An integrative system biology approach to delineate complex genotype - phenotype associations in Autism Spectrum Disorder

Asif M[a, b], Hugo F. Martiniano[b, c], Celia Rasga[a, b], Ana R. Marques[a, b], João X. Santos[a, b], Francisco M. Couto[c] and Astrid M. Vicente[a, b, d]

[a]Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; [b]BioISI - Biosystems & Integrative Sciences Institute, Lisboa; [c]Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal; [d]Instituto Gulbenkian de Ciência, Oeiras. masif@fc.ul.pt

Autism Spectrum Disorder (ASD) is characterized by a wide spectrum of behavioral presentation. While many genetic factors are implicated in ASD, the architecture of genotype/phenotype correlations is very unclear. Using data mining-based integrative approaches, we seek to identify patterns of association between ASD phenotypic subgroups and altered biological processes inferred from Copy Number Variants (CNVs) targeting brain genes. Analysis of clinical data from 2529 ASD patients collected by the Autism Genome Project, using Agglomerative Hierarchical Clustering, identified two distinct phenotypic clusters that differed in overall adaptive behavior profiles, verbal status and cognitive abilities, defining more severe and less severe phenotypes. Clusters were highly stable for 1000-bootstrap iterations. In the same ASD subjects, enrichment analysis of rare CNVs targeting brain genes, followed by removal of redundant biological processes using Gene Ontology hierarchy, identified 18 statistically significant biological processes generally consistent with reported literature for ASD, including nervous system development. We further used the Support Vector Machine (SVM) machine learning method to predict phenotypic clustering of patients from biological process disrupted by CNVs in brain genes, using 10-fold stratified cross-validation to train and test the SVM method on this patient dataset. SVM achieved an accuracy of 66.3% to differentiate between less and more severely affected individuals, thus allowing a reasonable prediction of clinical outcome from biological processes defined by genetic alterations. We are now performing validation of the classifier in an independent dataset. This approach seeks to support the clinical prediction of disease progression from genetic information in very young children, allowing earlier and more personalized intervention, and contributing to understanding the genetic basis of ASD clinical heterogeneity.