Impact of Different Copy Number Variation (CNV) Loci Definitions on Copy Number Variation Association Analysis with Whole Genome Sequencing (WGS) Data — Lessons learned from the Alzheimer's Disease Sequencing Project (ADSP)

Albert Tucci [1] Hui Wang [2] Yaqin Si [1] Yuhuan Cheng [1] Li-San Wang [2] Gerard Schellenberger [2] Jung-Ying Tzeng [1] Wan-Ping Lee [2]

[1]North Carolina State University [2]Penn Neurodegeneration Genomics Center

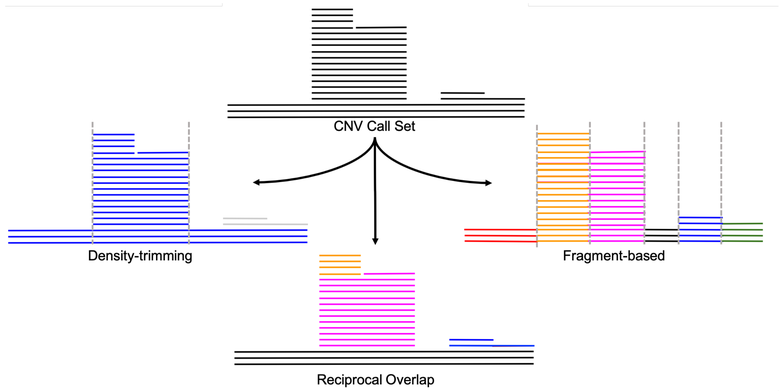
Abstract

Background: Currently, the heritability of Alzheimer's disease (AD) attributed to single nucleotide variants (SNVs) accounts for less than half of the total estimated heritability. CNVs are another form of genetic variation that could explain some of the remainder of AD heritability estimates. Selecting a method for defining CNV loci for association analyses is non-trivial because it affects downstream analysis and result interpretation. We show how different methods for defining CNV loci can influence association analysis for AD.

Methods: CNV callsets were generated using WGS data from the ADSP R3 dataset. Three approaches (as defined by CNVRuler) were used to define CNV loci: density trimming (DT), reciprocal overlap (RO), and fragment-based (FR). AD association via logistic regression was performed using each set of defined loci. The number of CNV loci defined and the regions identified as significant were used to compare definitions and inform guidelines for future analyses.

Results and Discussion: The RO approach yielded informative CNV loci, but was substantially more computationally expensive and less tractable for large-scale analyses. DT and FR approaches identified genomic regions of interest as significant, though each had trade-offs. DT is highly sensitive to setting a proper density threshold value, and for large-sample studies is susceptible to masking regions of the genome to downstream analysis. FR approaches can result in loss of power due to the large number of CNV loci defined. We recommend FR with size filtering (omitting <50bp) and DT with a much lower trimming threshold than the commonly used default values.

CNV Loci Definitions



Copy Number Variant Loci Using Manta Smoove

CH10 CNV Loci	Density-trimming	Reciprocal Overlap	Fragment-based
Total CNV Loci	133	5421	21590
Min. len (bp)	52	51	50
Avg. len (kb)	2.7	4607.4	6.1
Compute time (h:m)	00:02	18641:18	01:08

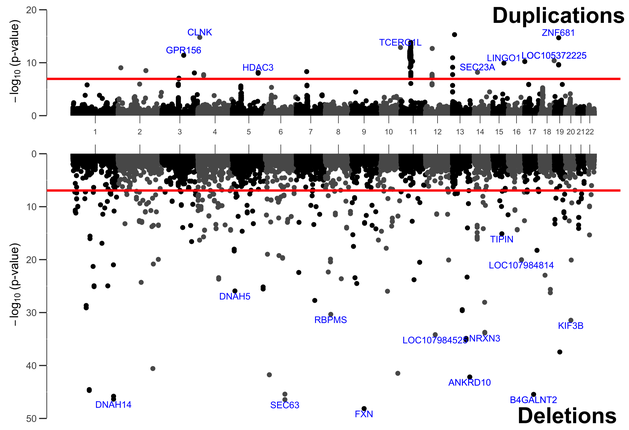
Chromosome 10 CNV loci defined using: DT, RO, and FR on a Manta-Smoove CNV call set.

Impact of CNV Caller

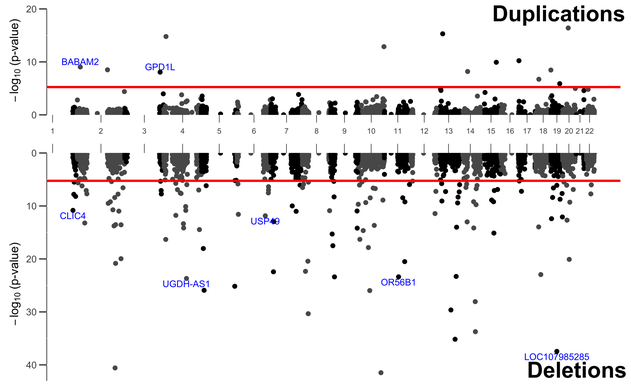
Manta Smoove		Graphtyper2	
Density-trimming	Fragment-based	Density-trimming	Fragment-based
8,843 total loci	438,502 total loci	86,641 total loci	29,615 total loci
Avg width = 187kb	Avg width = 6kb	Avg width = 4kb	Avg width = 2kb
14 significant dup	2813 significant dup	49 significant dup	132 significant dup
137 significant del	378 significant del	548 significant del	763 significant del

Graphtyper2 joint-genotyping method mitigates DT masking and FR overestimation.

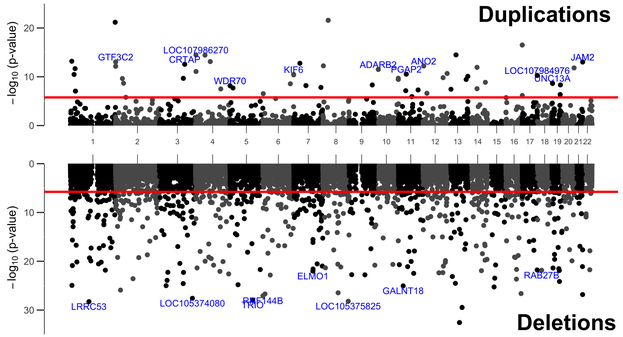
Fragment-based Loci Association (Manta-Smoove Call Set)



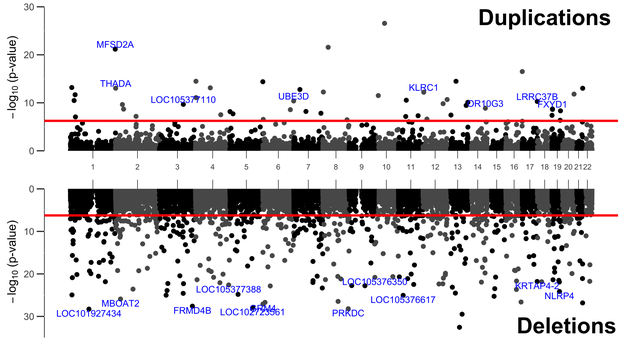
Density-trimming Loci Association (Manta-Smoove)



Fragment-based Loci Association (Graphtyper2 Call Set)



Density-trimming Loci Association (Graphtyper2)



Methods Comparison

- **Density trimming** is computationally efficient, but is highly sensitive to the nature of the call set data, and can result in loss of signal when long CNV calls overlap many smaller regions where shorter CNV calls are present.
- **Fragment-based** approaches do not exhibit computational nor masking draw backs observed with the other two methods, and are thus more robust for large-scale genome-wide CNV analyses. However, in cases where breakpoints in the CNV call set are not resolved, FR likely overestimates CNV loci.

References

atucci@ncsu.edu