IDENTIFICATION OF MISSING VARIANTS IN ALZHEIMER'S DISEASE, AND THE NEW STANDARDS FOR GENOMIC VARIANT IDENTIFICATION IN LARGE COHORTS

Allocation: Illinois/350 Knh
PI: Liudmila Mainzer[1]
Co-PIs: Yan Asmann[2], Matthew Hudson[1]

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

EXECUTIVE SUMMARY

Alzheimer's disease (AD) is the sixth-leading cause of death in the United States. While AD is genetically determined, no cure exists and onset remains difficult to predict, as the disease is influenced by a combination of rare and common genomic variants. Correct and complete identification of such variants is essential for advancing research. We have formulated new variant-calling procedures to recover variants previously missed by the community, and have applied this "integration" approach to reanalysis of 10,000 whole human exomes (protein-coding genes) from patients afflicted by AD. By combining two read aligners and several variant callers, we were able to recover 50% of variants that were missed by the standard protocol. This project has delivered a set of newly discovered AD variants for submission into public repositories, as well as new standards for scaling variant calling to cohorts containing tens or hundreds of thousands of samples.

RESEARCH CHALLENGE

Alzheimer's disease (AD) is a neurodegenerative dementia that affects more than five million Americans and more than 35 million people worldwide. It poses an increasing burden to healthcare due to the progressive aging of society. It is hypothesized that AD is shaped by genomic mutations that are highly diverse among afflicted individuals; we have shown that common analytic practice misses a substantial percentage of good-quality genomic variants. Getting the complete variant call set, especially the rare variants, is the critical prerequisite to successful identification of disease predisposition markers and druggable targets in human disease. Our 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 our understanding of the genomic underpinnings of AD, as well as drug development and treatment outcome prediction.

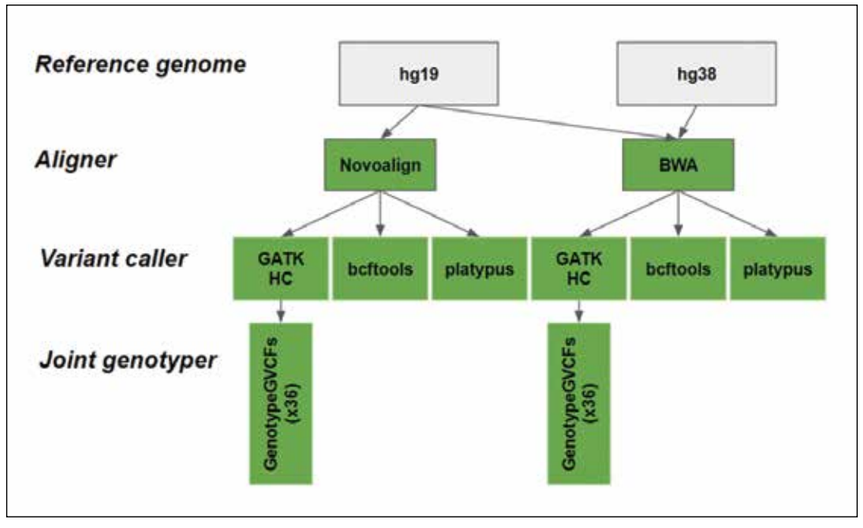

Figure 1: Workflow configurations utilized in this study. Two reference genomes, two aligners, and three variant callers have been included in various permutations to recover as many variants as possible. Joint genotyping was performed where possible.

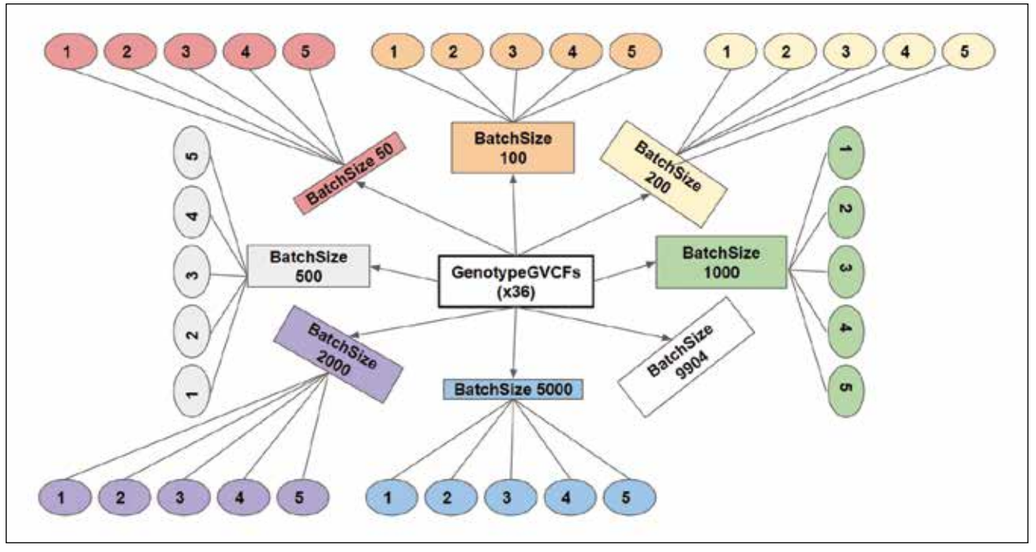

Figure 2: To study the effects of sample size, we performed joint genotyping on variable-size batches subsampled from 10,000 whole exomes. We subsampled each batch size five times, and ran joint genotyping to see how many variants could be recovered. See https://github.com/dpwickland/JointGenotyping_BW.

METHODS & CODES

We tested a select number of steps and parameters in the variant detection pipeline in the context of sample sizes. There were no published works to prioritize the set of steps or parameters to test. We grouped the Alzheimer Disease Sequencing Project (ADSP) samples into different samples sizes of 50, 500, 1,000, 2,000, 5,000, and 10,000. For each sample size, we 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

We constructed a flexible workflow using the Swift/T workflow management language that allowed easy swapping of tools, quality control, sample analysis parallelization, and job dependencies. By combining two read aligners and several variant callers into our workflow, we were able to recover 50% of the variants in the ADSP data that were missed by the standard protocol. Importantly, recovered variants had higher proportions of low-frequency variants, which are of most interest. We 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. We found 47 SNPs within 41 genes that resulted in a switch to a less-frequent codon and showed a greater percentage of alternate alleles in the cases rather than in the controls. 14 of these SNPs, and seven of the top 10 most significant, lay within genes previously reported to interact with Alzheimer's-related proteins or to function in the brain.

WHY BLUE WATERS

Our study utilizes data from the Alzheimer's Disease Sequencing Project, consisting of over 10,000 whole exome sequencing samples. For each sample, we tested multiple combinations of steps and parameters in the variant detection pipeline. Due to the need to test many different parameter combinations, this study requires a petascale resource. This work is unprecedented for human diseases and traits because the "workflow integration" approach necessary to overcome the inadequacy of individual variant-calling procedures is computationally prohibitive outside a petascale resource like Blue Waters. The total amount of time that would be required to complete this project on a single server would be 109 years. On Blue Waters, we were able to run a single workflow on the entire set of 10,000 AD samples by parallelizing across thousands of nodes. Further, we integrated results across all runs using the cleaned BAMs produced by multiple aligners, because Blue Waters is one of the very few systems that allows users to keep hundreds of terabytes of data in active storage for simultaneous processing.

PUBLICATIONS & DATA SETS

Ren, Y., et al., Identification of missing variants by combining multiple analytic pipelines. *BMC* 19:139 (2018), DOI:10.1186/s12859-018-2151-0.