Predicting Contribution of Genome-Wide Noncoding Mutations Disrupting Cell Type-Specific Regulatory Elements to Autism Risk Using CWAS-Plus

In Gyeong Koh[1,2], Yujin Kim[1,2,3], Minwoo Jeong[4], Ronald Yurko[5], Jae Hyun Kim[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]BK21FOUR R&E Center for Learning Health Systems, Korea University, Seoul, 02841, Republic of Korea; [3]Transdisciplinary Major in Learning Health Systems, Department of Healthcare Sciences, Graduate School, 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

The noncoding genome includes regulatory elements crucial for development. Recent advances in whole-genome sequencing (WGS) and the accumulating single-cell data offer opportunities to explore regulatory elements governing cell type-specific gene expression patterns and dynamics and to identify disease-associated noncoding mutations in these regions. In the past decades, understanding of the genetic factors of autism spectrum disorder (ASD) has established in de novo mutations (DNMs) disrupting protein-coding genes, accounting for ~30% of ASD cases. However, the association of noncoding DNMs in intergenic and intronic with ASD remains poorly understood. In this study, we present CWAS-Plus, a Python-based analytical tool for detecting disease-associated noncoding mutations utilizing cell type-specific regulatory elements. CWAS-plus integrates multilayered functional annotations involving regulatory elements, conservations, and genes of interest and evaluates genome-wide noncoding associations. We collected cell type-specific regulatory elements obtained from the human prefrontal cortex from fetal to adult. Then, we conducted CWAS-Plus for WGS data from 1,902 ASD cases and 1,902 controls. DNMs were categorized into combinations of functional annotations, called ‘category’ in CWAS-plus, serving as unique hypotheses for enrichment tests between phenotypes. Further, we evaluated the performance in predicting ASD risk and identified the categories contributing to ASD risk using a lasso regression. For feature selection, we employed two methods: 1) selecting the annotations with $R^2 > 0$ and 2) extracting contributing annotations from the coefficients. From categorization to category-wide association tests, CWAS-plus processed ~4,000 WGS data and diverse functional annotations within 3 hours. We identified noncoding categories enriched in ASD, particularly highlighting early excitatory neuron-specific regulatory elements. Furthermore, we predicted noncoding-associated ASD risk and identified the categories contributing to the risk through feature selections. Consequently, we suggest that CWAS-Plus processes large-scale WGS data and functional annotations and is applicable for investigating disease-associated noncoding mutations in future studies on a vast scale.