Inter-individual variation in symptoms and outcomes necessitates disease subtyping, yet most methods rely on phenotypic data, producing unstable groupings that often capture confounders like disease progression. As genetic variation is stable, delineating subtypes by genetic differences provides insights into the causal mechanistic differences. Here we present ΔOCCUR, a phenotype-agnostic, unsupervised framework that quantifies variant co-occurrence under the hypothesis that variants contributing to the same subtype will co-occur more frequently across affected individuals. This approach identifies clusters of genetic variants associated with distinct disease mechanisms that can be used to calculate partitioned polygenic risk scores (pPGS) to stratify individuals into disease subtypes.

We first validated ΔOCCUR using simulated heterogeneous phenotypes comprised of two distinct diseases, such as asthma and gout, where pPGS stratified individuals into their true phenotypes. Applied to type 2 diabetes, which has known heterogeneity, our phenotype-free approach recaptured clinical and genetic subtypes identified by benchmark studies using supervised subtyping. Notably, the identified genetic subtypes had opposing association profiles across key risk factors, such as BMI and total cholesterol, which were obscured when the genetic risk was aggregated. We also identify a novel genetic subtype of type 2 diabetes linked to impaired liver function with reduced hepatic insulin clearance, that carries excess genetic risk in individuals of African ancestries. These genetic subtypes were not driven by population structure and replicated in individuals of South Asian ancestry. Lastly, we highlight the translational utility of our framework for biomarker and drug-target discovery by identifying plasma proteins with differential associations to genetic subtypes, including known drug targets such as PCSK9 and IL5RA.

Taken together, we demonstrate that ΔOCCUR generalises across diseases as an unsupervised framework to enable the identification of mechanistic subtypes and biomarkers of any disease. The approach overcomes ongoing limitations associated with phenotype-driven subtyping methods by analysing genetic causes of complex traits and diseases.