Copy Number Variation Identification and Association Study on 3,800 Alzheimer's Disease Whole Genome Sequencing Data from the Alzheimer's Disease Sequencing Project (ADSP)

Albert Tucci [1] Jung-Ying Tzeng [1] Mitchell Conery [2] Yuk Yee Leung [2] Amanda Kuzma [2] Otto Vallardes [2]
Yi-Fan Chou [2] Wenbin Lu [1] Li-San Wang [2] Gerard D. Schellenberg [2] Wan-Ping Lee [2]

[1]Bioinformatics Research Center, North Carolina State University [2]Penn Neurodegeneration Genomics Center, University of Pennsylvania

Background

Earlier investigations into the association of CNVs and AD were performed using genotyping arrays. Though efficient, this method is limited due to the fact that it relies on specific predetermined probes for defining CNVs. A CNV needs to span multiple probes in order to be successfully detected, which can result in detection bias. To obtain more unbiased and comprehensive CNV profiles, the National Institute of Aging (NIA) has launched the Alzheimer's Disease Sequencing Project (ADSP), which is funding whole genome sequencing (WGS) of AD cases and cognitively normal controls.

This study made of the ADSP Umbrella R1 dataset (ng00067) released through the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS). After parsing with relatedness filters, the dataset contained 1,737 AD cases and 2,063 cognitively normal controls.

Objectives

- **Generate CNV analysis workflow** for calling CNVs from WGS data and performing downstream association analyses.
- **Characterize CNV callset of ADSP data** by calculating summary statistics for both the entire set of called CNVs and a subset of rare CNVs. All CNVs were characterized
- **Preliminary AD association analysis** with CNV regions which were defined using the density-trimming method with CNVRanger on the entire CNV callset. Association analysis was done via a logistic regression controlling for a list of covariates. These covariates included: sex, age, population substructure (determined through principal components analysis), and possession of an apolipoprotein E(APOE)-4 allele.

Analysis Workflow

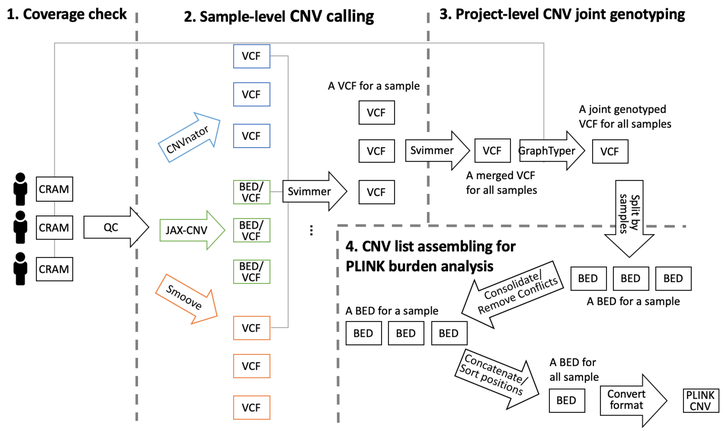

Figure 1. Diagram of novel bioinformatic pipeline used to prepare data and call CNVs.

Summary of CNV Callset

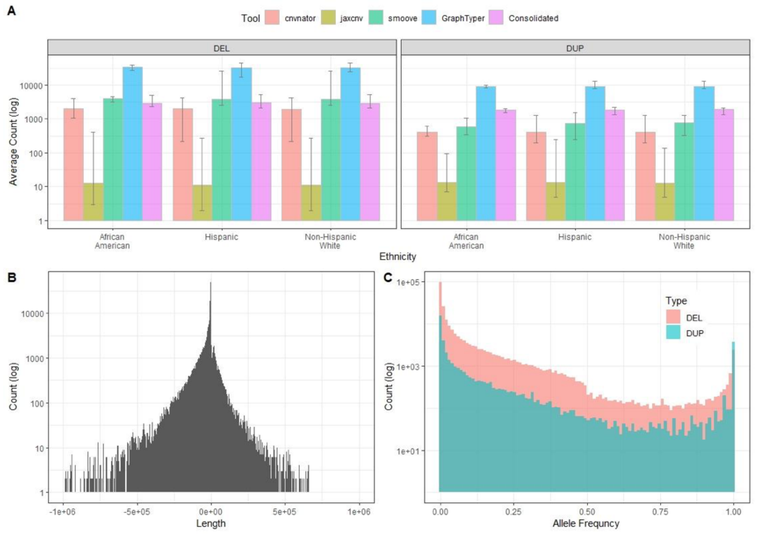

Figure 2. **A)** Average log count of CNV calls emerging from each step of the pipeline. CNVnator, JAX-CNV, and Smoove VCF results were merged ahead of GraphTyper step (Fig. 1). Consolidation was done by filtering based on relatedness and removing conflicts.**B)** CNV length histogram with deletions displayed as negative, and duplications as positive. **C)** Log count histogram of allele frequency.

Rare CNVs Across Autosomes

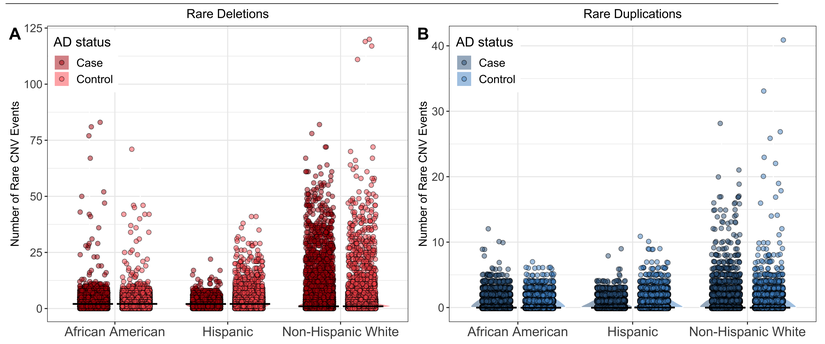

Figure 3. Violin plot overlaid with jitter-plot of rare deletions (**A**) and rare duplications (**B**). Broken down between AD cases vs controls by self-reported race and ethnicity. Black bars indicate median value.

Preliminary AD Association Analysis

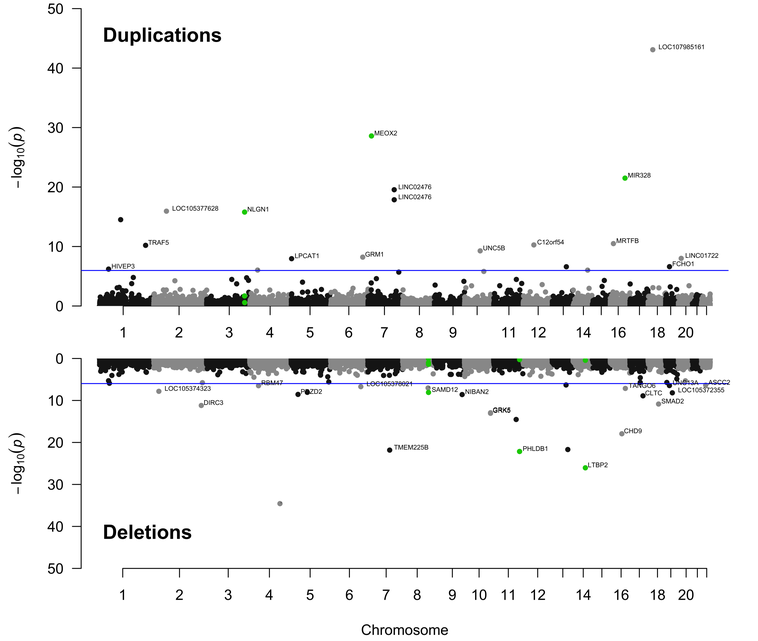

Figure 4. Chicago plot of results from association analysis of copy number variation and AD status. Each point represents a CNV region defined by CNVRanger. CNV regions that were identified as significant and overlapped with genes that have previously been linked to AD are highlighted in green.

Conclusions

- **We called 237,306 deletions and 42,767 duplications** using a novel bioinformatics pipeline for identifying CNVs from WGS data. Though the false discovery rate is still uncertain despite efforts to cross-validate with other CNV calling projects.
- **Preliminary AD association analysis identified 6 CNV regions** that overlap genes which have previously been linked to AD status in the literature.

References