DIRECT OLIGONUCLEOTIDE SEQUENCING WITH NANOPORES

SACHIN CHALAPATI, CONOR A CROSBIE, DIXITA LIMBACHIYA, NIMESH C PINNAMANENI

Helixworks Technologies, Environmental Research Institute, University College Cork, Cork, Ireland. T23 XE10

ABSTRACT

Third-generation DNA sequencing has enabled users to sequence long, unamplified DNA fragments with minimal sample and library preparation steps. Sequencing single-stranded nucleic acids directly without amplification or by ligating a spacer strand are challenging, as the single-strand species are poor templates to add the sequencing adapters. Sequencing ssDNA or RNA directly gives valuable insights like base-level modifications and degradation levels along with saving valuable time and resources. Biological nanopores used by Oxford Nanopore Technologies process the target strands at a single-strand level, although the typical samples sequenced are double-stranded or converted into double-strand. We have identified that the MinION platform from Oxford Nanopore can perform sequencing of short, single-strand oligonucleotides directly without amplification or second-strand synthesis by performing an annealing step before library preparation. Short 5' phosphorylated oligos when annealed to an adapter sequence can be directly sequenced in the 5' to 3' direction via nanopores, the adapters were designed to bind to the 5' end of the oligos and leave a 3' adenosine overhang after binding to their target. The 3' adenosine overhang of the adapter and the terminal phosphate makes the 5' end of the oligo to be analogous to an end-prepared dsDNA, rendering it compatible with ligation-based library preparation for sequencing. An oligo-pool containing 42,000 orthogonal sequences of 120 bp length were sequenced using the method and 37,265 of the total sequences were recovered with high accuracy.

While analyzing the raw data, we had interesting observations. In our raw data, we have identified that empty signals can be wrongly identified as a valid read by the MinION platform and sometimes multiple signals containing several strands can be fused into a single read by the platforms segmentation faults.

We believe that this method could enable novel applications of nanopore sequencing in DNA data-storage systems where short oligonucleotides function as the primary information carriers.

Keywords - ssDNA, RNA, oligonucleotide, oligo, single-stranded DNA, phosphorylation, T4 PNK, ligation, AMX, T4 DNA ligase, sequencing, ONT, MinION, Guppy, BLASTN, nanopore, helicase, dsDNA, ssDNA, basecalling, squiggle

INTRODUCTION

DNA sequencing has become a staple tool in biology and is getting affordable and more accessible to small labs and individual researchers in the past decade. Oxford Nanopore Technologies, with its biological nanopore-based sequencing technology, has opened the market wide open by releasing a $1000 sequencing platform – The MinION [1]. The MinION platform has the capability to sequence both amplified and non-amplified double-stranded DNA (dsDNA) [2] and direct RNA in the 3' to 5' direction using Poly-A tail capture [3]. Direct sequencing of short, single-stranded oligonucleotides has been regarded as a challenge due to nanopore chemistry, pore design and basecalling [4], however, attempts have been made to overcome these challenges by performing circularization [5]. Direct sequencing short, single-strand nucleic acid species without a polymerase or ligation step has not been evident in previous research. In this article we propose a method to perform direct sequencing of short single strand

oligonucleotides that can be leveraged by different applications such as DNA based data storage systems and direct RNA sequencing.

In DNA based information systems, the oligonucleotides serve the purpose of the information carriers and to rapidly sequence the oligos to extract the encoded data [6]. Several encoding and compressions techniques are widely used during the design of the oligonucleotides for DNA data storage purposes to increase the data capacity and to deal with amplification and sequencing issues. The possibility to sequence oligonucleotides directly without performing a PCR step or performing a ligation step enables users to design the oligos to have only one priming region on the 5' end and free up the reverse priming region. This increases the encoding space available to the users and opens the possibilities for new encoding schema and DNA data storage architectures. In this work, we propose a method which is incredibly fast when compared to PCR-based sequencing strategies [7] with hands-on time as low as 5 minutes and sequencing time of just 20 minutes for pre-phosphorylated oligos like INS3 and EINS3 shown in Figure 1. We have identified that the Oxford Nanopores MinION platform is capable of sequencing single-strand templates directly without the need for a complementary strand or to have a spacer strand to increase the strand length.

In this work, we solve this challenge of direct oligonucleotide sequencing by performing a simple annealing step before the library preparation step. The setup starts with a phosphorylation step using T4 Polynucleotide Kinase (PNK) that adds a 5' phosphate to the target oligos. The phosphorylated oligos are annealed to an adapter sequence that binds to their 5' end. The adapter sequences are designed to have a melting temperature of ~65°C to the 5' end of the target oligos; and when annealed to their targets, the adapter strands have an adenosine overhang at their 3' end as detailed in Figure 2. In the annealed state with the adapter sequence, the 5' end of the oligos are analogous to an end-prepared dsDNA and are compatible with the AMX sequencing adapters form the ligation sequencing kit (LSK-109) offered by the Oxford Nanopore Technologies [8]. The sequences used to implement the method and related protocols are discussed in the next section.

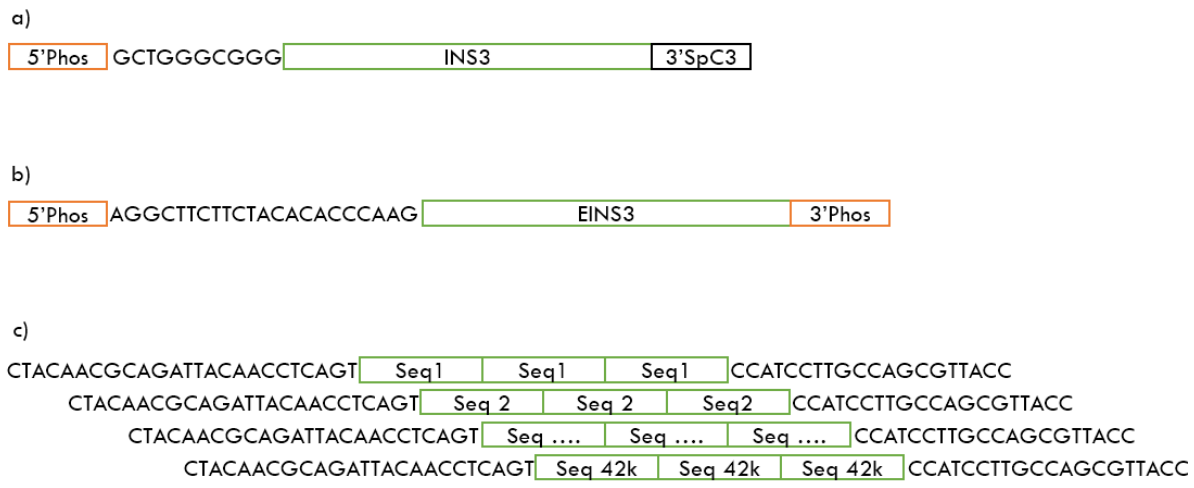

Figure 1. Sequencing templates a) INS3 oligo (175 bp) containing a 5' phosphate and 3' C3 spacer b) EINS3 oligo (200 bp) containing a 5' phosphate and 3' terminal phosphate c) 3xr6 oligo-pool containing 42,000 unique oligonucleotides of 120 bp length, each of these oligo contain a 3x repetitive region of a 25-bp orthogonal sequence. All sequences in the 3xr6 oligo-pool contain 5' and 3' priming regions.

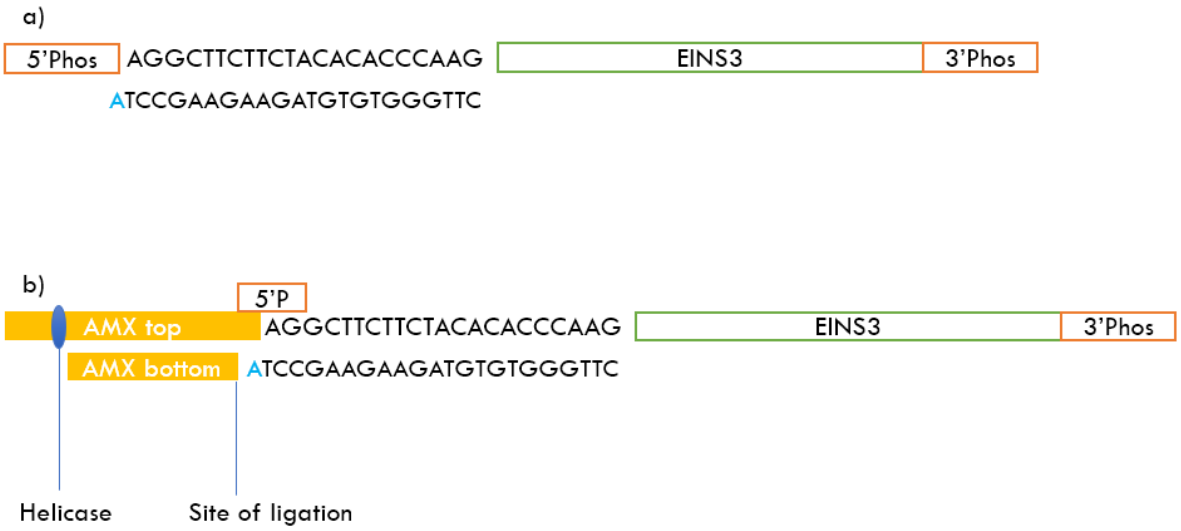

Figure 2. Annealing and ligation steps a) Annealing of the adapter to the 5' end of the EINS3 strand and leaving an adenosine overhang b) Ligation of the AMX sequencing adapter to the adapter + EINS3 strand.

MATERIALS AND METHODS

OLIGONUCLEOTIDES

INS3

/5Phos/GCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCG
TGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCC
GCAGGCAGCCCCACACCCGCCGCCTCCTGCACC/3SpC3/

INS3 is procured from IDT (Ultramer) – Normalized to 100 uM concentration.

INS3 RC

CCCGCCCAGCA

INS3 RC is procured from IDT (Standard desalting) – Normalized to 100 uM concentration.

EINS3

/5Phos/AGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTG
GGCGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTG
GAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGTGAA/3Phos/

EINS3 is procured from IDT (Ultramer) – Normalized to 100 uM concentration

EINS3 RC

CTTGGGTGTGTAGAAGAAGCCTA

EINS3 RC is procured from IDT (Standard desalting) – Normalized to 100 uM concentration.

3XR6 - 42,000 OLIGOS - 120 BP LENGTH

Supplementary table 1.

3xr6 oligo-pool is procured from Twist Biosciences – Normalized to 10 ng/ul concentration.

ARCFP_MOSS5

ACTGAGGTTGTAATCTGCGTTGTAGA

ArcFP_MoSS5 is procured from IDT (Standard desalting) – Normalized to 100 uM concentration.

ENZYMES AND BUFFERS

T4 PNK - NEB - M0201S

Blunt/TA Ligase Master Mix - NEB - M0367S

IDTE - IDT - 11-01-02-02

Qubit ssDNA Assay Kit - ThermoFisher - Q10212

Nuclease Free Water (NFW) – IDT - 11-04-02-01

WASH-KIT

Monarch PDR & DNA Cleanup Kit - NEB - T1030S

SEQUENCING KIT

Ligation Sequencing Kit - Nanopore - SQK-LSK109

INSTRUMENTS

Open qPCR (Single Channel) - ChaiBio - E013101

Microcentrifuge - Dlab - D3024

Qubit 4 - ThermoFisher - Q33226

SEQUENCER

MinION - Oxford Nanopore Technologies (ONT) - MIN-101B

NORMALIZATION AND QUBIT ANALYSIS

1. 3xr6 oligo-pool is normalized to 0.25 uM and verified with Qubit 4 using Qubit ssDNA Assay Kit

2. INS3 and EINS3 are diluted from their stock concentrations to 0.5 uM and verified with Qubit 4 using Qubit ssDNA Assay Kit

3. INS3 RC, EINS3 RC and ArcFP oligos are normalized by the supplier at 100 uM concentration and are diluted to 1 uM concentration.

PHOSPHORYLATION OF 3XR6

1. 3xr6 oligo-pool is normalized to 0.25 uM overall concentration

2. 8 ul of 0.25 uM 3xr6 oligo-pool is phosphorylated in combination with 2.5 ul T4 PNK reaction buffer, 2.5 ul of 10 mM ATP, 0.5 ul (5 units) of T4 PNK and 11.5 ul NFW

3. The phosphorylation reaction is carried out at 37° for 30 minutes

4. Heat inactivation at 65°C for 20 minutes

5. The mixture is washed using Monarch spin columns with the standard oligonucleotide cleanup protocol and the purified DNA is eluted into 9 ul of IDTE

LIBRARY PREPARATION OF INS3

1. A triplicate of the following reaction is performed

2. 0.5 ul of 0.5 uM INS3 is added to 0.5 ul of 1 uM INS3 RC along with 2 ul of NFW

3. The reaction mixtures are heated to 94°C for 2 minutes and gradually cooled to room temperature for annealing

4. 5 ul of AMX from the ligation sequencing kit is added to each of the tubes

5. 5 ul of Blunt/TA mastermix is added to each of the tubes

6. Tubes are incubated at room temperature for 10 minutes

7. Three new MinION flow cells are used to sequence each reaction mix of the triplicate and the sequencing is performed with the standard parameters for 20 minutes

LIBRARY PREPARATION OF EINS3

1. A triplicate of the following reaction setup is performed

2. 0.5 ul of 0.5 uM EINS3 is added to 0.5 ul of 1 uM EINS3 RC along with 2 ul of NFW

3. The reaction mixtures are heated to 94°C for 2 minutes and gradually cooled to room temperature for annealing

4. 5 ul of AMX from the ligation sequencing kit is added to each of the tubes

5. 5 ul of Blunt/TA mastermix is added to each of the tubes

6. Tubes are incubated at room temperature for 10 minutes

7. Three MinION flow cells from the INS3 run are used to sequence each reaction mix of the triplicate and the sequencing is performed with the standard parameters for 20 minutes

LIBRARY PREPARATION OF 3XR6

1. 9 ul elute from the phosphorylation step is 3-way split for triplicate sequencing runs. A triplicate of the following reactions is performed

2. 3 ul of washed, phosphorylated 3xr6 oligo pool is added to 1 ul of 1 uM ArcFP

3. The reaction mixtures are heated to 94°C for 2 minutes and gradually cooled to room temperature for annealing

4. 5 ul of AMX from the ligation sequencing kit is added to each of the tubes

5. 5 ul of Blunt/TA mastermix is added to each of the tubes

6. Tubes are incubated at room temperature for 10 minutes

7. Three new MinION flow cells are used to sequence each reaction mix of the triplicate and the sequencing is performed with the standard parameters for 4 hours

TROUBLESHOOTING

A triplicate of Qubit readings is recommended for reliability. The reaction mixtures are recommended to be washed with magnetic beads after the Blunt/TA ligation step by ONT, we avoided this step to improve yields as the strands are short in length.

TIME TAKEN

PHOSPHORYLATION OF 3XR6

Hands-on-time - 5 minutes

Total reaction time - 60 minutes

LIBRARY PREPARATION OF INS3

Hands-on-time - 5 minutes

Ligation - 10 minutes

Sequencing - 20 minutes

LIBRARY PREPARATION OF EINS3

Hands-on-time - 5 minutes

Ligation - 10 minutes

Sequencing - 20 minutes

LIBRARY PREPARATION OF 3XR6

Hands-on-time - 5 minutes

Ligation - 10 minutes

Sequencing - 4 hours

RESULTS

SEQUENCING OF INS3

The flowcells 1, 2 and 3 yielded 23248, 15917 and 21052 reads respectively after 20 minutes of sequencing. Basecalling was performed using guppy_hac (ver. 3.5.2) [9] model with a CUDA compatible GPU. A total of 289, 483 and 255 reads from flowcells 1 to 3 have passed the default q-score filter. The sequencing and BLASTN results with e-value are shown in Table 1.

Table 1. INS3 sequencing results

	Total reads	*Total reads passing q-score filter*	*Filter pass %*	*All significant matches*	*Significant matches to INS3 (e value < 1e-65)*	*Significant matches to AMXpINS3 (ACGTATTGCTGCTGGGCGGG) "blastn-short"*
Flowcell 1	23248	289	1.24	281	4	693
Flowcell 2	15917	483	3.03	414	11	820
Flowcell 3	21052	255	1.21	228	4	635

SEQUENCING OF EINS3

The flowcells 1, 2 and 3 yielded 23174, 30238, and 30051 reads respectively after 20 minutes of sequencing. Basecalling was performed using guppy_hac (ver. 3.5.2) model with a CUDA compatible GPU. A total of 6489, 5041 and 3826 reads from flowcells 1 to 3 have passed the default q-score filter. The sequencing and BLASTN results with e-value filter are shown in Table 2.

Table 2. EINS3 sequencing results

	Total reads	*Total reads passing q-score filter*	*Filter pass %*	*All significant matches*	*Significant matches to eins3 (e value < 1e -75)*	*Significant matches to AMXpEINS3 (ACGTATTGCTAGGCTTCTTC) "blastn-short"*
Flowcell 1	23174	6489	28.00	5922	143	10363
Flowcell 2	30238	5041	16.67	4767	53	10264
Flowcell 3	30051	3826	12.73	3699	34	8966

SEQUENCING OF 3XR6

The flowcells 1, 2 and 3 yielded 60299, 48432 and 42434 reads respectively after 4 hours of sequencing. Basecalling was performed using guppy_hac (ver. 3.5.2) model with a CUDA compatible GPU. A total of 28869, 13299 and 10268 reads from flowcells 1 to 3 have passed the default q-score filter. The sequencing and BLASTN results with e-value filter are shown in Table 3.

Table 3. 3xr6 sequencing results

	Total reads	*Total reads passing q-score filter*	*Filter pass %*	*At least 1 significant match vs 42k targets (e value < 1e -15)*	*At least 1 significant match vs 42k addresses (e value < 1e -02) "blastn-short"*	*Significant matches to AMXpFP (ACGTATTGCTCTACAACGCA) "blastn-short"*	*Potential fusion matches FPpRP (CAGCGTTACCCTACAACGCA) "blastn-short"*
Flowcell 1	60299	28869	47.877	25177	31436	88862	78
Flowcell 2	48432	13299	27.46	12146	19128	35789	39
Flowcell 3	42434	10268	24.2	10374	16678	30605	25

Combined	151165	52436	34.69	32028	37265	155256	142

DATA ANALYSIS

BLASTN analysis of INS3, EINS3 and 3xr6 are shown in Figure 4, Figure 5 and Figure 6 respectively, which show the total number of reads that pass the quality threshold and the number of significant matches that are found. For INS3 and EINS3, the input query is their full sequence and the total number of significant matches are plotted. The significant matches are searched with default BLASTN parameters. High-quality matches are also plotted with a 99-percentile label in the figures, these matches show very-high identity to the search query. Refer to the BLASTN output files available in the GitHub repository for the identity details given in supplementary data section.

The data from 3xr6 sequencing run is analyzed with BLASTN to search for the full-length sequences along with the short 25-bp orthogonal sequence. The 25-bp orthogonal sequences used during the design of 3xr6 oligo-pool are taken directly from a published source [13]. Each of the 42,000 (120-bp) oligos in the 3xr6 oligo pool contain a unique 25-bp orthogonal sequence that is repeated three times within the same strand (3x25-bp = 75-bp). All sequences in the 3xr6 oligo pool contain the same forward and reverse priming regions for PCR-compatibility. The 3xr6 oligo-pool was not amplified prior to sequencing in this study.

DISCUSSION

The technique of modifying the 5' end of an oligo to make it compatible for nanopore sequencing has resulted in some interesting insights into the sequencing mechanism. The helicase-bound sequencing adapter (AMX) has a thymine (T) overhang on its top-strand and an oligo with a 5' phosphate and a short (10-bp) 5'-end double-strand region with an adenosine (A) overhang can facilitate sequencing. The biological nanopore used by ONT's MinION system can process a single-stranded template without the impeding force provided by the complementary strand. Although the helicase modifications and the nanopore mechanism is proprietary, we believe that the voltage gradient is primarily the driving force behind the strand translocation through the R9.4.1 nanopore flowcells. The helicase may or may not be functional but could be impeding the strand translocation and slowing it enough to perform a high-resolution scan through a base.

We have also identified that the oligonucleotides can even be 3' unblocked if they do not contain a terminal thymine (T), which can lead to circularization or concatenation products. The INS3 and EINS3 sequences are designed using human insulin gene template [10] and manufactured by phosphoramidite process. Both INS3 and EINS3 strands have a 5' phosphate and 3' blocker molecules added during their synthesis. The oligos in 3xr6 oligo-pool are phosphorylate using T4 PNK before annealing to its adapter sequence. The 3' end of the oligos in 3xr6 are unprotected, unlike INS3 and EINS3.

We have observed that several of the reads that are generated by the sequencer are empty without any viable signal data as visualized in Figure 3 using HDFView (Ver 3.1.0) [11], these reads contain several stall events and helicase dissociation events that are evident from the spike signals. Reads that pass the Guppy quality check are basecalled and analyzed using BLASTN [12].

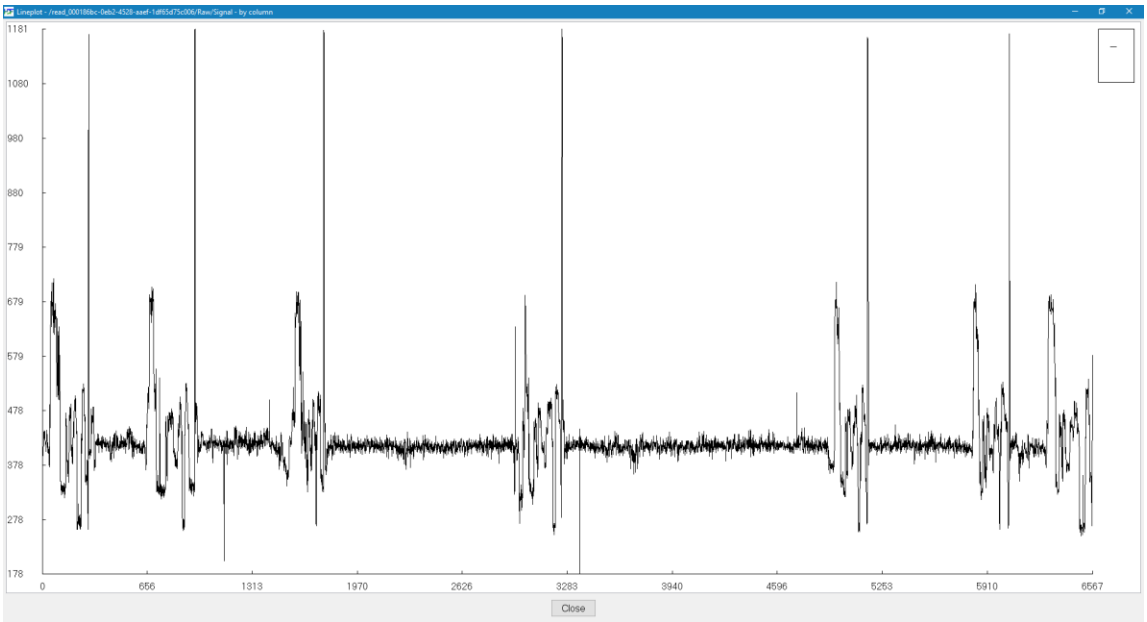

Figure 3. Visualization of a low-scoring INS3 (flowcell 1) read.

LIGATION JUNCTIONS AND MULTI-STRAND READS

The ligation junctions where the AMX adapter binds to the INS3, EINS3 and 3xr6 oligos are searched using BLASTN with 'blastn-short' flag. The junction sequences AMXpINS3, AMXpINS3 and AMXpFP as shown in the Table 1, Table 2 and Table 3 respectively are plotted in the figures below. The ligation junctions are higher in number than the actual number of reads due to the read artifacts where each read may contain more than one strands. We have identified several of these multi-strand reads, and visualization of such a read is provided in Figure 7.

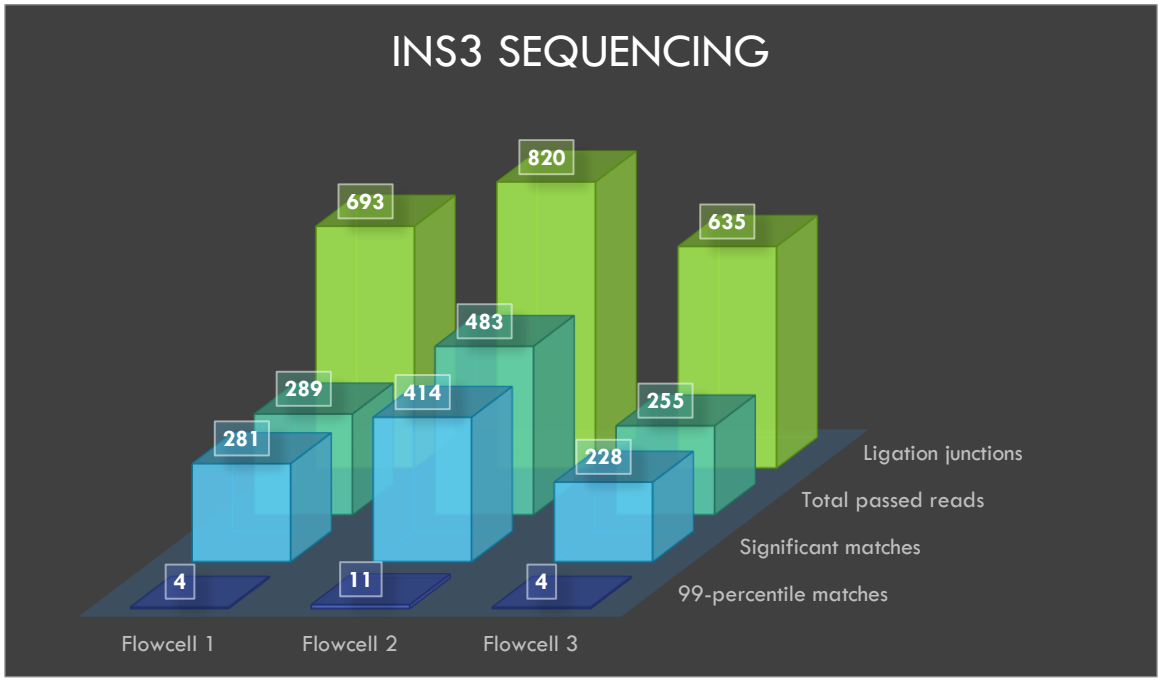

Figure 4. Visualization of INS3 sequencing results

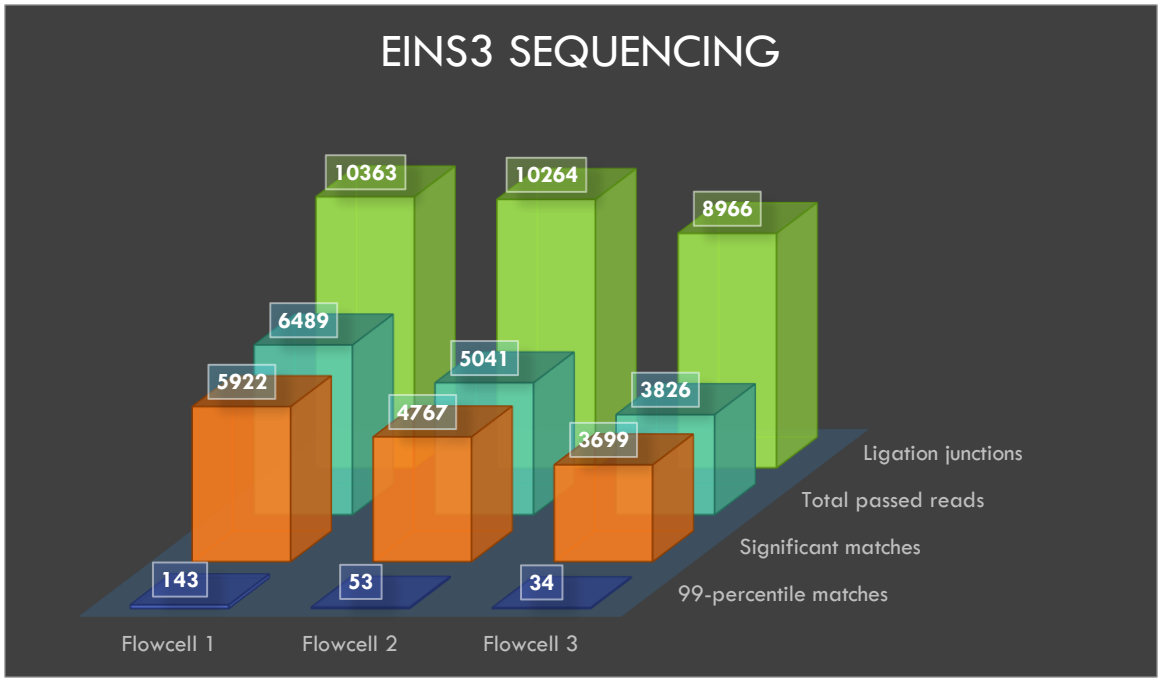

Figure 5. Visualization of EINS3 sequencing results

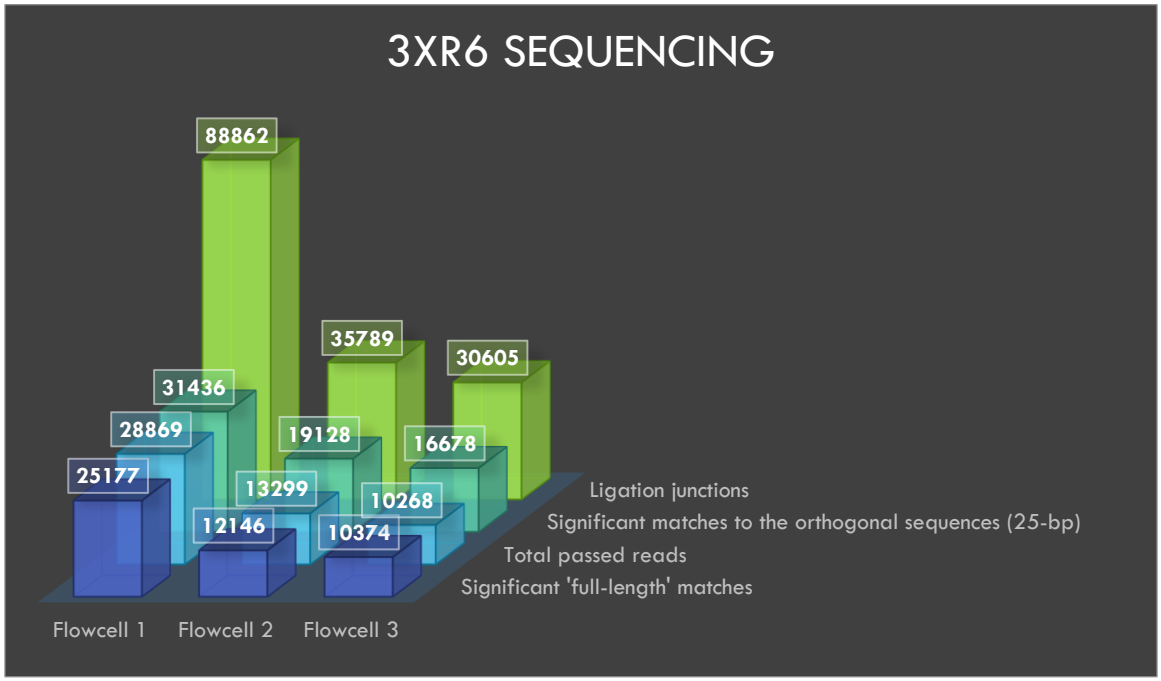

Figure 6. Visualization of 3xr6 sequencing results

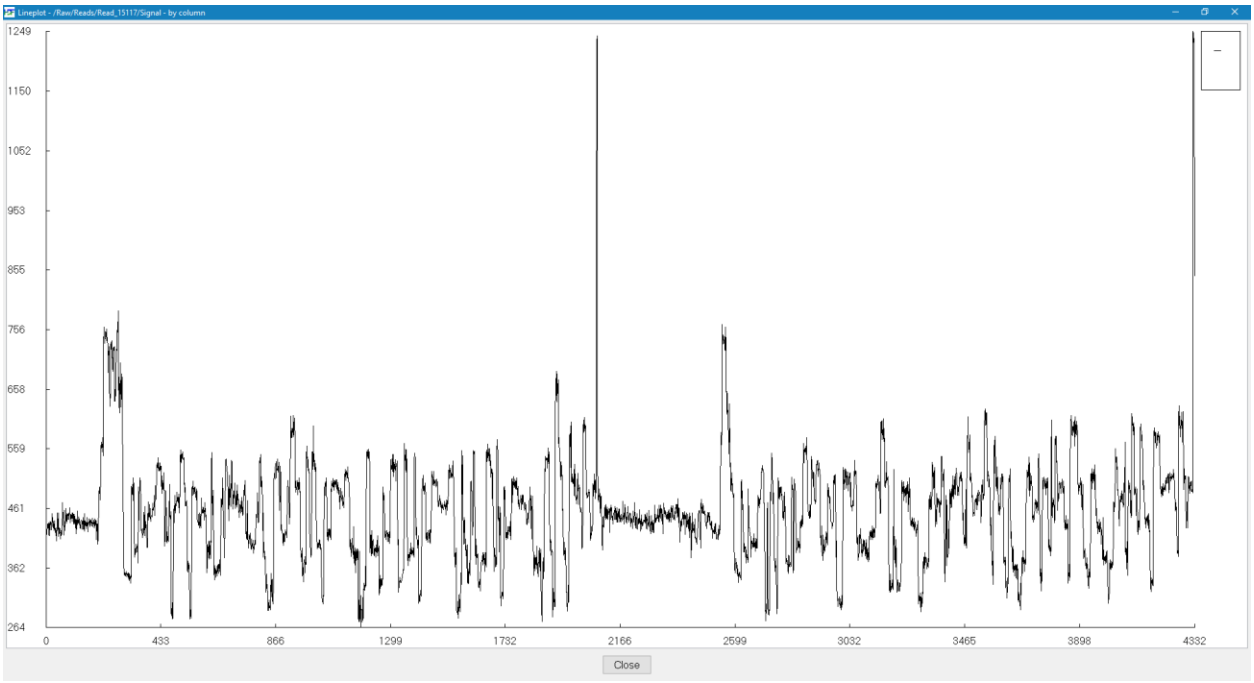

Figure 7. Visualization of a fusion read – [7a6c8a08-3d36-4fcf-a765-ef9cf3e28188] from 3xr6 sequencing (flowcell 1)

CONCLUSION

We have proposed and implemented a method to sequence single-strand nucleic acid species without performing amplification, second-strand synthesis or a spacer ligation step with helicase-based biological nanopores offered by Oxford Nanopore Technologies. Oligonucleotides with a free 3'-OH can also be sequenced successfully with the described method. We have identified sequencing artifacts during our data analysis. Some empty signals can be earmarked as valid reads by the MinION data processing system and reads containing several strands per read are also found, which could be because of segmentation errors.

We believe that this sequencing approach might open new avenues for DNA-based information storage systems and lead to improvements in signal-level data analysis for nanopore sequencing.

CONFLICT OF INTEREST

No conflict of interest.

SUPPLEMENTARY DATA

All the data used for sequencing and related results are available at GitHub repository - https://github.com/helixworks-technologies/dos

REFERENCES

[1] Y. . Wang, Q. . Yang and Z. . Wang, "The evolution of nanopore sequencing.," *Frontiers in Genetics,* vol. 5, no. , pp. 449-449, 2015.

[2] M. . Jain, H. E. Olsen, B. . Paten and M. . Akeson, "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community," *Genome Biology,* vol. 17, no. 1, p. 239, 2016.

[3] N. Harel, M. Meir, U. Gophna and A. Stern, "Direct sequencing of RNA with MinION Nanopore: detecting mutations based on associations.," *Nucleic Acids Research,* vol. 47, no. 22, p. 148, 2019.

[4] D. . Branton, D. W. Deamer, A. . Marziali, H. . Bayley, S. A. Benner, T. . Butler, M. . Di Ventra, S. . Garaj, A. . Hibbs, X. . Huang, S. B. Jovanovich, P. S. Krstic, S. . Lindsay, X. S. Ling, C. H. Mastrangelo, A. . Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. . Riehn, G. V. Soni, V. . Tabard-Cossa, M. . Wanunu, M. . Wiggin and J. A. Schloss, "The potential and challenges of nanopore sequencing," *Nature Biotechnology,* vol. 26, no. 10, p. 1146–1153, 2008.

[5] B. D. Wilson, M. Eisenstein and H. T. Soh, "High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets.," *Analytical Chemistry,* vol. 91, no. 10, pp. 6783-6789, 2019.

[6] S. M. H. T. Yazdi, R. . Gabrys and O. . Milenkovic, "Portable and Error-Free DNA-Based Data Storage," *Scientific Reports,* vol. 7, no. 1, pp. 5011-5011, 2017.

[7] M. . Blawat, K. . Gaedke, I. . Hütter, X. . Chen, B. M. Turczyk, S. A. Inverso, B. W. Pruitt and G. M. Church, "Forward Error Correction for DNA Data Storage," *Procedia Computer Science,* vol. 80, no. 80, pp. 1011-1022, 2016.

[8] T. . Karamitros, T. . Karamitros and G. . Magiorkinis, "Multiplexed Targeted Sequencing for Oxford Nanopore MinION: A Detailed Library Preparation Procedure.," *Methods of Molecular Biology,* vol. 1712, no. , pp. 43-51, 2018.

[9] R. R. Wick, L. M. Judd, K. E. Holt and K. E. Holt, "Performance of neural network basecalling tools for Oxford Nanopore sequencing," *Genome Biology,* vol. 20, no. 1, p. 129, 2019.

[10] G. I. Bell, R. . Pictet, W. J. Rutter, B. . Cordell, E. . Tischer and H. M. Goodman, "Sequence of the human insulin gene," *Nature,* vol. 284, no. 5751, pp. 26-32, 1980.

[11] N. J. Loman and A. R. Quinlan, "Poretools: a toolkit for analyzing nanopore sequence data," *bioRxiv,* vol. , no. , p. 007401, 2014.

[12] C. . Camacho, G. . Coulouris, V. . Avagyan, N. . Ma, J. S. Papadopoulos, K. . Bealer and T. L. Madden, "BLAST+: architecture and applications.," *BMC Bioinformatics,* vol. 10, no. 1, pp. 421-421, 2009.

[13] Q. . Xu, M. R. Schlabach, G. J. Hannon and S. J. Elledge, "Design of 240,000 orthogonal 25mer DNA barcode probes," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 106, no. 7, pp. 2289-2294, 2009.