[BIOINFO 2025]

Integrated assembly- and alignment-based long-read sequencing identifies novel de novo variants in Korean ASD families

Hyeji Lee[1,2], Seoyeon Kim[3], Nahyeon Kwon[1,2], Taejun Park[4], Wooheon Kim[1,2], Yujin Kim[1,2], Soo-Whee Kim[1,2], Chaewon Min[1,2], Yeojin Ryu[1,2], Guiyoung Bong[5], Sooyeon Lee[5], Jae Hyun Han[5,6], Paul Valdmanis[7], Jun Kim[3,8,*], Il Bin Kim[9,*], Hee Jeong Yoo[5,6,*], and Joon-Yong An[1,2,3,*]

[1]*Department of Integrated Biomedical and Life Science, Korea University, Seoul 02841, Republic of Korea*

[2]*L-HOPE Program for Community-Based Total Learning Health Systems, Korea University, Seoul 02841, Republic of Korea*

[3]*Department of Convergent Bioscience and Informatics, College of Bioscience and Biotechnology, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea*

[4]*School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul 02841, Republic of Korea*

[5]*Department of Psychiatry, Seoul National University Bundang Hospital, Seongnam 463707, Republic of Korea*

[6]*Department of Psychiatry, Seoul National University College of Medicine, Seoul 03080, Republic of Korea*

[7]*Department of Neurology, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, University of Montreal, Quebec H2L 4M1, Canada*

[8]*Graduate School of Life Sciences, College of Bioscience and Biotechnology, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea*

[9]*Department of Psychiatry, CHA Gangnam Medical Center, CHA University School of Medicine, Seoul, Republic of Korea*

**Corresponding author: joonan30@korea.ac.kr*

Short-read whole-genome sequencing (srWGS) often misses small indels and structural variants (SVs) within repetitive genomic regions, limiting diagnostic yield for complex disorders such as autism spectrum disorder (ASD). Long-read whole-genome sequencing (lrWGS) enables diploid genome assemblies and has revealed numerous variants invisible to srWGS. Leveraging this advantage, we reanalyzed unresolved ASD cases in Korean families to uncover overlooked variants. We performed PacBio HiFi sequencing at 30× coverage on 15 samples from four Korean ASD families, including five unresolved probands. Haplotype-resolved de novo assemblies were generated with Hifiasm and

Yak, followed by variant calling using both assembly-based (Minigraph-Cactus) and alignment-based pipelines. SNVs/indels were called with DeepVariant, SVs with Sniffles2, pbsv, and cuteSV, and STRs with TRGT. De novo variants (DNVs) were identified with Hail functions for SNVs/indels and stringent filters for SVs/STRs, and coverage titration with Seqtk was performed to evaluate detection sensitivity. Alignment-based calling detected more SNVs, while assembly-based calling identified more indels, including Korean-specific variants, though with higher Mendelian error rates. In DNV discovery, lrWGS detected substantially more variants than srWGS, particularly in repetitive and centromeric regions. Reducing parental coverage from 30× to 10× decreased concordance with srWGS and increased false positives; supplementing with high-coverage srWGS improved concordance but not specificity. Notably, we identified a 60 bp de novo deletion in intron 3 of *PRDM16*, a known ASD risk gene, which was detected by both assembly- and alignment-based lrWGS but invisible to srWGS, with Hi-C analysis suggesting long-range regulatory interactions. Overall, our study demonstrates that lrWGS can reveal clinically relevant indels and de novo variants in unresolved ASD cases, highlighting its value for identifying population-specific variants and improving diagnostic yield.