EXPERIMENTAL RESEARCH

THE IMPACT OF SEQUENCING DEPTH ON THE NUMBER OF TRANSCRIPT SPLICE VARIANTS REVEALED BY MinION NANOPORE SEQUENCING

K.G. Ptitsyn, A.S. Kozlova, S.A. Khmeleva, L.K. Kurbatov, S.P. Radko*, E.V. Ilgisonis, A.V. Lisitsa, E.A. Ponomarenko

Institute of Biomedical Chemistry, 10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: radkos@yandex.ru

Alternative splicing (AS) of a precursor mRNA is a fundamental regulatory process implicated in physiology and pathology. The long-read RNA sequencing with a nanopore sequencer such as ONT MinION allows for direct AS profiling. In this study the impact of sequencing depth on the number of transcribed genes and the overall number of transcripts (splice variants), revealed by MinION-based sequencing, has been investigated. This is of importance in AS profiling for the issue of comparability for biospecimens analyzed in different MinION runs. The sequencing depth was described in terms of the output of high-quality mapped reads produced by a MinION sequencer. The human liver tissue samples and hepatocyte-derived cell lines HepG2 and Huh7 were employed as model objects. It has been found that the yield of detected genes and transcripts substantially depends on the sequencing depth. While the number of transcribed genes levelled off at about 12 thousand when the reads output exceeded 1.2 million, the number of revealed transcripts steadily increased up to about 20 thousand splice variants at the highest reads output of 2.3 million, achieved in the study. At that reads output, the ratio of the number of revealed transcripts to that of genes was slightly below 1.7. The yield of more than 2.3 million high-quality mapped reads would be required in the MinION-based nanopore sequencing to approach the level of 1.8 transcripts (splice variants) per gene, expected from the known numbers of annotated genes and transcripts for human genome. The sequencing data used were produced for human liver tissue and hepatocyte-derived cells and it is still to be seen whether the findings are general and valid for other types of cells and tissues.

Key words: nanopore sequencing, transcribed genes, transcript splice variants, sequencing depth

DOI: 10.18097/BMCRM00300

INTRODUCTION

The processing of a precursor messenger RNA (pre-mRNA) molecule into several mRNA transcripts (transcript's isoforms or splice variants), known as "alternative splicing" (AS) [1], significantly enhances the diversities of transcriptome and proteome [2, 3]. In pre-mRNAs of higher eukaryotes, the splice sites can be differentially selected. This results in a number of single gene transcripts, varying in an exon composition and/or by the presence of intron sequences fully or partially retained [3, 4]. AS is a highly regulated process [5, 6]. The aberrations in AS have been associated with some pathological states which include hereditary diseases and cancer [5, 6].

The emergence and wide spread of next generation sequencing have brought the analysis of AS to a whole-transcriptome level [2, 7]. RNA sequencing employing both the short-read and long-read technologies was used to study AS. In the latter case, the nanopore-based sequencing developed by Oxford Nanopore Technologies (ONT) has become most common, especially with the use of MinION, a portable ONT sequencer (e.g., [8-12]). The short-read RNA sequencing (RNA-seq) is known to face difficulties when determining the exon connectivity. This makes this type of sequencing prone to errors when short reads are to be assembled into splice variants of a given transcript [2, 13, 14]. In contrast, the ONT-based sequencing is free from the problems associated with the reads assembling and allows for direct splice variant profiling [15]. Transcriptome-wide AS profiles can be described either as an assembly of all detected splice variants characterized by their abundances [16-18] or as an assembly of genes characterized by the number of splice variants which were detected for each gene [19, 20].

Clearly, both the number of transcribed genes and the overall number of identified splice variants, revealed by RNA-seq, depend on the depth of sequencing [21]. In RNA-seq, the total amount of sequenced nucleotides (or the total amount of reads) can be preset that eventually determines the sequencing depth. Yet, for ONT nanopore sequencing, the reads output may substantially vary in different sequencing runs. The number of reads produced from a MinION run heavily depends on an actual quality of a sequencer's flow cell. This quality is determined by a number of factors, including the initial quality of a flow cell, as well as duration and conditions of shipment and storage. In many instances, these factors cannot be completely controlled.

The aim of the study was to evaluate to which extent the variations in reads output influenced the number of splice variants revealed in MinION-based sequencing. The answer to this question is important for comparability of biospecimens analyzed in different MinION runs when performing AS profiling. The splice variants with modest and low abundances are thought to be mostly affected by variations in the sequencing depth from run to run [22]. It can make the revealed differences in AS profiles a technical artefact rather than the true differences between biospecimens. To evaluate the influence, the number of transcribed genes and identified splice variants was analyzed as a function of the number of high-quality mapped reads produced by MinION sequencing of mRNA extracted from human liver specimens and hepatocyte-derived cell lines HepG2 and Huh7.

METHODS

The human liver tissue samples (n = 3) were collected post-mortem. The collection was approved by the Ethical

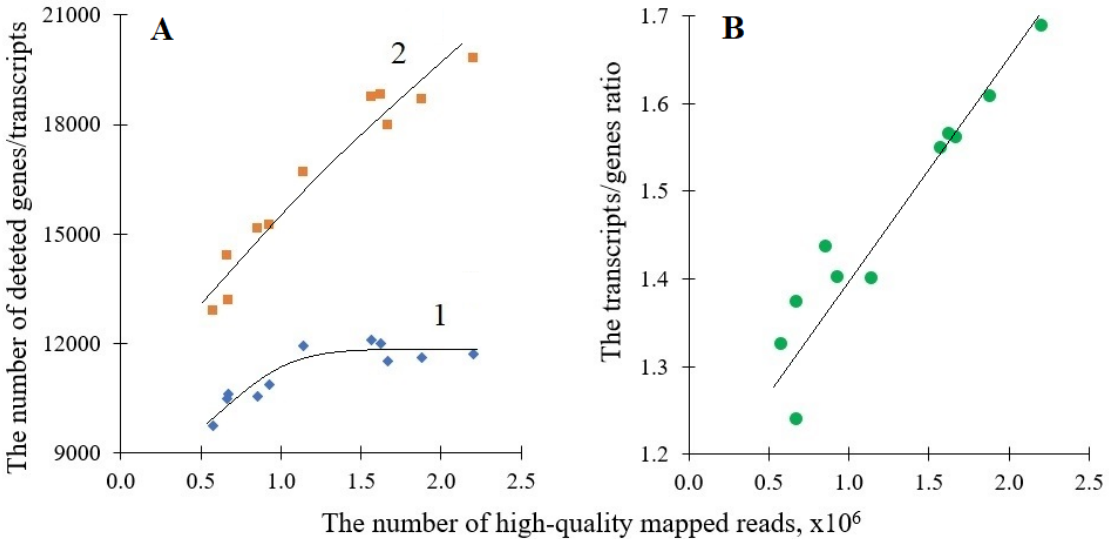

Figure 1. The numbers of detected genes and transcripts and the ratio of these numbers as a function of sequencing depth. The sequencing depth is presented as the number of high-quality mapped reads produced in single MinION runs. Panel A: the number of revealed genes – curve 1; the number of revealed transcripts – curve 2. Panel B: the transcript/gene ratio is a ratio of the number of revealed transcripts (splice variants) to the number of revealed genes.

Committee of the N.I. Pirogov Russian State Medical University. The cell lines HepG2 and Huh-7 were received from Merck (Germany) and Thermo Fisher Scientific (USA), respectively. All details on the tissue sample collection and the culturing of HepG2 and Huh-7 cells can be found elsewhere [19, 20]. Total RNA was isolated with a RNeasy Mini Kit (Qiagen, Germany) and its quality was evaluated with a Bioanalyzer 2100 System (Agilent Technologies, USA). The values of 7.8 or higher were obtained for RNA integrity numbers (RINs) in all cases. The isolation of mRNA was done using a Dynabead mRNA Purification Kit (Thermo Fisher Scientific). The mRNA concentration was measured on a Qubit 4 fluorometer, using a Qubit RNA HS Assay Kit (Thermo Fisher Scientific).

A kit for direct RNA sequencing (SQK-RNA002) from ONT (UK) was used to prepare sequencing libraries, in a strict accordance with the protocol of the manufacturer. Each sequencing was performed in a single run for 48 h with a flow cell FLO-MIN106 (ONT) on a nanopore sequencer MinION (ONT). The raw sequencing data obtained were processed post-run with the software "guppy_basecaller" (v. 3.1.5, ONT) as described earlier [23]. The quality score parameter for data filtering was set as >7.0 (the default value recommended by the software manufacturer). The additional control of reads quality was carried out by using the MinIONqc.R script [24]. The long-read aligner "minimap2" (v. 2.17) [25] and the release GRCh38 of the genome assembly Gencode38 (https://www.gencodegenes.org/human/release_38.html, last accessed on August 5, 2025) were employed to map the reads. The mapping was performed in the mode "-ax splice junc-bed". The files with sequencing data are deposited to the NCBI Sequence Read Archive (PRJNA765908, PRJNA893571, PRJNA635536).

RESULTS AND DISCUSSION

Figure 1A shows dependencies for the number of identified transcribed genes and the overall number of identified transcripts (splice variants) on the amount of high-quality mapped reads. Here transcribed genes are the genes for which at least one splice variant has been identified based on the MinION sequencing data produced in a given run. The data were plotted regardless of their origin (wherever they come from human liver tissue, HepG2 or Huh7 cells). As seen, all points fall into two master curves in the studied range of mapped reads (from about 0.58 to about 2.3 million reads): one master curve for the number of genes and another one – for the number of transcripts (Figure 1A). The increase in the number of genes appears to level off when the number of mapped read slightly exceeds a million. Apparently, about 12 thousand genes are actively transcribed in human normal liver tissue and malignantly transformed hepatocytes and about 1.2 million reads allow for detecting practically all of them in the MinION sequencing run with the yield of about 1.2 million or more high-quality mapped reads (Figure 1A).

In contrast to the genes, the number of detected transcripts keeps growing with the number of mapped reads in almost a linear fashion over the entire range studied (Figure 1A). When the average numbers of transcripts per gene were calculated (by dividing the overall number of transcripts revealed in a given MinION sequencing run by the corresponding number of genes), they demonstrated a steady growth over the range of mapped reads: the average number of transcripts per gene has risen from about 1.3 to about 1.7 (Figure 1B). Clearly, with the increase in sequencing depth, the low expressed splice variants for transcripts of already identified genes become detected. Since the average number of transcripts per gene is obviously a function of sequencing depth (Figure 1B), the present data did not allow us to estimate the actual average number of splice variants for a transcript in human liver tissue and hepatocyte-derived cells However, for example, one of the Ensemble annotations reported 21,694 protein-coding genes and 39,106 corresponding transcripts (https://www.ensembl.org/info/genome/genebuild/2014_07_human_genebuild.pdf, last accessed on September 2, 2025). These numbers give about 1.8 transcript per gene. The ratio is quite close to 1.7, achieved in our study with the output of 2.3 million mapped reads. One may suggest that practically all transcript splice variants can be detected for transcribed genes in human liver tissue and hepatocyte-derived cells in a single MinION run with the output of 2.3 million or more mappable reads.

The results obtained evidently demonstrate that sequencing depth has to be matched in the MinION-based nanopore sequencing to correctly compare AS profiles between various tissues and cells or their states for finding biologically related differences. Since the sequencing output can vary manyfold between MinION runs, the simplest way to match the sequencing depth may be to equal the number of mapped reads taken into bioinformatic analysis. It can easily be done with the Picard DownsampleSam tool (https://broadinstitute.github.io/picard/, accessed on September 7, 2025) which allows for randomly selecting the desired number of reads for further analysis. The disadvantage of such approach is that all sequencing outputs have to be adjusted to the output with the minimal number of high-quality mapped reads.

It should be noted that in the present study all identified transcripts were taken into consideration, regardless of their abundance. The transcript abundance can be taken into account by setting a threshold for transcript abundance, below which the detection of transcripts is considered as statistically insignificant. It is commonly to use the value of 0.1 as a threshold for transcript abundance expressed in TPM (transcripts per million) [26]. When transcripts with abundances below the threshold of 0.1 were removed from consideration, the number of identified genes practically did not change. At the same time, the number of identified transcripts and, consequently, the average number of transcripts per gene slightly decreased. Nonetheless, the shapes of curves describing the relationships between the numbers of genes, transcripts, and the mapped reads (data not shown) were similar to those of curves 1 and 2 in Figure 1A.

CONCLUSIONS

In MinION-based nanopore sequencing, the yield of detected genes and transcripts substantially depends on the sequencing depth presented as the sequencing output in high-quality mapped reads. While practically all of actively transcribed genes appear to be detected when the reads output exceeds 1.2 million, the number of revealed transcripts steadily increased over the studied range of reads output. The yield of more than 2.3 million high-quality mapped reads would be required in the MinION-based nanopore sequencing in order to approach the level of 1.8 transcripts (splice variants) per gene, expected from the known numbers of annotated genes and transcripts for human genome. The sequencing data used have been produced for human liver tissue and hepatocyte-derived cells and it is still to be seen whether the made conclusions can be generalized to other types of cells and tissues.

COMPLIANCE WITH ETHICAL STANDARDS

The collection of human liver tissue samples was approved by the Ethical Committee of the N.I. Pirogov Russian State Medical University (Protocol #3; March 15, 2018).

FUNDING

The study was performed within the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021–2030) (No. 122030100170-5).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. *Gilbert, W.* (1978) Why genes in pieces? Nature, **271**(5645), 501. DOI: 10.1038/271501a0
2. *Su, T., Hollas, M.A.R., Fellers, R.T., Kelleher, N.L.* (2023) Identification of splicevariants and isoforms in transcriptomics and proteomics. Annual Review of Biomedical Data Science, **6**, 357-376. DOI: 10.1146/annurev-biodatasci-020722-044021
3. *Wright, C.J., Smith, C.W.J., Jiggins, C.D.* (2022) Alternative splicing as a source of phenotypic diversity. Nature Reviews Genetics, **23**(11), 697-710. DOI: 10.1038/s41576-022-00514-4
4. *Montes, M., Sanford, B.L., Comiskey, D.F., Chandler, D.S.* (2019) RNA splicing and disease: animal models to therapies. Trends in Genetics, **35**(1), 68-87. DOI: 10.1016/j.tig.2018.10.002
5. *Nilsen, T.W., Graveley, B.R.* (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature, **463**(7280), 457-63. DOI: 10.1038/nature08909
6. *Jiang, W., Chen, L.* (2020) Alternative splicing: human disease and quantitative analysis from high-throughput sequencing. Computational and Structural Biotechnology Journal, **19**, 183-195. DOI: 10.1016/j.csbj.2020.12.009
7. *Stark, R., Grzelak, M., Hadfield, J.* (2019) RNA sequencing: the teenage years. Nature Reviews Genetics, **20**(11), 631-656. DOI: 10.1038/s41576-019-0150-2
8. *Zhang, Z., So, K., Peterson, R., Bauer, M., Ng, H., Zhang, Y., Kim, J.H., Kidd, T., Miura, P.* (2019) Elav-mediateexon skipping and alternative polyadenylation of the Dscam1 gene are required for axon outgrowth. Cell Reports, **27**(13), 3808-3817.e7. DOI: 10.1016/j.celrep.2019.05.083
9. *Moldovan, N., Torma, G., Gulyas, G., Hornyak, A., Zadori, Z., Jefferson, V.A., Csabai, Z., Boldogkoi, M., Tombacz, D., Meyer, F., Boldogkoi, Z.* (2020) Time-course profiling of bovine alphaherpesvirus 1.1 transcriptome using multiplatform sequencing. Scientific Reports, **10**(1), 20496. DOI: 10.1038/s41598-020-77520-1
10. *Xin, H., He, X., Li, J., Guan, X., Liu, X., Wang, Y., Niu, L., Qiu, D., Wu, X., Wang, H.* (2022) Profiling of the full-length transcriptome in abdominal aortic aneurysm using nanopore-based direct RNA sequencing. Open Biology, **12**(2), 210172. DOI: 10.1098/rsob.210172
11. D*eynichenko, K.A., Ptitsyn K.G., Radko, S.P., Kurbatov, L.K., Vakhrushev, I.V., Buromski, I.V., Markin, S.S., Archakov, A.I., Lisitsa, A.V., Ponomarenko, E.A.* (2022) Splice variants of mRNA of cytochrome P450 genes: analysis by the nanopore sequencing method in human liver tissue and HepG2 cell line. Biomeditsinskaya Khimiya, 68(2), 117-125. DOI: 10.18097/PBMC20226802117
12. *Wu, H., Lu, Y., Duan, Z., Wu, J., Lin, M., Wu, Y., Han, S., Li, T., Fan, Y., Hu, X., Xiao, H., Feng, J., Lu, Z., Kong, D., Li, S.* (2023) Nanopore long-read RNA sequencing reveals functional alternative splicing variants in human vascular smooth muscle cells. Communications Biology, 6(1), 1104. DOI: 10.1038/s42003-023-05481-y
13. *Soneson, C., Love, M.I., Robinson, M.D.* (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research, **4**, 1521. DOI: 10.12688/f1000research.7563.2
14. *Zhang, C., Zhang, B., Lin, L.L., Zhao, S.* (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics, **18**(1), 583. DOI:10.1186/s12864-017-4002-1
15. *Hussain, S.* (2018) Native RNA-sequencing throws its hat into the transcriptomics ring. Trends in Biochemical Sciences, **43**(4), 225-227. DOI: 10.1016/j.tibs.2018.02.007
16. *Wadsworth, M.E., Page, M.L., Aguzzoli Heberle, B., Miller, J.B., Steely, C.,* Ebbert, M.T.W. (2025) Sequencing the gaps: dark genomic regions persist in CHM13 despite long-read advances. bioRxiv [Preprint]. DOI: 10.1038/s41587-024-02245-9
17. *Yao, T., Zhang, Z., Li, Q., Huang, R., Hong, Y., Li, C., Zhang, F., Huang, Y., Fang, Y., Cao, Q., Jin, X., Li, C., Wang, Z., Lin, X.J., Li, L., Wei, W., Wang, Z., Shen, J.* (2023) Long-Read Sequencing Reveals Alternative Splicing-Driven, Shared Immunogenic Neoepitopes Regardless of SF3B1 Status in Uveal Melanoma. Cancer Immunology Research, **11**(12), 1671-1687. DOI: 10.1158/2326-6066.CIR-23-0083
18. *Halstead, Islas-Trejo, M.M., Goszczynski, D.E., Medrano, J.F., Zhou, H., Ross, P.J.* (2021) Large-Scale Multiplexing Permits Full-Length Transcriptome Annotation of 32 Bovine Tissues From a Single Nanopore Flow Cell. Frontiers in Genetics, **12**, 664260. DOI: 10.3389/fgene.2021.664260
19. *Sarygina, E., Kozlova, A., Deinichenko, K., Radko, S., Ptitsyn, K., Khmeleva, S., Kurbatov, L.K., Spirin, P., Prassolov, V.S., Ilgisonis, E., Lisitsa, A., Ponomarenko, E.* (2023) Principal component analysis of alternative splicing profiles revealed by long-read ONT sequencing in human liver tissue and hepatocyte-derived HepG2 and Huh7 cell lines. International Journal of Molecular Sciences, **24**(21), 15502. DOI: 10.3390/ijms242115502
20. *Kozlova, A. Sarygina, E., Deinichenko, K., Radko, S., Ptitsyn, K., Khmeleva, S., Kurbatov, L., Spirin, P., Prassolov, V., Ilgisonis, E., Lisitsa, A., Ponomarenko, E.* (2023) Comparison of alternative splicing landscapes revealed by long-read sequencing in hepatocyte-derived HepG2 and Huh7 cultured cells and human liver tissue. Biology (Basel), **12**(12), 1494. DOI: 10.3390/biology12121494
21. *Zhao, S., Ye, Z., Stanton, R.* (2020) Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA, **26**(8), 903-

909. DOI: 10.1261/rna.074922.120
22. *Schwenk, V., Leal Silva, R.M., Scharf, F., Knaust, K., Wendlandt, M., Häusser, T., Pickl, J.M.A., Steinke-Lange, V., Laner, A., Morak, M., Holinski-Feder, E., Wolf, D.A.* (2023) Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. Journal of Medical Genetics, **60**(8), 747-759. DOI: 10.1136/jmg-2022-108931
23. *Shapovalova, V., Radko, S., Ptitsyn, K., Krasnov, G., Nakhod, K., Konash, O., Vinogradina, M., Ponomarenko, E., Druzhilovskiy, D., Lisitsa, A.* (2020) Processing oxford nanopore long readsusing amazon web services. Biomedical Chemistry: Research and Methods, **3**(4), e00131. DOI: 10.18097/BMCRM00131
24. *Lanfear, R., Schalamun, M., Kainer, D., Wang, W., Schwessinger, B.* (2019) MinIONQC: fast and simple quality control for MinION sequencing data. Bioinformatics, **35**(3), 523-525. DOI: 10.1093/bioinformatics/bty654
25. *Li, H.* (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, **34**(18), 3094-3100. DOI: 10.1093/bioinformatics/bty191
26. *Pyatnitskiy M.A., Arzumanian V.A., Radko S.P., Ptitsyn K.G., Vakhrushev I.V., Poverennaya E.V., Ponomarenko E.A.* (2021) Oxford nanopore MinION direct RNA-Seq for systems biology. Biology (Basel), **10**(11), 1131. DOI: 10.3390/biology10111131

Received: 14.10.2025
Revised: 03.12.2025
Accepted: 04.12.2025

ВЛИЯНИЕ ГЛУБИНЫ СЕКВЕНИРОВАНИЯ НА КОЛИЧЕСТВО ВАРИАНТОВ СПЛАЙСИНГА ТРАНСКРИПТОВ, ВЫЯВЛЕННОЕ С ПОМОЩЬЮ НАНОПОРОВОГО СЕКВЕНАТОРА MinION

К.Г. Птицын, А.С. Козлова, С.А. Хмелева, Л.К. Курбатов, С.П. Радько*, Е.В. Ильгисонис, А.В. Лисица, Е.А Пономаренко

Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича, 119121, Москва, ул. Погодинская, 10; e-mail: radkos@yandex.ru

Альтернативный сплайсинг (АС) первичной мРНК является фундаментальным регуляторным процессом, связанным с физиологией и патологией. Секвенирование РНК длинными прочтениями с использованием нанопорового секвенатора, такого как ONT MinION, позволяет проводить прямое профилирование АС. В настоящей работе было исследовано влияние глубины секвенирования на количество транскрибируемых генов и общее количество вариантов транскриптов (вариантов сплайсинга), выявляемых с помощью секвенирования с использованием MinION. Это важно для профилирования АС с точки зрения сопоставимости данных, полученных для разных биообразцов в отдельных секвенированиях. Глубина секвенирования выражалась как количество картированных «прочтений», полученных в каждом секвенировании с использованием секвенатора MinION. В качестве модельных объектов были использованы образцы ткани печени человека и клеточные линии гепатоцитарного происхождения HepG2 и Huh7. Было обнаружено, что количество обнаруженных генов и транскриптов существенно зависит от глубины секвенирования. В то время как количество детектируемых генов достигало плато на уровне примерно 12 тысяч, когда количество «прочтений» превышало 1.2 миллиона, количество выявленных транскриптов неуклонно росло до примерно 20 тысяч сплайс-вариантов при самом высоком показателе в 2.3 миллиона «прочтений», достигнутом в ходе исследования. При данном количестве «простений» отношение числа выявленных транскриптов к числу генов было немного ниже 1.7. Для достижения уровня в 1.8 транскриптов (вариантов сплайсинга) на ген, ожидаемого из известного количества аннотированных генов и транскриптов для генома человека, при проведении секвенирования с использованием MinION потребуется получение более 2.3 миллионов высококачественных картированных «прочтений». Данные секвенирования, использованные в исследовании, были получены для гепатоцитов и клеток гепатоцитарного происхождения и для их обобщения потребуется анализ данных нанопорового секвенирования для других типов клеток и тканей.

Ключевые слова: нанопоровое секвенирование; транскрибируемые гены; сплайс-варианты транскриптов; глубина секвенирования

ФИНАНСИРОВАНИЕ

Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021 - 2030 годы) (№ 122030100170-5).

Поступила: 14.10.2025, после доработки: 03.12.2024, принята к публикации: 04.12.2025