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

Allocation: Illinois/280 Knh
PI: Matthew E. Hudson[1]
Co-PIs: Yan W. Asmann[2], Liudmila Sergeevna Mainzer[1]

[1]University of Illinois at Urbana–Champaign
[2]Mayo Clinic

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 likely contributing factors to vastly different 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 large sample sizes required to identify disease-associated variants in genomic sequencing data often introduce batch effects and other confounding variables or biases from study design. If not adequately addressed in the analysis, batch effects will reduce statistical power and increase false associations, and biases in study design, which may further hinder 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 and two 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 analysis of this data set.

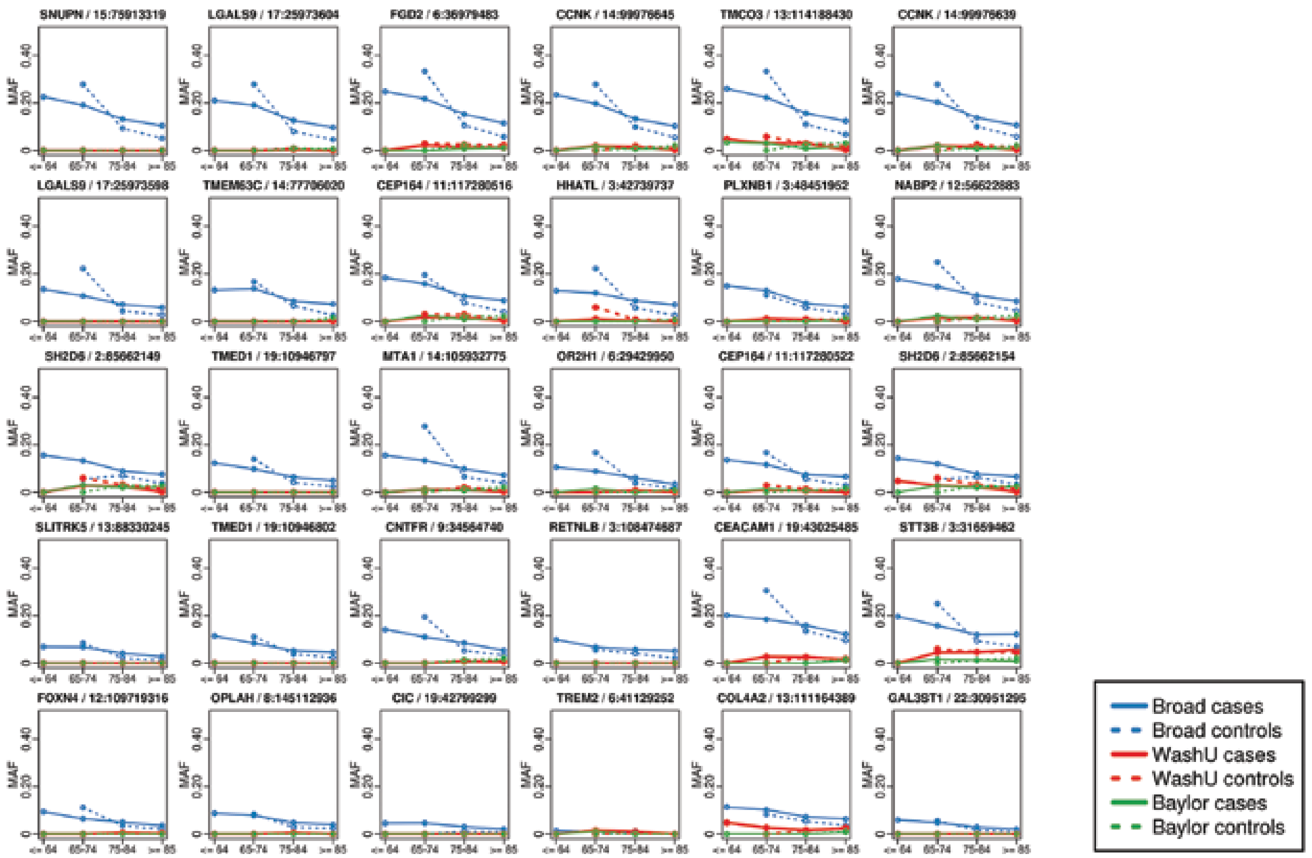

Figure 1: Minor allele frequencies in cases and controls stratified by sequencing facility and age group. Only the samples from Broad showed appreciable minor allele frequencies at the 30 variants identified as significant in the full-cohort analysis.

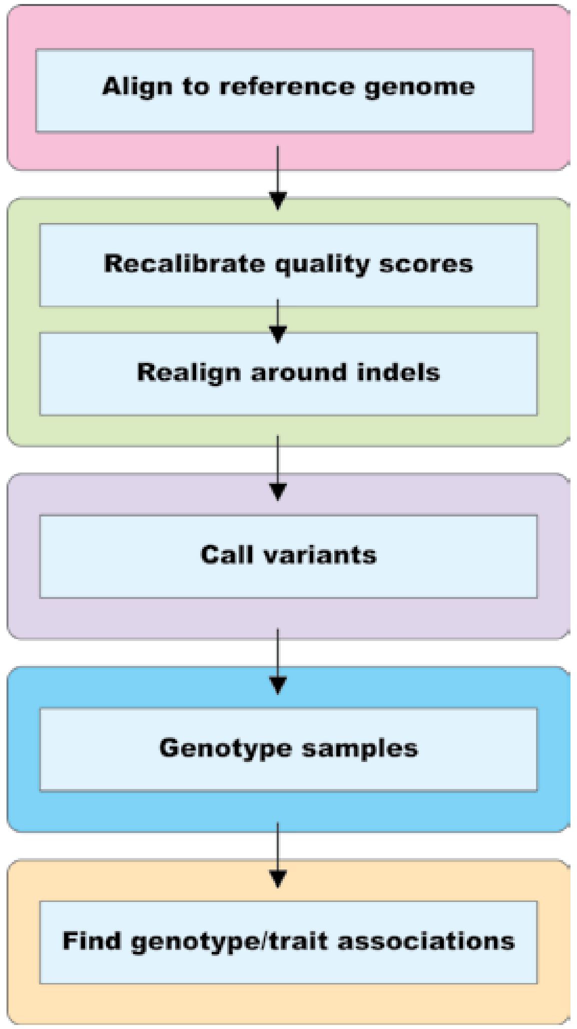
Align to reference genome
Recalibrate quality scores
Realign around indels
Call variants
Genotype samples
Find genotype/trait associations

Figure 2: Workflow configuration used in 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, the first four principal components underlying population substructure, and apolipoprotein E (APOE) genotype (a genetic risk factor for dementia).

RESULTS & IMPACT

The research team identified 30 novel Alzheimer's-associated genomic variants with exome-wide significance. Further examination showed that the significance of these variants originated entirely from samples processed at Broad, which used the Illumina capture kit. To investigate the cause of Broad-exclusive significance, the researchers compared multiple variant quality parameters including genotype quality, sequencing depth, and alternative allele concentration, and identified significant batch differences in genotype quality values of the variants from Broad (Illumina kit) compared to Washington University and Baylor (Nimblegen kit). 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 cases. Finally, the minor allele frequencies of the 30 variants declined with age in both Broad cases and controls, suggesting that the variants are associated with age.

WHY BLUE WATERS

The Alzheimer's Disease Sequencing Project data set used in this study consists of over 9,000 whole-exome sequencing samples. Processing this immense quantity of genomic data and conducting the downstream analysis required 250,000 node hours on Blue Waters. By parallelizing computational jobs across thousands of nodes, the research group was able to accomplish what would take over 100 years on a single server in months. In addition, Blue Waters is one of the few systems that allows users to keep hundreds of terabytes of data in active storage for simultaneous processing, an important step in data integration across genomic workflows.

PUBLICATIONS & DATA SETS

Y. Ren *et al.*, "Identification of missing variants by combining multiple analytic pipelines," *BMC Bioinform.*, vol. 19, 2018, doi: 10.1186/s12859-018-2151-0.

D. P. Wickland *et al.*, "Impact of batch effect and study design biases on identification of genetic risk factors in sequencing data," in preparation, 2019.