Gene ontology analysis of gene-gene interactions in two genome-wide association studies of sporadic ALS

Nora Kim¹, Peter Andrews¹, Folkert W. Asselbergs³, H. Robert Frost¹, Scott M. Williams¹, Brent T. Harris⁴, Cynthia Read², Kathleen Askland², Jason H. Moore¹,²

¹Institute for Quantitative Biomedical Sciences, Department of Genetics, Dartmouth Medical School, One Medical Center Dr., Lebanon, NH 03756, United States of America
²Department of Psychiatry and Human Behavior, Butler Hospital, Brown University, 345 Blackstone Blvd, Providence, RI 02906, United States of America
³Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht, the Netherlands
⁴Department of Neurology, 4000 Reservoir Rd., Georgetown University Medical Center, Washington, DC 20057, United States of America

Corresponding author: jason.h.moore@dartmouth.edu

ABSTRACT

Background: It is increasingly clear that common human diseases have a complex genetic architecture characterized, in part, by epistasis or gene-gene interactions. We have previously identified statistically significant gene-gene interactions that replicated across two independent genome-wide association studies (GWAS) of sporadic Amyotrophic Lateral Sclerosis (ALS). The goal of the present study was to determine whether patterns of gene-gene interactions aggregate in specific functional groups as defined by the Gene Ontology (GO).

Results: We first estimated all pairwise gene-gene interactions using the multifactor dimensionality reduction (MDR) method and evaluated statistical significance using permutation testing in two genome-wide association studies of ALS. The detection data consisted of 276 subjects with ALS and 271 healthy controls while the replication data consisted of 221 subjects with ALS and 211 healthy controls. Both studies included genotypes from approximately 550,000 single-nucleotide polymorphisms (SNPs). Each SNP was mapped to a gene if it was within 500kb of the start or end. Each SNP was assigned a p-value based on its strongest interaction with the other SNPs. We then used the Exploratory Visual Analysis (EVA) method and software to assign a p-value to each gene based on the overabundance of significant SNPs at the α=0.05 level in the gene. We also used EVA to assign p-values to each GO group based on the overabundance of significant genes at the α=0.05 level. A GO category was determined to replicate if that category was significant at the α=0.05 level in both studies. We found two GO categories that replicated in both studies. The first, ‘Regulation of Cellular Component Organization and Biogenesis’, a GO Biological Process, had p-values of 0.010 and 0.014 in the detection and replication studies, respectively. The second, ‘Actin Cytoskeleton’, a GO
Cellular Component, had p-values of 0.040 and 0.046 in the detection and replication studies, respectively.

**Conclusions:** Pathway analysis of gene-gene interactions in two GWAS of sporadic ALS revealed a set of genes involved in cellular component organization and actin cytoskeleton, more specifically, that were not reported by prior GWAS. However, prior biological studies have implicated actin cytoskeleton in ALS and other motor neuron diseases. This study supports the idea that pathway-level analysis of GWAS data may discover important associations not revealed using conventional one-SNP-at-a-time approaches.

**FINDINGS**

Amyotrophic Lateral Sclerosis (ALS) is a neurological disease that causes motor neuron degeneration, leading to paralysis and eventually death. Around 5,600 people are diagnosed with ALS each year with the incidence rate of two per 100,000 a year [1]. Despite the relatively low incidence rate compared to more prevalent diseases such as Alzheimer’s and Parkinson’s, ALS is a devastating disease with the average life expectancy of only two to five years from the time of diagnosis. Unfortunately, genome-wide association studies (GWAS) across multiple cohorts have not revealed replicable, genome-wide significant single-nucleotide polymorphisms (SNP) associations that could provide additional clues about the etiology of ALS [2-7]. It is important to note that the prior studies assumed a simple genetic architecture with an analytical approach that was designed to only detect single SNPs with large effects independent of the genomic background or the ecological context of the subjects being studies. It is our working hypothesis that the genetic architecture of sporadic ALS is complex and likely to be influenced by gene-gene interactions [8]. We further hypothesize that patterns of gene-gene interactions influence ALS susceptibility at the pathway level with individual gene effects playing a smaller role. The goal of the present study was to determine whether patterns of gene-gene interactions aggregate in specific functional groups as defined by the Gene Ontology (GO). We briefly outline here our bioinformatics approach and then summarize the findings.

Figure 1 provides a flowchart for our bioinformatics analysis. The analyses are split into three phases. The first phase is a SNP-level analysis that consists of data processing and gene-gene interaction analysis. The second phase is a gene-level analysis that determines whether each gene has more statistically significant SNPs than expected by chance. The third phase is a pathway-level analysis that determines whether each GO category has more statistically significant genes than expected by chance. This is followed by an assessment of replication between the independent GWAS studies of sporadic ALS. We briefly summarize the approach for each of these phases.

**Phase I: SNP-level analysis.** We ascertained and processed each GWAS dataset for sporadic ALS as described previously by Greene et al. [8]. The detection data consisted of 276 subjects with ALS and 271 healthy controls [3] while the replication data consisted of 221 subjects with ALS and 211 healthy controls [4]. Both studies included genotypes from approximately 550,000 single-nucleotide polymorphisms (SNPs). We
then carried out a Multifactor Dimensionality Reduction (MDR) analysis to estimate the pairwise effect of SNPs on ALS susceptibility in each study. MDR is a machine learning method that was designed specifically to detect and characterize gene-gene interactions using constructive induction [9-12]. The MDR approach combines multiple SNPs at a time and then estimates the accuracy of a naïve Bayes classifier for predicting case-control status. We repeated the entire MDR analysis on 1000 datasets for each study that randomized case-control status. This permutation-based approach to determining significance corrects for false-positives due to multiple testing. Permutation testing revealed a critical value of the MDR classification accuracy of 0.629 for the detection data and 0.640 for the replication data at a genome-wide $\alpha=0.05$ significance level. For each SNP, we picked the strongest interaction and assigned the p-value for that interaction to that SNP. We then assigned SNPs to genes if they were within 500kb upstream of a gene start or 500kb downstream of a gene end. This window was picked to ensure inclusion of regulatory SNPs.

**Phase II: Gene-level analysis.** The gene-level enrichment analysis was performed using the Exploratory Visual Analysis (EVA) methodology and software. EVA was designed specifically for the visualization and analysis of statistical results from gene expression studies [13]. We have more recently expanded EVA for the analysis of SNP data [14,15]. The goal of phase II was to determine whether a gene has an overabundance of statistically significant SNPs. Here, we used EVA to assess whether SNPs with MDR-based p-values of 0.05 or less are overrepresented given the size of the gene. Statistical significance is determined in EVA using a Fisher’s exact test based on the hypergeometric distribution. P-values are then assigned to genes and genes assigned to GO categories. For the GO assignment, we used the annotations from the Molecular Signatures Database or MSigDB, version 3.0 [16].

**Phase III: Pathway-level analysis.** The pathway-level GO enrichment analysis was also performed using the EVA methodology and software as described above. Here, we used EVA to assess whether genes with p-values of 0.05 or less are overrepresented in a given pathway accounting for its size. A GO category was considered statistically significant at an $\alpha=0.05$ level. Gene Ontology categories that had p-values less than or equal to 0.05 in both the detection and replication datasets were considered replicated and carried forward for presentation and discussion. We explicitly used replication to control for false-positives due to multiple testing across pathways.

**Results.** The bioinformatics analysis strategy described above and in Figure 1 revealed two replicated and statistically significant GO categories. The first, ‘Regulation of Cellular Component Organization and Biogenesis’, a GO Biological Process, had p-values of 0.010 and 0.014 in the detection and replication studies, respectively. The second, ‘Actin Cytoskeleton’, a GO Cellular Component, had p-values of 0.040 and 0.046 in the detection and replication studies, respectively. We focus here on interpretation of the role of actin cytoskeleton in sporadic ALS because it is a subcomponent of the set of cellular component organization and biogenesis genes.
**Discussion.** Actin is a highly conserved protein that forms microtubules in cells. Actin is an important part of the cytoskeleton and participates in or makes possible a number of cellular functions including protein trafficking and cell motility. The biological role of actin cytoskeleton in motor neuron diseases has long been recognized [17]. A recent review by Julien et al. [18] provides an overview of the many studies that have implicated cytoskeletal defects in ALS. Interestingly, SNPs in actin cytoskeleton genes have not replicated across the published GWAS for sporadic ALS [2-7]. They have, however, been implicated in other diseases such as multiple sclerosis [19]. Given the known biological basis for actin cytoskeleton in motor neuron diseases, including ALS, and the results presented here, we conclude that genes in this GO category should be examined more carefully in GWAS for sporadic ALS. In particular, such a careful examination should entail an analytical approach that is robust to the detection of pathway replication in the absence of gene replication. The results of this study support the idea that a bioinformatics approach to GWAS analysis that embraces the complexity of the genotype-phenotype map has the potential to reveal interesting new associations that have not been discovered using one-SNP-at-a-time approaches.

**AVAILABILITY**

The EVA and MDR software packages are freely available from the authors. More information can be found at http://www.epistasis.org.

**ACKNOWLEDGEMENTS**

This work was supported by NIH R01 grants LM010098, LM009012 and AI59694. NK was supported by a William H. Neukom Institute Fellowship at Dartmouth College and Howard Hughes Medical Institute Fellowship. FA was supported by a clinical fellowship from the Netherlands Organisation for Health Research and Development (ZonMw grant 90700342). The genotyping of samples was provided by the NINDS. The datasets used for the analyses described in this manuscript were obtained from the NINDS Database through dbGaP accession numbers phs000006.v1.p1 and phl000649.v1.p1. Permission to use these datasets was obtained through application to dbGAP by B.T.H.

**COMPETING INTERESTS**

The authors declare they have no competing interests.

**REFERENCES**

1. **ALS Association** [http://www.alsa.org/about-als/facts-you-should-know.html]

FIGURE LEGENDS

Figure 1. Flowchart summarizing the three phases of our Gene Ontology analysis strategy.