An Integrated Data Explorer for Genomic Studies on Mental Disorders

A web-based query interface service is established to allow integrated access to heterogeneous data sources from the NIH Center for Collaborative Genetic Studies of Mental Disorders (CGSMD). These data sources are collected from decades of large-scale, high density psychiatric genetic studies for autism, bipolar disorder, schizophrenia, depression and Alzheimer's disease.

The National Institute of Mental Health Center for Collaborative Genetic Studies of Mental Disorders (CGSMD) established in 1998 has accumulated a large collection of clinical and genotype data from a variety of high density genotyping studies. Access to data from multiple studies offers the possibility of significantly increasing statistical power by allowing investigators to conduct meta-studies that combine existing data in novel ways. Access to raw data is not enough; the data in different studies must be integrated so that the concepts, variables and values from different studies are compatible both semantically and syntactically. We have employed state-of-the-art data integration and synthesis approaches to permit uniform, semantically-consistent query access for investigators to heterogeneous data sources in CGSMD. A web-based query interface allows an investigator to combine data across studies in order to increase sample size, combine across diseases to search for common genetic associations, or query for data on individuals who have been evaluated for a particular phenotype. Currently, the data explorer allows access to 63,843 samples from studies of autism, bipolar disorder, schizophrenia, depression and Alzheimer's disease with ~7,046 phenotypic variables and ~2,317,995 genotype data. Such integrated data available may be among the largest collections of both genetic and phenotypic data available to the broader scientific community. We expect that our proposed data integration system will empower scientists to more fully explore the genetic architecture of mental disorders with much greater flexibility and richness than previously feasible, and that it will enable them to efficiently test novel scientific hypotheses.