

1 **An integrative systems medicine approach to delineate complex genotype-**
2 **phenotype associations in Autism Spectrum Disorder**

3 Muhammad Asif^{1,2,3}, Hugo F.M.C. Martiniano^{1,2}, Ana Rita Marques^{1,2}, João Xavier Santos^{1,2},
4 Joana Vilela^{1,2}, Celia Rasga^{1,2}, Guiomar Oliveira^{4,5,6}, Francisco M. Couto³, Astrid M. Vicente^{1,2*}

5 * astrid.vicente@insa.min-saude.pt

6 ¹Instituto Nacional de Saúde Doutor Ricardo Jorge, Avenida Padre Cruz, 1649-016 Lisboa,
7 Portugal

8 ²BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de
9 Lisboa, Lisboa, Portugal

10

11 ³LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

12 ⁴Unidade de Neurodesenvolvimento e Autismo (UNDA), Serviço do Centro de Desenvolvimento
13 da Criança, Centro de Investigação e Formação Clínica, Hospital Pediátrico, Centro Hospitalar e
14 Universitário de Coimbra, Coimbra, Portugal

15 ⁵Institute for Biomedical Imaging and Life Sciences, Faculty of Medicine, Universidade de
16 Coimbra, Coimbra, Portugal

17 ⁶University Clinic of Pediatrics, Faculty of Medicine, University of Coimbra, Portugal

18

19

20

21

22

23

24

Abstract

25 **Background:** The heterogeneous phenotype and complex genetic architecture of Autism
26 Spectrum Disorder (ASD) has thus far limited our understanding of genotype-phenotype
27 correlations, hindering early diagnosis and patient prognosis. Copy Number Variants (CNVs)
28 targeting a diversity of genes have been implicated in ASD, however correlations with clinical
29 patterns are unclear.

30 **Methods:** In this study, we developed a novel machine learning integrative approach that seeks
31 to delineate associations between ASD clinical profiles and disrupted biological processes
32 inferred from CNVs spanning brain-expressed genes.

33 **Results:** Clustering analysis of relevant clinical measures from 2446 ASD cases, retrieved from
34 the Autism Genome Project (AGP) database, identified two distinct phenotypic subgroups, with
35 a milder and a more severe phenotype. Patients in the two clusters differed significantly in verbal
36 status, ADOS-defined severity, adaptive behaviour profiles and intellectual ability, with verbal
37 status contributing the most for cluster stability and cohesion. In the clustered ASD cases,
38 functional enrichment analysis of brain-expressed genes disrupted by rare CNVs identified 15
39 statistically significant biological processes. These biological processes included cell adhesion,
40 nervous system development, cognition and protein polyubiquitination and were in line with
41 previous ASD findings. Random Forest feature importance analysis showed a positive
42 contribution of all disrupted biological processes to the classification of ASD cases in the
43 identified clusters. A Naive Bayes classifier was generated to predict the ASD phenotype from
44 the identified disrupted biological processes. For a subset of patients with higher Information
45 Content scores calculated for the disrupted biological processes, the classifier achieved
46 predictions with a high precision but low recall (Precision: 0.82, Recall: 0.39).

47 **Conclusions:** This study highlights the usefulness of machine learning approaches to reduce
48 clinical heterogeneity by taking advantage of multidimensional clinical measures. Furthermore, it
49 shows that phenotype-genotype correlations can be established in ASD, and that milder and more
50 severe clinical presentations have distinct underlying biological mechanisms. However, precise
51 predictions of the phenotype from genetic data were only achieved for the subset of patients with
52 higher biological information content. These findings are therefore a first step towards the

53 translation of genetic information into clinically useful applications, while emphasizing the need
54 for larger datasets with complete clinical and biological information.

55 **Keywords**

56 Autism Spectrum Disorder (ASD), machine learning, integrative systems medicine,
57 genotype/phenotype associations, ASD heterogeneity, integrating data, CNVs

58 **Background**

59 Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that manifests with
60 persistent deficits in social communication and interaction, and unusual or repetitive behaviour
61 and/or restricted interests [1]. ASD presents a highly heterogeneous clinical phenotype and
62 frequently co-occurs with other comorbidities, such as Intellectual Disability (ID), epilepsy and
63 Attention Deficit Hyperactivity Disorder (ADHD) [2–6]. Heritability estimates indicate a strong
64 genetic influence in ASD aetiology [7–9], however reliable genetic markers for the disease are
65 unavailable. ASD is diagnosed through neurodevelopmental assessment, which can be
66 challenging especially in the case of very young children. Improving early diagnosis and
67 prognosis using biological markers with a robust predictive power would provide an advantage
68 to young patients, who benefit the most from an early start of specific intervention [10].

69 Copy Number Variant (CNV) screening is nowadays widely used for etiological diagnosis, with
70 causative genetic alterations identified in approximately 25% of ASD cases [11]. A large number
71 of rare genetic variants have been implicated in ASD, and the wide genetic heterogeneity that
72 characterizes this disorder likely contributes to phenotypic variability in ASD patients [12].
73 Integrative pathway and network analysis of large scale ASD genomic studies have advanced
74 significantly the identification of disrupted biological processes [13–17]; however, our
75 understanding of the biological meaning of the large number of putative pathogenic variants,
76 their phenotypic manifestations, and the reliable interpretation of many genetic findings for
77 clinical application is still lagging.

78 To improve our ability to infer clinical meaning from rare CNVs in ASD, for eventual
79 application as biological markers, we developed a machine learning-based approach involving
80 the integration of gene functional annotations and clinical phenotypes. Our approach was

81 developed in four steps, namely: 1) definition of clinically distinct subgroups in ASD cases; 2)
82 discovery of functionally enriched biological processes defined by rare CNVs disrupting brain-
83 expressed genes in the same ASD cases; 3) assessment of the contribution of disrupted biological
84 processes for classification of ASD phenotypes; 4) design and predictive effectiveness
85 characterization of a machine learning classifier for clinical outcome in ASD patients.

86

87

Methods

88 Figure 1 shows the graphical representation of the overall methodology, described in detail
89 below.

90

91 Figure 1: Integrative systems medicine approach to identify complex genotype-phenotype
92 associations. Clinical and genetic data from the Autism Genome Project (AGP) was used in this
93 study **(A)** Clinical data analysis processing: clinical data comprises reports of ASD diagnosis and
94 neurodevelopmental assessment instruments. Agglomerative Hierarchical Clustering (AHC) was
95 used to identify clinically similar subgroups of individuals in stable, validated clusters, defined
96 by multiple clinical measures. **(B)** CNV data processing: rare high confidence CNVs previously
97 identified by the AGP, targeting brain-expressed genes, were retained for analysis. CNV data
98 was merged with clinical data from clustered ASD subjects for a final list of CNVs targeting
99 brain genes. **(C)** Functional annotation analysis: Biological processes defined by brain-expressed
100 genes targeted by CNVs were obtained using g:Profiler. **(D)** Classifier design: A Naive Bayes
101 machine learning classifier was trained and tested on patient's data, to predict the phenotypic clustering of
102 patients from biological processed disrupted by rare CNVs targeting brain-expressed genes.

103 • Participants

104 The ASD dataset used in this study was obtained from the Autism Genome Project (AGP) [18]
105 database, and comprises CNV data and clinical information from 2446 ASD patients. The AGP
106 was an international collaborative effort from over 50 different institutions to identify risk genes
107 for ASD. The group of individuals with phenotypic information from clustering and rare CNV
108 data, used in final analysis included 1213 males (83.4%) and 144 females (10.6%).

109 • **ASD diagnosis, clinical assessment instruments and clinical features**

110 Individuals meeting criteria defined by the Diagnostic and Statistical Manual of Mental
111 Disorders IV (DSM-IV) [19] and the thresholds for Autism or ASD from the Autism Diagnostic
112 Interview-Revised (ADI-R) [20] and the Autism Diagnostic Observation Schedule (ADOS) were
113 classified as ASD cases [21]. The AGP defined a phenotypic classification system based on the
114 combined ADI-R and ADOS diagnosis, categorizing subjects into Strict (meeting thresholds for
115 Autism by the ADI-R and ADOS), Broad (meeting thresholds for Autism from one instrument
116 and ASD from the other) and Spectrum (meeting thresholds for Autism from at least one
117 instrument or ASD from both). Individuals with an ASD diagnosis from only one instrument and
118 no information from the other, or not meeting thresholds for Autism or ASD from one of the
119 instruments, regardless from the classification from the other, were not included in the study.
120 Clinical measures used in this study were retrieved from the AGP database, including the ADIR
121 verbal status, ADOS severity score, Vineland Adaptive Behaviour Scales (VABS) [22] subscales
122 and an Intelligence Quotient (IQ).

123 The ADI-R verbal status is a dichotomized measure indicating the verbal status of the individual
124 at evaluation. The ADOS severity metric ranges from 1 to 10 and is calculated from ADOS
125 modules 1 to 3 raw scores [23]. As there is no algorithm available to calculate ADOS severity
126 score for ADOS module 4 reports, which is applied only to adolescents and adults, subjects with
127 the ADOS module 4 (N= 149) were dropped from further processing. The severity score
128 distribution is collapsed into three categories, namely Autism (severity scores ranging from 6 to
129 10), ASD (severity scores ranging from 4 to 5) and Non-Spectrum (severity scores from 1 to 3),
130 which reflect the mapping of the severity metric onto raw ADOS scores. The ADOS Non-
131 spectrum category includes individuals with a mild phenotype, and in this study 125 individuals
132 with a Non-spectrum ADOS severity score fell within the Spectrum phenotypic class from the
133 AGP, meaning they met thresholds for autism from the ADI-R, and were thus included.

134 The VABS is used to assess adaptive functioning of individuals and consists of three subscales,
135 namely, socialization, communication and daily living skills scores, and also computes a
136 composite score. Subjects with VABS scores ≤ 70 were classified in a dysfunctional adaptive
137 behavior category, for all subscales. IQ scores of ASD cases were also retrieved from the AGP

138 database, and categorized with the following thresholds: $IQ > 70$ normal, $50 < IQ < 70$ mild
139 intellectual disability, $IQ < 50$ severe intellectual disability.

140 Clinical reports from the ASD patients were examined for missing values, and clinical features
141 with more than 70% information were retained for the analysis. To minimise missing value
142 imputation bias, individuals with missing values above this threshold for more than two clinical
143 features were also excluded. Completeness of each clinical feature is reported in Table S1
144 (Additional file 1). Missing values were imputed using the missForest [24] R package that
145 implements the Random Forest [25] algorithm, a decision tree-based supervised machine
146 learning method. Imputation error was assessed using the normalised global Proportion of
147 Falsely Classification (PFC), and the missing values imputation error was 0.12.

148 • **Clustering analysis of ASD clinical data**

149 To focus on core domains of ASD symptoms, verbal skills, disease severity, adaptive behavior
150 and intellectual levels, which strongly condition prognosis, were selected for further analysis.
151 Verbal status was obtained from the ADI-R, ASD severity scored from the ADOS, adaptive
152 functioning from the VABS, using its three subdomains, and a performance IQ category from the
153 IQ assessment contributed by participating sites to the AGP database. Other IQ domains had too
154 many missing values to be used. The Agglomerative Hierarchical Clustering (AHC) [26] method
155 was used to identify independent phenotypic subgroups from the selected clinical features.
156 Correlations between clinical features were assessed using the Pearson method, and features with
157 a correlation value of > 0.75 were considered correlated. The Gower [27] metric was used to
158 calculate the distance matrix from the patient's clinical data. To normalise the effect of highly
159 correlated variables on clustering, the weight for correlated variables (VABS subscales of
160 socialisation, communication, and daily living skills) was reduced to half during distance matrix
161 calculation. To identify phenotypic subgroups, the AHC method using Ward2 [28] criteria was
162 applied to the distance matrix.

163 To assess the contributions of each clinical feature in defining the clusters, we excluded one
164 feature at a time, re-performed the clustering and observed the changes in Silhouette values of
165 both clusters. For this purpose, we selected Silhouette value as an evaluation metric because it

166 was also used to define outliers in clinical data. A decrease in the Silhouette value of a cluster
167 after removing one feature indicates its importance in defining this cluster and vice versa.

168 • **Goodness of clustering assessment**

169 A Silhouette method [29] was employed to estimate the goodness of the clustering results. The
170 Silhouette value for each individual shows how well the individual is clustered, and ranges from
171 -1 to 1, with individuals scoring below 0 considered as wrongly clustered. In addition, the
172 Silhouette value for each cluster was derived, and clusters with Silhouette value of > 0.25 were
173 considered as true clusters. Bootstrapping with 1000 iterations was used to measure the stability
174 of clusters, where a boot mean value above 0.85 corresponds to stable clusters. All clustering
175 analysis was performed in R environment, using Cluster [30] and FPC packages.

176 • **Functional enrichment analysis**

177 Genotyping and CNV calling methods for the AGP ASD subjects (N=2446) were previously
178 described [18]. CNVs called by any two algorithms (high confidence CNVs) and above 30kb in
179 size were retained for further analysis. To screen for rare CNVs ($<1\%$ in control population)
180 CNV frequencies in control populations were estimated using the genotypes from the studies by
181 Sheikh et al. [31] (N = 1320) and Cooper et al. [32] (N = 8329), identified using the same
182 genotyping platform [18]. Control genotypes were obtained from the Database of Genomic
183 Variants (DGV) [33].

184 To focus CNV selection on variants spanning brain-expressed genes, avoiding *a priori*
185 hypotheses from ASD candidate gene assumptions, an extensive list comprising 15585 brain-
186 expressed genes was obtained from Parikshak et al. [34]. The brain-expressed gene list was
187 prepared from brain RNA-seq data, collected at thirteen different developmental stages,
188 including genes expressing during early brain developmental phase. The full criteria and
189 parameters used to define the brain-expressed gene list were previously described [34]. .

190 The g:Profiler [35] tool was employed to identify biological processes enriched for brain-
191 expressed genes spanned by rare CNVs in ASD individuals. g:Profiler implements a
192 hypergeometric test to estimate the statistical significance of enriched biological processes,
193 followed by multiple corrections for the tested hypotheses using the Benjamini-Hochberg

194 procedure. g:Profiler uses Gene Ontology (GO) data to find the biological annotations for input
195 genes.

196 The GO tool contains a Directed Acyclic Graph (DAG) structure with a clear hierarchical parent-
197 to-child relationship between GO terms. Because of this DAG structure, functional enrichment
198 analysis can result in redundant GO terms, which may lead to high correlations between GO
199 terms. To minimise the correlations between GO terms, the Revigo [36] tool was employed to
200 redundant GO results. Revigo uses the methods of semantic similarity to measure similarities
201 between GO terms. The SimRel [37] method was used to calculate similarities between GO
202 terms, and terms with a similarity score of > 0.7 were grouped.

203 • **Feature importance assessment**

204 The mean decrease in accuracy of the Random Forest algorithm was used to compute the
205 importance score of each disrupted biological process for categorizing ASD subjects into defined
206 phenotypic clusters. A stratified ten-fold cross-validation quan

207 tifies the importance of all features. The importance score of all disrupted biological processes
208 was recorded at each fold. A final importance score for each biological process was calculated by
209 averaging their importance score values across all the ten folds. Random Forest was
210 implemented using randomForst R package [38].

211 • **Classifier learning and cross-validation**

212 A Naive Bayes [39] machine learning method was employed to predict the ASD phenotypic
213 group, defined by the clustering analysis, from biological processes disrupted by rare CNVs.
214 This method employs the Bayes theorem of probability for training and testing of the model, and
215 the algorithm was implemented using the klaR R package with default parameters. Precision,
216 recall, specificity and F-score were used as evaluation measures. To train and test the Naive
217 Bayes, a stratified five-fold cross-validation approach was used, in which data was first split into
218 five equal subsets with equal class probabilities; a Naive Bayes model was trained on any four
219 subsets, and the remaining subset was used as the test set. This process was repeated five times
220 and each time a different subset was used as test set. For each repetition, the model performance
221 was estimated and mean values for precision, recall, specificity and F-score were reported. The

222 Naive Bayes classifier was trained on patient’s data by using the “more severe” cluster as the
223 positive class and the “less severe” cluster as the negative class.

224 The Information Content (IC) from each individual represents the level of specificity of
225 biological processes disruption, and was derived by summing the IC values of all the biological
226 processes disrupted in each individual. IC is a numerical value that describes the specificity of a
227 GO term using its position in the GO DAG structure.

228

229

Results

• Identification of ASD clusters defined by clinical phenotype

230
231 A total of 1817 ASD subjects from the AGP were retained for analysis after assessment of
232 missing values in clinical features. Agglomerative hierarchical clustering analysis of clinical
233 observations from these patients initially identified two phenotypically independent clusters. To
234 minimise the phenotypic complexity and define the most stable and cohesive clusters, weakly
235 clustered individuals with a Silhouette value less than 0.300 (representing a balance between
236 number of individuals lost and goodness of clustering) were excluded from the clustering
237 analysis. After removal of weakly or wrongly clustered individuals, cluster 1 contained 903 ASD
238 cases, while cluster 2 comprised 494 patients (Table 1). Elimination of the loosely clustered
239 individuals resulted in more stable and cohesive clusters, with high values for clusters stability
240 and reduced average distance between the two individuals in a cluster (Table 1).

241

Table 1: Clustering validation, after removal of weakly clustered individuals.

Clusters validation measures	Cluster 1	Cluster 2
Clusters size (N)	903	494
Average distance between two patients	0.235	0.231
Silhouette value	0.567	0.579
Average Silhouette of both clusters	0.571	
Cluster stability	0.998	0.996

242

243 Overall, the cluster validation through the Silhouette method and bootstrapping showed that both
244 clusters were true and consistent.

245 • **Clinical interpretations of the clusters**

246 All clinical measures differed significantly between the two clusters, as shown in Table 2.
247 Cluster 1 (Additional file 1: black circles in Figure S1) includes a higher number of individuals,
248 who generally exhibited a milder clinical phenotype, while Cluster 2 (Additional file 1: red
249 triangles in Figure S1) included a higher percentage of subjects with severe dysfunction. All
250 individuals in Cluster 1 were verbal according to the ADI-R, while Cluster 2 included only non-
251 verbal cases. The mean age of ADI-R assessment was 7.7 years, an age when verbal status is
252 generally well established. Furthermore, the mean age of individuals in Cluster 1 (mean age
253 8.02) and Cluster 2 (mean age 7.01) did not significantly differ.

254 For all VABS sub-domains, roughly half of the subjects in Cluster 1 were in the normal range;
255 conversely, over 97% of individuals belonging to Cluster 2 showed dysfunctional adaptive
256 behaviour. Consistent with the other clinical measures, over 96% of cases from Cluster 1, but
257 less than one third in Cluster 2, scored at the normal level in performance IQ, while a much
258 higher percentage of ASD cases from Cluster 2 than from Cluster 1 presented with a
259 performance IQ in the range of severe intellectual disability.

260 Regarding the ADOS severity score, approximately 14% of the individuals in Cluster 1 were
261 assigned to the milder category of the ADOS severity score (“Non-spectrum” for ADOS, but
262 scoring positive for “Autism” in the ADI-R, and therefore classified in the AGP “Spectrum”
263 phenotypic class, see methods). Conversely, none of the individuals in Cluster 2 scored in this
264 category. On the other hand, a significantly higher percentage of cases in Cluster 2 (20.65%)
265 than individuals in Cluster 1 (7.09%) scored in the intermediate ASD severity category. It is
266 noteworthy that both clusters show a similarly high percentage of individuals scoring in the
267 “Autism” ADOS severity category. This is not surprising since this broad category (scores
268 ranging from 6 to 10) comprises all subjects classified in the Strict AGP phenotype class but also
269 a large proportion of individuals in the AGP Broad phenotype class. The “Autism” ADOS
270 severity score therefore targets a subset of the study population that can be quite heterogeneous
271 in phenotypic presentation. Corroborating this, we found that the “Autism” category of the

272 ADOS severity score is not significantly associated with the clusters ($\chi^2 = 0.15$, $p = 0.901$, $df =$
 273 2), even though overall there is a significant association of the overall ADOS severity scores
 274 (Table 2).

275 Table 2: Clusters 1 and 2 statistics for each clinical measure.

Clinical measure	Clinically defined categories	Cluster 1 N (%)	Cluster 2 N (%)	<i>p</i> -value
ADIR verbal status	ADI-R-non verbal	0 (0)	494 (100)	<0.00001 ^a
	ADI-R-verbal	903 (100)	0 (0)	
ADOS severity score	ADOS severity score Autism (score 6-10)	714 (79.07)	392 (79.35)	<0.00001 ^b
	ADOS severity score ASD (score 4-5)	64 (7.09)	102 (20.65)	
	ADOS severity score Non-spectrum (score 1-3)	125 (13.84)	0 (0)	
VABS communication	Dysfunctional VABS communication (score ≤ 70)	307 (34)	493 (99.8)	<0.00001 ^a
	Normal VABS communication (score > 70)	596 (66)	1 (0.2)	
VABS daily living skills	Dysfunctional VABS daily living skills (score ≤ 70)	478 (52.94)	484 (97.98)	<0.00001 ^b
	Normal VABS daily living skills (score > 70)	425 (47.07)	10 (2.02)	
VABS socialization	Dysfunctional VABS socialization (score ≤ 70)	497 (55.04)	490 (99.19)	<0.00001 ^a
	Normal VABS socialization (score > 70)	406 (44.96)	4 (0.81)	
Performance IQ Scale	Severe disability (score <50)	2 (0.22)	218 (44.13)	<0.00001 ^b
	Moderate disability (score ≥ 50 and ≤ 70)	31 (3.43)	125 (25.3)	
	Normal ability (score > 70)	870 (96.35)	151 (30.57)	
Gender	Male	830 (91.92)	417 (84.41)	0.000015 ^b
	Female	73 (8.08)	77 (15.59)	

276 ^aFisher Exact Test, ^bChi-Square test

277 Both clusters were strongly dominated by the male gender, partly because of the high percentage
 278 of males in the dataset after the elimination of weakly or wrongly clustered individuals.
 279 However, the percentage of males was higher in cluster 1, representing the milder phenotype,
 280 consistent with general observations that male to female ratios are higher in datasets that
 281 comprise more high- function ASD individuals.

282 Analysis of the contribution of each clinical feature in defining clusters showed that the main
283 contributor was the ADIR verbal status variable (Additional file 1: Table S2). The VABS
284 subscales had a strong effect on Cluster 1 but a modest role in defining Cluster 2. Performance
285 IQ also contributed more to Cluster 1 whereas for Cluster 2 it has the least effect. The ADOS
286 severity score did not have a major role in defining either cluster, as indicated by the similar high
287 percentage of subjects scoring within the range of “Autism” in the ADOS severity scale in both
288 clusters. Similarly, gender was not an important contributor to the definition of either cluster.

289 • **Disrupted biological processes from brain-expressed genes targeted by rare CNVs**

290 CNVs (N=129754) identified in 2446 subjects with ASD were filtered to select rare, high
291 confidence CNVs, over 30 Kb in size and that contained complete or partial brain-expressed
292 gene sequences. The selected high confidence, rare CNVs (N=12683) disrupted 4025 brain-
293 expressed genes in 2414 subjects with ASD (86.8% males and 13.2% females).

294 Phenotypic cluster and rare CNV data was complete for 1357 individuals with ASD, and
295 available for integration. Functional enrichment analysis of rare CNVs targeting brain-expressed
296 genes (N=2738) in 1357 patients identified 17 statistically significant biological processes
297 (Additional file 1: Table S3). g:Profiler did not recognize 187 genes from the input list.

298 The redundancy of GO terms in functional enrichment analysis, caused by overlapping
299 annotations in ancestors and descendent terms in the DAG structure of GO, was reduced by
300 grouping the terms that had a semantic similarity score higher than 0.7 (Additional file 1: Table
301 S3). The Revigo tool used to reduce redundancy did not recognise one biological process
302 (*Plasma membrane bounded cell projection organization*). After redundancy reduction, 16
303 biological processes remained (Table 3), with the *Calcium-dependent cell-cell adhesion via*
304 *plasma membrane cell adhesion molecules* biological process merged with *Homophilic cell*
305 *adhesion via plasma membrane adhesion molecules* (similarity score = 0.76).

306 The most significant biological process identified in this dataset was *Homophilic cell adhesion*
307 *via plasma membrane adhesion molecules*, which includes 53 brain-expressed genes disrupted
308 by the selected CNVs. The ten most significant biological processes were related to cell adhesion
309 and cellular organization, and also included nervous system development and protein

310 poliubiquitination (Table 3). Moreover, two significant biological processes were related to
311 behavior and cognition.

312 Table 3: Statistically significant enriched biological processes for CNVs spanning brain-
313 expressed genes (N=2738). FDR: False Discovery Rate

Biological processes	Enriched genes (N)	FDR <i>p</i> -value
Homophilic cell adhesion via plasma membrane adhesion molecules	53	6.30E-09
Cell-cell adhesion via plasma-membrane adhesion molecules	66	1.70E-07
Cellular component organization or biogenesis	944	5.70E-05
Cellular component organization	915	7.00E-05
Cellular component biogenesis	475	0.00066
Cellular component assembly	434	0.00177
Nervous system development	363	0.00215
Organelle organization	562	0.00475
Protein polyubiquitination	64	0.00592
Cell projection organization	231	0.00836
Cellular localization	418	0.0091
Single-organism behavior	83	0.0196
Regulation of cellular component organization	364	0.0257
Plasma membrane bounded cell projection organization	223	0.0282
Cognition	56	0.0364
Single-organism organelle organization	263	0.044

314

315 • **Biological process importance for prediction of ASD clinical phenotype**

316 The enriched biological processes and phenotypic cluster information for ASD cases were
317 combined in a matrix to assess the predictive value of the biological processes for categorization
318 in one of the two phenotypic clusters, broadly characterized by a milder and a more severe
319 phenotypic presentation. The 57 individuals containing both rare CNV and cluster information

320 that did not present any enriched biological process were excluded, so further analysis comprised
321 1300 ASD patients.

322 Table 4 shows the ranking in importance of disrupted biological processes for categorization of
323 subjects into ASD phenotypic clusters, computed using the Random Forest importance score
324 function.

325 Table 4: Importance of each biological process from Random Forest in classifying ASD subjects
326 into defined phenotypic clusters.

Random Forest rank	Biological process	Mean Decrease in Accuracy
1	Regulation of cellular component organization	0.052
2	Cell projection organization	0.025
3	Cellular component assembly	0.025
4	Single organism behaviour	0.020
5	Organelle organization	0.018
6	Single organism organelle organization	0.017
7	Cellular component biogenesis	0.014
8	Cognition	0.013
9	Nervous system development	0.010
10	Cellular localization	0.009
11	Cellular component organization	0.006
12	Protein polyubiquitination	0.005
13	Homophilic cell adhesion via plasma membrane adhesion molecules	0.005
14	Cell adhesion via plasma membrane adhesion molecules	0.005
15	Cellular component organization or biogenesis	0.003

327

328 The importance of each biological process was calculated using the mean decrease in accuracy,
329 computed by permuting each biological process. The feature importance analysis using Random
330 Forest, which was trained and tested using stratified 10-fold cross-validation over the integrated
331 dataset, revealed positive values for all features, indicating that all of the biological processes are
332 positively contributing for classification. The most important biological process for the
333 classification was *Regulation of cellular component organization*, with a mean decrease in
334 accuracy of 0.052. The most significantly enriched biological process in the overall ASD dataset,
335 *Homophilic cell adhesion via plasma membrane adhesion molecules* was ranked at position 14,

336 indicating it is not a top contributor to phenotypic categorization of ASD subjects into the
337 phenotypic clusters, in this population.

338 • **Predicting clinical phenotype from the biological processes disrupted by rare CNVs in**
339 **ASD patients**

340 The Naive Bayes supervised machine learning method was trained and tested using phenotypic
341 clustering information and the 15 biological processes inferred from rare CNVs targeting brain-
342 expressed genes in ASD patients. The classifier was trained with the assumption that ASD
343 subjects with a more dysfunctional clinical phenotype, subgrouped in Cluster 2, would present a
344 different pattern of disrupted biological processes from the individuals with a milder expression
345 of ASD phenotype in Cluster 1.

346 The Naive Bayes classifier trained on data from 1300 patients did not perform well in predicting
347 the more dysfunctional clinical phenotype from disrupted biological processes (Table 5), with
348 scores indicating a low accuracy of the predictive model.

349 To further dissect the information available, the biological process Information Content (IC) for
350 each individual was calculated by summing the IC values for all the biological processes
351 disrupted in that individual. ASD subjects in the first IC quantile (N = 325) had highest IC
352 scores, while ASD cases belonging to fourth quantile (N = 326) contained lowest IC scores. The
353 performance of the Naive Bayes classifier improved when only ASD subjects with higher IC
354 were selected for analysis. Analysis of the group of individuals with highest IC (first quantile)
355 resulted in a higher predictability of ASD clinical outcome (Table 5). The classifier trained and
356 tested on individuals from the first two (1st and 2nd) and first three (1st, 2nd, and 3rd) quantiles also
357 performed better than the classifier designed using the whole dataset of clusters and biological
358 processes (Table 5). The Naive Bayes classifier was thus able to make reasonably good
359 predictions of ASD severity, but only for a subset of ASD individuals with higher IC. This
360 indicates that improved GO information, as well as larger datasets with more GO information
361 available, are needed to usefully integrate clinical and biological data.

362

363

364

365 Table 5: Naive Bayes performance in predicting the severe phenotype of ASD

Data used for classification	N	Precision	Recall	Specificity	F-score
All ASD cases	1300	0.221	0.379	0.655	0.279
ASD cases from 1st quantile with highest IC	325	0.816	0.389	0.699	0.526
ASD cases from 1st and 2nd quantiles of IC	649	0.23	0.384	0.65	0.284
ASD cases from first three quantiles of IC	974	0.29	0.389	0.672	0.329

366

367

Discussion

368 The discovery of diagnostic and prognostic biomarkers for ASD has the potential to improve the
369 reliability of diagnosis at earlier stages of development, as well as the phenotypic categorization
370 for prognosis, eventually informing personalized intervention that is particularly beneficial for
371 very young children. However, in spite of the enormous volume of genetic information generated
372 by genomic approaches in the past decade, the clinical diagnosis of ASD patients is still solely
373 based on neurodevelopmental assessment. The results of many genomic tests, including CNV
374 arrays and clinical exomes, still leave about 80% of the cases without any explanation regarding
375 the biological pathways underlying their disease and their personal clinical presentation.

376 In this study, we developed a novel integrative approach to predict ASD phenotypes from
377 biological processes defined by genetic alterations. Overall, our approach sought to exploit
378 multidimensional clinical measures to define subgroups of ASD patients presenting similar
379 clinical profiles, and then to identify the biological processes disrupted by CNVs that might
380 predict these more homogeneous clinical patterns. For the sake of eventual clinical utility, we
381 chose clinical measures with well-established relevance and frequently used in clinical settings,
382 but established no other restrictions. Further, we did not set any *a priori* hypothesis for gene
383 selection, besides being expressed in the brain.

384 The clustering of clinical data from ASD cases resulted in two subgroups that were clearly
385 distinguishable in terms of severity of phenotype, defined by multiple clinically relevant
386 measures including verbal status, ASD severity, adaptive function and cognitive ability. The
387 identification of only two clusters for the clinical phenotype, with an important proportion of
388 individuals in the AGP dataset that could not be adequately clustered was expected, as it reflects
389 the high clinical heterogeneity of ASD. The identification of these subgroups was in line with

390 previous results by Veatch et al. [40], who also identified two clusters differing in severity using
391 two independent population samples, including the Autism Genetic Resource Exchange (AGRE)
392 and also the AGP dataset. While clinical variables were not fully coincidental between the two
393 studies, we confirmed that the verbal status, ADOS-based severity, VABS-based
394 communication, socialization and daily living skills, as well as gender, were all significantly
395 different between clusters. We noted an unequal contribution of each clinical measure to
396 definition of each cluster, with verbal status the main contributor and the ADOS severity score a
397 low contributor for both clusters, while Performance IQ was mainly important for Cluster 1.
398 In our study, the larger Cluster 1 was characterized by a generally milder phenotype, with all
399 individuals being verbal, a large proportion in the normal IQ range and significantly higher
400 numbers of subjects scoring better in adaptive behavior subscales. Cluster 1 also showed a higher
401 male to female ratio, as expected given the general observation that higher functioning ASD
402 subgroups have a larger proportion of males. The smaller Cluster 2 included only non-verbal
403 subjects, and had a higher percentage of subjects with a more dysfunctional phenotype in terms
404 of adaptive behavior, as well as lower IQ scores. Because cognitive ability is such an important
405 variable for prognosis, we included performance IQ as a clinical variable, in spite of the
406 limitations related to the heterogeneity of IQ measurement tools used for patient assessment by
407 AGP contributing sites. For the AGP dataset, an effort was previously made to rationalize the
408 tests used, and cognitive level was established using a categorical classification provided by
409 AGP sites in three categories, namely severe intellectual disability, mild intellectual disability
410 and normal IQ, for verbal, performance and full scale IQ scores. Limitations were also
411 introduced by the proportions of missing data; given the adopted control of the validity of
412 imputation procedures, only performance IQ met the criteria for reliable imputation, so only this
413 measure was used.
414 Because our main goal was to improve the power for phenotypic subgroup prediction by
415 genetically defined biological processes, we focused on obtaining compact and stable clusters by
416 using strict criteria for cluster stability to assess the goodness of clustering, at the expense of
417 population sample dimension. As expected, the weakly clustered individuals tended to have more
418 divergent scores across clinical measures (data not shown), and therefore were more difficult to
419 cluster with high confidence. It is intriguing that a higher proportion of females than males was
420 removed, suggesting that this divergence of scores is more frequent in girls. This observation

421 generally supports recent debates on the lower adequateness of assessment criteria to the female
422 autism phenotype [41].

423 To test the hypothesis that phenotypic subgroups have specific underlying pathological
424 mechanisms, we first sought to identify the biological processes enriched in the gene sets
425 disrupted by rare CNVs detected in the AGP dataset. The functional enrichment analysis
426 conducted in this study was independent of any prior assumptions or weighting criteria of genes
427 relative to ASD risk. To make functional enrichment analysis hypothesis-free and to let the data
428 speak, we screened for CNVs disrupting any brain-expressed genes. The objective was to obtain
429 a complete picture of the convergence of rare CNVs, targeting any brain-expressed genes, into
430 biological processes relevant for brain function.

431 The biological processes identified in the functional enrichment analysis showed an overlap with
432 putative core biological mechanisms of ASD defined by previous studies. For example, 363
433 brain genes spanned by rare CNVs were enriched for neurodevelopment biological process and
434 56 genes were associated with cognition process. Enrichment of nervous system development
435 and cognition processes in ASD has been previously reported by studies using different
436 approaches, including transcriptome analysis and co-expression networks [15] and is supported
437 by the function of genes most consistently implicated in ASD, like *PTEN*, *RELN*, *SYNGAP1*,
438 *ANK2*, *SCN2A* and *SHANK3* [42]. Noh et al. analysis of *de novo* CNVs spanning ASD genes
439 also implicated cognitive processes, and showed a convergence in cellular component
440 organization or biogenesis, cellular component assembly, and organelle organization biological
441 processes [16]. Other studies implicated cell adhesion processes in ASD as important
442 components of synapse formation and function (46, 47). Dysregulation of polyubiquitination was
443 also in line with previous studies reporting an excess of variants in genes involved in
444 ubiquitination processes, which regulate neurogenesis, neuronal migration and synapse
445 formation, and are thus essential for brain development [43–46].

446 This biological heterogeneity parallels the extensive phenotypic heterogeneity that characterizes
447 ASD. For this reason, we sought to identify the biological processes underlying the more
448 homogeneous phenotypic subgroups defined by the clusters. The Random Forest algorithm was
449 used to assess the importance of each enriched biological process in discriminating the two ASD
450 phenotype subgroups. Feature importance analysis showed that all the biological process
451 contributed positively to the classification of ASD severity. However, the feature importance

452 ranking was different from the significance ranking of enriched biological processes. Despite
453 their relevance for ASD, the top three statistically significant biological processes identified by
454 functional enrichment analysis were least important for the classification of subjects into the
455 phenotypic milder and more dysfunctional subgroups. These findings support the concept that
456 the integration of datasets with multidisciplinary information, including genomic and clinical
457 data, is necessary to discover the biological mechanisms that lead to specific clusters of
458 symptoms.

459 The Naive Bayes classifier was able to make useful predictions of ASD phenotypic subgroups
460 from disrupted biological processes, but only for a subset of individuals for whom annotations
461 had higher information content for the biological processes defined by the CNVs. Currently, GO
462 contains more than 40,000 biological concepts, which are rapidly evolving with the increasing
463 knowledge of biological phenomena and with our ability to structure this knowledge. Therefore,
464 it is expected that the performance of the proposed classifier will improve with the progress in
465 GO annotations.

466 Given the high clinical heterogeneity of ASD, clustering of individuals according to a
467 multidimensional phenotype will result in subgroups with more homogeneous clinical patterns
468 and for whom the causes of this disease are more likely to have the same underlying biological
469 mechanism. The clustering of individuals according to multidimensional clinical symptoms *per*
470 *se* is likely to have implication for prognosis and outcomes, as concurrent symptoms may have a
471 synergistic effect on disease progression, and may thus also help guide clinical practice and
472 intervention. However, thus far this perspective has been insufficiently explored, and not enough
473 datasets are yet available with detailed clinical information that can be merged for large scale
474 analysis. The alterations in diagnostic criteria over time and the changes in versions of
475 instruments like the ADI-R and the ADOS create important challenges for data merging across
476 population samples, which are needed so that sufficient statistical power is achieved for definite
477 conclusions. This study is clear in this limitation, as the number of subjects with important
478 missing data in multiple clinical features was high in the AGP dataset, reducing analytical power,
479 and thus only two stable clusters could be defined. The next research steps will necessarily have
480 to involve overcoming limited clinical information and merging challenges between available
481 datasets, like AGRE and the Simons Foundation Autism Research Initiative (SFARI), so that
482 models established for biological predictions can be useful in clinical settings. On the other hand,

483 while genomic information gets easier and cheaper to collect, improvements are also necessary
484 regarding GO annotations; a large number of subjects with phenotypic subgroup data did not
485 have sufficient GO information content to be useful for classifier predictions.

486 **Conclusion**

487 Overall, the present approach is proof of concept that genotype-phenotype correlations can be
488 established in ASD, and that biological processes can predict multidimensional clinical
489 phenotypes. Importantly, it highlights the usefulness of machine learning approaches that take
490 advantage of multidimensional measures for the construction of more homogeneous clinical
491 profiles. It further stresses the need to overcome the limitations of analyzing individual gene
492 variants in favor of considering biological processes disrupted by an heterogeneous set of gene
493 variants. The results stress two major requisites for translation of genomic information into
494 useful clinical applications: that study datasets include detailed and complete clinical
495 information, and that databases containing biological process information are rigorously and
496 extensively curated. Identification of biological processes for specific clinical subgroups will be
497 important to discover physiological targets for pharmacological therapy that can be efficient for
498 subgroups of patients. This strategy can equally become very useful in clinical settings, for
499 predicting outcomes and planning interventions for subgroups of patients whose specific patterns
500 of clinical presentation are defined by the genes disrupted by specific genetic variants.

501

502

503

504

505

506

507

508

509

510

511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used in this study are from Autism Genome Project (AGP), which are available at dbGaP database.

Competing interests

The authors declare that they have no competing interests.

Funding

Work was supported by UID/MULTI/04046/2013 centre grant from FCT, Portugal (to BioISI) and MA, is recipient of a fellowship from BioSys PhD programme (Ref: SFRH/BD/52485/2014) from FCT (Portugal). Patients and parents were genotyped in the context of the Autism Genome Project (AGP), funded by NIMH, HRB, MRC, Autism Speaks, Hilibrand Foundation, Genome Canada, OGI, and CIHR.

Acknowledgements

We acknowledge the families who participated in these projects.

Authors' contributions

All the authors consented to the submission. MA performed all the analysis and drafted the manuscript. AMV and FMC conceived the study and review the manuscript. Other authors, HFMCM, ARM, JXS, JV, CR and GO proofread the manuscript and helped in understanding the data.

535

References

- 536 1. American Psychiatric Association. Cautionary Statement for Forensic Use of DSM-5.
537 Diagnostic Stat Man Ment Disord 5th Ed. 2013;:280.
538 doi:10.1176/appi.books.9780890425596.744053.
- 539 2. Christensen DL, Baio J, Braun KVN, Bilder D, Charles J, Constantino JN, et al. Prevalence
540 and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and
541 Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR Surveill*
542 *Summ.* 2016;65:1–23. doi:10.15585/mmwr.ss6503a1.
- 543 3. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. *Curr Opin Genet*
544 *Dev.* 2012;22:229–37. doi:10.1016/j.gde.2012.03.002.
- 545 4. Croen LA, Zerbo O, Qian Y, Massolo ML, Rich S, Sidney S, et al. The health status of adults
546 on the autism spectrum. *Autism.* 2015;19:814–23.
- 547 5. Matson JL, Cervantes PE. Commonly studied comorbid psychopathologies among persons
548 with autism spectrum disorder. *Research in Developmental Disabilities.* 2014;35:952–62.
- 549 6. Oliveira G, Ataíde A, Marques C, Miguel TS, Coutinho AM, Mota-Vieira L, et al.
550 Epidemiology of autism spectrum disorder in Portugal: prevalence, clinical characterization, and
551 medical conditions. *Dev Med Child Neurol.* 2007;49:726–33. doi:10.1111/j.1469-
552 8749.2007.00726.x.
- 553 7. Tick B, Bolton P, Happé F, Rutter M, Rijdsdijk F. Heritability of autism spectrum disorders: A
554 meta-analysis of twin studies. *J Child Psychol Psychiatry Allied Discip.* 2016;57:585–95.
- 555 8. Colvert E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E, et al. Heritability of
556 autism spectrum disorder in a UK population-based twin sample. *JAMA Psychiatry.*
557 2015;72:415–23.
- 558 9. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, et al. Common genetic
559 variants, acting additively, are a major source of risk for autism. *Mol Autism.* 2012;3:9.
560 doi:10.1186/2040-2392-3-9.
- 561 10. Chawarska K, Macari S, Shic F. Decreased spontaneous attention to social scenes in 6-
562 month-old infants later diagnosed with autism spectrum disorders. *Biol Psychiatry.* 2013;74:195–
563 203.
- 564 11. Geschwind DH, State MW. Gene hunting in autism spectrum disorder: On the path to
565 precision medicine. *The Lancet Neurology.* 2015;14:1109–20.
- 566 12. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic
567 Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. *N Engl J Med.*
568 2012;367:1321–31. doi:10.1056/NEJMoa1200395.
- 569 13. Liu L, Lei J, Roeder K. Network assisted analysis to reveal the genetic basis of autism. *Ann*
570 *Appl Stat.* 2015;9:1571–600.

- 571 14. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide
572 prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat*
573 *Neurosci.* 2016;19:1454–62. doi:10.1038/nn.4353.
- 574 15. Mahfouz A, Ziats MN, Rennert OM, Lelieveldt BPF, Reinders MJT. Shared Pathways
575 Among Autism Candidate Genes Determined by Co-expression Network Analysis of the
576 Developing Human Brain Transcriptome. *J Mol Neurosci.* 2015;57:580–94. doi:10.1007/s12031-
577 015-0641-3.
- 578 16. Noh HJ, Ponting CP, Boulding HC, Meader S, Betancur C, Buxbaum JD, et al. Network
579 Topologies and Convergent Aetiologies Arising from Deletions and Duplications Observed in
580 Individuals with Autism. *PLoS Genet.* 2013;9.
- 581 17. Correia C, Oliveira G, Vicente AM. Protein interaction networks reveal novel autism risk
582 genes within GWAS statistical noise. *PLoS One.* 2014;9:1–11.
- 583 18. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional Impact of
584 Global Rare Copy Number Variation in Autism Spectrum Disorder. *Nature.* 2010;466:368–72.
585 doi:10.1038/nature09146.Functional.
- 586 19. APA. American Psychiatric Association Diagnostic and Statistical Manual of Mental
587 Disorders (DSM-IV). SpringerReference. 2000;:Fifth Edition. Arlington, VA.
588 doi:10.1007/SpringerReference_179660.
- 589 20. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: A revised version of
590 a diagnostic interview for caregivers of individuals with possible pervasive developmental
591 disorders. *J Autism Dev Disord.* 1994;24:659–85.
- 592 21. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, et al. Autism
593 diagnostic observation schedule: A standardized observation of communicative and social
594 behavior. *J Autism Dev Disord.* 1989;19:185–212.
- 595 22. Sparrow S, Balla D, Cicchetti D. The Vineland Adaptive Behavior Scales: Interview edition,
596 survey. In: *Major psychological assessment instruments.* 1984. p. 199–231.
- 597 23. Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in
598 autism spectrum disorders. *J Autism Dev Disord.* 2009;39:693–705.
- 599 24. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-
600 type data. *Bioinformatics.* 2012;28:112–8.
- 601 25. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
- 602 26. Rokach L, Maimon O. Clustering Methods. *Data Min Knowl Discov Handb.* 2010;:321–52.
603 doi:10.1007/0-387-25465-X_15.
- 604 27. Gower JC. A General Coefficient of Similarity and Some of Its Properties. *Biometrics.*
605 1971;27:857. doi:10.2307/2528823.
- 606 28. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which

- 607 Algorithms Implement Ward's Criterion? *J Classif.* 2014;31:274–95.
- 608 29. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster
609 analysis. *J Comput Appl Math.* 1987;20 C:53–65.
- 610 30. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster Analysis Basics and
611 Extensions. R package version 2.0.1. 2015. [http://cran.r-](http://cran.r-project.org/web/packages/cluster/index.html)
612 [project.org/web/packages/cluster/index.html](http://cran.r-project.org/web/packages/cluster/index.html).
- 613 31. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping
614 and analysis of copy number variations in the human genome: A data resource for clinical and
615 research applications. *Genome Res.* 2009;19:1682–90. doi:10.1101/gr.083501.108.
- 616 32. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number
617 variation morbidity map of developmental delay. *Nat Genet.* 2011;43:838–46.
618 doi:10.1038/ng.909.
- 619 33. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic
620 Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.*
621 2014;42.
- 622 34. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative functional
623 genomic analyses implicate specific molecular pathways and circuits in autism. *Cell.*
624 2013;155:1008–21. doi:10.1016/j.cell.2013.10.031.
- 625 35. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web
626 server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*
627 2016;44:W83–9.
- 628 36. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of
629 gene ontology terms. *PLoS One.* 2011;6.
- 630 37. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional
631 similarity of gene products based on gene ontology. *BMC Bioinformatics.* 2006;7:302.
632 doi:10.1186/1471-2105-7-302.
- 633 38. Liaw A, Yan J, Li W, Han L, Schroff F, Criminisi A, et al. Package “randomForest.” *R news.*
634 2015;XXXIX:54.1-54.10.
- 635 39. Kuncheva LI. On the optimality of Naïve Bayes with dependent binary features. *Pattern*
636 *Recognit Lett.* 2006;27:830–7. doi:10.1016/j.patrec.2005.12.001.
- 637 40. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically
638 meaningful phenotypic subgroups in autism spectrum disorders. *Genes, Brain Behav.*
639 2014;13:276–85.
- 640 41. Rynkiewicz A, Schuller B, Marchi E, Piana S, Camurri A, Lassalle A, et al. An investigation
641 of the “female camouflage effect” in autism using a computerized ADOS-2 and a test of
642 sex/gender differences. *Mol Autism.* 2016;7:10. doi:10.1186/s13229-016-0073-0.

- 643 42. Wen Y, Alshikho MJ, Herbert MR. Pathway network analyses for autism reveal multisystem
644 involvement, major overlaps with other diseases and convergence upon MAPK and calcium
645 signaling. *PLoS One*. 2016;11:1–23.
- 646 43. O’Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de
647 novo mutations implicate novel genes underlying simplex autism risk. *Nat Commun*. 2014;5.
- 648 44. Iossifov I, O’Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution
649 of de novo coding mutations to autism spectrum disorder. *Nature*. 2014;515:216–21.
- 650 45. Kawabe H, Brose N. The role of ubiquitylation in nerve cell development. *Nature Reviews*
651 *Neuroscience*. 2011;12:251–68.
- 652 46. Nava C, Lamari F, Héron D, Mignot C, Rastetter A, Keren B, et al. Analysis of the
653 chromosome X exome in patients with autism spectrum disorders identified novel candidate
654 genes, including TMLHE. *Transl Psychiatry*. 2012;2:e179. doi:10.1038/tp.2012.102.

A. Clinical data processing pipeline

Clinical reports:
ADOS, ADI-R,
IQ, and VABS

Clinical reports
with imputed
missing values

Agglomerative
hierarchical
clustering

Clusters validation

Silhouette
analysis

Internal and
external clusters
validation

Clusters stability

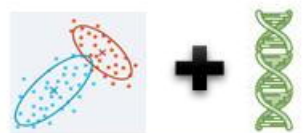
Stable and
validated
multidimensional
clusters



B. CNVs data processing and functional annotation pipeline

High confidence
and rare CNVs
targeting brain
genes

Merging clinical and
CNV information

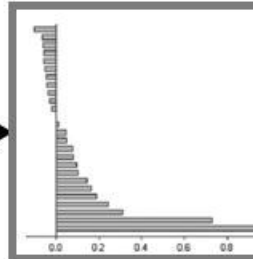


Functional
annotation analysis

Disrupted
biological
processes

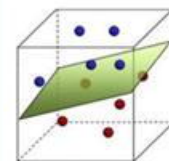


C. Feature importance analysis

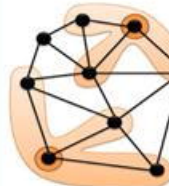


D. Machine learning pipeline to predict clinical outcome

Classifier training
and testing



Genotype-phenotype
associations



Prediction of ASD
multidimensional
phenotype

