Research Challenge

Alzheimer’s disease (AD) is a neurodegenerative dementia that affects more than five million Americans and more than 35 million people worldwide. This project will deliver novel genomic variants that have remained hitherto undetected by the standard workflow in AD. These variants will be posted into public databases for use by researchers and clinicians worldwide to improve understanding of the genomic underpinnings of AD, as well as drug development and treatment outcome prediction.

Methods & Codes

The team tested a select number of steps and parameters in the variant detection pipeline in the context of sample sizes. They grouped the Alzheimer Disease Sequencing Project (ADSP) samples into different samples sizes of 50, 500, 1000, 2000, 5000, and 10,000. For each sample size, they tested two different aligners, three different variant callers, multi- vs. single-sample variant calling, and five different parameter settings in the variant calling and quality control process. The goal was to build a new set of guidelines for variant discovery based on project size.

Results & Impact

By combining two read aligners and several variant callers into the workflow, they were able to recover 50% of the variants in the ADSP data that were missed by the standard protocol. They further annotated SNPs, or genetic variations in a single DNA building block, as synonymous or nonsynonymous, and assessed the proportion of alternate alleles between cases and controls. They discovered SNPs within genes previously reported to interact with Alzheimer's-related proteins or to function in the brain.

Why Blue Waters

The team tested multiple combinations of steps and parameters using data from the Alzheimer's Disease Sequencing Project consisting of over 10,000 whole exome sequencing samples. The total amount of time that would be required to complete this project on a single server would be 109 years. On Blue Waters, they were able to run a single workflow on the entire set of 10,000 AD samples by parallelizing across thousands of nodes.