Abstract:

SARS-CoV-2, the causative agent of COVID-19, has exhibited extensive evolutionary change since its emergence in late 2019. While inter-host transmission and the emergence of major variants have been widely studied, relatively little attention has been given to intra-host diversity (IHD) at population scales. Intra-host diversity—defined as the presence of multiple viral variants within an individual—can influence viral adaptability, immune evasion, and clinical outcomes. In this study, we analyzed thousands of publicly available SARS-CoV-2 genomes from the United States to investigate temporal variation in intra-host diversity and to identify factors contributing to this variability. Our results show that clade richness, reflecting the number of distinct viral lineages circulating in the population, is significantly associated with increased intra-host diversity ($R^2 = 0.589$, $p < 0.001$). By contrast, national case counts and sequencing depth were not significant predictors of intra-host diversity. These findings suggest that intra-host diversity is shaped more by the ecological and evolutionary dynamics of circulating strains than by overall case burden or technical factors. Expanding this work to include global datasets and incorporating additional epidemiological variables will provide deeper insights into the drivers of intra-host diversity and its implications for viral evolution, transmission, and treatment outcomes.