Robust Foundations for Host-Microbiome Systems Medicine

Systems medicine seeks to address questions about human health in a holistic manner, modeling disease progression or treatment intervention at multiple physiological levels. In practice, this involves measurement and integration of high-throughput data (e.g. gut microbiota composition, gene expression, circulating metabolites) in order to capture a complex biological signal. Confidently linking these signals to disease states remains challenging due to underpowered clinical studies and a poor knowledge of covariates, both of which potentiate confounding in the context of biomarker identification. Furthermore, there is a longstanding lack of consensus regarding one of the most fundamental statistical tasks performed in clinical microbiome research: differential abundance analysis. This project was a dual effort to (i) comprehensively evaluate and (ii) apply computational methods that address these challenges.

In order to establish points of reference for method behavior and make recommendations, benchmarking studies typically rely on simulated data to assess method performance under varied yet controlled conditions. Yet, for insights to translate to real-world applications, it is critical that simulations reflect the complexity of real experimental data. Thus, the first project in this dissertation addressed the even more fundamental lack of consensus regarding how best to simulate microbiome data, where *both* a ground truth and maximally realistic data are needed. A novel simulation framework was implemented and quantitatively verified to reproduce realistic disease signals, as well as signals resembling confounding covariates. Across an unprecedented parameter space, these simulations were used to evaluate 18 differential abundance methods, and found linear (mixed-effect) models to be a robust and flexible solution when applied to human-associated microbiota.

As a research target, the gut microbiota has exploded in recent years owing to a deeper understanding of its relevance in host physiology, metabolism, and immunity. Subsequently, many chronic or acute inflammatory diseases implicate gut bacterial community composition and dynamics in their etiologies or treatment responses. In collaboration with clinicians, the second and third projects in this dissertation explored (i) the gut-joint axis in a mixed cohort of spondyloarthritis, Crohn's disease, and acute anterior uveitis, as well as (ii) the integrated multi-omics of hospitalized COVID-19 patients. Differential abundance analyses were performed as informed by the first aim of this dissertation, and potentially confounding covariates, such as clinical or lifestyle factors, were mitigated wherever possible. Both projects generated mechanistic hypotheses that must still be validated in further computational and especially experimental studies. In order to summarize recommendations for rational study design in the context of clinical host-microbiome analysis, this dissertation also consolidates insights from all three projects.