IMPACT OF BATCH EFFECT AND STUDY DESIGN BIASES ON IDENTIFICATION OF GENETIC RISK FACTORS IN SEQUENCING DATA

EXECUTIVE SUMMARY

To explore how systematic biases within a genomic data set can impact downstream statistical analysis of genetic variants, the research team conducted stratified association analysis on Alzheimer’s disease genomic data. The researchers profiled a set of variants with highly significant, novel associations with Alzheimer’s disease that were impacted by heterogeneity in subcohort composition and exome capture. The team identified genotype quality, age, and population stratification as factors contributing to differences in minor allele frequencies as well as a batch effect across sequencing center cohorts. These findings highlight important considerations for analysis of this data set and for the design of future studies.

RESEARCH CHALLENGE

The research team conducted stratified association analysis on Alzheimer’s disease genomic data. The researchers profiled a set of variants with highly significant, novel associations with Alzheimer’s disease that were impacted by heterogeneity in subcohort composition and exome capture. The team identified genotype quality, age, and population stratification as factors contributing to differences in minor allele frequencies as well as a batch effect across sequencing center cohorts. These findings highlight important considerations for analysis of this data set and for the design of future studies.

RESULTS & IMPACT

The team also found several age-related differences between Broad and the other two centers. First, the Broad samples had higher minor allele frequency in both cases and controls, possibly indicating population stratification. Second, the Broad cohort consisted of a disproportionately large number of younger individuals, which may affect the ability to detect true genotype–trait associations. Common batch effects include different sequencing centers, different sample collection protocols, and different exome capture kits. For example, the Alzheimer’s Disease Sequencing Project sequenced exomes of more than 10,000 cases and controls using three sequencing centers, different sample collection protocols, and different exome capture kits. In addition, the controls were intentionally older than cases in an effort to increase the confidence of the Alzheimer’s variants because “true” disease-causal variants should be absent in older but cognitively normal individuals. This design introduced an age variable confounded with disease status. The research team studied both batch effects and confounding variables and demonstrated that both significantly impacted the association data of this study.

METHODS & CODES

Samples in this public data set were sequenced by three centers using two exome capture kits: Broad Institute (Illumina Rapid Capture Exome kit), Washington University (Nimblegen VCRome v2.1 kit), and Baylor College of Medicine (Nimblegen VCRome v2.1 kit). The research team aligned paired-end reads to the hg19 and hg38 human reference genomes using Novoalign and BWA, and called variants using the Genome Analysis Toolkit. Following variant calling, the association analysis included only variants located in the common capture regions of the two exome kits. The variant associations with Alzheimer’s disease were adjusted for sequencing center, gender, and population stratification, which are important for accurate interpretation of genetic associations.

PUBLICATIONS & DATA SETS
