[BIOINFO 2023]

Comprehensive benchmarking of variant calling on X chromosome from large-scale whole genome sequencing data

Gang-Hee Lee[1,2], Jung Woo Park[3], Soowhee Kim[1,2], Hyeji Lee[1,2], Hee Jeong Yoo[4,5],

Junehawk Lee[3], Joon-Yong An[1,2, *]

[1] *Department of Integrated Biomedical and Life Science, Korea University*
[2] *BK21FOUR R&E Center for Learning Health Systems, Korea University*
[3] *Center for Supercomputing Applications, Division of National Supercomputing, Korea Institute of Science and Technology Information (KISTI)*
[4] *Department of Psychiatry, Seoul National University Bundang Hospital*
[5] *Department of Psychiatry, Seoul National University College of Medicine*
**Corresponding author: joonan30@korea.ac.kr*

The human X chromosome comprises more than 800 genes, making it the 8th largest chromosome in the human genome. Despite this, genome-wide association studies have overlooked the examination of variants on the X chromosome in relation to diseases and traits. This can be attributed in part to the complexities inherent in data analysis, as well as the challenges surrounding variant calling. In this study, we benchmarked whole genome sequencing (WGS) pipelines for X chromosome analyses. To achieve this, we conducted a comparative analysis of variant calling outcomes obtained using GATK and Illumina DRAGEN pipelines. We compared variant calling results by GATK and Illumina DRAGEN for large-scale WGS dataset of Korean autism families and generated joint-genotype VCF for ~2,300 individuals. Our investigation extended to evaluating the distribution of both common and rare variants on the X chromosome, alongside the precision of imputation for determining sex based on genotypic information. Furthermore, we delved into the examination of variant calls and heterozygosity within PAR (pseudoautosomal region) and non-PAR loci. We were able to delineate factors contributing to the lower quality of genetic variants within PAR. Leveraging the advantages of pedigree data, we further explored the patterns of maternal and paternal transmission of genetic variants and identified potential Mendelian violations resulting from de novo variants or inaccuracies in variant calling quality. Finally, our study enabled us to conduct genetic association tests specifically targeting de novo variants situated on the X chromosome. Through this effort, we successfully prioritized X-linked genes associated with autism spectrum disorder. Our findings offer an encompassing benchmark for the analysis of WGS data as applied to the X chromosome and provide insights into the characteristics of variant calls and transmission patterns.