EPISTATIC INTERACTIONS FOR BRAIN EXPRESSION GWAS IN ALZHEIMER'S DISEASE

Allocation: Private sector/31 Knh
PI: Nilufer Ertekin-Taner[1]
Collaborators: Mariet Allen[1]; Liudmila Mainzer[2,3]; Curtis Younkin[1]; Victor Jongeneel[2]; Thierry Schüpbach[4]; Gloria Rendon[2]; Julia Crook[1]; Julie Cunningham[5]; Summit Middha[5]; Chris Kolbert[5]; Dennis Dickson[1]; Steven Younkin[1]

[1]Mayo Clinic in Jacksonville
[2]University of Illinois at Urbana-Champaign
[3]National Center for Supercomputing Applications
[4]Swiss Institute of Bioinformatics
[5]Mayo Clinic in Rochester

EXECUTIVE SUMMARY:

It is well established that the risk for Alzheimer's disease (AD) is under substantial genetic control, and is thought to arise from multiple genetic variants. Such disease-associated variants can be identified using expression-based genome-wide association studies (eGWAS), based on the rationale that some variants will influence AD risk via their effects on brain gene expression. We hypothesize that some of the risk for AD may be due to the interaction of two or more genetic variants (epistasis). The aim of our Blue Waters project was to test for the presence of epistatic interactions that influence brain gene expression levels using data from 359 temporal cortex samples (181 AD, 178 non-AD), 223,632 SNP genotypes and ~24,526 transcripts that were measured using an expression array. We also ran an analysis on 343 cerebellum samples (173 AD and 170 non-AD). The analysis of epistatic effects in studies of this size would not be possible without the unique computing capabilities of Blue Waters. All the planned computational runs have now been completed, along with the analyses of scalability and performance.

INTRODUCTION

We have previously collected gene expression measures from pathologically confirmed AD subjects (test group) and those with non-AD pathologies (control group, Table 1) from two brain regions: temporal cortex (TCX) and cerebellum (CER). Identifying genetic variants that associate with altered gene expression levels in these subjects may pinpoint novel risk factors for AD. We investigated single genetic variants for association with these gene expression measures and found significant expression quantitative trait loci [1]. We also determined that some of the known AD risk variants likewise associate with expression of nearby gene(s) thus implicating the potential mechanism of action and the affected gene at these loci [2].

For our study on Blue waters, we hypothesized that pairs of variants may likewise influence gene expression through an interaction known as epistasis. Identifying additional genetic factors that influence AD risk can provide further insights into the pathophysiology of this disease and may have a significant impact on the development of novel therapeutic targets, identification of potential, premorbid biomarkers, and generation of in vivo disease models, much needed for pre-clinical development and testing of novel therapies.

METHODS AND RESULTS

Three groups of subjects with temporal cortex measures were analyzed: AD's-only, Non-AD's only and the two combined (AD+nonAD); the combined set only (AD+non-AD) was assessed for cerebellum (Table1, Figure 1). Prior to launching our analysis, we implemented conservative quality control measures (Figure 1) and LD pruned our dataset in order to capture the maximum genetic data whilst minimizing the multiple testing penalty, similar to a protocol described elsewhere [3].

The currently available epistasis approaches are unable to efficiently incorporate covariates into regression models. To address this we generated gene expression residuals using R for all 24,526 expression measures to account for the following key covariates: Age, Gender, #ApoEε4 alleles, PCR plate, RIN, RINsqAdj (RIN-RINmean)2 and diagnosis when appropriate, (AD=1, Non-AD=0), as described previously [4].

Three different software programs were considered for detecting the epistatic interactions: PLINK [5], EpiGPU [6] and FastEpistasis [7]. It was determined that FastEpistasis performed most optimally for testing of multiple quantitative phenotypes using the computational architecture of Blue Waters. FastEpistasis builds on the analysis paradigm used in PLINK, but is multithreaded and runs up to 75 times faster by splitting the analysis into three phases.

All the planned computational runs have now been completed, along with the analyses of scalability and performance. Results will be organized into a database to facilitate efficient filtering to identify relevant significant results. The most significant findings will be tested for validation in additional subjects using a targeted approach.

WHY BLUE WATERS

The Blue Waters supercomputer was instrumental for the success of this work. Our project simply would not have been possible without this supercomputer and its dedicated team.

First, its sheer size allows us to run all phenotypes in parallel, at the same time. While FastEpistasis is not MPI-enabled, we were still able to pack up to 32 phenotypes per compute-node. Additionally, we utilized the MPI launcher software developed by the Blue Waters staff, to run all 24,526 phenotypes in parallel in a single 787-node reservation. This configuration was used for all four experiments: TCX-AD, TCX-Control, TCX-ALL, CER-ALL. As a result, each of these tests took only 5-10 hours to run, as opposed to the two years of walltime predicted to be necessary for PLINK.

Second, the analysis presented certain challenges that required a close collaboration between the Mayo team and the Blue Waters support group:

- We have learned to monitor our jobs using resource profiling software developed by NCSA staff. This helped us detect whether jobs were progressing normally.
- FastEpistasis generated millions of files across the project, and we worked with the Blue Waters team to manage data storage and transfers in efficient ways.
- Heterogeneity of data across phenotypes also resulted in uneven walltimes within each multi-node job, and NCSA staff helped us study the impact of this property on computational cost of the jobs, so as to better plan future analyses.

Both the hardware access and staff support were required to complete this project.

Tissue, Diagnosis	N	Female (%)	Mean Age (SD)	ApoE4+ (%)	Mean RIN (SD)
TCX, AD*	181	94 (52)	73.6 (5.6)	108 (60)	6.3 (0.8)
TCX, Non-AD*	178	67 (38)	71.5 (5.6)	46 (26)	6.9 (1.0)
TCX, ALL*	359	161 (45)	72.5 (5.6)	154 (43)	6.6 (1.0)
CER, AD	173	88 (51)	73.5 (5.7)	108 (63)	7.1 (1.0)
CER, Non-AD	170	60 (35)	71.6 (5.5)	45 (26)	7.2 (0.9)
CER, ALL*	343	148 (43)	73.0 (5.7)	153 (45)	7.2 (0.9)

TABLE 1: Demographics and characteristics of the samples with expression measures. *Indicates subjects groups analyzed using Fast Epistasis. TCX: Temporal cortex, CER: Cerebellum, AD: Subjects with Alzheimer's disease, non-AD: Subjects without Alzheimer's disease, Age: Age at death, ApoE4+: Number (percent) subjects with ApoE4 allele, RIN: RNA integrity number, SD (standard deviation).