Use of machine learning approaches to explore genetic and phenotypic associations for autism

Asif M [1,2] , Conceição, IC[1,2,3]; Café, C[4]; Almeida, J[4]; Mouga, S[4]; Oliveira, G[4, 5, 6, 7] , Couto, F[8] Vicente, AM[1,2,3]

[1]Instituto Nacional de Saúde Doutor Ricardo Jorge, Lisboa; [2]Center for Biodiversity, Functional & Integrative Genomics (BioFIG), Lisboa; [3]Instituto Gulbenkian de Ciência, Oeiras; [4]Unidade Neurodesenvolvimento e Autismo, Centro de Desenvolvimento, Hospital Pediátrico (HP), Centro Hospitalar e Universitário de Coimbra (CHUC), Coimbra; [5]Instituto Biomédico de Investigação em Luz e Imagem, Faculdade de Medicina da Universidade de Coimbra, Coimbra; [6]Faculdade de Medicina da Universidade de Coimbra, Coimbra; [7]Centro de Investigação e Formação Clinica do HP-CHUC, Coimbra, [8] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal

Autism spectrum disorder (ASD) is neurodevelopmental disorder of well known complexity. ASD is characterized by impaired social interaction, communication and stereotyped behavior, and a high heterogeneity in clinical and genetic presentation. It is hypothesized that such complex heterogeneous phenotypic behaviors are associated with genetic factors.

To further dissect the complex correlations between phenotype and genotype in ASD, in the current study we integrate clinical information (from diagnostic instruments ADI-R and ADOS as well as adaptive behavior and cognitive scales VABS and WISC) and genetic data (Copy Number Variants, CNVs) of 335 Portuguese individuals, using powerful machine learning algorithms like decision trees.

We confirmed direct associations of clinical observations with CNVs by designing a 10 fold cross validation model. The relationship among physical and genetic data was tested using interpretable decision tree algorithms (J48, REP and Random tree) with a high accuracy, 67.66%. This was strongly improved relative to results from the same algorithms using randomized data (53%).

We also identified a phenotypic behavioral signature of socialization dysfunctionality, social interaction problems and dysfunctionality in communication, coupled with slight a level intellectual disability which is directly linked with inherited CNVs. Computational model predicted this signature for 123 male individuals. Moreover, we found that CNV deletions were more frequent in females with more severe intellectual disability, as compared to males. This result confirms previous separate observations of certain deletions being more pathogenic than duplications of the same genomic regions, and of a more severe clinical presentation, with lower cognitive levels, in females with ASD.

We believe that enhanced understanding of genetic and behavioral data association will be useful to assist in ASD diagnosis. However, for that we need more data sets to increase the power to detect true effects and to improve accuracy and its soundness (statistical significance). Therefore we expect to expand our analysis to a larger dataset comprising 980 individuals with ASD.