# Development and validation of a genome-wide SNP model that stratifies *Helicobacter pylori* strains across Correa's cascade

Xiuling Song

Department of Laboratory Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, Guangdong Province, China

## Abstract

**Background:** While *Helicobacter pylori* (*H. pylori*) infection is the primary cause of gastric cancer, only a minority of infected individuals progress along Correa's cascade from gastritis to malignancy. Classical virulence markers do not completely explain this heterogeneity, suggesting a role for broader, strain-specific genomic variation.

**Methods:** We conducted a bacterial genome-wide association study (GWAS) on 528 high-quality H. pylori whole-genome sequences, encompassing non-atrophic gastritis (NAG), atrophic gastritis (AG), intestinal metaplasia (IM), and gastric cancer (GC). A machine-learning framework was used to integrate associated variants into a random forest–based *H. pylori* Genomic Risk Score (HpRS) model, designed to distinguish strains associated with advanced (IM/GC) versus early (NAG/AG) lesions. Model performance was rigorously evaluated by cross-validation and an internal held-out test set, and externally validated in independent cohorts.

**Results:** Phylogenetic analyses showed that genome-wide clustering was primarily driven by geographic lineage rather than disease stage, indicating that stage association is not explained by lineage alone. HpRS achieved strong discrimination in repeated cross-validation (mean AUC = 0.902, 95% CI: 0.892–0.912), remained discriminatory in an internal held-out set (AUC = 0.780), and generalized to two independent cohorts (BV-BRC: AUC = 0.871; hospital cohort distinguishing IM vs NAG/AG: AUC = 0.843). Top predictive variants mapped predominantly to genes involved in core cellular functions—including DNA repair (nth), translation (ileS), central metabolism (tkt), and lipid metabolism (fabX)—rather than classical virulence factors, with multiple high-impact variants localized to conserved protein domains.

**Conclusion:** We developed and validated a genome-wide SNP-based risk model that quantifies the propensity of *H. pylori* strains to be associated with advanced stages of Correa's cascade. These findings support a polygenic architecture involving core metabolic and homeostatic pathways, and HpRS provides a foundation for strain-aware risk stratification in gastric cancer prevention.