1   **Identification of biological mechanisms underlying a multidimensional ASD**

2   **phenotype using machine learning**

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. Couto[3], Astrid M. Vicente[1,2*]

5   [*]Correspondence: Astrid M. Vicente, Instituto Nacional de Saúde Doutor Ricardo Jorge, Avenida

6   Padre Cruz, 1649-016 Lisboa, Portugal. Email: astrid.vicente@insa.min-saude.pt

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

8   Portugal

9   [2]BioISI: Biosystems & Integrative Sciences Institute, Faculdade de Ciências, Universidade de

10   Lisboa, Lisboa, Portugal

11

12   [3]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

13   [4]Unidade de Neurodesenvolvimento e Autismo (UNDA), Serviço do Centro de Desenvolvimento

14   da Criança, Centro de Investigação e Formação Clínica, Hospital Pediátrico, Centro Hospitalar e

15   Universitário de Coimbra, Coimbra, Portugal

16   [5]Institute for Biomedical Imaging and Life Sciences, Faculty of Medicine, Universidade de

17   Coimbra, Coimbra, Portugal

18   [6]University Clinic of Pediatrics, Faculty of Medicine, University of Coimbra, Portugal

19

20

21

22

23

Abstract

The complex genetic architecture of Autism Spectrum Disorder (ASD) and its heterogeneous phenotype make molecular diagnosis and patient prognosis challenging tasks. To establish more precise genotype-phenotype correlations in ASD, we developed a novel machine learning integrative approach, which seeks to delineate associations between patients' clinical profiles and disrupted biological processes inferred from their Copy Number Variants (CNVs) that span brain genes. Clustering analysis of relevant clinical measures from 2446 ASD cases in the Autism Genome Project identified two distinct phenotypic subgroups. Patients in these clusters differed significantly in ADOS-defined severity, adaptive behaviour profiles, intellectual ability and verbal status, the latter contributing the most for cluster stability and cohesion. Functional enrichment analysis of brain genes disrupted by CNVs in these ASD cases identified 15 statistically significant biological processes, including cell adhesion, neural development, cognition and polyubiquitination, in line with previous ASD findings. A Naive Bayes classifier, generated to predict the ASD phenotypic clusters from disrupted biological processes, achieved predictions with a high Precision (0.82) but low recall (0.39), for a subset of patients with higher biological Information Content scores. This study shows that milder and more severe clinical presentations can have distinct underlying biological mechanisms. It further highlights how machine learning approaches can reduce clinical heterogeneity using multidimensional clinical measures, and establish genotype-phenotype correlations in ASD. However, predictions are strongly dependent on patient's information content. Findings are therefore a first step towards the translation of genetic information into clinically useful applications, but emphasize the need for larger datasets with very complete clinical and biological information.

Keywords

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

Introduction

52

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

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

72 To improve our ability to infer clinical meaning from rare CNVs in ASD, for eventual
73 application as biological markers, we developed a machine learning-based approach involving
74 the integration of gene functional annotations and clinical phenotypes. Our approach was
75 developed in four steps, namely: 1) definition of clinically distinct subgroups in ASD cases; 2)
76 discovery of functionally enriched biological processes defined by rare CNVs disrupting brain-
77 expressed genes in the same ASD cases; 3) assessment of the contribution of disrupted biological
78 processes for classification of ASD phenotypes; 4) design and predictive effectiveness
79 characterization of a machine learning classifier for clinical outcome in ASD patients.

80

3

Methods

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

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

- **Participants**

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

- **ASD diagnosis, clinical assessment instruments and clinical features**

Individuals meeting criteria defined by the Diagnostic and Statistical Manual of Mental Disorders IV (DSM–IV) [19] and the thresholds for Autism or ASD from the Autism Diagnostic Interview-Revised (ADI-R) [20] and the Autism Diagnostic Observation Schedule (ADOS) were classified as ASD cases [21]. The AGP defined a phenotypic classification system based on the combined ADI-R and ADOS diagnosis, categorizing subjects into Strict (meeting thresholds for

4

109   Autism by the ADI-R and ADOS), Broad (meeting thresholds for Autism from one instrument

110   and ASD from the other) and Spectrum (meeting thresholds for Autism from at least one

111   instrument or ASD from both). Individuals with an ASD diagnosis from only one instrument and

112   no information from the other, or not meeting thresholds for Autism or ASD from one of the

113   instruments, regardless from the classification from the other, were not included in the study.

114   Clinical measures used in this study were retrieved from the AGP database, including the ADIR

115   verbal status, ADOS severity score, Vineland Adaptive Behaviour Scales (VABS) [22] subscales

116   and an Intelligence Quotient (IQ).

117   The ADI-R verbal status is a dichotomized measure indicating the verbal status of the individual

118   at evaluation. The ADOS severity metric ranges from 1 to 10 and is calculated from ADOS

119   modules 1 to 3 raw scores [23]. As there is no algorithm available to calculate ADOS severity

120   score for ADOS module 4 reports, which is applied only to adolescents and adults, subjects with

121   the ADOS module 4 (N= 149) were dropped from further processing. The severity score

122   distribution is collapsed into three categories, namely Autism (severity scores ranging from 6 to

123   10), ASD (severity scores ranging from 4 to 5) and Non-Spectrum (severity scores from 1 to 3),

124   which reflect the mapping of the severity metric onto raw ADOS scores. The ADOS Non-

125   spectrum category includes individuals with a mild phenotype, and in this study 125 individuals

126   with a Non-spectrum ADOS severity score fell within the Spectrum phenotypic class from the

127   AGP, meaning they met thresholds for autism from the ADI-R, and were thus included.

128   The VABS is used to assess adaptive functioning of individuals and consists of three subscales,

129   namely, socialization, communication and daily living skills scores, and also computes a

130   composite score. Subjects with VABS scores ≤70 were classified in a dysfunctional adaptive

131   behavior category, for all subscales. IQ scores of ASD cases were also retrieved from the AGP

132   database, and categorized with the following thresholds: IQ>70 normal, 50<IQ<70 mild

133   intellectual disability, IQ<50 severe intellectual disability.

134   Clinical reports from the ASD patients were examined for missing values, and clinical features

135   with more than 70% information were retained for the analysis. To minimise missing value

136   imputation bias, individuals with missing values above this threshold for more than two clinical

137   features were also excluded. Completeness of each clinical feature is reported in Table S1

138   (Additional file 1). Missing values were imputed using the missForest [24] R package that

139 implements the Random Forest [25] algorithm, a decision tree-based supervised machine
140 learning method. Imputation error was assessed using the normalised global Proportion of
141 Falsely Classification (PFC), and the missing values imputation error was 0.12.

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

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

157 To assess the contributions of each clinical feature in defining the clusters, we excluded one
158 feature at a time, re-performed the clustering and observed the changes in Silhouette values of
159 both clusters. For this purpose, we selected Silhouette value as an evaluation metric because it
160 was also used to define outliers in clinical data. A decrease in the Silhouette value of a cluster
161 after removing one feature indicates its importance in defining this cluster and vice versa.

162 • **Goodness of clustering assessment**

163 A Silhouette method [29] was employed to estimate the goodness of the clustering results. The
164 Silhouette value for each individual shows how well the individual is clustered, and ranges from
165 -1 to 1, with individuals scoring below 0 considered as wrongly clustered. In addition, the
166 Silhouette value for each cluster was derived, and clusters with Silhouette value of > 0.25 were

6

167 considered as true clusters. Bootstrapping with 1000 iterations was used to measure the stability

168 of clusters, where a boot mean value above 0.85 corresponds to stable clusters. All clustering

169 analysis was performed in R environment, using Cluster [30] and FPC packages.

170 • **Functional enrichment analysis**

171 Genotyping and CNV calling methods for the AGP ASD subjects (N=2446) were previously

172 described [18]. CNVs called by any two algorithms (high confidence CNVs) and above 30kb in

173 size were retained for further analysis. To screen for rare CNVs (<1% in control population)

174 CNV frequencies in control populations were estimated using the genotypes from the studies by

175 Sheikh et al. [31] (N = 1320) and Cooper et al. [32] (N = 8329), identified using the same

176 genotyping platform [18]. Control genotypes were obtained from the Database of Genomic

177 Variants (DGV) [33].

178 To focus CNV selection on variants spanning brain-expressed genes, avoiding *a priori*

179 hypotheses from ASD candidate gene assumptions, an extensive list comprising 15585 brain-

180 expressed genes was obtained from Parikshak et al. [34]. The brain-expressed gene list was

181 prepared from brain RNA-seq data, collected at thirteen different developmental stages,

182 including genes expressing during early brain developmental phase. The full criteria and

183 parameters used to define the brain-expressed gene list were previously described [34]. **.**

184 The g:Profiler [35] tool was employed to identify biological processes enriched for brain-

185 expressed genes spanned by rare CNVs in ASD individuals. g:Profiler implements a

186 hypergeometric test to estimate the statistical significance of enriched biological processes,

187 followed by multiple corrections for the tested hypotheses using the Benjamini-Hochberg

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

189 genes.

190 The GO tool contains a Directed Acyclic Graph (DAG) structure with a clear hierarchical parent-

191 to-child relationship between GO terms. Because of this DAG structure, functional enrichment

192 analysis can result in redundant GO terms, which may lead to high correlations between GO

193 terms. To minimise the correlations between GO terms, the Revigo [36] tool was employed to

194 redundant GO results. Revigo uses the methods of semantic similarity to measure similarities

195   between GO terms. The SimRel [37] method was used to calculate similarities between GO

196   terms, and terms with a similarity score of > 0.7 were grouped.

197   • **Feature importance assessment**

198   The mean decrease in accuracy of the Random Forest algorithm was used to compute the

199   importance score of each disrupted biological process for categorizing ASD subjects into defined

200   phenotypic clusters. A stratified ten-fold cross-validation quan

201   tifies the importance of all features. The importance score of all disrupted biological processes

202   was recorded at each fold. A final importance score for each biological process was calculated by

203   averaging their importance score values across all the ten folds. Random Forest was

204   implemented using randomdForst R package [38].

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

206   A Naive Bayes [39] machine learning method was employed to predict the ASD phenotypic

207   group, defined by the clustering analysis, from biological processes disrupted by rare CNVs.

208   This method employs the Bayes theorem of probability for training and testing of the model, and

209   the algorithm was implemented using the klaR R package with default parameters. Precision,

210   recall, specificity and F-score were used as evaluation measures. To train and test the Naive

211   Bayes, a stratified five-fold cross-validation approach was used, in which data was first split into

212   five equal subsets with equal class probabilities; a Naive Bayes model was trained on any four

213   subsets, and the remaining subset was used as the test set. This process was repeated five times

214   and each time a different subset was used as test set. For each repetition, the model performance

215   was estimated and mean values for precision, recall, specificity and F-score were reported. The

216   Naive Bayes classifier was trained on patient's data by using the "more severe" cluster as the

217   positive class and the "less severe" cluster as the negative class.

218   The Information Content (IC) from each individual represents the level of specificity of

219   biological processes disruption, and was derived by summing the IC values of all the biological

220   processes disrupted in each individual. IC is a numerical value that describes the specificity of a

221   GO term using its position in the GO DAG structure.

222

8

Results

223

- **Identification of ASD clusters defined by clinical phenotype**

224

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

235 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

236

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

239

- **Clinical interpretations of the clusters**

240 All clinical measures differed significantly between the two clusters, as shown in Table 2.
241 Cluster 1(Additional file 1: black circles in Figure S1) includes a higher number of individuals,
242 who generally exhibited a milder clinical phenotype, while Cluster 2 (Additional file 1: red
243 triangles in Figure S1) included a higher percentage of subjects with severe dysfunction. All
244 individuals in Cluster 1 were verbal according to the ADI-R, while Cluster 2 included only non-

9

245   verbal cases. The mean age of ADI-R assessment was 7.7 years, an age when verbal status is
246   generally well established. Furthermore, the mean age of individuals in Cluster 1 (mean age
247   8.02) and Cluster 2 (mean age 7.01) did not significantly differ.

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

254   Regarding the ADOS severity score, approximately 14% of the individuals in Cluster 1 were
255   assigned to the milder category of the ADOS severity score ("Non-spectrum" for ADOS, but
256   scoring positive for "Autism" in the ADI-R, and therefore classified in the AGP "Spectrum"
257   phenotypic class, see methods). Conversely, none of the individuals in Cluster 2 scored in this
258   category. On the other hand, a significantly higher percentage of cases in Cluster 2 (20.65%)
259   than individuals in Cluster 1 (7.09%) scored in the intermediate ASD severity category. It is
260   noteworthy that both clusters show a similarly high percentage of individuals scoring in the
261   "Autism" ADOS severity category. This is not surprising since this broad category (scores
262   ranging from 6 to 10) comprises all subjects classified in the Strict AGP phenotype class but also
263   a large proportion of individuals in the AGP Broad phenotype class. The "Autism" ADOS
264   severity score therefore targets a subset of the study population that can be quite heterogeneous
265   in phenotypic presentation. Corroborating this, we found that the "Autism" category of the
266   ADOS severity score is not significantly associated with the clusters ($\chi 2= 0.15$, $p = 0.901$, df =
267   2), even though overall there is a significant association of the overall ADOS severity scores
268   (Table 2).

269

270

271

272

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

Clinical measure	Clinically defined categories	Cluster 1 N (%)	Cluster 2 N (%)	*p-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)	

274   [a]Fisher Exact Test, [b]Chi-Square test

275   Both clusters were strongly dominated by the male gender, partly because of the high percentage

276   of males in the dataset after the elimination of weakly or wrongly clustered individuals.

277   However, the percentage of males was higher in cluster 1, representing the milder phenotype,

278   consistent with general observations that male to female ratios are higher in datasets that

279   comprise more high- function ASD individuals.

280   Analysis of the contribution of each clinical feature in defining clusters showed that the main

281   contributor was the ADIR verbal status variable (Additional file 1: Table S2). The VABS

282   subscales had a strong effect on Cluster 1 but a modest role in defining Cluster 2. Performance

283   IQ also contributed more to Cluster 1 whereas for Cluster 2 it has the least effect. The ADOS

11

284 severity score did not have a major role in defining either cluster, as indicated by the similar high

285 percentage of subjects scoring within the range of "Autism" in the ADOS severity scale in both

286 clusters. Similarly, gender was not an important contributor to the definition of either cluster.

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

288 CNVs (N=129754) identified in 2446 subjects with ASD were filtered to select rare, high

289 confidence CNVs, over 30 Kb in size and that contained complete or partial brain-expressed

290 gene sequences. The selected high confidence, rare CNVs (N=12683) disrupted 4025 brain-

291 expressed genes in 2414 subjects with ASD (86.8% males and 13.2% females).

292 Phenotypic cluster and rare CNV data was complete for 1357 individuals with ASD, and

293 available for integration. Functional enrichment analysis of rare CNVs targeting brain-expressed

294 genes (N=2738) in 1357 patients identified 17 statistically significant biological processes

295 (Additional file 1: Table S3). g:Profiler did not recognize 187 genes from the input list.

296 The redundancy of GO terms in functional enrichment analysis, caused by overlapping

297 annotations in ancestors and descendent terms in the DAG structure of GO, was reduced by

298 grouping the terms that had a semantic similarity score higher than 0.7 (Additional file 1: Table

299 S3). The Revigo tool used to reduce redundancy did not recognise one biological process

300 (*Plasma membrane bounded cell projection organization*). After redundancy reduction, 16

301 biological processes remained (Table 3), with the *Calcium-dependent cell-cell adhesion via*

302 *plasma membrane cell adhesion molecules* biological process merged with *Homophilic cell*

303 *adhesion via plasma membrane adhesion molecules* (similarity score = 0.76).

304 The most significant biological process identified in this dataset was *Homophilic cell adhesion*

305 *via plasma membrane adhesion molecules*, which includes 53 brain-expressed genes disrupted

306 by the selected CNVs. The ten most significant biological processes were related to cell adhesion

307 and cellular organization, and also included nervous system development and protein

308 poliubiquitination (Table 3). Moreover, two significant biological processes were related to

309 behavior and cognition.

310

12

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

Biological processes	Enriched genes (N)	FDR *p*-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

313

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

315 The enriched biological processes and phenotypic cluster information for ASD cases were
316 combined in a matrix to assess the predictive value of the biological processes for categorization
317 in one of the two phenotypic clusters, broadly characterized by a milder and a more severe
318 phenotypic presentation. The 57 individuals containing both rare CNV and cluster information
319 that did not present any enriched biological process were excluded, so further analysis comprised
320 1300 ASD patients.

13

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

324   Table 4: Importance of each biological process from Random Forest in classifying ASD subjects
325   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

326

327   The importance of each biological process was calculated using the mean decrease in accuracy,
328   computed by permuting each biological process. The feature importance analysis using Random
329   Forest, which was trained and tested using stratified 10-fold cross-validation over the integrated
330   dataset, revealed positive values for all features, indicating that all of the biological processes are
331   positively contributing for classification. The most important biological process for the
332   classification was *Regulation of cellular component organization*, with a mean decrease in
333   accuracy of 0.052. The most significantly enriched biological process in the overall ASD dataset,
334   *Homophilic cell adhesion via plasma membrane adhesion molecules* was ranked at position 14,
335   indicating it is not a top contributor to phenotypic categorization of ASD subjects into the
336   phenotypic clusters, in this population.

14

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

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

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

348 To further dissect the information available, the biological process Information Content (IC) for
349 each individual was calculated by summing the IC values for all the biological processes
350 disrupted in that individual. ASD subjects in the first IC quantile (N = 325) had highest IC
351 scores, while ASD cases belonging to fourth quantile (N = 326) contained lowest IC scores. The
352 performance of the Naive Bayes classifier improved when only ASD subjects with higher IC
353 were selected for analysis. Analysis of the group of individuals with highest IC (first quantile)
354 resulted in a higher predictability of ASD clinical outcome (Table 5). The classifier trained and
355 tested on individuals from the first two (1st and 2nd) and first three (1st, 2nd, and 3rd) quantiles also
356 performed better than the classifier designed using the whole dataset of clusters and biological
357 processes (Table 5).

358 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

359

15

360   The Naive Bayes classifier was thus able to make reasonably good predictions of ASD severity,

361   but only for a subset of ASD individuals with higher IC. This indicates that improved GO

362   information, as well as larger datasets with more GO information available, are needed to

363   usefully integrate clinical and biological data.

Discussion

365   The discovery of diagnostic and prognostic biomarkers for ASD has the potential to improve the

366   reliability of diagnosis at earlier stages of development, as well as the phenotypic categorization

367   for prognosis, eventually informing personalized intervention that is particularly beneficial for

368   very young children. However, in spite of the enormous volume of genetic information generated

369   by genomic approaches in the past decade, the clinical diagnosis of ASD patients is still solely

370   based on neurodevelopmental assessment. The results of many genomic tests, including CNV

371   arrays and clinical exomes, still leave about 80% of the cases without any explanation regarding

372   the biological pathways underlying their disease and their personal clinical presentation.

373   In this study, we developed a novel integrative approach to predict ASD phenotypes from

374   biological processes defined by genetic alterations. Overall, our approach sought to exploit

375   multidimensional clinical measures to define subgroups of ASD patients presenting similar

376   clinical profiles, and then to identify the biological processes disrupted by CNVs that might

377   predict these more homogeneous clinical patterns. For the sake of eventual clinical utility, we

378   chose clinical measures with well-established relevance and frequently used in clinical settings,

379   but established no other restrictions. Further, we did not set any *a priori* hypothesis for gene

380   selection, besides being expressed in the brain.

381   The clustering of clinical data from ASD cases resulted in two subgroups that were clearly

382   distinguishable in terms of severity of phenotype, defined by multiple clinically relevant

383   measures including verbal status, ASD severity, adaptive function and cognitive ability. The

384   identification of only two clusters for the clinical phenotype, with an important proportion of

385   individuals in the AGP dataset that could not be adequately clustered was expected, as it reflects

386   the high clinical heterogeneity of ASD. The identification of these subgroups was in line with

387   previous results by Veatch et al. [40], who also identified two clusters differing in severity using

388   two independent population samples, including the Autism Genetic Resource Exchange (AGRE)

389   and also the AGP dataset. While clinical variables were not fully coincidental between the two

390 studies, we confirmed that the verbal status, ADOS-based severity, VABS-based
391 communication, socialization and daily living skills, as well as gender, were all significantly
392 different between clusters. We noted an unequal contribution of each clinical measure to
393 definition of each cluster, with verbal status the main contributor and the ADOS severity score a
394 low contributor for both clusters, while Performance IQ was mainly important for Cluster 1.

395 In our study, the larger Cluster 1 was characterized by a generally milder phenotype, with all
396 individuals being verbal, a large proportion in the normal IQ range and significantly higher
397 numbers of subjects scoring better in adaptive behavior subscales. Cluster 1 also showed a higher
398 male to female ratio, as expected given the general observation that higher functioning ASD
399 subgroups have a larger proportion of males. The smaller Cluster 2 included only non-verbal
400 subjects, and had a higher percentage of subjects with a more dysfunctional phenotype in terms
401 of adaptive behavior, as well as lower IQ scores. Because cognitive ability is such an important
402 variable for prognosis, we included performance IQ as a clinical variable, in spite of the
403 limitations related to the heterogeneity of IQ measurement tools used for patient assessment by
404 AGP contributing sites. For the AGP dataset, an effort was previously made to rationalize the
405 tests used, and cognitive level was established using a categorical classification provided by
406 AGP sites in three categories, namely severe intellectual disability, mild intellectual disability
407 and normal IQ, for verbal, performance and full scale IQ scores. Limitations were also
408 introduced by the proportions of missing data; given the adopted control of the validity of
409 imputation procedures, only performance IQ met the criteria for reliable imputation, so only this
410 measure was used.

411 Because our main goal was to improve the power for phenotypic subgroup prediction by
412 genetically defined biological processes, we focused on obtaining compact and stable clusters by
413 using strict criteria for cluster stability to assess the goodness of clustering, at the expense of
414 population sample dimension. As expected, the weakly clustered individuals tended to have more
415 divergent scores across clinical measures (data not shown), and therefore were more difficult to
416 cluster with high confidence. It is intriguing that a higher proportion of females than males was
417 removed, suggesting that this divergence of scores is more frequent in girls. This observation
418 generally supports recent debates on the lower adequateness of assessment criteria to the female
419 autism phenotype [41].

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

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

443   This biological heterogeneity parallels the extensive phenotypic heterogeneity that characterizes
444   ASD. For this reason, we sought to identify the biological processes underlying the more
445   homogeneous phenotypic subgroups defined by the clusters. The Random Forest algorithm was
446   used to assess the importance of each enriched biological process in discriminating the two ASD
447   phenotype subgroups. Feature importance analysis showed that all the biological process
448   contributed positively to the classification of ASD severity. However, the feature importance
449   ranking was different from the significance ranking of enriched biological processes. Despite
450   their relevance for ASD, the top three statistically significant biological processes identified by

18

451   functional enrichment analysis were least important for the classification of subjects into the
452   phenotypic milder and more dysfunctional subgroups. These findings support the concept that
453   the integration of datasets with multidisciplinary information, including genomic and clinical
454   data, is necessary to discover the biological mechanisms that lead to specific clusters of
455   symptoms.

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

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

481   regarding GO annotations; a large number of subjects with phenotypic subgroup data did not

482   have sufficient GO information content to be useful for classifier predictions.

Conclusion

484   Overall, the present approach is proof of concept that genotype-phenotype correlations can be

485   established in ASD, and that biological processes can predict multidimensional clinical

486   phenotypes. Importantly, it highlights the usefulness of machine learning approaches that take

487   advantage of multidimensional measures for the construction of more homogeneous clinical

488   profiles. It further stresses the need to overcome the limitations of analyzing individual gene

489   variants in favor of considering biological processes disrupted by an heterogeneous set of gene

490   variants. The results stress two major requisites for translation of genomic information into

491   useful clinical applications: that study datasets include detailed and complete clinical

492   information, and that databases containing biological process information are rigorously and

493   extensively curated. Identification of biological processes for specific clinical subgroups will be

494   important to discover physiological targets for pharmacological therapy that can be efficient for

495   subgroups of patients. This strategy can equally become very useful in clinical settings, for

496   predicting outcomes and planning interventions for subgroups of patients whose specific patterns

497   of clinical presentation are defined by the genes disrupted by specific genetic variants.

498

Acknowledgements

Conflict of Interest

508   The authors declare that they have no conflict of interest.

509

References

510

511   1. American Psychiatric Association. Cautionary Statement for Forensic Use of DSM-5.

512   Diagnostic Stat Man Ment Disord 5th Ed. 2013;:280.

513   doi:10.1176/appi.books.9780890425596.744053.

514   2. Christensen DL, Baio J, Braun KVN, Bilder D, Charles J, Constantino JN, et al. Prevalence

515   and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and

516   Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. MMWR Surveill

517   Summ. 2016;65:1–23. doi:10.15585/mmwr.ss6503a1.

518   3. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. Curr Opin Genet

519   Dev. 2012;22:229–37. doi:10.1016/j.gde.2012.03.002.

520   4. Croen LA, Zerbo O, Qian Y, Massolo ML, Rich S, Sidney S, et al. The health status of adults

521   on the autism spectrum. Autism. 2015;19:814–23.

522   5. Matson JL, Cervantes PE. Commonly studied comorbid psychopathologies among persons

523   with autism spectrum disorder. Research in Developmental Disabilities. 2014;35:952–62.

524   6. Oliveira G, Ataíde A, Marques C, Miguel TS, Coutinho AM, Mota-Vieira L, et al.

525   Epidemiology of autism spectrum disorder in Portugal: prevalence, clinical characterization, and

526   medical conditions. Dev Med Child Neurol. 2007;49:726–33. doi:10.1111/j.1469-

527   8749.2007.00726.x.

528   7. Tick B, Bolton P, Happé F, Rutter M, Rijsdijk F. Heritability of autism spectrum disorders: A

529   meta-analysis of twin studies. J Child Psychol Psychiatry Allied Discip. 2016;57:585–95.

530   8. Colvert E, Tick B, McEwen F, Stewart C, Curran SR, Woodhouse E, et al. Heritability of

531   autism spectrum disorder in a UK population-based twin sample. JAMA Psychiatry.

532   2015;72:415–23.

533   9. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, et al. Common genetic

534   variants, acting additively, are a major source of risk for autism. Mol Autism. 2012;3:9.

535   doi:10.1186/2040-2392-3-9.

536   10. Chawarska K, Macari S, Shic F. Decreased spontaneous attention to social scenes in 6-

537   month-old infants later diagnosed with autism spectrum disorders. Biol Psychiatry. 2013;74:195–

538   203.

539   11. Geschwind DH, State MW. Gene hunting in autism spectrum disorder: On the path to

540   precision medicine. The Lancet Neurology. 2015;14:1109–20.

541   12. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic

542   Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N Engl J Med.

543   2012;367:1321–31. doi:10.1056/NEJMoa1200395.

544   13. Liu L, Lei J, Roeder K. Network assisted analysis to reveal the genetic basis of autism. Ann

545   Appl Stat. 2015;9:1571–600.

546   14. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide

547   prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat

548   Neurosci. 2016;19:1454–62. doi:10.1038/nn.4353.

549   15. Mahfouz A, Ziats MN, Rennert OM, Lelieveldt BPF, Reinders MJT. Shared Pathways

550   Among Autism Candidate Genes Determined by Co-expression Network Analysis of the

551   Developing Human Brain Transcriptome. J Mol Neurosci. 2015;57:580–94. doi:10.1007/s12031-

552   015-0641-3.

553   16. Noh HJ, Ponting CP, Boulding HC, Meader S, Betancur C, Buxbaum JD, et al. Network

554   Topologies and Convergent Aetiologies Arising from Deletions and Duplications Observed in

555   Individuals with Autism. PLoS Genet. 2013;9.

556   17. Correia C, Oliveira G, Vicente AM. Protein interaction networks reveal novel autism risk

557   genes within GWAS statistical noise. PLoS One. 2014;9:1–11.

558   18. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, et al. Functional Impact of

559   Global Rare Copy Number Variation in Autism Spectrum Disorder. Nature. 2010;466:368–72.

560   doi:10.1038/nature09146.Functional.

561   19. APA. American Psychiatric Association Diagnostic and Statistical Manual of Mental

562   Disorders (DSM-IV). SpringerReference. 2000;:Fifth Edition. Arlington, VA.

563   doi:10.1007/SpringerReference_179660.

564   20. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: A revised version of

565   a diagnostic interview for caregivers of individuals with possible pervasive developmental

566   disorders. J Autism Dev Disord. 1994;24:659–85.

567   21. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, et al. Austism

568   diagnostic observation schedule: A standardized observation of communicative and social

569   behavior. J Autism Dev Disord. 1989;19:185–212.

570   22. Sparrow S, Balla D, Cicchetti D. The Vineland Adaptive Behavior Scales: Interview edition,

571   survey. In: Major psychological assessment instruments. 1984. p. 199–231.

572   23. Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in

573   autism spectrum disorders. J Autism Dev Disord. 2009;39:693–705.

574   24. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-

575   type data. Bioinformatics. 2012;28:112–8.

576   25. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

577   26. Rokach L, Maimon O. Clustering Methods. Data Min Knowl Discov Handb. 2010;:321–52.

578   doi:10.1007/0-387-25465-X_15.

579   27. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics.

580   1971;27:857. doi:10.2307/2528823.

581   28. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which

582   Algorithms Implement Ward's Criterion? J Classif. 2014;31:274–95.

583   29. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster

584   analysis. J Comput Appl Math. 1987;20 C:53–65.

585   30. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster Analysis Basics and

586   Extensions. R package version 2.0.1. 2015. http://cran.r-

587   project.org/web/packages/cluster/index.html.

588   31. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, et al. High-resolution mapping

589   and analysis of copy number variations in the human genome: A data resource for clinical and

590   research applications. Genome Res. 2009;19:1682–90. doi:10.1101/gr.083501.108.

591   32. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number

592   variation morbidity map of developmental delay. Nat Genet. 2011;43:838–46.

593   doi:10.1038/ng.909.

594   33. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic

595   Variants: A curated collection of structural variation in the human genome. Nucleic Acids Res.

596   2014;42.

597   34. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative functional

598   genomic analyses implicate specific molecular pathways and circuits in autism. Cell.

599   2013;155:1008–21. doi:10.1016/j.cell.2013.10.031.

600   35. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web

601   server for functional interpretation of gene lists (2016 update). Nucleic Acids Res.

602   2016;44:W83–9.

603   36. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of

604   gene ontology terms. PLoS One. 2011;6.

605   37. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional

606   similarity of gene products based on gene ontology. BMC Bioinformatics. 2006;7:302.

607   doi:10.1186/1471-2105-7-302.

608   38. Liaw A, Yan J, Li W, Han L, Schroff F, Criminisi A, et al. Package "randomForest." R news.

609   2015;XXXIX:54.1-54.10.

610   39. Kuncheva LI. On the optimality of Naïve Bayes with dependent binary features. Pattern

611   Recognit Lett. 2006;27:830–7. doi:10.1016/j.patrec.2005.12.001.

612   40. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically

613   meaningful phenotypic subgroups in autism spectrum disorders. Genes, Brain Behav.

614   2014;13:276–85.

615   41. Rynkiewicz A, Schuller B, Marchi E, Piana S, Camurri A, Lassalle A, et al. An investigation

616   of the "female camouflage effect" in autism using a computerized ADOS-2 and a test of

617   sex/gender differences. Mol Autism. 2016;7:10. doi:10.1186/s13229-016-0073-0.

618   42. Wen Y, Alshikho MJ, Herbert MR. Pathway network analyses for autism reveal multisystem

619   involvement, major overlaps with other diseases and convergence upon MAPK and calcium

620   signaling. PLoS One. 2016;11:1–23.

621   43. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de

622   novo mutations implicate novel genes underlying simplex autism risk. Nat Commun. 2014;5.

623   44. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution

624   of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–21.

625   45. Kawabe H, Brose N. The role of ubiquitylation in nerve cell development. Nature Reviews

626   Neuroscience. 2011;12:251–68.

627   46. Nava C, Lamari F, Héron D, Mignot C, Rastetter A, Keren B, et al. Analysis of the

628   chromosome X exome in patients with autism spectrum disorders identified novel candidate

629   genes, including TMLHE. Transl Psychiatry. 2012;2:e179. doi:10.1038/tp.2012.102.

630   **Figure legends**

631   Figure 1: Integrative systems medicine approach to identify complex genotype-phenotype

632   associations. Clinical and genetic data from the Autism Genome Project (AGP) was used in this

633   study (A) Clinical data analysis processing: clinical data comprises reports of ASD diagnosis and

634   neurodevelopmental assessment instruments. Agglomerative Hierarchical Clustering (AHC) was

635   used to identify clinically similar subgroups of individuals in stable, validated clusters, defined

636   by multiple clinical measures. (B) CNV data processing: rare high confidence CNVs previously

637   identified by the AGP, targeting brain-expressed genes, were retained for analysis. CNV data

638   was merged with clinical data from clustered ASD subjects for a final list of CNVs targeting

639   brain genes. (C) Functional annotation analysis: Biological processes defined by brain-expressed

640   genes targeted by CNVs were obtained using g:Profiler. (D) Classifier design: A Naive Bayes

641   machine learning classifier was trained and tested on patient's data, to predict the phenotypic

642   clustering of patients from biological processed disrupted by rare CNVs targeting brain-

643   expressed genes.

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
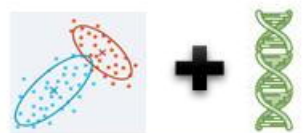
Clusters stability

Stable and validated multidimensional clusters

B. CNVs data processing and functional annotation pipeline

High confidence and rare CNVs targeting brain genes
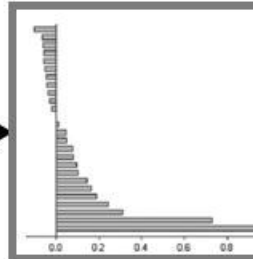
Merging clinical and CNV information

Functional annotation analysis
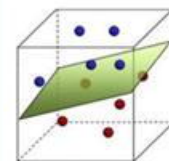
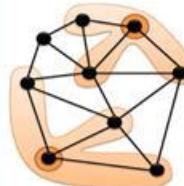Disrupted biological processes

C. Feature importance analysis

D. Machine learning pipeline to predict clinical outcome

Classifier training and testing

Genotype-phenotype associations

Prediction of ASD multidimensional phenotype