TRANSLATING THE COMPLEX ASD GENETIC ARCHITECTURE INTO CLINICAL PHENOTYPE USING AN INTEGRATIVE SYSTEM BIOLOGY APPROACH

Asif M[a, b], Hugo F. Martiniano[b, c], Celia Rasga[a, b], Ana R. Marques[a, b], João X. Santos[a, b], Guiomar Oliveira, 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

Background: Autism Spectrum Disorder (ASD) is characterized by a wide spectrum of behavioral presentation, rendering ASD difficult to diagnose particularly in very young children. While many genetic factors are implicated in ASD, the architecture of genotype/phenotype correlations is still very unclear. Delayed diagnose leads to delay in applying behavioral therapies that may help to reduce symptoms, particularly when applied at young age.

Objective: The aim of the study was to develop a novel machine learning-based integrative system biology approach to predict the clinical outcome from biological processes defined by rare Copy Number Variants (CNVs) in ASD children.

Methods: Agglomerative Hierarchical Clustering (AHC) was used to identify ASD phenotypic subgroups from the clinical reports from 2529 ASD patients recruited by the Autism Genome Project. Altered biological processes in the same ASD patients were inferred from rare CNVs targeting brain genes, by employing functional annotation analysis. To predict phenotypic clustering of patients from biological process disrupted by rare CNVs in brain genes, four different machine learning methods were trained and tested on the clustered patient and disrupted biological processes datasets, and performance of implemented methods were compared using “accuracy” measure.

Results: Analysis of clinical data using AHC 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 and clusters validation through the Silhouette method also indicated that both clusters were true and consistent. Cluster 1 represented the subgroup with the most severe clinical presentation, with all patients being non-verbal and presenting dysfunctional adaptive behavior profiles for all VABS subscales, as well as lower performance IQ. Cluster 2 individuals presented less severe ASD symptoms and milder deficits for all clinical variables. 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. Support Vector Machine (SVM) outperformed the

other three methods by achieving highest 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.

Conclusion: To address the ASD heterogeneous phenotype which has hindered the identification of genotype/phenotype associations, ASD patients were clustered into two clusters with more and less severe phenotype that is consistent with a previously reported analysis. Functional annotation analysis showed that rare CNVs targeting brain genes from ASD subjects tend to aggregate in common biological processes that have been previously associated to ASD, such as nervous system development and protein polyubiquitination. The presented approach seeks to enhance our knowledge on ASD diagnosis and prognosis by elucidating the complex genotype/phenotype associations in patients, allowing earlier and more personalized intervention, and contributing to understanding the genetic basis of ASD clinical heterogeneity.