[BIOINFO 2025]

Genomic Language Model Fine-tuning for Rice Phenotype-Genotype Interpretation

Dasol Kim[1], Hee-Jong Koh[2*], and Tae In Ahn[1*]

[1,2] *Department of Agriculture, Forestry and Bioresources, Seoul National University*
**Co-corresponding author: Hee-Jong Koh (heejkoh@snu.ac.kr), Tae In Ahn (tiahn@snu.ac.kr)*

Plant breeding researchers commonly employ genome-wide association studies (GWAS) to screen genotype-phenotype relationships across species. However, determining causal variants requires experimental confirmation, a time-consuming process. Genomic language models (gLMs) offer new opportunities for interpreting genetic sequences and examining how genetic elements influence complex traits, potentially complementing traditional GWAS.

We aimed to identify high-effect variants on cooked rice texture traits using genomic sequence data. Texture Profile Analysis investigated phenotypes (hardness, adhesiveness, resilience) in 291 rice accessions, including temperate and tropical japonica subpopulations.

The Genomic Pre-Trained Network (GPN), a publicly available genomic language model (Benegas et al., 2023), was trained on the rice reference genome (Nipponbare, IRGSP v1.0), capturing genomic patterns influencing gene expression and phenotypes. For fine-tuning, we incorporated genetic variation data from whole-genome resequencing and phenotypic mapping of 191 diverse rice varieties. The fine-tuning focused on learning associations between genomic variants and texture phenotypes through supervised learning. Model validation used 100 rice varieties, comparing predictions to GWAS results. GWAS employed high-quality SNPs with stringent quality control, and linkage disequilibrium analysis assessed association panel structure. GPN successfully identified statistically significant variants associated with rice texture, showing meaningful correlation with GWAS.

These findings suggest genomic language models can serve a role analogous to GWAS in identifying genotype-phenotype associations, while offering advantages in capturing sequence context. This approach provides complementary insights into the genetic architecture of rice texture and supports precision breeding for improving texture quality.