gc,fill=library))+geom_histogram(binwidth=0.001,alpha=.5)+xlim(0.35,.6)+theme_bw()+scale_fill_manual(values=c("red","steelblue"))+geom_vline(xintercept=GC(lambda_ref[[1]]),color="red")+geom_vline(xintercept=snake_ref_gc,color="steelblue")+annotate("text",x=GC(lambda_ref[[1]])+.02,y=310,label="lambda GC")+annotate("text",x=snake_ref_gc+.025,y=310,label="snake assembly GC")
```

```
## Warning: position_stack requires constant width: output may be incorrect
```



Read length summary statistics

```
summary(sapply(lambda,length))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1150    3760    4770    5030    6030   21200
```

```
summary(sapply(snake,length))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5     198     427     688     673   29400
```

Supplementary Methods

Lambda phage genomic DNA preparation and sequencing

Library preparation was accomplished as per manufacturer instructions provided for the Oxford Nanopore MAP cycle 1 Burn-in Genomic DNA Sequencing Kit. In brief, lambda genomic DNA (NEB) was sheared using a Covaris G-tube to fragments approximately 7kb in length. The fragments were end-repaired, A-tailed, and ligated to sequencing adaptors. The ligation reactions were purified using 0.4x Agencourt AMPure XP beads. Subsequent steps involved tether annealing and a 30-minute library conditioning step, which attaches the motor protein. Note: the most recent version of the protocol (Version MN001_1115_revI_06May2014) recommends overnight conditioning to increase the number of 2D reads. Finally, ‘fuel mix’ was added to the conditioned mixture prior to sequencing.

Protobothrops flavoviridis venom gland cDNA preparation and sequencing

Total RNA isolation and first strand cDNA synthesis

Total RNA was isolated from venom gland of *Protobothrops elegans* using Qiagen RNeasy Plus Mini Kit. One µg total RNA was used for first strand cDNA synthesis following Aird *et al.* (Aird et al. 2013) with slight modifications: (1) ERCC92 ExFold spike-ins (Life Technologies) were added according to the manufacturer’s instructions (2) oligo dT primer: 5’-AATTGCAGTGGTATCAACGCAGAGCGGCCGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN and template switching RNA oligo: 5’-

AAGCAGUGGUAUCAACGCAGAGUACAUGGG were used instead; (3) 20 µl reaction volume was used, (4) 80 µl water was added to dilute the first strand cDNA.

The first PCR amplification using specific primers tailed with Nanopore universal sequences

The 50 µl PCR reaction consisted of 1x Phusion HF buffer (Thermo Scientific), 200 µM dNTP (Promega), 0.5 µM forward primer: 5'-GCCATCAGATTGTGTTTGTTAGTCGCTTCCGGAGAAATTGCAGTGGTATCAACGCAGAG, 0.5 µM reverse primer: 5'-GCTTACGGTTCACTACTCACGACGATGATAGAGGCAAGCAGTGGTATCAACGCAGAGTACA, 0.5 µl of 2 U/µl Phusion DNA polymerase (Thermo Scientific) and 10 µl diluted first strand cDNA. Eight 50 µl PCRs were set up. Note that the primers include a template sequence, an adaptor sequence, and, rather optimistically, a barcode. The PCR was carried out with the following conditions: initial denaturation at 98°C for 30 seconds, with 10 cycles of denaturation at 98°C for 10 seconds, 68°C for 6 minutes, followed by final extension at 72°C for 10 minutes. The PCRs were concentrated with Amicon Ultra-0.5 column (Millipore). The concentrated PCR product was purified by solid phase reversible immobilization using 17% PEG/NaCl/Tris and Dynabeads MyOne Carboxylic Acid (Tin et al. n.d.). The concentration of the DNA was measured with Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen).

Developer kit I-Amplicon sequencing preparation of snake cDNA

The second PCR was carried out using supplied Oxford Nanopore primers according to the manufacturer's manual of Nanopore Developer Kit I-Amplicon Sequencing (DEV-MAP001) except that Phusion DNA polymerase was used.

The rest of the library preparation procedures were followed according to the protocol of the same user manual. In brief, deoxyadenosine triphosphate (dATP) was added to the 3'ends of the purified PCR products to make them compatible with the hairpin adaptor in the later part of the process. The PCR fragments were then tethered within a binding buffer to the motor protein. Hairpin adaptors were ligated to the PCR fragments. Excess adaptors were removed by exonuclease digestion. After that, the steps were the same as in the genomic library preparation. The library was conditioned and mixed with a 'fuel mix', before loading into the sequencer.

Data analysis

FAST5 files produced by the MinION were base called using Oxford Nanopore's cloud-based Metrichor software using 2D workflow (v 1.9.2). The scripts used in the analysis and instructions for obtaining the raw data are available at https://github.com/mikheyev/MinION-review, but are summarized below. The base called files were converted to fastq format using a custom script and mapped to the reference using BLASR (Chaisson & Tesler 2012) and BLASTN. For the lambda genome reference we used the NCBI accession number NC_001416, and for *P. flavoviridis*, we used the assembly from Aird *et al.* (Aird et al. 2013). Reads were not trimmed for quality or possible adaptor contamination prior to mapping. In addition to mapping the MinION reads to reference sequence, we used stampy aligner in sensitive mode, permitting a 12% divergence (Lunter & Goodson 2011) to map illumina PE100 GAII reads from the snake transcriptome (Aird et al. 2013) Oxford Nanopore data. In addition to reference-based mapping, we used BLASTN to search MinION-generated reads against a local copy of NCBI's nr nucleotide database.

References

Aird, S.D. et al., 2013. Quantitative high-throughput profiling of snake venom gland transcriptomes and proteomes (*Ovophis okinavensis* and *Protobothrops flavoviridis*). *BMC Genomics*, 14(1), p.790.

Chaisson, M.J. & Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1), p.238.

Lunter, G. & Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6), pp.936–939.

Tin, M.M.Y. et al., Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype-calling accuracy. *Molecular Ecology Resources*.