cnvGSA Package Introduction

Joseph Lugo
The Centre for Applied Genomics
joseph.r.lugo@gmail.com

May 29, 2026

`cnvGSA` is an `R` package for testing the gene-set rare variant burden in case-control studies of copy number variation (CNV).

Only rare CNV (e.g. at frequency <1%) should be present in the input data. Gene-sets need to be pre-compiled based on user-curated data or publicly available gene annotations like **Gene Ontology** or pathways.

"Competitive" gene-set over-representation tests are commonly used to analyze differentially gene expressed genes (e.g. Fisher's Exact Test, GSEA), but they are not suitable for rare CNV; the most appropriate choice for rare CNV is a "self-contained" burden test with global burden correction implemented by `cnvGSA` or other tools.

Global burden correction is very important. For many disorders (including autism and schizophrenia), the disease-affected subjects (i.e. cases) are enriched in large, recurrent CNV. Those CNV are not observed (or observed at very low frequency) in controls and only a minority of their genes may contribute to disease risk. In the absence of global burden correction, many gene-sets would present a biologically unspecific burden, uniquely driven by those larger and recurrent CNV. Global burden correction thus helps identifying specific pathways and functional categories implicated in disease risk by rare CNVs.

In `cnvGSA`, subjects are treated as statistical sampling units. Subject-level covariates that may act as confounders can be provided by the user (e.g. sex, ethnicity, CNV genotyping platform, CNV genotyping site, array quality metrics, etc.). The gene-set burden is tested using a logistic regression approach. Two logistic regression models are fit: model A includes the subject-level covariates and a variable quantifying global CNV burden for each subject (total CNV length, or total number of CNV-overlapped genes per subject, etc.); model B includes all variables present in model A, plus the number of CNV-overlapped genes that are members of the gene-set being tested. Presence of significantly higher burden in cases compared to controls for the gene-set of interest is then tested by comparing the two models using a deviance chi-square test, as implemented by `anova.glm`.