[BIOINFO 2025]

Benchmarking Diagnostic Performance of Genomic Language Model

Woojong Sim[1], Jiho Lee[1], Jayoung Park[2], Yujin Kim[3,4], Soowhee Kim[3,4], Yeojin Ryu[3,4] , Hwanseok Sim[5] , Joon-Yong An[3,4,5,*] and Minji Jeon[1,6,**]

[1]*Department of Biomedical Sciences, Korea University College of Medicine*
[2]*Department of Biomedical Informatics, Korea University College of Medicine*
[3]*Department of Integrated Biomedical and Life Science, Korea University*
[4]*L-HOPE Program for Community-Based Total Learning Health Systems*
[5]*School of Biosystems and Biomedical Sciences, Korea University College of Health Sciences*
[6]*Department of Biomedical Informatics, Biomedical Research Center, Korea University Anam Hospital*
**Corresponding author: joonan30@korea.ac.kr*
***Corresponding author: mjjeon@korea.ac.kr*

Genomic language models (gLMs) have emerged as powerful tools for interpreting noncoding DNA, beyond the limitations of traditional variant effect predictors that rely on fixed annotations or sequence alignments. While these deep learning models have shown promise for recognition of functional regions through learning on regulatory grammar from raw sequence, recent reviews and benchmarks suggest a significant performance gap exists between the zero-shot capabilities of pre-trained gLMs and highly specialized biological tasks such as variant effect prediction and gene expression regulation by non-coding elements. This disparity arises due to the fact that language modeling is a separate field from understanding biological roles of genomic elements, highlighting the need for rigorous benchmarking on clinically relevant tasks. Here, we present the first systematic benchmarking of genomic language models, including DNABERT, DNABERT-2, Nucleotide Transformer, PhyloGPN, HyenaDNA and Evo 2 for whole-genome disease prediction. Using whole-genome sequencing data from 14,606 individuals across three major ASD cohorts, we trained a Set Transformer on variant-centered embeddings from each gLM and predict ASD diagnosis. DNABERT-2 consistently outperformed other models(AUROC = 0.5568), though overall predictive power remained modest. Fine-tuning marginally improved model performance, while integrating variant-level functional annotations such as conservation scores and regulatory impact predictions significantly enhanced prediction. Notably, high-impact variants prioritized by the models were enriched in ultra-conserved noncoding regions, implicating regulatory regulatory disruption as a

mechanism in ASD pathogenesis. These findings provide the systematic evaluation of gLMs for complex disease prediction and emphasize the need for integrative approaches to realize their full potential in complex trait diagnosis.