[BIOINFO 2023]

CWAS-Plus: Estimating genome-wide association of noncoding variation from whole genome sequencing data

Yujin Kim[1,2,3], Minwoo Jeong[4], Ronald Yurko[5], Jae Hyun Kim[1,2,3], In Gyeong Koh[1,2,3], Hyeji Lee[1,2,3], Donna M. Werling[6,7], Stephan J. Sanders[8,9,10], Joon-Yong An[1,2,3,4*]

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

[2] *Transdisciplinary Major in Learning Health Systems, Department of Healthcare Sciences, Graduate School, Korea University, Seoul, 02841, Republic of Korea*

[3] *BK21FOUR R&E Center for Learning Health Systems, Korea University, Seoul, 02841, Republic of Korea*

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

[5] *Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA*

[6] *Waisman Center, University of Wisconsin-Madison, Madison, WI, 53705, USA*

[7] *Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, 53706, USA*

[8] *Department of Psychiatry and Behavioral Sciences, Weill Institute for Neuroscience, University of California, San Francisco, CA 94143, USA*

[9] *Institute for Human Genetics, University of California, San Francisco, CA, 94158, USA*

[10] *Institute for Developmental and Regenerative Medicine, Old Road Campus, Roosevelt Dr, Headington, Oxford OX3 7TY, UK*

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

The noncoding genome contains regulatory elements that play a critical role in human development. Due to advancements in whole genome sequencing (WGS) technologies, we are now able to explore mutations within these regulatory elements. In the meantime, the recently accumulated single-cell data allows the investigation of cell-type-specific regulatory elements, facilitating the identification of cell-type-specific noncoding mutations associated with diseases. To perform a genome-wide evaluation of noncoding mutations using WGS data, an analytic framework that enables fast and easy integration of diverse functional annotations to the WGS data and empowers multiple testing comparisons is essential. Our study aims to develop CWAS-Plus, a statistical framework to perform a category-wide association test for noncoding variants and provides an efficient analysis of genome-wide noncoding associations. CWAS-Plus conducts genome-wide assessment of noncoding associations using WGS data by integrating functional annotation datasets, including cell-type-specific enhancers and promoters. Variants are categorized into functional annotation combinations, referred to as categories in CWAS-Plus, allowing an enrichment test for qualifying variants. To

evaluate the performance of CWAS-Plus, a thorough assessment was conducted using WGS data obtained from 1,991 families with autism spectrum disorder (ASD). We developed CWAS-Plus, a fast and user-friendly Python package for efficiently detecting risks associated with diverse functional annotations through effective multiple hypotheses testing (https://cwas-plus.readthedocs.io/en/latest/). From annotation to noncoding association testing, CWAS-Plus can process WGS data from approximately 4,000 individuals, along with various functional annotations, including single-cell epigenome data, within 2 hours. Our findings successfully identified noncoding categories enriched for ASD, particularly highlighting regulatory elements of excitatory neurons. Our findings showed that CWAS-Plus efficiently processes large-scale WGS data and functional annotations, enabling multiple comparisons of hypotheses. Hence, we present CWAS-Plus as an analytic framework applicable for investigating diverse genomic disorders in future studies.