A novel data-driven approach reveals gene networks and biological processes underlying autism

Leonardo Emberti Gialloreti (leonardo.emberti.gialloreti@uniroma2.it)
Universita degli Studi di Roma Tor Vergata https://orcid.org/0000-0002-3575-1192
Roberto Enea
IMME Research Centre
Valentina Di Micco
University of Rome Tor Vergata
Daniele Di Giovanni
University of Rome Tor Vergata
Paolo Curatolo
University of Rome Tor Vergata

Methodology

Keywords: Autism spectrum disorder (ASD), cluster analysis, gene networks, patient similarity analytics, genome sequencing, neurite morphogenesis, cell adhesion assembly, synapse assembly, connectivity

Posted Date: May 28th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-31108/v1

Abstract

Background:

Developments in gene-hunting techniques identified several ASD associated genes. The considerable significance of cluster analysis associated with gene network studies has led to reveal many disrupted key pathways in ASD, even if its genetic underpinnings remain a challenging task. This study aims to determine, through a novel data-driven approach, how networks of mutated genes impact biological processes underlying autism.

Methods:

We analyzed the VariCarta dataset, which presents more than 200,000 genomic variant events collected from 13,069 people with ASD. Firstly, we created a whole-genome and an exome sequencing subset. Then, for each subset we compared pairwise patients of each group to build “patient similarity matrices”. Hierarchical-agglomerative-clustering and heatmap were performed to identify clusters of patients with common occurrences of gene networks within these matrices. The subsequent enrichment analysis (EA) highlighted biological processes that might be impacted by the mutated genes of each subgroup.

Results:

Considering the whole-genome matrix, we identified three main genetic clusters of ASD patients, each one characterized by a network of shared genetic variants. We isolated 11,609 genetic variants shared by at least two subjects in each cluster; 4,187 of these variants (36.1%) were common to the three clusters. Only 331 patients (2.5%) shared none or very few mutated genes with anyone else. The EA highlighted common or cluster-specific biological processes related to the variants. Most of the common abnormal processes were involved in neuron projections guidance and morphogenesis, cell junctions and synapse assembly. Exome sequencing alone was not effectual in identifying ASD subgroups.

Limitations:

Caution is warranted when interpreting our results, as we did not compare them with a control group and did not verify if the identified subgroups where actually associated with different phenotypes. Future work will have to ascertain the strength and reproducibility of these results.

Conclusions:

Itemizing not just single mutated genes, but also gene networks and specific biological processes that characterize different ASD subpopulations might allow to better understand which networks of genetic variants play a major role in the etiopathology of ASD. The proposed methodology may represent a novel

approach to help disentangle ASD complexity and an instrument to boost more focused genotype-phenotype studies.

Background

Autism spectrum disorder (ASD) is a heterogeneous group of neurodevelopmental disorders characterized by impaired social communication, repetitive behaviors and restrictive interests [1]. Genetic [2] [3] and epigenetic [4] factors have been identified as leading actors in ASD pathophysiology, as twin studies confirm. A meta-analysis on 6,413 twins including affected twins showed a heritability in families with an autistic patient of 64–91% [5]. Autism is the final outcome of a complex genetic architecture [3]. Thousands of genes may contribute to this disorder [6], indicating a heterogeneous etiology for ASD [7].

During the last decades, developments in gene-hunting techniques identified several ASD associated genes, including genes that code for proteins involved in synaptic functions [6] [8]. Such new technologies, including exome-wide and genome-wide interrogation, are considered an effective methodology to detect links between a common variant located in a specific DNA region and the risk of developing ASD [2].

In order to better assess the strength of evidence associated with ASD candidate genes [8], large systematic databases are needed. There are already several databases collecting ASD variant data as, for example, SFARI [9], AutismKB [10] and the one developed within the Autism Speaks' MSSNG project [11].

Notwithstanding these efforts, no major causative gene has been isolated so far and, to this day, genetic alterations can be identified only in 20 to 25% of ASD cases [12] [13]. Some advancements in the detection of rare genetic variants have been made by studying genetic syndromes related to ASD [7], which have shed a new light on the pathophysiology of the disorder [14]. By studying the syndromic ASD, some evidence is emerging that even heterogeneous ASD phenotypes might possibly present with a convergent pathophysiology [6].

Considering this complex contest, network-based analyses have been suggested as possible powerful tools to discover interaction patterns in ASD pathophysiology and provide a functional explanation to genetic heterogeneity and non-overlapping genes in ASD [15] [16]. Several authors have investigated possible methods to produce network–based prioritization of ASD genes [15]. In particular, machine learning algorithms have been employed in order to delineate the ASD architecture [17] [18] [19]. Also cluster analysis has seen increasing applications in biomedicine to further the understanding of ASD [20]. However, even though advances in technologies and widespread databases are continuing to disentangle the genetic aspects of ASD, the pathogenesis of the disorder is still a matter of speculation.

Moving from this background, the present study aimed to look at gene-networks rather than at specific gene variations and to identify and categorize those biological processes that might act as a pathophysiological substrate for ASD. Hence, instead of focusing on single genes or DNA segments, we prioritized the disrupted processes that occur in possibly genetically related clusters of patients. The availability of VariCarta [21], an ASD specific database, allowed us to execute our analysis on a significant amount of patients.

We firstly defined a metric to measure the genetic similarity between patients according to their mutations and then we applied hierarchical clustering in order to identify groups of genetically similar individuals [22]. Afterwards, we proceeded with the enrichment analysis upon each cluster of patients, as to identify ASD-related biological pathways, which might have been disrupted by the mutations. Finally, we discussed the implications of autism gene networks knowledge for clinical practice.

Methods

Gene Database

For this research work we used the VariCarta dataset from British Columbia University, a web-based database of human DNA genetic variants identified in individuals with an ASD diagnosis [21]. It also presents also a list of genes both with a set of mutational events and the reference to the patient affected by the mutation. This information was fundamental for the cluster analysis we carried out.

VariCarta was developed with the aim to identify rare, possibly causative genomic variants in ASD individuals. To address this challenge, it is necessary to collect a large number of subject information also through the aggregation of data, with the risk of methodological inconsistencies and subject overlap across studies. VariCarta developers tackled this demanding task by collecting and cataloging literature-derived genomic variants found in ASD subjects, through the use of an ongoing semi-manual curation and with a robust data import pipeline. Thus, while developing the database, it was possible to find and correct errors, to convert variants into a standardized format, to identify and harmonize cohort overlaps and to document data origin. The database is constantly updated with new relevant gene-targeted ASD research papers. The current version contains 184,212 variant events from 13,069 subjects, collected across 69 publications. The version used in the present paper is the one dated 12/11/2019. It consists of 211,669 records, each one containing a mutative event, as reported in the paper from where it was retrieved. Since a single mutative event can be reported by more than one paper, we removed duplicated events during the analysis.

Analysis of the Dataset

The dataset was accessible both using a web interface or by downloading the whole dataset in csv format. As the web interface allows limited research, we downloaded the whole dataset in csv format. Each row of the dataset is a variant event including, among others, the symbol of the affected gene, the category of mutation (synonymous SNV and nonsynonymous SNV, frameshift insertion, etc.), the adopted sequencing type (whole genome sequencing, exome sequencing, targeted sequencing) and the subject id that is a unique identifier of the patient presenting the variant. The dataset is also provided with reference information allowing to trace the paper where the information has been gathered from. Since the number of variants detected in each patient might be affected by the used sequencing type, we handled whole-genome and exome sequencing separately. In the current study we did not consider targeted analysis sequencing, because this technique is focused on identifying specific genes highly related to a disease, assuming that these are known. Such an assumption can be possibly made only for well documented causes of ASD, such as tuberous sclerosis or Fragile X syndrome, diagnosed in about 15% of individuals with ASD [23]. This is not the case for the “idiopathic ASD”, which represents the majority of all ASD diagnoses. Therefore, limiting the

analysis to some genes, while ignoring the others, could lead to a limited detection of gene mutations in a single subject. Furthermore, in the database the number of mutative events revealed by targeted sequencing is only about 2% of all events (3,698/184,212 mutations).

From the remaining variant events we created two subsets, one for whole-genome sequencing and one for exome sequencing; they were composed, respectively, by 84.6% (155,799/184,212 variants) and 13.4% (24,715/184,212) of all mutative events collected in the dataset. For each subset we selected the two features “Gene Symbol” and “Subject id” and made a pairwise comparison of the patients of each group to build a “patient similarity matrix” defined as follows:

Let A be a matrix NxN where N is the number of patients; let G_i be the set of genes affected by a mutation in the subject i and G_j the set of genes affected by a mutation in the subject j; we defined each element $a_{i,j}$ of A as the intersection between G_i and G_j.. A Log2 transformation has been applied to each element ai,j of the matrix A in order to normalize the results.

The whole-genome matrix numbered 2,062 patients and the exome matrix 7,427. In order to sort rows and columns to highlight clusters underlying the common occurrence of genes’ networks in patients, we used the clustermap function of the Python library Seaborn [24]. This function leverages hierarchical clustering [25] and heatmaps [26] to identify clusters inside the rows and columns of the input matrix that can be either a rectangular observation matrix or a square distance matrix. Hierarchical clustering is a widely used clustering algorithm that is able to identify hierarchical relations between groups [27]. It is particularly effective when some hierarchical structure (like a taxonomy) is expected to be identified and when the number and nature of groups and subgroups are not known in advance. Hierarchical clustering has been used in biology [28] since the 70 s. Nowadays it is applied to genetics, combined with heatmaps for microarray analysis [29], and, recently, to psychiatry to identify subgroups of patients with ASD based on comorbidity [30] or phenotype analysis [31] [32].

Hierarchical Agglomerative Clustering (HAC)

The specific algorithm we used is the Hierarchical Agglomerative Clustering (HAC) [33], a bottom up iterative process. It tries to progressively identify sets of similar subsets of data leveraging a distance function, also called hierarchical linkage. In the HAC process items are iteratively aggregated using the distance function to evaluate similarities between subgroups.

The HAC we used during the analysis was the Python implementation provided by the Seaborn library in clustermap function. Seaborn’s clustermap combines HAC to sort rows and columns of its heatmaps and shows dendgrograms on the axes to highlight the hierarchical structure. The implementation of HAC included in clustermap begins with a forest of clusters that have yet to be used in the hierarchy being formed. When two clusters B and C from the forest are combined into a single BC cluster, B and C are removed from the forest, and BC is added to the forest. When only one cluster remains in the forest, the algorithm stops, and this cluster becomes the root. A distance matrix is maintained at each iteration. The d[i,j] entry corresponds to the distance between cluster i and j in the original forest. At each iteration, the algorithm has to update the distance matrix to reflect the distance of the newly formed cluster BC with the remaining clusters in the forest.

The computation of the distance d[i,j] depends on the method used. In that event we adopted the “complete” method that applies the following function:

$$d(u, v) = \max(dist(u[i], v[j]))$$

This function, also called Farthest Point Algorithm or Voor Hees Algorithm, implies that the distance between two clusters u and v is the maximum distance between the farthest points. We applied this method in order to try to maximize the differences between clusters. The function used to measure the distance between the points is the Euclidean distance (2-norm). We then used the partitions to try to identify genes networks that could be characteristic of each subgroup.

Enrichment Analysis

In order to identify biological processes that could be impacted by mutated genes of each subgroup we used enrichment analysis. Gene Set Enrichment Analysis [34] (GSEA) (also called functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may be associated to disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. This can be done by comparing the input gene set to each of the bins (terms) in the gene ontology. A statistical test can be performed for each bin to evaluate if it is enriched for the input genes. Results for each pathway are expressed in terms of Fold Enrichment (FE), ie the ratio between the number of genes present in the cluster list belonging to that pathway and the number of genes expected to belong to that pathway in a random set of genes of the same size. The setting used for the enrichment analysis was the GO Ontology database (Released 2020-02-21) [35]. The applied reference list of expected genes was the one of homo sapiens.

Statistical Analysis

PantherDB [36] has been applied for the Enrichment analysis (PANTHER Overrepresentation Test; Released 07/04/2020). To verify the statistical significance of the submitted set of genes we applied the Chi-square/Fisher’s Exact Test. The obtained raw p-values were adjusted for multiple comparisons by means of the False Discovery Rate method (FDR) [37]. A statistical significance threshold of $p < 0.005$ (two-tailed) was applied for all analyses. As both raw and FDR-adjusted p-values are strongly dependent on sample size, once the statistically significant terms were identified, we ranked the biological processes by Fold Enrichment, a measure of the effect-size [38]. In order to highlight interactions between the set of identified genes we used String DB [39] Gene Network View. In the produced network interaction graph, the number of edges represents the existing interactions between the submitted genes. It was compared with the expected number of edges, ie the expected interactions between a set of random genes of the same size. STRING DB was used also to compute protein-protein interaction (PPI) enrichment p-values on partitions identified by the intersection of genes sets identified by the clustering.

Results

Cluster heatmap for whole-genome sequencing

The cluster heatmap of the dissimilarity matrix of the patients whose DNA has been analyzed using whole-genome sequencing is presented in Fig. 1a. All 2,062 patients belonging to the group are distributed along the two axes. Each number is the subject identification (id) assigned to a specific patient as it is presented in the VariCarta database. Each cell of the matrix is the log2 transformation of the number of the common genes mutated between two patients. The usual diagonal line that should show the comparison of each patient with himself was suppressed by the clustermap function.

Further analyses were performed on the macro-clusters represented in Fig. 1b. The dendrogram showing their relationships is extracted from the one visible in the cluster heatmap axes limiting it to the third hierarchy level.

The whole-genome matrix presented a high number of shared variations between patients. Three clusters have been identified (A, B, and C). We labeled the three clusters according to their density - ie the amount of shared variations between the patients - from the highest to the lowest. We considered as a density factor (DF) the average value of all the elements of a cluster. Each value of the matrix is the log2 transformation of the number of mutated genes in common between the two patients representing, respectively, the row and the column of the matrix. Cluster A is composed by 574 patients sharing 8,357 mutated genes (DF: 1.429); Cluster B is composed by 507 patients sharing 6,818 mutated genes (DF: 1.014); Cluster C is composed by 650 patients sharing 7,704 mutated genes (DF: 0.720). Beyond the three clusters, we also identified 331 patients who shared none or very few mutated genes (all together 41 genes). We included them in the “mixed group” D, whose DF was just 0.001.

Cluster heatmap for exome sequencing

The same similarity matrix computed only on exome data is presented in Fig. 2. In this case the similarity matrix is almost empty, meaning that most of the patients do not share any mutations or few mutations between each other. Only a single dense cluster (Cluster E) is detectable in the top left corner of the matrix. This cluster is composed of 218 patients, which is 2.9% of all the patients whose DNA has been analyzed using exome sequencing (218/7,427). The density of this identified cluster is very high (DF: 2.551).

Enrichment analysis for whole-genome sequencing

For each of the clusters shown in Fig. 1b we extracted the list of common mutated genes - genes mutated in at least two patients of the cluster - and then we executed the enrichment analysis. The results of the 40 biological processes with the highest FE values are presented in Table 1. Only clusters A, B, and C have been considered, as the mixed group D did not return any statistically significant result. The involved biological processes are ranked by Fold Enrichment to highlight the most specific processes involved.

Table 1

Enrichment Analysis on Clusters A, B and C ranked by Fold Enrichment. Whole-genome sequencing data.

Cluster A			Cluster B			Cluster C		
GO biological process	**FE**	**FDR**	**GO biological process**	**FE**	**FDR**	**GO biological process**	**FE**	**FDR**
neuron projection guidance (GO:0097485)	1.75	6.88E-06	dendrite morphogenesis (GO:0048813)	2.38	3.03E-03	neuron recognition (GO:0008038)	2.46	2.19E-03
axon guidance (GO:0007411)	1.74	9.84E-06	neuron projection guidance (GO:0097485)	2.07	7.85E-10	ventricular septum development (GO:0003281)	2.18	2.85E-03
regulation of axonogenesis (GO:0050770)	1.71	9.35E-04	axon guidance (GO:0007411)	2.07	1.29E-09	cardiac septum development (GO:0003279)	2.01	5.15E-04
axonogenesis (GO:0007409)	1.68	3.30E-07	synapse assembly (GO:0007416)	2.07	1.28E-03	negative regulation of developmental growth (GO:0048640)	1.89	4.10E-03
neuron projection morphogenesis (GO:0048812)	1.68	2.13E-09	action potential (GO:0001508)	2.06	2.60E-03	cell-cell junction assembly (GO:0007043)	1.87	3.43E-03
plasma membrane bounded cell projection morphogenesis (GO:0120039)	1.68	1.95E-09	developmental growth involved in morphogenesis (GO:0060560)	2.05	4.71E-04	neuron projection morphogenesis (GO:0048812)	1.86	1.45E-13
regulation of JNK cascade (GO:0046328)	1.68	2.23E-03	regulation of synapse assembly (GO:0051963)	2.02	1.89E-03	cell morphogenesis involved in neuron differentiation (GO:0048667)	1.85	5.59E-12
cell morphogenesis involved in neuron differentiation (GO:0048667)	1.68	3.49E-08	cell morphogenesis involved in neuron differentiation (GO:0048667)	2.00	1.31E-14	neuron projection guidance (GO:0097485)	1.84	5.84E-07

FE: Fold Enrichment; FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes of each cluster are shown. An additional table file shows the full list of the 58 biological processes of Cluster A, the 135 processes of Cluster B, and the 87 processes of Cluster C with a FE ≥ 1.5 [Additional file 1]. The [illegible]al file 2.

Cluster A			Cluster B			Cluster C		
cell projection morphogenesis (GO:0048858)	1.67	2.25E-09	axonogenesis (GO:0007409)	2.00	2.11E-12	plasma membrane bounded cell projection morphogenesis (GO:0120039)	1.84	3.30E-13
regulation of cell junction assembly (GO:1901888)	1.67	1.34E-03	negative regulation of cell morphogenesis involved in differentiation (GO:0010771)	2.00	4.30E-03	axon guidance (GO:0007411)	1.84	6.70E-07
axon development (GO:0061564)	1.67	1.76E-07	neuron projection morphogenesis (GO:0048812)	2.00	3.64E-16	axonogenesis (GO:0007409)	1.84	6.78E-10
cell part morphogenesis (GO:0032990)	1.66	2.93E-09	plasma membrane bounded cell projection morphogenesis (GO:0120039)	1.99	4.61E-16	cell projection morphogenesis (GO:0048858)	1.83	4.04E-13
renal system development (GO:0072001)	1.62	1.87E-04	regulation of cell junction assembly (GO:1901888)	1.99	2.52E-06	axon development (GO:0061564)	1.82	1.99E-10
telencephalon development (GO:0021537)	1.62	4.40E-04	cell projection morphogenesis (GO:0048858)	1.97	1.10E-15	regulation of axonogenesis (GO:0050770)	1.82	1.27E-04
kidney development (GO:0001822)	1.62	3.42E-04	cell junction assembly (GO:0034329)	1.95	7.79E-08	cell part morphogenesis (GO:0032990)	1.80	8.91E-13
regulation of small GTPase mediated signal transduction (GO:0051056)	1.61	6.55E-05	cell-cell junction assembly (GO:0007043)	1.93	3.16E-03	regulation of small GTPase mediated signal transduction (GO:0051056)	1.78	3.01E-07
regulation of neuron projection development (GO:0010975)	1.61	5.54E-08	synapse organization (GO:0050808)	1.93	2.32E-08	cell junction assembly (GO:0034329)	1.76	8.13E-06

FE: Fold Enrichment; FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes of each cluster are shown. An additional table file shows the full list of the 58 biological processes of Cluster A, the 135 processes of Cluster B, and the 87 processes of Cluster C with a FE ≥ 1.5 [Additional file 1]. The [illegible] al file 2.

Cluster A			Cluster B			Cluster C		
regulation of cell morphogenesis involved in differentiation (GO:0010769)	1.60	1.26E-04	axon development (GO:0061564)	1.92	6.41E-12	cell-cell junction organization (GO:0045216)	1.76	1.54E-03
positive regulation of neuron differentiation (GO:0045666)	1.60	1.03E-05	cell part morphogenesis (GO:0032990)	1.92	5.66E-15	neuron projection development (GO:0031175)	1.75	8.64E-15
cell morphogenesis involved in differentiation (GO:0000904)	1.60	1.97E-08	cell morphogenesis involved in differentiation (GO:0000904)	1.90	1.35E-15	regulation of cell junction assembly (GO:1901888)	1.73	5.78E-04

FE: Fold Enrichment; FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes of each cluster are shown. An additional table file shows the full list of the 58 biological processes of Cluster A, the 135 processes of Cluster B, and the 87 processes of Cluster C with a FE ≥ 1.5 [Additional file 1]. The occurrences of mutated genes are listed in the Additional file 2.

Cluster comparisons

Most of the processes of the three clusters are common, even though some differences can be detected either because of the different order in the list that indicates a different weight of that process in the cluster or because of the presence of cluster-specific processes. Therefore, for each cluster we extracted mutations shared by at least two subjects in the same cluster. As shown in Fig. 3, by intersecting the three clusters we identified seven partitions. We executed enrichment analysis on all the partitions separately. None of the partitions, except the intersection between the three clusters returned significant results. Table 2 shows the enrichment analysis on the ABC intersection.

Table 2

Enrichment Analysis on the intersection among clusters A, B, and C. Whole-genome sequencing data.

GO biological process	FE	FDR
cell-cell adhesion mediated by cadherin (GO:0044331)	3.85	1.74E-03
outflow tract septum morphogenesis (GO:0003148)	3.62	4.27E-03
synaptic transmission, glutamatergic (GO:0035249)	3.36	8.03E-04
neuron recognition (GO:0008038)	3.30	1.50E-04
glutamate receptor signaling pathway (GO:0007215)	3.18	4.96E-04
dendrite morphogenesis (GO:0048813)	3.05	1.98E-04
receptor localization to synapse (GO:0097120)	3.05	2.83E-03
heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules (GO:0007157)	2.99	2.58E-03
regulation of cell-substrate junction organization (GO:0150116)	2.87	2.72E-04
synapse assembly (GO:0007416)	2.84	2.45E-06
heart valve morphogenesis (GO:0003179)	2.82	2.78E-03
regulation of focal adhesion assembly (GO:0051893)	2.81	6.54E-04
regulation of cell-substrate junction assembly (GO:0090109)	2.81	6.52E-04
retina morphogenesis in camera-type eye (GO:0060042)	2.79	1.71E-03
neuron projection guidance (GO:0097485)	2.75	2.15E-15
axon guidance (GO:0007411)	2.73	5.03E-15
negative regulation of axonogenesis (GO:0050771)	2.67	8.24E-04

FE: Fold Enrichment (FE); FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes related to the intersection among clusters A, B, and C are reported. An additional table file shows the full list of the 359 biological processes with a FE ≥ 1.5 [Additional file 1]. The occurrences of mutated genes are

GO biological process	FE	FDR
adherens junction organization (GO:0034332)	2.66	2.70E-03
regulation of glutamate receptor signaling pathway (GO:1900449)	2.64	1.56E-03
cardiac septum morphogenesis (GO:0060411)	2.64	6.87E-04
negative regulation of cell morphogenesis involved in differentiation (GO:0010771)	2.59	6.52E-05
positive regulation of synapse assembly (GO:0051965)	2.59	2.49E-03
protein localization to synapse (GO:0035418)	2.56	4.75E-03
regulation of neurotransmitter receptor activity (GO:0099601)	2.52	1.13E-03
mechanoreceptor differentiation (GO:0042490)	2.52	4.82E-03
regulation of synaptic transmission, glutamatergic (GO:0051966)	2.50	3.75E-03
cell morphogenesis involved in neuron differentiation (GO:0048667)	2.48	8.42E-20
neuron projection morphogenesis (GO:0048812)	2.47	1.37E-21
axonogenesis (GO:0007409)	2.47	1.72E-16
regulation of synapse assembly (GO:0051963)	2.46	1.41E-04
plasma membrane bounded cell projection morphogenesis (GO:0120039)	2.45	2.14E-21
cell projection morphogenesis (GO:0048858)	2.43	4.51E-21
positive regulation of phosphatidylinositol 3-kinase signaling (GO:0014068)	2.43	1.35E-03
dendrite development (GO:0016358)	2.42	1.01E-04

FE: Fold Enrichment (FE); FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes related to the intersection among clusters A, B, and C are reported. An additional table file shows the full list of the 359 biological processes with a FE ≥ 1.5 [Additional file 1]. The occurrences of mutated genes are

GO biological process	FE	FDR
neural crest cell differentiation (GO:0014033)	2.42	1.85E-03
regulation of sodium ion transport (GO:0002028)	2.39	2.01E-03
axon development (GO:0061564)	2.38	2.47E-16
regulation of potassium ion transmembrane transport (GO:1901379)	2.38	3.48E-03
cell part morphogenesis (GO:0032990)	2.38	1.50E-20
regulation of potassium ion transport (GO:0043266)	2.37	6.70E-04

FE: Fold Enrichment (FE); FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes related to the intersection among clusters A, B, and C are reported. An additional table file shows the full list of the 359 biological processes with a FE ≥ 1.5 [Additional file 1]. The occurrences of mutated genes are listed in the Additional file 2.

The resulting network interaction graph, generated by String-DB, is presented in Fig. 4. For the sake of clarity, the analysis has been limited to genes, which were present in at least 50 patients. The number of nodes of the network was 316 and the number of edges 990. The expected number of edges was 293: 3.4 times less than the number of edges of the submitted network.

Enrichment analysis for exome sequencing

For the cluster shown in Fig. 2 we extracted the list of genes mutated in at least two patients and then we executed the enrichment analysis. The results of the biological processes with the highest FE values are presented in Table 3. The involved biological processes are ranked by Fold Enrichment.

Table 3
Enrichment Analysis on Cluster E. Exome sequencing data.

GO biological process	FE	FDR
actin filament capping (GO:0051693)	9.41	2.58E-03
negative regulation of actin filament depolymerization (GO:0030835)	8.63	3.98E-03
neuromuscular junction development (GO:0007528)	7.89	2.20E-03
regulation of actin filament depolymerization (GO:0030834)	6.96	4.32E-03
dendrite morphogenesis (GO:0048813)	6.65	2.22E-03
negative regulation of protein depolymerization (GO:1901880)	6.05	3.94E-03
dendrite development (GO:0016358)	5.35	1.88E-04
cell junction assembly (GO:0034329)	4.65	4.37E-08
negative regulation of supramolecular fiber organization (GO:1902904)	4.25	1.50E-03
multicellular organismal signaling (GO:0035637)	4.24	2.89E-03
protein localization to plasma membrane (GO:0072659)	4.23	1.49E-04
cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)	4.16	3.98E-06
synapse organization (GO:0050808)	4.09	8.30E-07
negative regulation of cytoskeleton organization (GO:0051494)	4.06	2.32E-03
cell junction organization (GO:0034330)	3.99	6.33E-11
positive regulation of cell morphogenesis involved in differentiation (GO:0010770)	3.88	3.41E-03
regeneration (GO:0031099)	3.83	3.85E-03
positive regulation of neuron projection development (GO:0010976)	3.79	1.03E-05
homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156)	3.76	4.46E-03
positive regulation of cell projection organization (GO:0031346)	3.72	1.75E-07
protein localization to cell periphery (GO:1990778)	3.63	4.79E-04
cell part morphogenesis (GO:0032990)	3.57	3.50E-09
actin filament organization (GO:0007015)	3.55	3.84E-04
neuron projection morphogenesis (GO:0048812)	3.52	2.82E-08
neuron projection guidance (GO:0097485)	3.49	2.97E-04

FE: Fold Enrichment (FE); FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes of cluster E are reported. An additional table file shows the full list of the 137 biological processes with a FE [illegible] are listed in the Additional file 2.

GO biological process	FE	FDR
plasma membrane bounded cell projection morphogenesis (GO:0120039)	3.49	3.31E-08
cell projection morphogenesis (GO:0048858)	3.46	4.01E-08
neuron projection development (GO:0031175)	3.39	1.10E-10
cell-cell adhesion (GO:0098609)	3.34	1.38E-07
epithelial tube morphogenesis (GO:0060562)	3.29	2.51E-04
actin cytoskeleton organization (GO:0030036)	3.20	6.36E-07
axon guidance (GO:0007411)	3.19	2.13E-03
actin filament-based process (GO:0030029)	3.18	7.31E-08
positive regulation of neuron differentiation (GO:0045666)	3.14	8.72E-05
regulation of cell morphogenesis involved in differentiation (GO:0010769)	3.12	7.73E-04
cell morphogenesis involved in neuron differentiation (GO:0048667)	3.10	1.68E-05
neuron development (GO:0048666)	3.07	2.80E-10
positive regulation of nervous system development (GO:0051962)	3.04	1.26E-06
positive regulation of neurogenesis (GO:0050769)	3.01	1.33E-05
cell morphogenesis (GO:0000902)	3.00	1.78E-08

FE: Fold Enrichment (FE); FDR: False Discovery Rate p-value. Biological processes are identified by their reference numbers (GO:XXXXXX) in the Gene Ontology. The first 40 FE ranked biological processes of cluster E are reported. An additional table file shows the full list of the 137 biological processes with a FE ≥ 1.5 [Additional file 1]. The occurrences of mutated genes are listed in the Additional file 2.

The full list of biological processes with a FE ≥ 1.5, as well as the number of reference genes, the expected and observed numbers of genes, and the raw p-values for each biological process of the clusters and of the intersection are provided in the Additional file 1, while the list of occurrences of mutated genes can be retrieved from the Additional file 2.

Discussion

We used the VariCarta dataset to apply a metric to measure the genetic similarity between patients, followed by a hierarchical clustering analysis. We identified three main genetic clusters of ASD patients, each one characterized by a set of common mutated genes. The subsequent enrichment analysis (EA), performed upon the clusters' genes, allowed us to pinpoint disrupted biological processes both common to the three clusters and cluster-specific.

Most of the processes that were common to the three clusters were involved in neuron projections guidance and morphogenesis. Proper plasticity of axon and dendrites is a highly dynamic process leading functional [redacted] ages of brain development [41] [42]. These

processes involve numerous ASD related genes associated with disrupted synaptic connectivity and function [43] [44]. Neuronal migration and morphogenesis defects in ASD contribute to an altered cortical connectivity in different brain regions [45] [46], as the well-known prefrontal area [47]. Therefore, in the developing brain, a premature alteration in neuronal plasticity, affecting mostly cortical regions involved in cognitive and behavioral functions, might trigger the autistic phenotype [48] [49]. Furthermore, also fMRI [50] and electrophysiological [51] studies have emphasized the role of an atypical interconnection of specific brain areas in ASD.

Considering these findings, the biological processes we underlined for each cluster give further support to the theory that neurogenesis and its molecular microenvironment are associated with autism. This pathogenetic background was pointed out also by the enrichment analysis performed on the intersection of the three clusters, which confirmed the role in the ASD etiopathology of atypical gene networks related to biological processes affecting neuron development as, for example, “axon guidance”, “neuron projection guidance”, or “dendrite morphogenesis”.

As further evidence of a disrupted connectivity in ASD, the EA on the single clusters and on their intersection showed the importance of processes like “cell junction assembly” and “synapse assembly”. The integrity of synaptic proteins and cell adhesion molecules is crucial for synaptic formation, signal transduction and transmission [52]. Thus, synaptic dysfunction due to molecular damage and the consequently altered signal conduction are consistent with ASD etiopathology [53]. Overall, the presence of shared genetic networks between the three clusters and, consequently, of common atypical biological processes involving both neurite and synaptic formation and function might be the underlying cause of comparable ASD phenotypes among different individuals.

Besides the shared genetic networks, the enrichment analysis signaled also some processes that differentiate the three clusters and that could possibly bring to phenotypic dissimilarities. Actually, the analysis brought to light the presence of cluster-specific processes, which were not shared with the other clusters; this is, for example, the case of the biological process “kidney development” in Cluster A, or “cardiac septum development” in Cluster C. This observation should be considered with caution, as the relative FDR p-values - though statistically significant - are close to the threshold of < 0.005. Nevertheless, it is noteworthy that even if ASD-related processes regarding heart and kidney are still poorly characterized in literature, more recently they have become a feature of scientific and clinical interest [54] [55].

Our findings might prompt to speculate that shared mutations could underlie the core ASD-symptoms, while the cluster-specific ones may be one of the possible explanations for the ASD phenotype heterogeneity and autism continuum. Therefore, both cluster-specific and common atypical biological processes might interact to shape the specific phenotype of each individual. At any rate, these observations point to the fact that when considering the disrupted processes of the different ASD phenotypes we should not only look at the CNS but at other systems as well. Thus, this is an important evidence supporting the polygenic model that assumes that ASD is the result of rare and common variants combination [56].

It is noteworthy to mention that we also identified a group of 331 people composed of individual patients that have either no or just a few mutated genes in common with the rest of the sample. They could not be

included in any cluster. These findings further support an ASD heterogeneity, as it may indicate that the clinical phenotype might be also the outcome of genetic common variants that could interact with a main mutation and/or with epigenetic [56] or environmental factors [57].

We also performed an analysis of VariCarta exome data, collected from more than 7,400 patients, in order to evaluate if there is a difference in terms of information provided by whole-genome compared to exome sequencing. Even though a single dense cluster - involving less than 3% of the patients - has been identified, data resulting from exome sequencing seems not to be sufficient to identify significant similarities between ASD patients. The comparison of the results we obtained by analyzing the whole-genome with those of the exome draw the attention to the major contribution of noncoding mutations to the development of autistic traits; a finding in line with previous studies [58].

A relevant element of our work was the use of patient similarity analytics, followed by a hierarchical clustering analysis. In this way, we identified the genetic clusters and the related biological processes. As already highlighted in scientific literature [59], the use of a similarity metrics between patients in order to identify “patient similarity networks” is becoming a significant aspect of big data analytics in healthcare systems. By assisting patients’ clustering, it allows researchers to detect homogeneous subgroups of individuals sharing similar characteristics. This data-driven approach has been already applied in medicine to the study of different diseases, including behavioral disorders [60] and other diseases of the CNS [61]. The main condition to develop patient similarity algorithms is the existence and availability of large datasets containing detailed information of each individual. VariCarta is an example of this kind.

Specifically, in ASD research and clinical medicine patient similarity algorithms could play a role in structuring autism heterogeneity, i.e., in identifying ASD patient subgroups who share the same etiopathology. Subgroups of patients can be then assessed by additional fine-grained stratifications, based - for example - on their genetic characteristics or biological processes. Once subsets of patients have been isolated and characterized, it becomes possible to perform systematic individualized analyses, by evaluating the distance of a single patient from each subgroup.

Enumerating both the common and specific processes that characterize different ASD subpopulations might allow understanding which one, among a large number of autism genetic variants, are those that play a major role in the etiopathology of ASD. This could help to bridge the crucial gap between the detection of new genetic risk variants for ASD [62] and the clinical translation of these discoveries.

Although patient similarity analytics is still in development, it promises to be helpful in predicting patients' trajectory over time [63], in providing clinical decision support [64], and in tailoring individual treatments [65]. Therefore, stratifying ASD patients into clinical subgroups through the identification of common and specific disrupted biological processes might be beneficial both for a more precise prognosis and for choosing tailored therapeutic approaches to ASD. This is because the identification of pathways underlying a specific subpopulation could give us the right information about the natural history of the disorder and, consequently, the specific clinical intervention for that subpopulation. Thus, this methodological approach could provide a fresh insight into precision medicine applied to ASD [65], as it can help guide clinical management and support the best therapeutic choice that fits a specific patient.

Limitations

Our findings should be viewed with some limitations in mind. Firstly, as the database we used does neither contain genetic data of family members of the ASD patients nor of neurotypical subjects we did not perform the same analysis on a group of non-affected individuals. Therefore, we did not compare our results with a control group. It will be important for future research to replicate our findings, comparing them with genetic data of neurotypical subjects. Secondly, the absence in the dataset of the description of each phenotype, including the IQ, did not allow us to confirm our clustering results. We could not verify if the biological processes impacted by the mutation networks of each cluster actually produced different phenotypes. A comparison of the identified gene networks and related biological processes with associated phenotypes will be necessary to confirm the clinical validity and usefulness of our results. Finally, we included in the analysis all mutations, without differentiating their type (base substitution, deletion, or insertions), category of nucleotide mutation, or sequence variation (exonic, intronic). However, at this stage, we preferred to focus on the identification of networks of mutated genes characterizing ASD subpopulations, rather than on the impact of mutations on a specific ASD patient.

Conclusions

To the best of our knowledge, this is the first time that clustering analysis has been applied to patient similarity in an ASD study. We deem that an important asset of the proposed methodology is determined by the fact that identifying ASD subpopulations using genetic sequencing, and not only phenotypic clustering, might support the diagnostic process by more precisely depicting and differentiating the specific autistic traits and phenotypic characteristics of each person.

To improve our understanding of ASD heterogeneity, neurobiology and genomic architecture, study designs have to increasingly consider innovative methodologies and newly developed biomedical informatics. Integrated genomic approaches, supported by advanced mathematical modeling, might lead to a better comprehension of the etiology and of the pathogenetic mechanisms of ASD. The proposed methodology might represent a novel approach to help disentangle ASD complexity and an instrument to foster more focused genotype-phenotype studies.

Declarations

Ethics approval and consent to participate

This study was conducted exclusively on anonymized data, and the study was approved by the Ethical Review Board of our institution. As the study was based on a free-available and anonymous dataset, informed consent was waived by the Board.

Consent for publication

Availability of data and materials

The VariCarta dataset is freely available at https://varicarta.msl.ubc.ca/index, both using a web interface or by downloading the whole dataset in csv format. All data generated or analyzed during the present study are included in this published article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Authors' contributions

LEG, RE, PC conceived and designed the study. LEG, RE conceptualized the statistical approach and analyzed the data. VDM, DDG performed the literature review and drafted the initial manuscript. LEG, RE, VDM, DDG, PC interpreted the results and edited the manuscript in its different stages. LEG, PC supervised and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge the Pavlidis Lab at the Michael Smith Laboratories at the University of British Columbia for their tremendous effort of building and maintaining the VariCarta web application and database. We are also thankful for allowing to freely use these data for academic purposes.

References

1. APA. Diagnostic and Statistical Manual of Mental Disorder (DSM-5®). Washington: American Psychiatric Association; 2013.

2. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D et al . The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-221.

3. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-1233.

4. Hamza M, Halayem S, Mrad R, Bourgou S, Charfi F, Belhadj A. Epigenetics' implication in autism spectrum disorders: A review. Encephale. 2017;43(4):374-381.

5. Tick B, Bolton P, Happé F, Rutter M, Rijsdijk F. Heritability of autism spectrum disorders: A meta-analysis of twin studies. J Child Psychol Psychiatry. 2016;57(5):585-595.

6. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-215.

7. Yuen RKC, Thiruvahindrapuram B, Merico D, Walker S, Tammimies K, Hoang N et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21(2):185-191.

8. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA et al. SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013;4(1):36.

9. Basu SN, Kollu R, Banerjee-Basu S. AutDB: A gene reference resource for autism research. Nucleic Acids Research. 2009;37(Database issue):D832-836.

10. Yang C, Li J, Wu Q, Yang X, Huang AY, Zhang J et al. AutismKB 2.0: A knowledge for the genetic evidence of autism spectrum disorder. Database (Oxford). 2018:bay106.

11. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-611.

12. Geschwind D, State M. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015;14(11):1109–1120.

13. Asif M, Martiniano HFMC, Marques AR, Santos JX, Vilela J, Rasga C et al. Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning. Translational Psychiatry. 2020;10(1):43.

14. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94(5):677-694.

15. Mosca E, Bersanelli M, Gnocchi M, Moscatelli M, Castellani G, Milanesi L et al. Network diffusion-based prioritization of autism risk genes identifies significantly connected gene modules. Front Genet. 2017;8:129.

16. Li J, Wang L, Guo H, Shi L, Zhang K, Tang M et al. Targeted sequencing and functional analysis reveal brain-size-related genes and their networks in autism spectrum disorders. Mol Psychiatry. 2017;22(9):1282-1290.

17. Duda M, Zhang H, Li H-D, Wall DP, Burmeister M, Guan Y. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. Transl Psychiatry. 2018;8(1):56.

18. Brueggeman L, Koomar T, Michaelson JJ. Forecasting risk gene discovery in autism with machine learning and genome-scale data. Sci Rep. 2020;10(1):4569.

19. Zhang Y, Chen Y, Hu T. PANDA: Prioritization of autism-genes using network-based deep-learning approach. Genet Epidemiol. 2020;44(4):382-394.

20. Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME et al. Gene expression profiling differentiates autism case controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. Autism Res. 2009;2(2):78-97.

21. Belmadani M, Jacobson M, Holmes N, Phan M, Nguyen T, Pavlidis P et al. VariCarta: A Comprehensive Database of Harmonized Genomic Variants Found in Autism Spectrum Disorder Sequencing Studies. Autism Res. 2019;12(12):1728-1736.

22. Brown SH. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. Front Physiol. 2016;7:561.

23. Casanova MF, Casanova EL, Frye RE, Baeza-Velasco C, LaSalle JM, Hagerman RJ et al. Editorial: Secondary vs. Idiopathic Autism. Frontiers in psychiatry. 2020;11:297.

24. Waskom M, Seaborn. [https://seaborn.pydata.org/index.html]. Accessed May 18, 2020.

25. Johnson SC. Hierarchical Clustering Schemes. Psychometrika. 1967;32(3):241-254.

26. Wilkinson L, Friendly M. The History of the Cluster Heat Map. The American Statistician. 2009;63(2):179-184.

27. Wallace M, Akrivas G, Stamou G. Automatic thematic categorization of documents using a fuzzy taxonomy and fuzzy hierarchical clustering. 2003. The 12th IEEE International Conference on Fuzzy Systems. FUZZ '03. St Louis. MO. USA.

28. Rohlf FJ. Adaptive Hierarchical Clustering Schemes. Systematic Zoology. 1970;19(1):58-82.

29. Seo J, Shneiderman B. Interactively Exploring Hierarchical Clustering Results. Computer. 2002;35(7):80-86.

30. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics. 2014;133(1):e53-e63.

31. Hu VW, Steinberg ME. Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. Autism Research. 2009;2(2):67-77.

32. Obafemi-Ajayi T, Lam D, Takahashi TN, Kanne S, Wunsch D. Sorting the Phenotypic Heterogeneity of Autism Spectrum Disorders: a Hierarchical Clustering Model. 2015. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Niagara Falls. ON. Canada.

33. Müllner D. Modern hierarchical, agglomerative clustering algorithms. Arvix Eprint. 2011;arXiv:1109.2378.

34. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545-50.

35. The G. O. Consortium. Gene Ontology. [http://geneontology.org]. Accessed May 18, 2020.

36. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acid Research. 2019;47(D1):D419-D426.

37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995;57(1):289-300.

38. Harrison PF, Pattison AD, Powell DR, Beilharz TH. Topconfects: a package for confident effect sizes in differential expression analysis provides a more biologically useful ranked gene list. Genome Biol. 2019;20(1):67.

39. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2019;47(D1):D607–D613.

40. Xu Q, Liu Y-Y, Wang X, Tan G-H, Li H-P, Hulbert SW et al. Autism-associated CHD8 deficiency impairs axon development and migration of cortical neurons. Mol. Autism. 2018;9:65.

41. Schafer ST, Paquola ACM, Stern S, Gosselin D, Ku M, Pena M et al. Pathological priming causes developmental gene network heterochronicity in autism patient-derived neurons. Nat. Neurosci. 2019;22(2):243–255.

42. Ciarrusta J, Dimitrova R, Batalle D, O'Muircheartaigh J, Cordero-Grande L, Price A et al. Emerging functional connectivity differences in newborn infants vulnerable to autism spectrum disorders. Translational Psychiatry. 2020;10(1):131.

43. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180(3):568-584.

44. García-Cabezas M, Barbas H, Zikopoulos B. Parallel Development of Chromatin Patterns, Neuron Morphology, and Connections: Potential for Disruption in Autism. Front Neuroanat. 2018;12:70.

45. Lai M, Lombardo M, Baron-Cohen S. Autism. Lancet. 2014;383(9920):896-910.

46. Hashimoto R, Nakazawa T, Tsurusaki Y, Yasuda Y, Nagayasu K, Matsumura K et al. Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. J Hum Genet. 2016;61(3):199-206.

47. Zikopoulos B, Barbas H. Changes in prefrontal axons may disrupt the network in autism. J Neurosci.

48. Bakos J, Bacova Z, Grant SG, Castejon AM, Ostatnikova D. Are Molecules Involved in Neuritogenesis and Axon Guidance Related to Autism Pathogenesis?. Neuromolecular Med. 2015;17(3):297-304.

49. Nishiyama J. Plasticity of dendritic spines: Molecular function and dysfunction in neurodevelopmental disorders. Psychiatry Clin Neurosci. 2019;73(9):541-550.

50. Gabrielsen TP, Anderson JS, Stephenson KG, Beck J, King JB, Kellems R et al. Functional MRI connectivity of children with autism and low verbal and cognitive performance. Mol Autism. 2018;9:67.

51. Mehdizadefar V, Ghassemi F, Fallah A. Brain Connectivity Reflected in Electroencephalogram Coherence in Individuals With Autism: A Meta-Analysis. Basic Clin Neurosci. 2019;10(5):409-417.

52. Wang J, Gong J, Li L, Chen Y, Liu L, Gu HT et al. Neurexin gene family variants as risk factors for autism spectrum disorder. Autism Res. 2018;11(1):37-43.

53. Baig D, Yanagawa T, Tabuchi K. Distortion of the normal function of synaptic cell adhesion molecules by genetic variants as a risk for autism spectrum disorder. Brain Res Bull. 2017;129:82-90.

54. Sigmon ER, Kelleman M, Susi A, Nylund CM, Oster ME. Congenital Heart Disease and Autism: A Case-Control Study. Pediatrics. 2019;144(5):e20184114.

55. Wang Y, Kou Y, Meng D. Network Structure Analysis Identifying Key Genes of Autism and its Mechanisms. Hindawi Computational and Mathematical Methods in Medicine. 2020. Volume 2020.

56. Wiśniowiecka-Kowalnik B, Nowakowska B. Genetics and epigenetics of autism spectrum disorder—current evidence in the field. Journal of Applied Genetics. 2019;60(1):37–47.

57. Emberti Gialloreti L, Mazzone L, Benvenuto A, Fasano A, Garcia Alcon A, Kraneveld A et al. Risk and Protective Environmental Factors Associated with Autism Spectrum Disorder: Evidence-Based Principles and Recommendations. J Clin Med. 2019;8(2):217.

58. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C et al. Whole-genome deep learning analysis identifies contribution of noncoding mutations to autism risk. Nat. Genet. 2019;51(6):973-980.

59. Brown S. Patient Similarity: Emerging Concepts in Systems and Precision Medicine. Front Physiol. 2016;7:561.

60. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput. Biol. 2011;7(8):e1002141.

61. Bolouri H, Zhao L, Holland E. Big data visualization identifies the multidimensional molecular landscape of human gliomas. Proc Natl Acad Sci U S A. 2016;113(19):5394-9.

62. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431-444.

63. Gallego B, Walter SR, O Day R, Dunn AG, Sivaraman V, Shah N et al. Bringing cohort studies to the bedside: framework for a ‘green button’ to support clinical decision-making. J. Comp. Eff. Res. 2015;4(3):191-197.

Figures

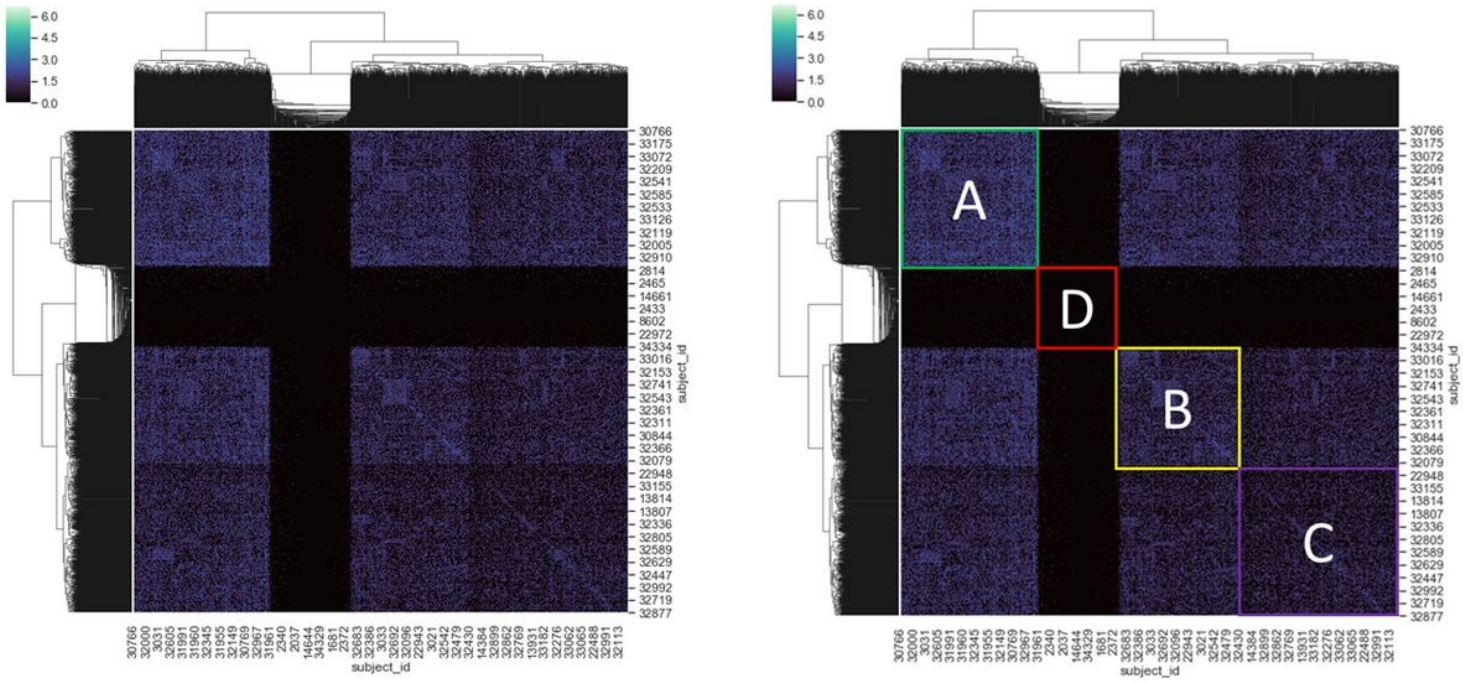

Figure 1

1a and Figure 1b: Dissimilarity matrix patients, representing pairwise similarity values. Whole-genome sequencing data. (a) Both axes are represented by patients. Each patient is uniquely identified by a subject id as it is identified in the VariCarta dataset. For the sake of readability, only 2% of the 2062 subjects are represented in the heatmap. Each value of the matrx is the result of a pairwise comparison of patients and it is computed calculating the log2 of the size of the intersection between the sets of mutated genes in the couple of patients. The values range from 0 (no mutated genes in common) to 6 (more than 60 mutated genes in common). The dendrograms show the hierarchy of clusters proposed by HAC algorithm. On the diagonal axis it is possible to identify four macro-areas with different densities, identified considering the first three levels of the hierarchy. (b) The four areas are highlighted in different colors. A, B, and C represent three clusters. Each cluster is characterized by a different density, expressed by a Density Factor (DF), i.e. the mean value of all the cells belonging to the cluster. Cluster A: DF=1.429; Cluster B: DF=1.014; Cluster C: DF= 0.720. Beyond the three clusters, also patients who shared none or very few mutated genes were identified and included in the “mixed group” (D): DF=0.001. The areas outside the highlighted squares represent overlapping

areas between clusters, implying the existence of a set of mutated genes that is common to all the subgroups.

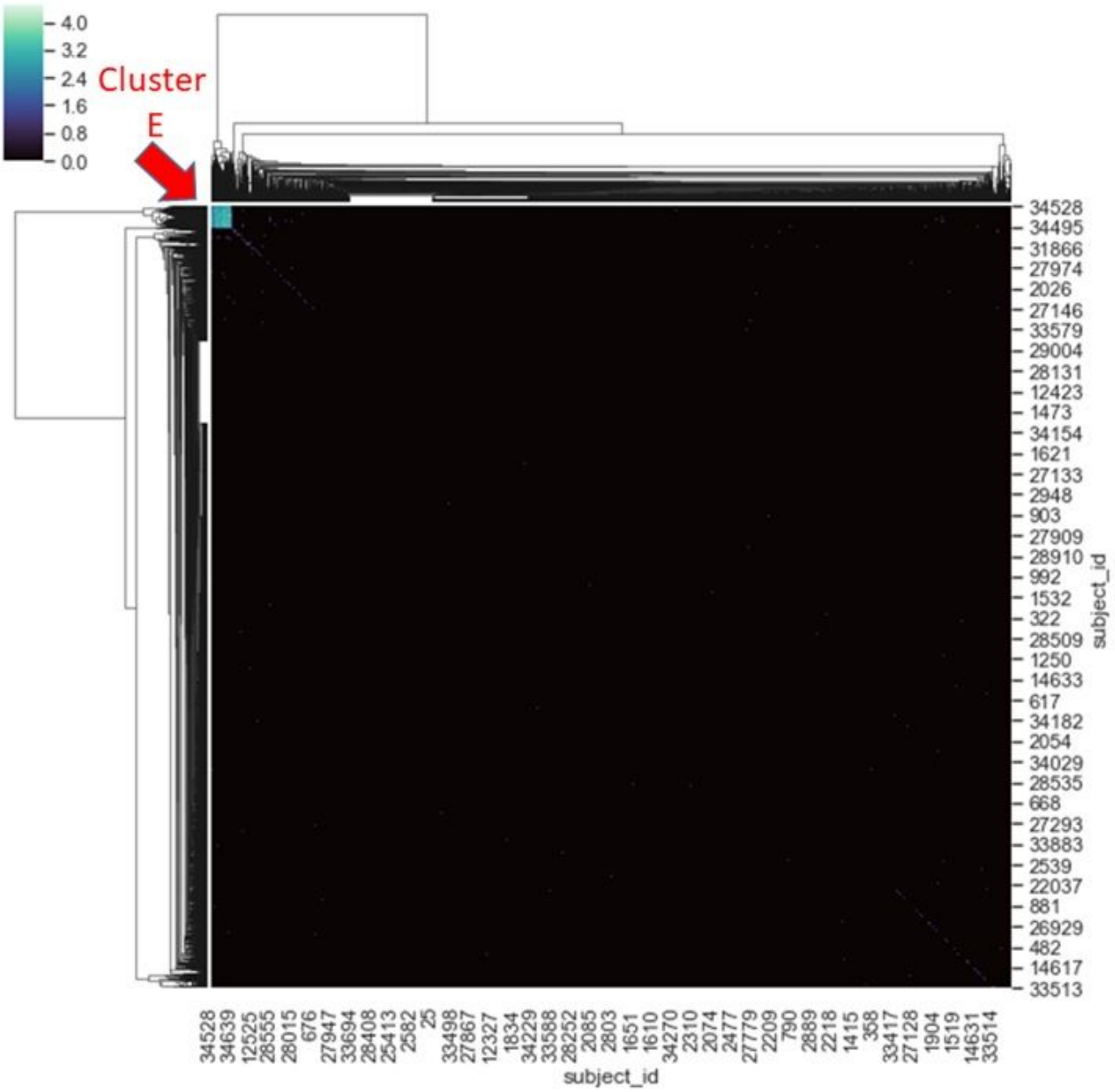

Figure 2

Dissimilarity matrix patients, representing pairwise similarity values. Exome sequencing data. Both axes are represented by patients. Each patient is uniquely identified by a subject ID as it is identified in VariCarta dataset. For the sake of readability not all the 7427 subject IDs are represented in the heatmap but about one in every 190 patients. Each value of the matrix is the result of a pairwise comparison of patients and it is computed calculating the log2 of the size of the intersection between the sets of mutated genes in the couple of patients. The values range from 0 (no mutated genes in common) to 4 (more than 10 mutated genes in common). The dendrograms show the hierarchy of clusters proposed by HAC algorithm. A single cluster (DF=2.551) is visible on the top left of the heatmap (cluster E; highlighted in blue).

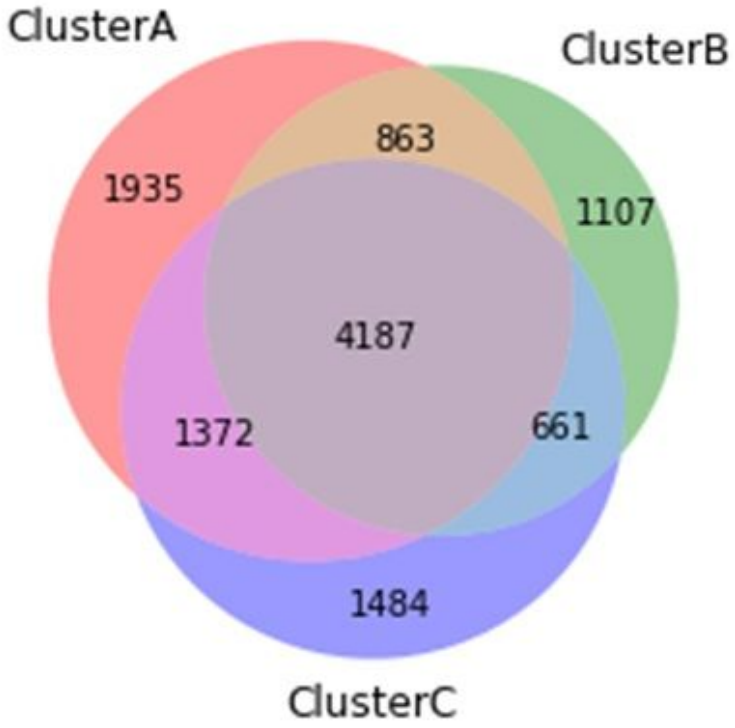

Figure 3

Venn Diagram of gene overlapping among clusters A, B, and C. Whole-genome sequencing data. By intersecting clusters A, B, and C, seven partitions can be identified. Figures overwritten on the seven partitions of the diagram show the number of genetic variants shared by at least two subjects of each intersection. Overall, 11609 genetic variants have been isolated: 4187 of these variants (36.1%) were common to the three clusters; 1935 (16.7%), 1107 (9.5%), and 1484 (12.8%) variants were specific to clusters A, B, and C, respectively. The subsequent enrichment analysis was executed on each partition separately. Only the intersection between the three clusters returned a significant result: FDR PPI enrichment p-value for the intersection of clusters A, B, C was $p < 1.0e\text{-}16$.

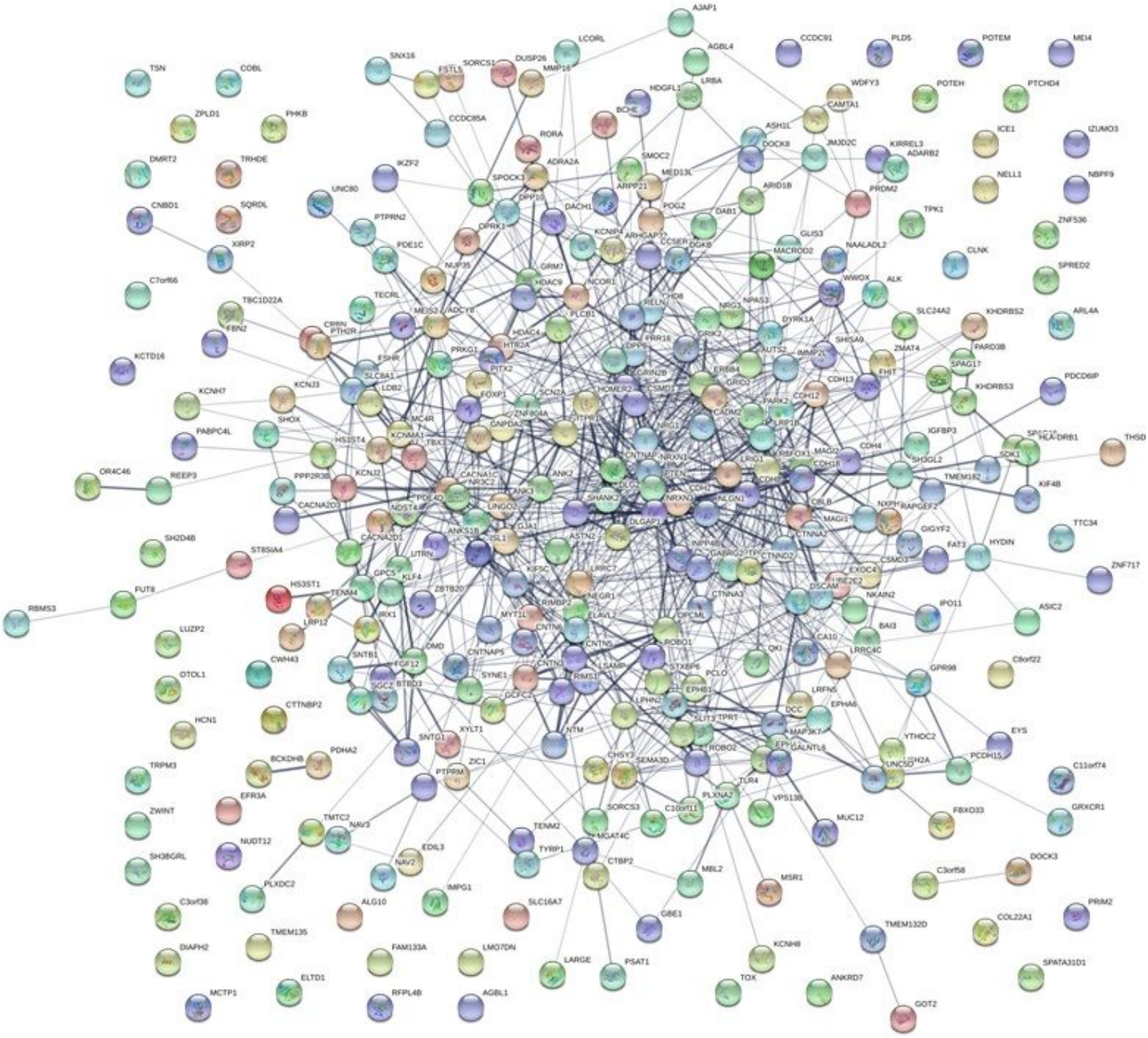

Figure 4

Network interaction graph of the intersection between clusters A, B, and C. Whole-genome sequencing data. Analysis of genes present in at least 50 patients. Nodes: 360. Edges: 990. Expected number of edges: 293. The number of edges represents the existing interactions between the submitted genes. It was compared with the expected number of edges, ie the expected interactions between a set of random genes of the same size.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile2Geneswithoccurrences.xlsx