[BIOINFO 2024]

Rapid precision diagnosis of repeat expansion disorders by Cas9-enrichment and long-read sequencing

Yoojung Han[1,2], Ja-Hyun Jang[3,*], and Hyeshik Chang[1,2,4,*]

[1] *Center for RNA Research, Institute for Basic Science (IBS), Seoul National University*
[2] *Interdisciplinary Program in Bioinformatics, Seoul National University*
[3] *Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine*
[4] *School of Biological Sciences, Seoul National University*
**Corresponding author: jahyun.jang@samsung.com, hyeshik@snu.ac.kr*

Huntington's disease, myotonic dystrophy, and fragile X syndrome are neurodegenerative disorders caused by abnormally expanded short tandem repeats in DNA. Accurate diagnosis of these repeat expansions is crucial for distinguishing these disorders from others with similar symptoms. Current diagnostic methods like Southern blotting and modified PCR have limitations: They can't identify repeats longer than 200 units and don't provide detailed genetic information such as repeat length, sequence interruptions, and methylation profiles—factors often linked to disease severity and onset age.

Nanopore sequencing offers a promising alternative, enabling rapid diagnosis with rich information. However, it has drawbacks in accuracy, ease of analysis, and cost-efficiency. To address these issues, we've developed a new diagnostic platform combining Cas9-targeted nanopore sequencing with an automated analysis pipeline called RepeatLab.

RepeatLab analyzes raw sequencing data to estimate repeat length, structure, and methylation profiles. It uses multi-step basecalling, adjusted basecall model parameters, and alternative repeat length calling strategies to improve accuracy, especially with low read coverage. This enhanced accuracy allows for cost-effective multiplexing of multiple samples and genes in a single assay.

We tested our platform using samples from 13 myotonic dystrophy type 1 (DM1) patients, 4 healthy individuals, and 4 reference cell lines. The results showed high accuracy, with a 0.98 correlation coefficient between actual and expected repeat lengths, even at a low sequencing depth of 12X. Methylation pattern analysis of individual alleles revealed detailed CpG site maps, showing associations between methylation levels and expanded repeat lengths.

RepeatLab is available on Google Colab for easy access and can also run on Linux command line interfaces for integration into larger systems. Our improved platform offers a cost-effective and comprehensive diagnostic tool for repeat expansion diseases, providing results in less than a day for under $200, suitable for various diagnostic settings.