BMC Research Notes

RESEARCH NOTE

Open Access

BED-Craft for nanopore adaptive sampling: a tool for generating bed files with gene names as input data for enrichment sequencing

Fuyuki Miya[1*] and Kenjiro Kosaki[1]

Abstract

Objective Adaptive sampling, a nanopore sequencing method that enriches regions of interest (ROI), is cost-effective and useful. However, the process of defining targeted regions and creating the corresponding definition file (.bed file) are time-consuming and laborious. To simplify this process, we have developed a tool to easily create a .bed file for adaptive sampling directly from gene names.

Results The tool is freely available on GitHub at https://github.com/medicalbioinfo/BED-Craft. The input is a text file containing one or more gene names (symbols), and even with a large number of input genes (e.g., thousands), the resulting .bed file is generated in less than a second. The length of the buffer region added upstream and downstream of the ROI is designed to account for genome strand orientation, ensuring efficient adaptive sampling. The buffer length can also be modified by the user. The tool supports the genomes of human hg19, hg38, T2T-CHM13, and other species. For researchers unfamiliar with command-line input, a GUI version of the tool is also available at https://keio-cmg.jp/BED-Craft/. This easy-to-use .bed file generation tool enables adaptive sampling by easily changing the target genes of interest in nanopore sequencing, and provide great benefits to researchers and diagnostic laboratories.

Keywords Adaptive sampling, Nanopore, Bed file, Enrichment

Introduction

Long-read sequencing surpasses short-read sequencing in its ability to accurately sequence complex genomic regions, such as repeat regions, and to detect genomic structural variations. It has significantly contributed to the complete sequencing of the human genome by the Telomere-to-Telomere (T2T) Consortium [1, 2]. Among long-read sequencing technologies, nanopore sequencing by Oxford Nanopore Technologies and Single Molecule Real-Time (SMRT) sequencing by Pacific Biosciences are the most widely used worldwide. In addition to standard nucleotide sequencing, long-read sequencing can also detect methylated base modifications. This capability holds promise for applications in disease research, including methylation analysis for genomic imprinting disorders and the detection of genomics structural abnormalities [3].

For enrichment sequencing of specific regions of interest (ROI), such as cancer genes targeted in cancer gene panel testing or coding regions of all genes (for whole-exome sequencing), the most widely used method is to use chemically synthesized capture probes with sequences complementary to the ROI. These probes are used to isolate targeted DNA, which is then sequenced using next-generation sequencing (NGS) [4]. Examples of such capture probes include Illumina's DNA Prep

*Correspondence:
Fuyuki Miya
fmiya@keio.jp
[1] Center for Medical Genetics, Keio University School of Medicine, Tokyo 160-8582, Japan

with Exome 2.5 Enrichment, Twist Bioscience's Exome 2.0, and Agilent" SureSelect Human All Exon probes. In contrast, nanopore sequencing allows the use of a technique called adaptive sampling, which electronically enriches ROIs without relying on capture technologies. This method involves sequencing DNA in real time as it passes through a nanopore protein and mapping the sequences to a reference genome. If the DNA contains an ROI, sequencing continues; otherwise, the DNA is rejected and expelled via reverse voltage, enriching for the ROI [5–7]. The advantages of adaptive sampling include the elimination of the need for experimental capture and retrieval of DNAs of ROI. Adaptive sampling requires only the electronic definition of ROIs, enabling flexible adaptation to different ROIs.

To perform adaptive sampling, a reference genome and a .bed file are required. The .bed file is a file that defines the sequencing targeted regions containing the ROIs. The .bed file follows the Browser Extensible Data (BED) format and must contain at least the following information: the chromosome number, the start position, and the end position of the targeted regions [8]. While the positional information for genes within an ROI can be obtained from databases such as the National Center for Biotechnology Information (NCBI) or the UCSC Genome Browser, manually retrieving this information for large numbers of genes is a labor-intensive and time-consuming task. Manually searching for gene names on the UCSC Genome Browser to identify gene regions and creating a BED-formatted text file is considered a relatively straightforward approach. However, this process takes approximately 15 s per gene, and performing this task manually for hundreds of genes each time the target gene set changes is impractical.

To simplify the time-consuming process of creating BED format files, we have developed a tool that allows users to easily generate .bed files directly from gene names.

Main text

Implementation

We have developed a tool for creating .bed files for nanopore adaptive sampling from gene names, which we have named 'BED-Craft', and have made it freely available on GitHub. This tool is written in the Perl programming language. For researchers unfamiliar with command-line input, a graphical user interface (GUI) version of the tool is also available at https://keio-cmg.jp/BED-Craft/.

The tool supports three versions of the human reference genome, hg19 (GRCh37), hg38 (GRCh38), and T2T-CHM13, and two version of mouse reference genome, mm10 (GRCm38) and mm39 (GRCm39). The corresponding gene annotation files for these genomes are included within the tool. Gene annotations for hg19 was obtained from NCBI RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and Ensembl (http://www.ensembl.org/). When genomic coordinates for the same gene were available in both databases, the RefSeq data was prioritized. For hg38, we prioritized the MANE transcripts as defined by the MANE project, a collaboration between NCBI RefSeq and Ensembl to unify transcript annotation [9]. If a MANE transcript was not available, the annotations from RefSeq and Ensembl were used in that order of priority to create the gene annotation file. The gene annotations for CHM13 were derived from the data on GitHub (https://github.com/marbl/CHM13), we used the "JHU RefSeqv110 + Liftoff v5.2" file for the CHM13 annotations. For mm10 and mm39, the annotation was obtained from Ensembl. These annotation files will be updated on the Bed-Craft GitHub website when new reference human/mouse builds are published. Details of the latest annotation file versions can be found on the tool's GitHub site. As well as human and mouse, the tool can be used for over 200 other various species. How to support species other than human and mouse is described below.

The BED-Craft tool automatically adds additional regions, called "buffer" regions, to both sides of the genomic positions of the ROI when generating a .bed file. By default, ± 50 kb are added. The length of the buffer can be customized by the user. More details about buffers, refer to the "Results and discussion" section. The sum of the ROI and the buffer length constitutes the targeted regions, as expressed in the following equation:

$$ROI + Buffer = Targeted\_regions.$$

Users can perform adaptive sampling on nanopore MinKNOW using the .bed files generated by the BED-Craft tool along with a reference genome file of the same version. Note, when using the GRCh38/hg38 reference genome, it is recommended to use a "no-alt" reference genome (e.g., "human_GRCh38_no_alt_analysis_set.fasta", available at "https://github.com/PacificBiosciences/reference_genomes") to avoid potential adaptive sampling problems caused by the inclusion of ALT sequences.

If any questions or issues arise while using the tool, users can contact us via the "Issues" section on the tool's GitHub repository.

Results and discussion

Usage and target genes size

We have developed a software tool named 'BED-Craft' that generates .bed files for nanopore adaptive sampling using gene symbols as an input file. Workflow for generating .bed files from gene names using BED-Craft is shown in Fig. 1. The BED-Craft tool enables easy exchange of

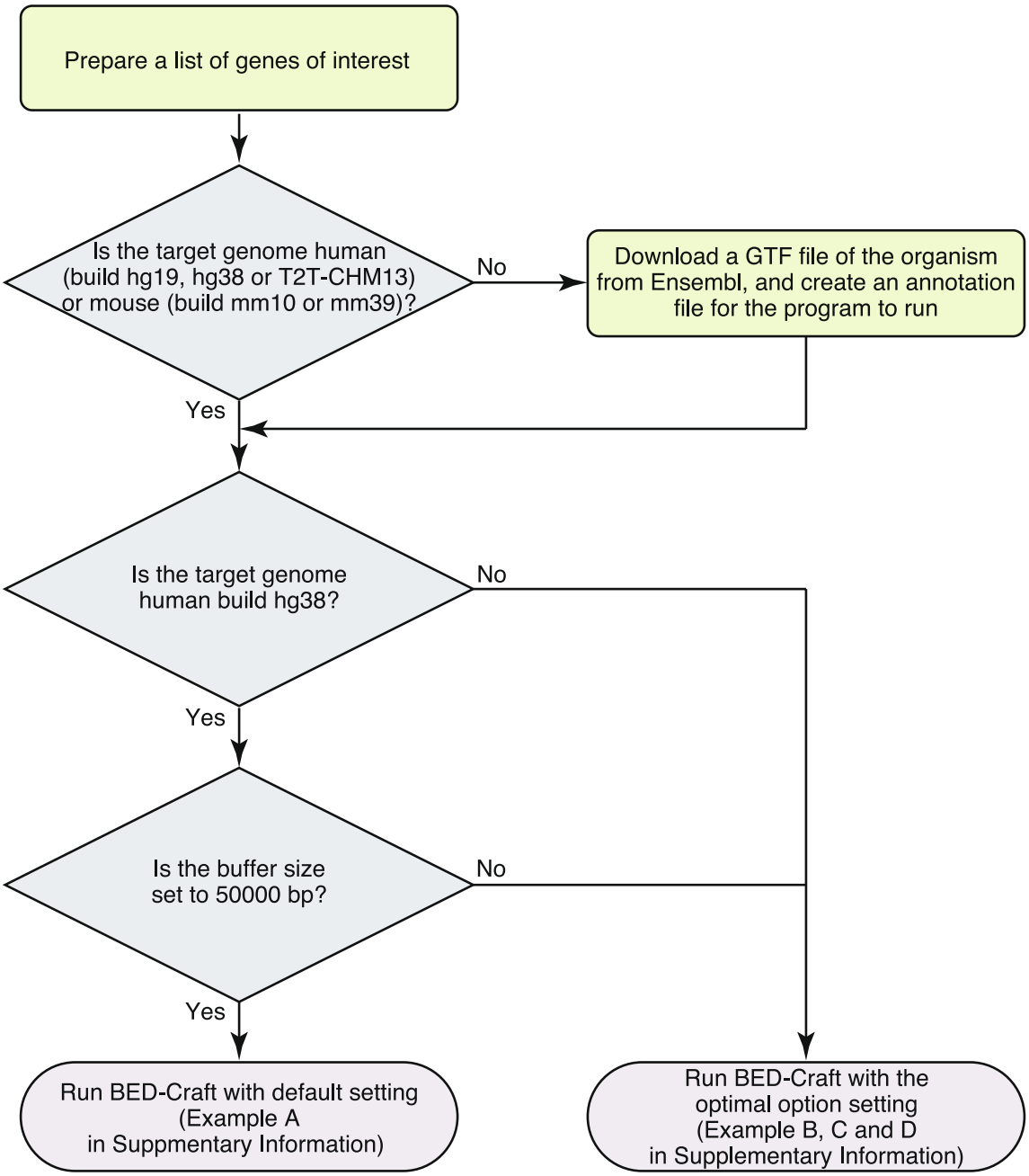

Fig. 1 Workflow for generating .bed files from gene names using BED-Craft tool

gene sets to perform adaptive sampling. For example, by referencing Genomics England's PanelApp [10] (https://panelapp.genomicsengland.co.uk), gene sets tailored to a patient's phenotype can be used for sequencing and analysis, facilitating rapid personalized genetic testing. There is no limit to the number of genes in the input file; however, Oxford Nanopore Technologies recommends that the total size of the target regions, including buffer regions, should not exceed 5% of the whole-genome length [7]. In our experience, when using the same flow cell under identical conditions, the enrichment efficiency (i.e., sequencing depth) was significantly higher when the target regions occupied 1% of the genome compared to 5%. However, when the target size ranged from approximately 1% to a single gene (~0.004% of the genome), the enrichment efficiency (sequencing depth) showed no significant differences (data not shown). The proportion (%) of the ROI and targeted regions, including buffers,

relative to the whole genome is displayed as part of the standard output when running the BED-Craft tool.

With the BED-Craft tool, users can generate a BED-formatted file (e.g., target_gene.bed) by simply typing the following command on the command line for a gene symbol input file (e.g., target_gene.txt):

```
perl BED_Craft.pl target_gene.txt
```

(Fig. 2a–c). The tool provides options to specify the genome build version (hg19, hg38, CHM13, mm10, mm39, or other), the buffer size (from 0 to infinity), and whether to include "chr" prefixes in the chromosome numbers in the output.bed file (Fig. 2b). By default, the genome build is set to hg38, the buffer size to 50,000 (± 50 kb), and the inclusion of "chr" prefixes to "yes". Further usage examples are described in Supplementary Information.

In addition to the command-line version, users can also utilize the GUI version (Fig. 2d, https://keio-cmg.jp/BED-Craft/). Detailed instructions for its usage are described in the Supplementary Information.

Buffer size

The buffer size refers to the additional regions added on both sides of the ROI, which is set to ± 50,000 by default (see Fig. 2c). As shown in the output example, the buffer extends the targeted size upstream by 50,000 bases for the plus (+, forward) strand and downstream by 50,000 bases for the minus (−, reverse) strand. In other words, the buffer is only added to the 5′-side of the ROI for both strands. This approach is based on the fact that DNA is always read from the 5'-end. If the starting point of a sequenced DNA is within the 5′-buffer region, the subsequent portion of the DNA is likely to contain the ROI. However, adding a buffer to the 3′-side is unnecessary because DNA reads starting within the 3′-buffer cannot include the ROI. This strand-aware and strategic buffer addition reduces unnecessary off-target sequencing and increases enrichment efficiency for the ROI. If the buffer size is set to 0, the targeted regions will only include the full-length gene regions defined as the ROI (from the 5′ UTR to the 3′ UTR of the genes). However, reducing the buffer size decreases the likelihood of sequencing near the edges of the ROIs. Oxford Nanopore Technologies recommends setting the buffer size to approximately the N10 value of the average sequencing library DNA length to optimize enrichment efficiency [7]. If the DNA for adaptive sampling is fragmented prior to library preparation using tools such as the g-TUBE (Covaris), a buffer size of less than 50 kb may be sufficient. In our experience, pre-fragmentation often increases the final sequencing yield by more than 1.5 times. However, fragmentation naturally results in shorter read lengths, which can negatively affect the length of haplotype phasing. For example, in genomic imprinting studies, such as the one we reported on previously [11], it is important to analyze methylation in each phased homologous chromosome separately. In such cases, avoiding DNA fragmentation and prioritizing longer haplotype phasing may be more advantageous. In summary, a default buffer size of 50 kb is generally recommended. However, when sequencing fragmented DNA, a smaller buffer size (e.g., 25 kb) may also be appropriate.

Performance characterization

To evaluate the performance of the software, we ran BED-Craft on the human genome (hg38) using two input file patterns: one containing 100 genes and the other containing 20,000 genes. As a result, there was no significant difference between the two cases. Therefore, we report the results for the 20,000-gene input file. The program took 0.12 s from the start of the program to the output generation, with a memory usage of 0.019 MB. While minor variations may occur depending on the target genome or reference version, it is estimated that the execution time will be less than 1 s and the memory usage stays below 1 MB for the expected usage range. This performance evaluation was conducted on a MacPro (3.3 GHz Intel Xeon, 12 cores, 2019 model).

Support for other species

For over 200 species other than humans and mouse, a program is provided to generate the desired .bed file using Ensembl GTF files. The detailed method is described in the Supplementary Information. The GUI version of this tool does not support other species, but the command line version does.

Verification of proper functioning of adaptive sampling using the generated .bed file

The functionality of adaptive sampling with the generated .bed file can be verified using the "Read length histogram" screen in the nanopore MinKNOW software. Typically, adaptive sampling will begin approximately 5 min after the start of the run with adaptive sampling enabled. At this point, DNAs containing the targeted regions (ROI + buffer regions) continue to be sequenced, while other DNAs are ejected from the nanopore. If adaptive sampling is working correctly, the read length histogram in MinKNOW will show a sharp increase in short reads (~ 700 bp or shorter) after 5 min or more. These short reads correspond to DNAs that does not contain the targeted region and are rejected after minimal sequencing (Fig. 3a). If 1% of the genome

Fig. 2 Overview of the BED-Craft tool. **a** Example of an input file. **b** Examples of commands. **c** Example of the output .bed file generated by the command. **d** Screenshots of the GUI version of the BED-Craft tool

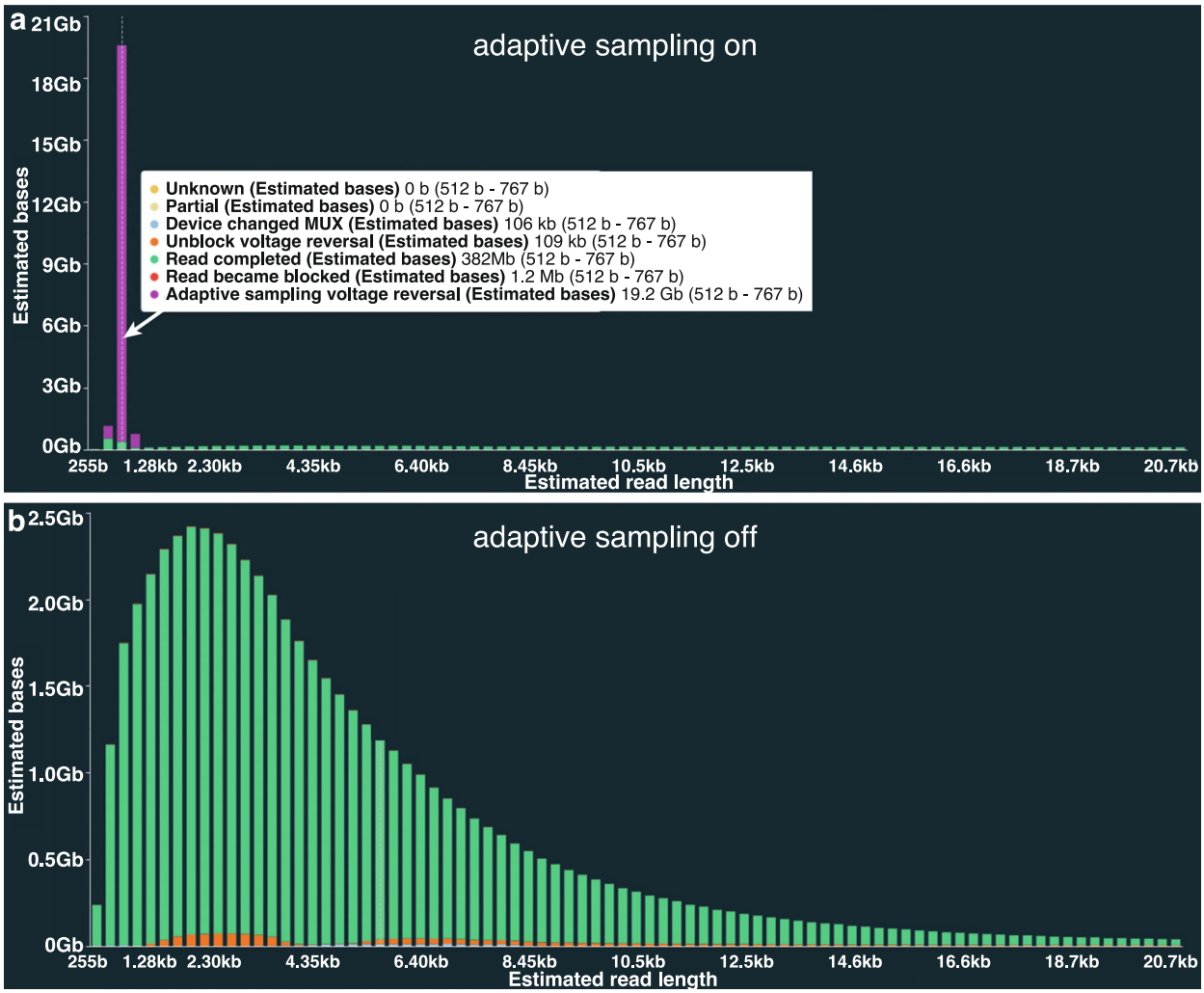

Fig. 3 A screenshot of "Read length histograms" displayed in the MinKNOW software, comparing cases with and without adaptive sampling. **a** Histogram of reads when adaptive sampling is working correctly. The purple bars represent reads rejected by adaptive sampling, most of which fall between 512 and 767 bp. **b** Histogram of reads when adaptive sampling is not applied. Typically, the peak appears at several kilobases, and the shape of the histogram is significantly different from the one shown above. If for some reason adaptive sampling is not working properly, the pattern will resemble this histogram. The screenshot image has been modified to improve readability by enlarging only the font portions

is targeted for adaptive sampling, 99% is rejected as off-target and the DNA is ejected after being sequenced for around 700 bp. For regular sequencing without adaptive sampling, the frequency of the sequence read length distribution does not suddenly drop off around 1000 bp, and the histogram will have a long tail towards long read lengths with a peak at several kilobases or more (Fig. 3b). In other words, when adaptive sampling is not applied, the read length distribution reflects the original library length, resulting in a typical distribution where reads of several kilobases are most frequently observed, as shown in Fig. 3b. You can check whether adaptive sampling works by performing a test run using the demo file included in the BED-Craft tool.

Limitations

Gene annotation information may vary depending on the database, and the annotation information compatible with this tool is limited. Users may need to create a BED format file containing gene location information as necessary.

While there is no upper limit on the number of input genes in the program, as mentioned above, a target size of approximately 1% of the genome is considered reasonable. However, while this program displays the proportion of the target size relative to the genome, it does not issue warnings. This is because the appropriateness of the target size depends on the user's specific requirements and perspectives, and it is not possible to define

an absolute optimal value. Furthermore, advancements in nanopore technology may alter enrichment efficiency, potentially leading to variations in the optimal target size for adaptive sampling.

This tool extracts the regions of genes registered in the public databases without considering the genomic sequences contained within the targeted locations. In complex genomic regions, such as those with multiple similar sequences (e.g., repeat regions or pseudogenes) or regions with incomplete reference genome information, there is no guarantee that adaptive sampling will always succeed.

This tool is specifically designed for nanopore sequencing. While we plan to provide follow-up updates in response to future changes in the nanopore MinKNOW version, permanent operational guarantees are not provided.

Conclusion

In this work, we have developed an easy-to-use tool to generate .bed files from gene names. This tool is expected to provide significant benefits to researchers and diagnostic laboratories performing nanopore adaptive sampling. In particular, when dealing with a wide variety of diseases or gene sets, the tool allows for easy customization of targeted genes to suit each sample, enabling adaptive sampling with great flexibility and efficiency.

Abbreviations

BED	Browser extensible data
bp	Base pairs
GTF	Gene transfer format
GUI	Graphical user interface
kb	Kilobases
ROI	Region of interest
UTR	Untranslated region

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-025-07152-z.

Supplementary Material 1.

Acknowledgements

We thank Y.Uemura, H.Yazaki, K.Tsukue, Y.Ohbayashi, Y.Wada for technical support in this study.

Author contributions

FM conceived the project, developed the software, performed validation experiments and wrote the manuscript. KK wrote the manuscript and supervised the work. All authors read and approved the final manuscript.

Funding

This work was supported by the Japan Agency for Medical Research and Development (AMED) under grant numbers grant number JP24ek0109672 (FM and KK).

Availability of data and materials

The data, tool programs, source code in this article can be freely and openly accessed at GitHub: https://github.com/medicalbioinfo/BED-Craft [12].

Declarations

Ethic approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 December 2024 Accepted: 14 February 2025
Published online: 21 February 2025

References

1. Method of the Year 2022: long-read sequencing. Nat Methods 2023; 20:1. https://doi.org/10.1038/s41592-022-01759-x
2. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53. https://doi.org/10.1126/science.abj6987.
3. Olivucci G, Iovino E, Innella G, Turchetti D, Pippucci T, Magini P. Long read sequencing on its way to the routine diagnostics of genetic diseases. Front Genet. 2024;15:1374860. https://doi.org/10.3389/fgene.2024.1374860.
4. Miya F, Kato M, Shiohama T, Okamoto N, Saitoh S, Yamasaki M, et al. A combination of targeted enrichment methodologies for whole-exome sequencing reveals novel pathogenic mutations. Sci Rep. 2015;5:9331. https://doi.org/10.1038/srep09331.
5. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat methods. 2016;13:751–4. https://doi.org/10.1038/nmeth.3930.
6. Weilguny L, De Maio N, Munro R, Manser C, Birney E, Loose M, Goldman N. Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. Nat Biotechnol. 2023;41:1018–25. https://doi.org/10.1038/s41587-022-01580-z.
7. Adaptive sampling (ADS_S1016_v1_revM_11Dec2024). https://nanoporetech.com/document/adaptive-sampling. Accessed 13 Feb 2025.
8. The Browser Extensible Data (BED) format. https://github.com/samtools/hts-specs/blob/master/BEDv1.pdf. Accessed 13 Feb 2025.
9. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604:310–5. https://doi.org/10.1038/s41586-022-04558-8.
10. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51:1560–5. https://doi.org/10.1038/s41588-019-0528-2.
11. Yamada M, Okuno H, Okamoto N, Suzuki H, Miya F, Takenouchi T, Kosaki K. Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing. Eur J Med Genet. 2023;66: 104690. https://doi.org/10.1016/j.ejmg.2022.104690.
12. Miya F. BED-Craft. GitHub. 2024. https://github.com/medicalbioinfo/BED-Craft.

Publisher's Note