ABSTRACT

Biobanks linked to massive, longitudinal electronic health record (EHR) data make numerous new genetic research questions feasible. Among these is the study of biomarker trajectories. For example, high blood pressure measurements over visits strongly predict stroke onset, and consistently high fasting glucose and Hb1Ac levels define diabetes. Recent research reveals that not only the mean level of biomarker trajectories but also their fluctuations, or within-subject (WS) variability, are risk factors for many diseases. Glycemic variation, for instance, is recently considered an important clinical metric in diabetes management. It is crucial to identify the genetic factors that shift the mean or alter the WS variability of a biomarker trajectory. Compared to traditional cross-sectional studies, trajectory analysis utilizes more data points and captures a complete picture of the impact of time-varying factors, including medication history and lifestyle. Currently, there are no efficient tools for genome-wide association studies (GWAS) of biomarker trajectories at the biobank scale, even for just mean effects. We propose TrajGWAS, a linear mixed model-based method for testing genetic effects that shift the mean or alter the WS variability of a biomarker trajectory. It is scalable to biobank data with 100,000 to 1,000,000 individuals, many longitudinal measurements and robust to distributional assumptions. Simulation studies corroborate that TrajGWAS controls the type I error rate and is powerful. Analysis of eleven biomarkers measured longitudinally and extracted from UK Biobank primary care data for more than 150,000 participants with 1,800,000 observations reveals novel loci that significantly alter the mean or WS variability.