An Implementation of Proactive Strategies for Clinical Genome and Transcriptome Data Utilization

Hyo Jeong Park[1], Jongkeun Park[1], and Dongwan Hong[1]
[1] *College of Medicine, Catholic University, Seoul 06591, South Korea*

The scope of genomic data utilization continues to expand, with various techniques such as sequencing panels, exome sequencing, and whole genome sequencing. These are actively employed for the diagnosis and treatment of various diseases both domestically and internationally. The proactive utilization of this data, in conjunction with state-of-the-art artificial intelligence technologies, holds the potential to advance accurate analysis, drug development, and industrial sectors. However, this data contains sensitive personal information, leading to hesitations in sharing among hospitals, companies, and research institutes over accuracy concerns.

Also, the present landscape is marked by restrictions in the utilization of genomic data due to legal and regulatory differences between countries, hampering its unrestricted use in Korea. Therefore, ongoing efforts that include international collaborations to establish standardized data-sharing protocols, advancements in privacy-preserving technologies such as homomorphic encryption and federated learning, and the development of ethical frameworks that balance data utilization with individual privacy rights are geared. Moreover, initiatives to harmonize regulatory frameworks across nations are gaining momentum, paving the way for a more unified and responsible global approach to genomic data utilization. Besides, we aim to capitalize on the distinct requirements for pseudonymization across various file types while acknowledging the need to specify and proactively utilize this approach. The diverse landscape of genomic data formats including Variant Call Format (VCF), Binary Alignment/Map (BAM), and Sequence Alignment/Map (SAM) necessitates tailored pseudonymization strategies and privacy considerations. Our study develops to enhance data reliability by validating the stability of diverse genomic data forms and providing practical methodologies for implementation. Additionally, we propose methodologies for the proactive utilization of genomic transcriptomic data in alignment with health and medical data utilization guidelines.