
**Leveraging Large Language Models (LLMs) for Disease Risk Prediction and
Personalized Gene Expression Modeling with Whole-Genome Sequencing Data**

Wanwen Zeng and Wing Hung Wong
Department of Statistics
Department of Biomedical Data Science
Bio-X Program
Stanford University

October 4, 2024

1:30PM-2:50PM

Clark S361

Abstract:

This talk introduces methods for advancing disease risk prediction and personalized gene expression modeling using whole-genome sequencing (WGS) data. In the first part, we will focus on utilizing large language models (LLMs) to enhance disease risk prediction by integrating rare and *de novo* variants from biobank-scale WGS datasets, such as the UK Biobank. This approach significantly improves predictive accuracy and biological interpretability by capturing the role of rare variants in disease susceptibility. In the second part, we will present a new genetic LLM designed to predict personal gene expression. Using a pretraining and fine-tuning strategy, this model learns individual-specific gene regulatory patterns with a limited number of training instances. This genetic LLM offers deeper insights into how genetic variation influences personal gene regulation. Together, these approaches underscore the potential of LLMs to bridge population-level disease risk assessment with individual-level gene regulation.

Reading List:

How to improve polygenic prediction from whole-genome sequencing data by leveraging predicted epigenomic features? Biorxiv.