Network Diffusion-Based Prioritization of Autism Risk Genes Identifies Significantly Connected Gene Modules

Ettore Mosca 1*, Matteo Bersanelli 2, Matteo Gnocchi 1, Marco Moscatelli 1, Gastone Castellani 2, Luciano Milanesi 1 and Alessandra Mezzelani 1

1 Bioinformatics Group, Institute of Biomedical Technologies, National Research Council of Italy, Segrate, Italy, 2 Applied Physics Group, Department of Physics and Astronomy, University of Bologna, Bologna, Italy

Autism spectrum disorder (ASD) is marked by a strong genetic heterogeneity, which is underlined by the low overlap between ASD risk gene lists proposed in different studies. In this context, molecular networks can be used to analyze the results of several genome-wide studies in order to underline those network regions harboring genetic variations associated with ASD, the so-called “disease modules.” In this work, we used a recent network diffusion-based approach to jointly analyze multiple ASD risk gene lists. We defined genome-scale prioritizations of human genes in relation to ASD genes from multiple studies, found significantly connected gene modules associated with ASD and predicted genes functionally related to ASD risk genes. Most of them play a role in synapsis and neuronal development and function; many are related to syndromes that can be in comorbidity with ASD and the remaining are involved in epigenetics, cell cycle, cell adhesion and cancer.

Keywords: autism spectrum disorder, biological networks, network diffusion, data integration, gene module

INTRODUCTION

“Autism spectrum disorder” (ASD) includes clinically and etiologically wide range of neurodevelopmental disorders such as the less severe disorders Asperger's syndrome and pervasive developmental disorder, not otherwise specified, as well as the most severe childhood disintegrative disorder. ASD symptoms are recognized mainly by the complex behavioral phenotype that manifests within the first 3 years of life: difficult in communication and social interaction, limited interests and repetitive behaviors (National Institute of Mental Health, 2013).

Genetics play a crucial role in autism pathogenesis (Devlin and Scherer, 2012). Indeed, ASD has a high-heritability index (0.85–0.92) (Monaco and Bailey, 2001), a significant sib recurrence risks (8.6%) and 64% concordance among monozygotic twins (Smalley, 1998). Thousands of causative or predisposing genetic variations have been found in ~30% of autistic patients (O’Roak et al., 2012), thus making autism a complex multifactorial disorder involving many genes and loci contributing to the phenotype. Genetic variations involved in ASD are chromosomal abnormalities (~5%), copy number variations (CNVs) (10–20%) and single-gene mutations (~5%) (Miles, 2011). Although the role of genetics in ASD etiology is recognized for ~70% of cases, the causative factor is still unknown.

The approaches currently used to disentangle the genetic complexity of ASDs include large genome-wide association studies (GWAS), CNV testing and genome sequencing. Interestingly, the
application of these different approaches yielded many non-overlapping genes, which may suggest different molecular mechanisms within connected pathways (Pinto et al., 2014). The analysis of molecular interactions and pathways is therefore crucial for the interpretation of the results emerging from genome-scale studies on a pathology marked by a significant genetic heterogeneity. Indeed biological pathways associated with a specific pathology are likely to be more conserved than individual genetic variations, because multiple combinations of variations might perturb each pathway (Barabási et al., 2011).

Network-based and pathway-based analyses can therefore provide a functional explanation to non-overlapping genes and narrow the targets for therapeutic intervention (Devlin and Scherer, 2012).

One of the problems that network-based analyses can solve is indeed the identification of the so-called disease modules, i.e., network regions associated with a disease (Barabási et al., 2011). Recently, molecular interaction networks have been used in the analysis of ASD genetic data to define gene networks associated with ASD. The identification of a subnetwork with desired properties from a large biological network (like the one formed by all protein-protein interactions) poses many challenges and therefore several approaches have been proposed (Mitra et al., 2013)

Regarding the integration of networks and ASD genetic data, Cristino et al. (2014) studied the interacting partners of genes known to be associated with ASDs and other related disorders; Noh et al. (2013) identified a significantly interconnected network of genes affected by CNVs; Li et al. (2014) studied the association between ASDs and genes forming topological communities (clusters of genes with a high density of connection between genes of the community and less connections with genes outside the community); Gilman et al. (2011) found functionally connected clusters of genes affected by CNVs.

Recently, network smoothing index (NSI) was proposed as a network-based quantity that allows to define a network region enriched with a priori information (Bersanelli et al., 2016). The NSI is based on network diffusion, a method that simulates the flow of a fluid throughout a network. NSI quantifies the network relevance of each gene in relation to a set of input genes (e.g., ASD genes), considering the whole network and mitigating the importance of hubs.

In this work, we use network diffusion and the NSI to propose a possible disease module for ASD, encompassing the network regions most frequently hit by molecular variations reported in several studies and collected in curated public databases. Moreover, our study introduces a network-based genome-wide prioritization of genes in relation to their known and predicted relevance for ASDs.

**MATERIALS AND METHODS**

**Molecular Interactions**

STRING interactions were collected from STRING (version 10), a database of direct and indirect PPIs (Szklarczyk et al., 2015). Native identifiers were mapped to Entrez Gene (Brown et al., 2015) identifiers. In case multiple proteins mapped to the same gene identifier, only the pair of gene identifiers with the highest STRING confidence score was considered. A total of 11,535 genes and 207,157 links with confidence score ≥ 700 was retained.

**ASD Risk Genes**

ASD risk genes were collected from The Simons Foundation Autism Research Initiative SFARI Gene database (Abrahams et al., 2013, version available in July 2015) and from Li et al. (2014).

SFARI Gene provides a publicly available database where genes are scored according to the strength of the evidence of gene’s association with autism. In particular, genes are assigned to 7 categories (Supplementary Table 1): “ syndromic” (1), “ high confidence” (1), “ strong candidate” (2), “ suggestive evidence” (3), “ Minimal evidence” (4), “ Hypothesized” (5), and “ Not supported” (6). SFARI genes were divided into to broad classes of high and low strength of association with ASD. Genes belonging to categories 1, 2, 3, 1S, 2S, 3S, 4S were included in SFARI list, while genes of categories 4 and 5 were grouped into SFARII. Native gene identifiers were converted to Entrez Gene (Brown et al., 2015) identifiers.

In addition to SFARIh genes, we considered 5 sources (namely dnCNVn, dnCNV3s, rCNV, dMUT, mMUT) of genes harboring CNVs and mutations associated with ASDs, proposed by large recent studies (Li et al., 2014). dnCNVn contains genes from an ASD-associated network composed of genes with de novo CNVs identified in 181 individuals and genes previously implicated in ASDs (Noh et al., 2013). dnCNV3s contains genes with de novo CNVs found in all three independent studies on more than 1,000 families (Levy et al., 2011; Sanders et al., 2011; Pinto et al., 2014). rCNV contains genes with rare CNVs found in a study involving approximately 1000 ASD individuals of European ancestry and matched controls (Pinto et al., 2010). dMUT and mMUT include genes with, respectively, disruptive and missense mutations (Neale et al., 2012; O’Roak et al., 2012; Sanders et al., 2012; Li et al., 2014).

Only genes occurring in STRING network were considered in network-based analyses (Table 1). Therefore, whenever appropriate, we will specifically refer to the original gene lists and the corresponding derived lists with only genes occurring in STRING network using suffixes “i” and “n0” respectively (e.g., SFARII_{i}, SFARII_{n0}).

**Network-Based Analysis**

Given an input gene list L and a gene network encoded as the n-by-n symmetrically normalized adjacency matrix W (Bersanelli et al., 2016), the n-sized vector X₀ was defined in order to have positive quantities only in its elements representing the genes in L, and null values for all the other genes. Network diffusion finds the vector X_n, in which the quantities initially available in X₀ are subject to smoothing according to the pattern of interactions W. The vector X_n was calculated using an iterative procedure (Zhou et al., 2004), as described in Bersanelli et al. (2016):

\[ X_{t+1} = \alpha W X_t + (1 - \alpha) X_0, \quad X_n = \lim_{t \to \infty} X_t, \]
TABLE 1 | Overlap between lists of genes harboring variations associated with ASDs.

|        | dnCNVn | dnCNV3s | rCNV  | dMUT  | mMUT  | SFARIh |
|--------|--------|---------|-------|-------|-------|--------|
| dnCNVn | 203    | 154     | 9     | 2     | 6     | 11     |
| dnCNV3s| 50     | 263     | 221   | 2     | 9     | 3      |
| rCNV   | 39     | 11      | 67    | 45    | 1     | 16     |
| dMUT   | 16     | 365     | 251   | 24    | 206   | 154    |
| mMUT   | 18     | 14      | 3     | 4     | 9     | 3      |
| SFARIh | 16     | 18      | 22    | 24    | 206   | 154    |

Gene list size (diagonal) and number of genes co-occurring between pairs of lists (off-diagonal elements); lower triangle and diagonal (bottom): original gene lists; upper triangle and diagonal (top): original gene lists with only genes occurring in STRING network.

where $\alpha$ [here set to 0.7 as in previous works (Bersanelli et al., 2016)] is a parameter that weights to which extent the initial information is retained or spread throughout the network. In the independent smoothing of each of the six ASD gene lists described above, genes belonging to the list were set to 1. In the joint analysis of all gene lists, genes belonging to SFARIh or mMUT were set to 1, while genes belonging to other lists were set to 0.5: this setting was chosen so that genes strongly associated with ASD had a higher priority.

For each gene $g$, the network smoothing index $S$ quantifies the network proximity of $g$ to genes marked by a positive value in $X_0$, i.e., associated with ASD, as ratio between gene values after and before network diffusion:

$$S(g) = \frac{X(g)}{X_0(g) + \epsilon}$$

where $\epsilon$ is a small positive quantity that weights the importance of the initial values $X_0$. In order to mitigate the tendency of hub genes to gather excessive amounts of information only because of their central position, the permutation-adjusted network smoothing index $S_p$ was introduced as,

$$S_p(g) = -\log_{10}(p_s(g)) \cdot S(g)$$

where $p_s(g)$ is an empirical $p$-value, computed using $K$ permutations of $X_0$, each one denoted as $X_0^K$, and the corresponding $S^K(g)$ (calculated using $X_0^K$):

$$p_s(g) = \frac{1 + \#(S^K(g) \geq S(g))}{K + 1}.$$

In the analysis of the six ASD risk gene lists, $\epsilon$ values were defined in order to predict genes in network proximity to the input genes. Given a gene set of size $N$, $\epsilon$ was set in order to obtain, among the first $N$ top ranking genes by $S_p$, a ratio of 1:1 between (i) the number of input genes and (ii) the number of genes in network proximity to input genes. The resulting values were 0.21 for dnCNVn and 0.19 for dnCNV3n. RcnV, mMUT, mMUTn and SFARIh. In the joint analysis of all gene lists $\epsilon$ was set equal to 1, because the analysis was mainly aimed at defining a network-based prioritization of the 956 input genes, rather than at predicting other genes in network proximity. In all these analyses we used $K = 999$.

Network resampling (NR) shows to what extent a network score, resulting from the combination of gene scores, is expected if links among genes are shuffled. Also in this case permutations are used to define the null model. Given a number $m$ of genes at the top of a ranked gene list, NR consists of two steps. First, a non-decreasing quadratic objective function $\Omega(m)$ is defined:

$$\Omega(m) = S_p^2(m) \cdot A_m \cdot S_p(m),$$

where $S_p(m)$ is the vector referring to the first $m$-scoring genes and $A_m$ is the adjacency matrix between such genes. In the second step, $q$ permutations of $A_m$ are defined keeping the same degree distribution. Lastly, an empirical $p$-value ($p_N$) is calculated to quantify the fraction of times the objective function calculated on a permuted network, $\Omega^b(m)$, is greater than or equal to $\Omega(m)$. The procedure is repeated for different $m$, providing an overview on whether gene links and gene scores determine significant network scores when moving down in the ranked gene list (see figures below).

Network resampling (NR) was applied to genes ranked by $S_p$ in descending order and using a total of 200 permutations, which was enough to underline the presence of significantly connected components (gene modules).

Pathway Analysis

Pathway analysis was carried out using over-representation analysis (ORA). ORA estimates the significance of a pathway in relation to an input gene list, calculating the hypergeometric probability of finding the observed number of input genes that are also members of the considered pathway, in the context of a background set of genes. As a background we considered all the genes occurring in original lists and all genes occurring in the gene network. Gene-pathway associations were downloaded from NCBI Biosystems (version: February 2017) (Geer et al., 2010); in particular, only pathways (gene sets) with a number of genes between 10 and 200 were considered. Hypergeometric probabilities were calculated using “phyper” and “dhyper” R functions, and were corrected for multiple hypotheses testing using the Benjamini-Hochberg method, implemented in “p.adjust” R routine. The similarity between two gene sets $(A, B)$ was calculated using the overlap coefficient:

$$\alpha = \frac{|A \cap B|}{\min(|A|, |B|)}.$$
considered direct and indirect protein-protein interactions (PPI) and quantified, via the permutation-adjusted NSI \((Sp)\) (Bersanelli et al., 2016), the network proximity of each human gene in relation to the network location of 154 genes reported as strongly associated with ASD (SFARI\(_{h0}\) list, Table 1).

We found several genes in significant network proximity to SFARI\(_{h0}\) genes, with high \(S\) and low \(p_S\) (Figure 1A and Supplementary Table 2). Interestingly, among these genes, we found a significant number of genes having minimal/hypothetical evidences of association with autism in SFARI (SFARI\(_{n0}\)) (Table 2).

In order to assess whether genes ranked by \(S_p\) formed a significantly connected gene module, we applied the NR approach (Bersanelli et al., 2016). We observed a significantly connected gene module (\(M_{SFARI}\)) resulting from the top 244 genes (Figures 1B–D). This module includes 142 (out of 154) SFARI\(_{h0}\) genes, 9 SFARI\(_{n0}\) genes and 93 genes not in SFARI.

These 93 genes include regulators of synaptic development and plasticity, are involved in syndromic conditions in comorbidity with ASD, regulate epigenetic mechanisms and a few are associated to cancer (Supplementary Table 2). For example, among the 93 genes that have a relevant position within the module (Figure 1C) we found cancer genes that control cell proliferation, e.g., Tumor Protein P53 (TP53), AKT Serine/Threonine Kinase 1 (AKT1), Mechanistic Target Of Rapamycin (MTOR), C-Terminal Binding Protein 1 (CTBP1)], a process that was recently proposed as a common denominator of cancer and ASDs (Crawley et al., 2016), and genes with relevant role for brain function e.g., histone deacetylase-1 (HDAC1), histone deacetylase-3 HDAC3 (Volmar and Wahlestedt, 2015) and contactin-2 (CNTN2) (Anderson et al., 2012). These 93

![Figure 1](image-url)
predicted genes act as a bridge between SFARIh genes that were not directly linked in STRING (Figure 1D).

Interestingly, while this manuscript was under review, 3 out of the 93 predicted genes were added to the SFARI database independently from our study. Namely, HADAC3 was added among “hypothesized” genes, while TANC2 and PPP2R1B among “minimal evidence” genes.

Genes in Network Proximity to Genes Harboring Variations Associated with ASD

In addition to the SFARIh gene list, we considered five other lists of genes found altered in ASD subjects by previous studies. These lists vary in size from 67 to 530 and from 45 to 263 genes after integration with STRING gene network. The percent overlap between lists is low (Table 1) and most of the genes occur only in one list (Figure 2). In this situation, as introduced earlier, the information on how gene products interact to regulate biological functions can be used to explain the heterogeneity of ASD risk gene lists. Firstly, we analyzed each list separately to underline the specificities and commonalities of each list. Subsequently, we used network information to define a prioritization among all ASD risk genes proposed in the considered studies (union of all gene lists).

We calculated the Sp of all genes in STRING network considering as input each of the six ASD gene lists (SFARIh, dCNVn0, dCNVn0, rCNVn0, dMUTn0, and mMUTn0) and selected the top 2n genes ranked by decreasing values of Sp, where n is list size (Supplementary Table 3). Note that almost all these genes are in significant network proximity (p < 0.05) to the corresponding input genes (Figure 3 and Supplementary Table 3), which are also ranked among the top 2n genes. For convenience we will refer to these network-based gene lists—which contains input and predicted genes—as SFARIh, dCNVn0, dCNVn0, rCNVn0, dMUTn0, and mMUTn0.

Only 14 genes occur in three or more input gene lists (Figure 2). Among those genes, at least DLGAP2 (Discs Large Homolog Associated Protein 2) and SYNGAPI (Synaptic Ras GTPase Activating Protein 1) are worth mentioning. In fact, DLGAP2, a post-synaptic density protein with probable implication in ASD pathogenesis (Chien et al., 2013) scored as “minimal evidence” in SFARI (SFARII), is part of all three CNV lists and was predicted as functionally related to SFARIh genes and dMUT genes. Furthermore, DLGAP1 (Discs Large Homolog Associated Protein-1) was predicted as functionally related to genes of 3 input lists, including SFARIh (Figure 2B). Similarly, SYNGAPI1, which codes an autism related brain-specific synaptic Ras GTP-activating protein (Berryer et al., 2013), occur in three input lists (dCNVh, dCNV3s and SFARIh) and was predicted as functionally related to genes harboring rCNVs.

Globally, a total of 913 genes were predicted as functionally related to at least one ASD risk gene list (Table 3 and Supplementary Table 4). Interestingly, 106 of these genes were already proposed as ASD risk genes in 1 or more

![Figure 2](image)

**Figure 2** A small number of ASD risk genes is found in more than one large study. (A) Number of risk genes found in 1 or more of the six gene lists on ASD; i: original gene lists; n0: original gene lists wit only genes occurring in the STRING network. (B) Permutation-adjusted network smoothing index of the 14 genes occurring in 3 or more original lists (the number is reported between parenthesis); the asterisk (*) indicates genes of the corresponding original list.
A total of 913 genes were predicted to be in network proximity to ASD risk genes of one or more studies (rows) where could appear as ASD risk genes (columns). For example, 15 genes were predicted in network proximity to ASD risk genes of 2 studies and were proposed as ASD risk genes in 1 study ($n_i = 1, n_p = 2$); $\#$: row or column sum.

A total of 913 genes were predicted to be in network proximity to ASD risk genes of one or more studies (rows) where could appear as ASD risk genes (columns). For example, 15 genes were predicted in network proximity to ASD risk genes of 2 studies and were proposed as ASD risk genes in 1 study ($n_i = 1, n_p = 2$); $\#$: row or column sum.

Many of the 29 genes predicted as functionally related to three or more lists—27 genes not included in any input list and 2 included in one study—take part in many PPIs, implicating they are central in the PPI network (e.g., TP53, AKT1). The significance of their $Sp$ suggests that these genes were not only selected in relation to their centrality, but also because their network distance to ASD risk genes is lower than expected by chance (Figure 3 and Supplementary Table 3). A further observation that supports this hypothesis is that these genes establish a number of interactions with ASD risk genes that is higher than expected ($p < 0.05$, hypergeometric test) (Table 4). From a network point of view, these 29 genes are “surrounded” by 369 ASD risk genes (first order neighbors).

It is also worth mentioning that several genes resulting with the highest network proximity score to each list tend to be list specific (Figure 5).

A total of 956 unique ASD risk genes occur in the 6 input lists. We calculated the $Sp$ of all genes relative to these 956 genes (Figure 6A) and, by means of NR, found a significantly connected component of 561 genes (Figure 6B and Supplementary Table 5). This gene module (M_{ASD}) includes all SFARIh genes, 70% of genes occurring in dnCNV_n list, approximately 50% of the other gene lists (Figure 6C), 26 genes in SFARIi and 8 genes that do not belong to the input list. Among these 8 genes, we find the already mentioned AKT1, TP53, and SRC, which occupy a central role in the PPI network and were also predicted during the independent analysis of each input list (Figures 4, 6D).
Pathway Analysis of Gene Lists Associated with ASDs

We carried out an over-representation analysis to characterize original gene lists and network-based gene lists in terms of pathways. We observed significant pathways (at adjusted $p < 0.01$) in only two of the six original gene lists (SFARIi and dnCNVn) and in five network-based lists (SFARIin, mMUTn, rCNVn, dnCNV3n, and dnCNVn) (Supplementary Table 6). Overall, we obtained a much higher number of pathways in network-based lists than in original ones, despite the number of tested genes was similar between the former and the latter ones. Therefore, the observed enrichment in pathways can be mainly brought back to the network-based analysis, since, by definition, it prioritizes genes functionally related to those considered in input. Further, apart from a single exception in SFARIi, pathways found only in original lists were similar, in terms of gene content, to pathways found also in network-based lists (Figures 7A–F). In other words, network-based analysis resulted in an enrichment at pathway level with a very limited loss of information. Interestingly, pathways found in original gene lists, and lost due to genes for which network-based analysis was not applicable, were recovered by network-based analysis (compare rows labeled with suffix “i”, “n0”, and “n*” in Figure 7F).

We summarized the significant pathways found in each lists in a unique pathway network, the so-called enrichment map (Merico et al., 2010). Specifically, we took into account up to 20 of the most significant pathways as representatives of each pathway cluster found in each list. This selection resulted in a total of 366 pathways, clustered in 11 groups: transmission across synapses (clusters 1 and 4), signal transduction (2), Rho GTPase and apoptosis (3), inositol phosphate metabolism (5), ER-associated apoptosis (6), transport (7), chromatin remodeling (8), oxygen transport (9), proteoglycan biosynthesis (10) and Wnt signaling (11). While the majority...
### Table 4 | Hub genes predicted in network proximity to ASD risk genes of three or more studies establish a significant number of interactions with ASD risk genes.

| Gene and function | Symbol | Band | # Studies | \(n_p\) | \(|A\cap I|\) | \(p\) |
|-------------------|--------|------|-----------|------|-------------|------|
| SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase | SRC | 20q11.23 | 0 | 5 | 532 | 82 | \(2.00 \times 10^{-8}\) |
| Nonreceptor tyrosine kinase, frequently implicated in cancer | | | | | | |
| Tumor Protein P53 | TP53 | 17p13.1 | 25 | 4 | 719 | 92 | \(1.36 \times 10^{-5}\) |
| Involved in cell cycle regulation where negatively regulate cell division. Mutations in this gene are associated with a variety of human cancers | | | | | | |
| AKT Serine/Threonine Kinase 1 | AKT1 | 14q32.33 | 25 | 4 | 589 | 73 | \(2.90 \times 10^{-4}\) |
| Implicated in the regulation of cell growth, proliferation, survival and differentiation (OMIM 164730) | | | | | | |
| Ras Homolog Family Member A | RHOA | 3p21.31 | 13 | 4 | 406 | 65 | \(1.53 \times 10^{-7}\) |
| Regulates remodeling of the actin cytoskeleton during cell morphogenesis and motility. Overexpression of this gene is associated with tumor cell proliferation and metastasis | | | | | | |
| Mitogen-Activated Protein Kinase 14 | MAPK14 | 6p21.31 | 11 | 4 | 333 | 44 | \(1.30 \times 10^{-3}\) |
| is a member of the MAP kinase family that are involved in cellular processes such as proliferation, differentiation, transcription regulation and development | | | | | | |
| Histone Deacetylase 3 | HDAC3 | 5q31.3 | 8 | 4 | 275 | 43 | \(3.54 \times 10^{-5}\) |
| belongs to the histone deacetylase family and represses transcription when tethered to a promoter; down-regulates p53 function and thus modulate cell growth and apoptosis | | | | | | |
| Heat Shock Protein 90 Alpha Family Class A Member 1 | HSP90AA1 | 14q32.31 | 10 | 3 | 589 | 65 | \(9.99 \times 10^{-3}\) |
| is a molecular chaperone involved in signal transduction, protein folding, protein degradation, and morphologic evolution. | | | | | | |
| Amyloid Beta Precursor Protein | APP | 21q21.3 | 15 | 3 | 363 | 42 | \(1.68 \times 10^{-2}\) |
| Is involved in promoting transcriptional activation; can participate in the formation of amyloid plaques of Alzheimer disease. | | | | | | |
| Histone Deacetylase 1 | HDAC1 | 1p35.2 | 12 | 3 | 351 | 48 | \(3.69 \times 10^{-4}\) |
| Is a histone deacetylase and represses transcription; interacts with retinoblastoma tumor-suppressor protein to control cell proliferation and differentiation; modulates p53 effect on cell growth and apoptosis. | | | | | | |
| Cell Division Cycle 42 | CDC42 | 1p36.12 | 9 | 3 | 336 | 56 | \(2.97 \times 10^{-7}\) |
| Acts in cell morphology, migration, endocytosis and cell cycle progression. | | | | | | |
| Growth Factor Receptor Bound Protein 2 | GRB2 | 17q25.1 | 21 | 3 | 279 | 42 | \(1.06 \times 10^{-4}\) |
| is involved in the signal transduction pathway. | | | | | | |
| Protein Kinase CAMP-Activated Catalytic Subunit Alpha | PRKACA | 19p13.12 | 14 | 3 | 278 | 41 | \(2.00 \times 10^{-4}\) |
| phosphorylates proteins and substrates, changing their activity; contributes to the control glucose metabolism, cell division, and contextual memory; developmental changes in synapse morphology. | | | | | | |
| SOS Ras/Rac Guanine Nucleotide Exchange Factor 1 | SOS1 | 2p22.1 | 14 | 3 | 260 | 50 | \(1.26 \times 10^{-8}\) |
| participates in signal transduction pathways. | | | | | | |
| SMAD Family Member 3 | SMAD3 | 15q22.33 | 1 | 3 | 241 | 33 | \(2.83 \times 10^{-3}\) |
| signal transducer and transcriptional modulator that mediates multiple signaling pathways probably involved in carcinogenesis | | | | | | |
| Protein Kinase CAMP-Activated Catalytic Subunit Beta | PRKACB | 1p31.1 | 22 | 3 | 213 | 28 | \(9.81 \times 10^{-3}\) |
| is a member of the serine/threonine protein kinase family involved in cell proliferation and differentiation | | | | | | |
| Cbl Proto-Oncogene | CBL | 11q23.3 | 12 | 3 | 203 | 37 | \(3.77 \times 10^{-6}\) |
| targets substrates for degradation by the proteasome; is mutated or translocated in many cancers | | | | | | |
| Protein Kinase CAMP-Activated Catalytic Subunit Gamma | PRKACG | 9q21.11 | 7 | 3 | 185 | 27 | \(2.70 \times 10^{-3}\) |
| is involved in the regulation of lipid and glucose metabolism and in the memory formation signaling cascade | | | | | | |
| Ras Homolog Family Member B | RHOB | 2p24.1 | 8 | 3 | 174 | 38 | \(2.19 \times 10^{-8}\) |
| Involved in intracellular protein trafficking of a number of proteins; plays a negative role in tumorigenesis | | | | | | |
| SOS Ras/Rho Guanine Nucleotide Exchange Factor 2 | SOS2 | 14q21.3 | 12 | 3 | 109 | 24 | \(7.36 \times 10^{-6}\) |
| is involved in the positive regulation of ras proteins | | | | | | |
| Dynamin 2 | DNMT2 | 19p13.2 | 19 | 3 | 84 | 21 | \(3.27 \times 10^{-6}\) |
| produces microtubule bundles and binds and hydrolyzes GTP; regulates neuron morphology, axon growth; vesicular trafficking processes and cytokinesis | | | | | | |

Gene and function from GeneCards (Salari et al., 2003) and OMIM (Amberger et al., 2015); Band: cytogenetic band associated with genetic variations in SFARI (Abrahams et al., 2013); # Studies: number of studies supporting the genetic variations observed in the cytogenetic bands; I: interactors; A: ASD risk genes, \(|A| = 986\); \(n_p\): number of ASD risk gene sets the gene is network proximity to; \(|A\cap I|\): interactors; \(p\): probability that \(|A\cap I|\) is greater than or equal to the observed value in a hypergeometric experiment; the total number of genes composing the network is 11,353.
of pathways were found in more than 1 list (Figure 7G, yellow circles), some pathway clusters were composed of pathways uniquely associated with rCNV, ion transport with mMUT, oxygen transport with dnCNV3s and dnCNVn.

**DISCUSSION**

Recently, the knowledge of molecular interactions has been used for the interpretation of genetic data on ASDs. In comparison to previous works, we analyzed multiple ASD risk gene lists proposed in large studies, for a total of approximately 1,000 genes. We observed a low overlap between ASD risk gene lists. Whether this heterogeneity reflects the biology of ASD or is the result of confounding factors, the analysis of network proximity between genes underlines the ASD risk genes that are also in functional relation and lead to the identification of modules of functionally related genes hit by genetic variations. The main limitation of a network-based analysis such as ours is the availability of *a priori* annotations required for the definition of the genome-scale network. In this work we considered both direct (physical) and indirect (functional) high confidence PPI from STRING, which allowed us to analyze 11,535 human genes. Note that the use of direct and indirect STRING interactions
showed good performances in prioritizing candidate disease genes (Köhler et al., 2008). Moreover, unlike other network-based works on ASD genetic data, we used network diffusion to quantify network proximity between ASD risk genes and other genes. Network diffusion (a global approach) considers the whole network topology in its full complexity and, therefore, has better performances than local approaches (e.g., direct neighborhood or shortest path length; Wang et al., 2011). Lastly, we underlined ASD risk gene modules without constraining the search to topological communities. In fact, there is no guarantee that topological communities are able to capture disease modules (Ghiassian et al., 2015). Hence, we quantified the significance of the observed network proximity scores in comparison to random networks of the same degree distribution (Bersanelli et al., 2016).

Our work provides a network-based prioritization of human genes associated with ASD by previous studies. We extracted a module of 244 genes in network proximity to genes reported in SFARI as strongly associated with ASD. Interestingly, the module contains a significant number of genes proposed as possibly involved in ASD (categorized as “minimal evidence” and “hypothesized” in SFARI) and another 93 genes not scored in SFARI (Supplementary Table 2). While this manuscript was under review, 3 of these 93 genes were included in SFARI independently from our study.

From the 93 genes, 16 genes are involved in synaptogenesis and synaptic plasticity or transmission, and alterations in structure and function of neuronal synapses are well known causes of ASD. Among these, APLP2 also regulates proper progression of neuronal differentiation program during cortical development (Shariati et al., 2013), is involved in Alzheimer disease and interacts with CNTN in neurodevelopment and diseases (Osterfield et al., 2008). Then again, the 3 genes, CACNA2D1, CACNB1, CACNG1 induce the repression of the downstream regulatory element antagonist modulator (DREAM) and the expression of the neuropeptide dynorphin (DYN). DREAM plays a role in synaptic plasticity and behavioral memory (Wu et al., 2010), while DYN is involved in behavioral symptoms characteristic of human depressive disorders (Knoll and Carlezon, 2010). Also CACNG3, a calcium channel protein, regulates the function of AMPA-selective glutamate
FIGURE 7 | Significant pathways enriched in genes associated with ASD. (A–E) Pathways found in the analysis of each original list (i) and corresponding predicted genes (n*). (F) Heatmap of all the pathways found in the analysis of the lists reported on the rows; also pathways found analyzing the original lists without genes not occurring in the STRING network (n) are represented; the color bar indicates the $-\log_{10}(p)$ of the hypergeometric p-value, adjusted for multiple hypothesis testing; the number 7 indicates $p \leq 10^{-7}$. (G) Enrichment map; vertex size is proportional to pathway significance (adjusted p-value); links are reported only for overlap coefficients $>0.5$; for each pathway, only the links with the top 5 most similar pathways are drawn. (A–G) See Supplementary Table 6.

receptors and mediates synaptic transmission in CNS while FLRT3 takes part in a trans-synaptic complex (Lu et al., 2015). Eight genes are involved in neuronal differentiation, neurodevelopment and neuronal function. More specifically, AKT1 is a downstream mediator of the PI3K pathway that regulates synaptic formation and plasticity and which imbalance leads to autism and schizophrenia (Enriquez-Barreto and Morales, 2016); genetic variations in contactins (CNTN) have been described in association with neurodevelopmental disorders, including autism. Specifically, CNTN1 and CNTN2 are members of the presynaptic NRXX superfamily and 13 rare non-synonymous variants of CNTN2 have been found in ASDs patients while mice with Ctnm5 mutations show an abnormal audiogenic response due to defects in the formation of synapse in auditory neurons (Cottrell et al., 2011; Chen et al., 2014).

Many of the 93 genes are involved in syndromic comorbidities, including auditory and visual senses deficit, epilepsy, mental retardation and psychiatric conditions that affect nearly three-quarters of children with ASD. For instance, CAMK2G and PDZD7 are involved in Usher Syndrome the most common condition leading to deafness and blindness, as well as DNMT1 has a role in DNMT1-Related Dementia, Deafness, and Sensory Neuropathy (Vernon and Rhodes, 2009) and LRTOMT in deafness. Again, MYL7 mutations are associated with Fechtner Syndrome which features include hearing loss and eye abnormalities. SP4 is involved in bipolar disorder and schizophrenia while ANK3, ACSL4, DLG3 are associated with mental retardation and, interestingly, recalling the high male prevalence of ASD, the latter two map on X-cromosome; NHS also maps on X-cromosome and mutations in this gene cause Nance-Horan Syndrome characterized by congenital
cataaract leading to vision loss; in males mild or moderate mental retardation may also occur and ASD have also been described in few patients (Toutain et al., 1997). Mutations in NAGLU and HGSNAT cause the Sanfilippo Syndrome (also called mucopolysaccharidosis Type III) often misdiagnosed with idiopathic developmental delay, attention deficit/hyperactivity disorder and/or ASD (Wijburg et al., 2013). QDPR mutations provoke hyperphenylalaninemia (Trujillano et al., 2014), (also called atypical phenylketonuria (PKU), a genetic metabolic disease provoking postnatal cognitive deficit due to the neurotoxic effect of hyperphenylalaninemia; interestingly, PKU could be a comorbid condition of ASD, although with low prevalence (Baieli et al., 2003). MKRN3 is associated with Prader Willy Syndrome, NPAP1 both with Prader Willy Syndrome and Angelman Syndrome while DSCAML1 with Down Syndrome. These syndromes are characterized by mental retardation and can have co-occurring ASDs (Peters et al., 2004; Capone et al., 2005; Dykens et al., 2011). KMT2D and WDR5 defects are involved in Kabuki Syndrome characterized by multiple congenital abnormalities, from mild to severe developmental delay and intellectual disability. People suffering from this syndrome may also manifest seizures, hypotonia, strabism, hear infections, hearing loss and autism (Parisi et al., 2015). The very rare mutations in MANBA results in β-mannosidosis with a severe neurological disorder that can include mental retardation, cerebellar ataxia along with visual and hearing deficits (Sabourdy et al., 2009). CACNG3 is involved in Childhood Absence Epilepsy and is also associated with some cases of ASD (Danielsson et al., 2005) while mutations of KATNB1 cause complex cerebral malformations (Mishra-Gorur et al., 2014).

The remaining genes (among the 93) are mostly involved in epigenetics, cell cycle and cell adhesion and some of them are also implicated in tumor development as already reported by Crespi (2011) and Crawley et al. (2016).

The network-based analysis of genes from SFARI and other 5 previous studies resulted in the definition of a gene module that involves 561 ASD risk genes in significant functional relation. The module contains all the considered SFARI genes (strongly associated with ASD) and from 40% to 70% of genes from each of the other lists of ASD risk genes. Therefore, this module can be seen as a further screening of the genes proposed by such studies, which underline those in significant functional relation from a network perspective.

More generally, the network-based scores that we calculated for every gene in the considered STRING network can be used to quantify the functional relation between any gene and ASD risk genes found in one or more previous studies.

Biological pathways enriched in genes in network proximity to ASD risk genes encompass several functions already proposed to be associated with ASD (see, for example, Pinto et al., 2010). Network-based analysis, through the prioritization of functionally related genes, enriched the number of significant pathways found by ORA in comparison to the analysis of original gene lists. Despite not all genes occurring in original lists underwent network-based analysis, the latter was not affected by a loss of information at pathway level.

The predicted genes in network proximity to ASD risk genes that have a central role in the PPI networks, but SRC, mapped in ASD risk loci. SFARI Gene database lists all the studies reporting CNV at the chromosome bands where predicted genes are localized (Table 4). In many reports, the CNV of interest was subsequently confirmed or validated by an independent method following its discovery. Additionally, from a functional point of view, most of the predicted genes are involved in epigenetics, cell cycle, growth-, proliferation- and differentiation-signaling and are often implicated in cancer development. This finding indicates pleotropic effects of some autism-associated genes on cancer risk and is supported by previous discussions that highlight a wide overlap in risk genes and pathways for cancer and autism (Crespi, 2011; Crawley et al., 2016). Advances in pharmacological therapies to ameliorate autism symptoms could be resulted from cancer drugs that target the same growth-signaling pathways (Crespi, 2011).

AUTHOR CONTRIBUTIONS

EM collected the ASD data from the literature, setup and run the analyses; MB curated the physical mathematical modeling of network diffusion; MG and MM managed the high performance computing infrastructure; GC supported the physical mathematical modeling; LM coordinated the research; AM analyzed the ASD literature, interpreted the biological results, coordinated the research. All authors discussed the results, contributed to manuscript writing and revision.

FUNDING

The work has been supported by: the EU FP7 project “MIMOmics” (305280); the Italian Ministry of Education, University and Research (MIUR) projects “INTEROMICS” (PB05) and “PRIN 2015”; the Lombardy Region Foundation FRRB project “LYRA” (2015-0010).

ACKNOWLEDGMENTS

We would like to thank John Hatton (CNR-ITB) for proofreading the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fgene.2017.00129/full#supplementary-material

REFERENCES

Abrahams, B. S., Arking, D. E., Campbell, D. R., Mefford, H. C., Morrow, E. M., Weiss, L. A., et al. (2013). SFARI gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism 4:36. doi: 10.1186/2040-2392-4-36

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). Omim.org: online mendelian inheritance in man (omim), an online
catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Anderson, G. R., Gallin, T., Xu, W., Aoto, J., Malenka, R. C., and Sdhof, T. C. (2012). Candidat autism gene screen identifies critical role for cell-adhesion molecule caspr2 in dendritic arborization and spine development. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18120–18125. doi: 10.1073/pnas.1216398109

Bailey, S., Pavone, L., Meli, C., Fiumara, A., and Coleman, M. (2003). Autism and phenylketonuria. *J. Autism Dev. Disord.* 33, 201–204. doi: 10.1023/A:1022997712639

Barabási, A. L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 889–907. doi: 10.1016/j.neuron.2011.05.021

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interconnectome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Knoll, A. T., and Carlezen, W. A. (2010). Dynorphin, stress, and depression. *Brain Res.* 1314, 56–73. doi: 10.1016/j.brainres.2009.09.074

Levy, D., Ronemus, M., Vawter, B., Lee, Y. H., Loetta, A., Kendall, J., et al. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* 70, 886–897. doi: 10.1016/j.neuron.2011.05.015

Li, J., Shi, M., Ma, Z., Zhao, S., Euskirchen, G., Ziskin, J., et al. (2014). Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.* 10:774. doi: 10.15252/msb.20145487

Lu, Y. C., Nazarko, O. V., Sando, R., Salzman, G. S., Sdhof, T. C., and Ara, (2015). Structural basis of latrophilin-flt-unc5 interaction in cell adhesion. *Structure* 23, 1678–1691. doi: 10.1016/j.str.2015.06.024

Merico, D., Isserlin, R., Steuver, O., Emili, A., and Bader, G. D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* 5:e13984. doi: 10.1371/journal.pone.0013984

Miles, J. H. (2011). Autism spectrum disorders—a genetics review. *Genet. Med.* 13, 278–294. doi: 10.1097/GIM.0b013e3181f67ba

Mishra-Gorur, K., Caglayan, A. O., Schaffer, A. E., Chabu, C., Henegariu, O., Vonhoff, F., et al. (2014). Mutations in KATNB1 cause complex cerebral malformations by disrupting asymmetrically dividing neural progenitors. *Neuron* 84, 1226–1239. doi: 10.1016/j.neuron.2014.12.014

Mitra, K., Carvunis, A. R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732. doi: 10.1038/nrg3552

Monaco, A. P., and Bailey, A. J. (2001). Autism. The search for susceptibility genes. *Lancet* 358(Suppl. S3). 503–507. doi: 10.1016/S0140-6736(00)08701-7

National Institute of Mental Health (2013). *Autism Spectrum Disorder*. Available online at: https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

Neale, B. M., Kow, Y., Liu, L., Mååyan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242–245. doi: 10.1038/nature11011

Noh, H. J., Ponting, C. P., Boulding, H. C., Meader, S., Betancur, C., Buxbaum, J. D., et al. (2013). Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism. *PLoS Genet.* 9:e1003523. doi: 10.1371/journal.pgen.1003523

O’Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coo, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Osterfeld, M., Egelund, R., Young, L. M., and Flanagan, J. G. (2008). Interaction system. *Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet.* 94, 677–694. doi: 10.1016/j.ajhg.2014.03.018

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Rengarajan, R., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466, 368–372. doi: 10.1038/nature09146

Saboludry, F., Labauge, P., Stensland, H. M. F. R., Nieto, M., Garcia, V. L., Renard, D., et al. (2009). A manba mutation resulting in residual beta-mannosidase activity of the 7q11.23 Williams syndrome region, are strongly associated with Autism. *PLoS Genet.* 5:e1000323. doi: 10.1371/journal.pgen.1000323

Sandberg, J., Jarl-Escéniecek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., et al. (2014). Multiple recurrent de novo cnvs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with Autism. *Neuron* 70, 863–885. doi: 10.1016/j.neuron.2011.05.002
