Network Propagation-Based Semi-supervised Identification of Genes Associated with Autism Spectrum Disorder

Hugo F. M. C. Martiniano[1,2(✉)], Muhammad Asif[1,2], Astrid Moura Vicente[1,2], and Luís Correia[1]

[1] Faculdade de Ciências, BioISI - Biosystems and Integrative Sciences Institute, Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal
hfmartiniano@ciencias.ulisba.pt, muhasif123@gmail.com
[2] Instituto Nacional de Saúde Doutor Ricardo Jorge, Avenida Padre Cruz, 1649-016 Lisbon, Portugal

Abstract. Autism Spectrum Disorder (ASD) is an etiologically and clinically heterogeneous neurodevelopmental disorder with more than 800 putative risk genes. This heterogeneity, coupled with the low penetrance of most ASD-associated mutations presents a challenge in identifying the relevant genetic determinants of ASD. We developed a machine learning semi-supervised gene scoring and classification method based on network propagation using a variant of the random walk with restart algorithm to identify and rank genes according to their association to know ASD-related genes. The method combines information from protein-protein interactions and positive (disease-related) and negative (disease-unrelated) genes. Our results indicate that the proposed method can classify held-out known disease genes in a cross-validation setting with good performance (area under the receiver operating curve ∼0.85, area under the precision-recall curve ∼0.8 and Matthews correlation coefficient 0.57). We found a set of top-ranking novel candidate genes identified by the method to be significantly enriched for pathways related to synaptic transmission and ion transport and specific neurotransmitter-associated pathways previously shown to be associated with ASD. Most of the novel candidate genes were found to be targeted by *denovo* single nucleotide variants in ASD patients.

Keywords: Semi-supervised learning · Network propagation · Autism Spectrum Disorder · Machine learning · Protein-protein interactions

1 Introduction

Autism Spectrum Disorder (ASD) is an etiologically and clinically heterogeneous neurodevelopmental disorder, with an approximate prevalence of 1% of the population [1]. It is characterized by early-onset of difficulties in social interaction

M. Raposo et al. (Eds.): CIBB 2018, LNBI 11925, pp. 239–248, 2020.
https://doi.org/10.1007/978-3-030-34585-3_21

and communication, and repetitive, restricted behaviors, interests, or activities, frequently associated with co-morbidities like intellectual disability, epilepsy and language disabilities [2].

Twin studies comparing the concordance in identical and fraternal twins provide evidence for a string contribution of genetic factors to ASD risk variance [3]. In the past decade, an intense effort to identify the genetic determinants of ASD has been undertaken by several international consortia. This produced several large-scale genomic datasets using high-throughput techniques, including Single Nucleotide Polymorphisms (SNPs) from genome-wide association studies, whole exome sequencing and whole genome sequencing [3–5].

However, despite the enormous amount of data generated over the past decade [3], a clear genetic cause can only be identified in 10% to 20% of affected individuals, and diagnostic still relies on clinical observation, rather than ASD-specific etiology. Genetic causes identified in ASD patients include large chromosomal abnormalities, Copy Number Variants (CNVs) and single-gene mutations, either transmitted or *de novo*. More than putative 800 ASD-risk genes have been identified [3], but most CNVs and Single Nucleotide Variants (SNVs) are characterized by incomplete penetrance [3,6].

An increasing body of evidence suggests that ASD risk variance is the result of a combination of common and rare variants, acting on specific biological pathways (for example, neuronal development and axonal guidance, synaptic function, and chromatin remodeling) [7], with various transmission modes, and possibly combined with gene-environment interactions [3].

This heterogeneity presents a challenge for translational approaches and therefore clinical application lags behind the research knowledge [6].

One of the consequences of the genetic heterogeneity of ASD is that the sample sizes necessary for establishing statistically significant genotype-phenotype associations are very large (of the order of tens of thousands of individuals).

Given this scenario, one possible alternative to large-scale genomic studies is to target genes involved in known ASD-related biological pathways to gain insight into the specific processes disrupted in each individual patient or group of patients.

One promissing approach for the identification of novel ASD-related genes and pathways is through the application of machine learning techniques.

2 Scientific Background

Krishnan *et* al. [8] used Support Vector Machines (SVM) to discriminate ASD-related from ASD-unrelated genes based on information from a human brain-specific gene interaction network. Despite obtaining a good (AUC = 0.8), they used an arbitrary candidate gene-weighing scheme.

Asif *et* al. [9] achieved an AUC of 0.8 for the identification of ASD-related genes using a Random Forest trained on gene semantic similarity measures calculated from Gene Ontology.

In this paper we describe a semi-supervised machine learning approach to identify and rank potentially relevant disease risk genes, based on prior information from publicly available databases. We present the results of the application of this method to the discovery of novel ASD risk genes.

We used a modification of the semi-supervised classification method proposed by Zhou *et* al. [10], which is closely related to Network Propagation (NP). NP is a family of methods that use the flow of information through network connections as a means to establish relationships between nodes. Several NP variants are widely used to identify genes and genetic modules that underlie a process of interest (see for instance [11] for a review). The main concept behind the technique, when applied to gene-gene interaction networks, is that genes underlying the same phenotype tend to interact. NP has been has been applied to the discovery of significantly connected gene modules associated with ASD [12].

To the best of our knowledge, reported applications of NP (for example [12]), are done within a positive-unlabeled learning framework. In this work we use a different approach, where the algorithm learns from both positive-labeled (disease-associated) and negative-labeled (non disease-associated) genes. Although it is not currently possible to define a true set of negative genes for ASD the inclusion of a negative gene set adds more information for the method to exploit, as genes thought to be unrelated to the disease (and their neighbors) are down-weighed.

3 Materials and Methods

3.1 Semi-supervised Learning Algorithm

We use a modified version of the semi-supervised classification method first described by Zhou *et* al. [10]. The algorithm uses as inputs a network $G(\mathcal{V}, \mathcal{E})$, where the nodes, $\mathcal{V}$, represent genes, and the edges, $\mathcal{E}$, represent gene-gene interactions, and two gene sets, $\mathcal{P}$ and $\mathcal{N}$, containing disease-related and non disease-related genes, respectively.

Using information from the predefined gene sets, we build an initial score vector (f^0) for all genes in the network, where the element f_i^0 corresponding to gene g_i is:

$$f_i^0 = \begin{cases} \frac{1}{|\mathcal{P}|} & g_i \in \mathcal{P} \\ -\frac{1}{|\mathcal{N}|} & g_i \in \mathcal{N} \\ 0 & g_i \notin \mathcal{P} \wedge g_i \notin \mathcal{N} \end{cases} \tag{1}$$

The initial scores are then propagated through the network, using the iterative formulation of the random walk with restart (RWR) algorithm [11]:

$$f^{t+1} = (1 - \lambda)\mathbf{W}f^t + \lambda f^0 \tag{2}$$

Where f^t is the vector of gene scores for step t, λ is the restart coefficient ($\lambda = 0$ corresponds to a random walk without restart) and $\mathbf{W}$ is a weight matrix

derived from the normalization of the network adjacency matrix: $\mathbf{W} = \mathbf{D}^{\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}$, where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ is the diagonal node degree matrix $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_j, \ldots, d_n)$, where $d_j = \sum_i A_{ij}$.

In the case of weighed networks, the A_{ij} matrix element is the weight of the edge connecting the i and j nodes, and the D_{ii} matrix element contains the sum of the weights of all edges connected to node i.

The iterative update (Eq. 2) is performed until convergence of f^t, that is until it verifies $||f^t - f^{t-1}||^2 < 1 \times 10^{-6}$.

In the context of the random walk with restart framework, this is equivalent to a random walk starting with equal probabilities from the positive and negative nodes. After convergence, the probabilities of the visits from negative nodes are subtracted from the probabilities of visits from positives nodes, yielding the final score vector.

Performance was evaluated using 10-fold cross-validation, with λ varying from 0.1 to 0.9 in 0.1 steps. As performance evaluation metric we used the area under the receiver operating curve (AUROC), the area under the precision-recall curve (AUPRC), the Matthews correlation coefficient (MCC), and report mean values over all folds for each λ parameter.

3.2 Gene Sets

The Simons Foundation for Autism Research (SFARI)[1] curates an authoritative list of ASD-associated genes. The Human Gene Module of the SFARI database lists 1007 genes, grouped into several categories, according to the level of evidence linking them to ASD (Table 1). The genes in the *Syndromic* category are implicated in syndromic forms of autism, in which subpopulations of patients with a specific genetic syndrome, such as Angelman syndrome or fragile X syndrome, present symptoms of autism. These can also be in one of the other categories, or just classified as *Syndromic*.

Table 1. Categories of Autism candidate risk genes from SFARI.

Category	Description	Number of genes
S	*Syndromic*	143
1	*High confidence*	25
2	*Strong candidate*	58
3	*Suggestive evidence*	176
4	*Minimal evidence*	405
5	*Hypothesized but untested*	157
6	*Evidence does not support a role*	21
–	*Uncategorized*	88

[1] https://gene.sfari.org/database/human-gene/, accessed 1 May 2018.

For this application we labeled as positive class the genes from SFARI categories 1, 2, 3 and 4, as well as those classified just as syndromic. We labeled as negatives the non mental genes used by Krishnan *et* al. [8], from which we removed the ones overlapping with genes from the positive class. All other genes in the network, including those from SFARI categories 5 to 6, are unlabeled and therefore are assigned a score of 0, according to Eq. 1.

For both labeled gene sets we converted the gene identifiers to the latest Hugo Gene Nomenclature Consortium (HGNC) symbols, discarding all genes for which there was no correspondence. This resulted in two sets with 739 positive and 1132 negative genes, which correspond to the $\mathcal{P}$ and $\mathcal{N}$ sets mentioned above.

3.3 Biological Network

As input network we used the STRING [13] (version 10.5) protein-protein interaction (PPI) database. The STRING database contains both experimental data and interactions inferred from text-mining the scientific literature, with an associated confidence score.

From the human subset of PPI interactions we converted all protein identifiers to the respective HGNC symbol using data obtained from the Ensembl BioMart [14], discarding those involving symbols for which conversion was not possible. Redundant interactions were removed, keeping the interaction with the highest confidence score.

We built a network with the entire set of edges, assigning the confidence score to edge weights. Only the largest connected component was selected and all other nodes were deleted. The final String-derived network contains 18003 genes and 5007158 edges.

3.4 Enrichment Analysis

To characterize the identified ASD-related gene sets we used enrichment analysis (hypergeometric test) for pathways from the Reactome database, using the *Enrichr* web Application Programming Interface[2]. All reported p-values are adjusted for multiple testing with the Benjamini-Hochberg correction method.

4 Results and Discussion

4.1 Performance Evaluation

To evaluate the capability of the method to identify disease-related genes we performed 10-fold cross-validation on the gene sets defined above. Generally, we observe that the method exhibits a good capability to classify genes in the held-out folds.

[2] http://amp.pharm.mssm.edu/Enrichr, accessed 1 Jun. 2018.

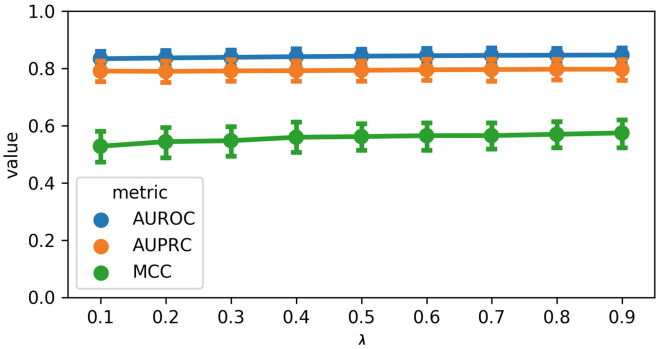

Fig. 1. Mean AUROC, AUPRC and MCC values from 10-fold cross validation for various values of lambda. Error bars represent 95% confidence intervals.

None of the evaluation metrics are very sensitive to the restart parameter (λ), as can be seen in Fig. 1. Maximum values for all metrics are obtained for $\lambda = 0.9$, with an AUROC values 0.85, an AUPRC of 0.8 and a MCC of 0.57.

For comparison we also tried applying the procedure to networks produced by setting edge cutoffs based on the interaction confidence, for values, 400 and 700, corresponding to medium and high confidence interactions, respectively (results not shown). Cross-validated mean AUROC values for these networks, with or without confidence scores as weights are substantially (∼5% and ∼15%, respectively) lower than for the full weighed network.

Treating the networks as weighed or binarizing interactions after applying cutoffs yields AUROC values differing by less than 1%. This indicates that the use of weighed networks has little impact on the performance of the method. In the following we analyze the scores obtained with $\lambda = 0.9$.

4.2 Selection and Characterization of Top Ranking Genes

The final score vector (f^{∞}) contains a score for each gene in the network, corresponding to the degree of association with the disease.

To assess the relevance of the top-ranking genes identified by the method we selected, from the cross-validation procedure, a threshold corresponding to the score of 95^{th} percentile. Genes with scores above the threshold are considered as strong candidates for association to ASD.

The composition of the set of top candidate genes is displayed in Fig. 2. As expected, the top ranking genes are constituted by genes in the SFARI categories 1 to 4 or Syndromic, which are defined beforehand as belonging to the positive class. In addition to those, the method identifies six genes (*SYN3*, *COP1*, *CBLN1*, *HTR2A*, *GABRB1*, *CLSTN3*) from SFARI category 5 (*Hypothesized but untested*).

Eight genes (*DLGAP3*, *GRM1*, *PPFIA1*, *SLC24A2*, *GRIK3*, *PHF8*, *PTPRT*, *CACNA1B*) added to the SFARI database but, as of now, not yet assigned to any category (*Unc* in Fig. 2), are also present in the candidate gene sets.

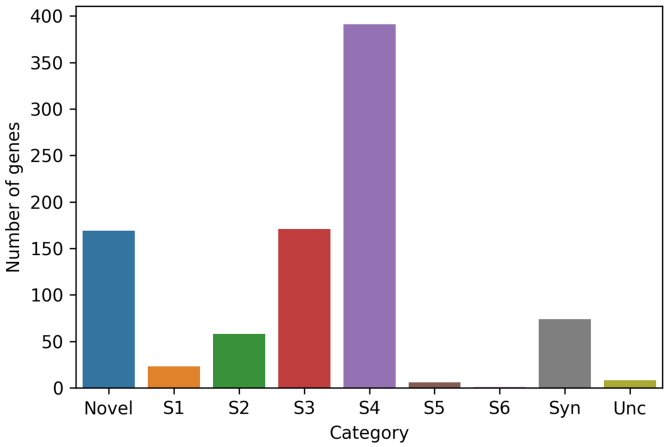

Fig. 2. Composition of the selected gene set. S1 to S6 correspond to SFARI gene categories 1 to 6, *Syn* denotes SFARI genes classified as Syndromic, *Unc* denotes genes in the SFARI gene list which are not yet categorized, *Novel* denotes genes not present in SFARI.

We consider the remaining 177 top-ranking genes as putative new candidate risk genes (*Novel* in Fig. 2). Of these, only some are directly implicated in human pathologies. The Online Mendelian Inheritance in Man (OMIM) database lists the *DLG3* gene as associated with *X-linked mental retardation-90* (MRX90, OMIM:300850), *CDH15* is associated to autosomal dominant mental retardation-3 (MRD3, OMIM:612580), *CHRNB2*, *CACNB4* and *CHRNA2* are associated with several types of nocturnal frontal lobe epilepsy and *CACNB4* is associated with type 5 episodic ataxia (EA5, OMIM:613855).

It is noteworthy that the method selects only one gene (*GRM8*) from SFARI category 6 (*Evidence does not support a role*), of which 21 are present in the network with initial weights set to zero. We take this as another indication of the accuracy of the method, as these genes have been shown not to be implicated with ASD and are not ranked as such.

For the subset of novel candidate genes we performed enrichment analysis to find significantly overrepresented biological pathways from the Reactome database. We report only highly significantly enriched pathways.

Results for pathway enrichment analysis (Table 2) show a significant enrichment in pathways related to synaptic transmission and calcium ion transport, with several other neurotransmitter-related pathways also being identified as significant.

The novel candidate risk gene set is enriched in genes involved in specific biological processes related to glutamate binding and AMPA receptor activity, both of which have been linked to ASD [3].

We also note a significant enrichment in genes related to the *cardiac conduction* pathway. These are genes of the *Voltage-dependent calcium channel*

Table 2. Subset of significantly enriched biological pathways (multiple correction adjusted p-values $< 1^{-10}$) for novel candidate risk genes.

Term name	Adjusted p-value
Transmission across Chemical Synapses (R-HSA-112315)	1.69×10^{-40}
Neuronal System (R-HSA-112316)	1.02×10^{-35}
Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell (R-HSA-112314)	3.31×10^{-21}
Phase 2 - plateau phase (R-HSA-5576893)	4.42×10^{-19}
Phase 0 - rapid depolarisation (R-HSA-5576892)	4.42×10^{-19}
Phase 1 - inactivation of fast Na+ channels (R-HSA-5576894)	5.92×10^{-19}
Dopamine Neurotransmitter Release Cycle (R-HSA-212676)	5.59×10^{-16}
Neurotransmitter Release Cycle (R-HSA-112310)	8.85×10^{-15}
Glutamate Binding, Activation of AMPA Receptors and Synaptic Plasticity (R-HSA-399721)	1.51×10^{-12}
Trafficking of AMPA receptors (R-HSA-399719)	1.51×10^{-12}
Cardiac conduction (R-HSA-5576891)	1.39×10^{-11}
Developmental Biology (R-HSA-1266738)	1.52×10^{-11}
Serotonin Neurotransmitter Release Cycle (R-HSA-181429)	1.79×10^{-11}
Norepinephrine Neurotransmitter Release Cycle (R-HSA-181430)	1.79×10^{-11}

(*CACN*) family. Calcium signaling disregulation is reported as being involved in ASD [3] and the genes in this pathway overlap with those in the AMPA-receptor-related pathways.

For confirmation we searched for *denovo* variants from genome or exome sequencing of ASD patients obtained from the Simons Simplex Collection (SSC) subset of denovo-db (version 1.6.1)[3]. Of the 177 candidate genes identified, 70% (123) were found to harbor at least one *denovo* variant, with more than half of these (88) having a variant in at least two samples.

5 Conclusion

We present a semi-supervised machine learning method for systematic discovery of novel candidate disease risk genes for ASD. The method consists on network propagation using a variant of the random walk with restart algorithm, combining information from biological networks and prior information on both positive (disease-related) and negative (disease-unrelated) genes.

Our results indicate that the method can classify held-out known disease genes in a cross-validation setting with very good performance (AUROC ~ 0.85, AUPRC ~ 0.8, MCC ~ 0.57).

[3] http://denovo-db.gs.washington.edu/denovo-db/, accessed 1 February 2019.

The identified putative novel candidate genes for association with ASD were analyzed for significant enrichment in gene sets from pathways from the Reactome database, and found to be significantly enriched for pathways related to synaptic transmission and ion transport and specific neurotransmitter-associated pathways previously shown to be associated with ASD.

A large percentage of the identified novel genes were found to be targeted by *denovo* variants in patients from the Simons Simplex Collection (SSC) dataset.

We expect that the outcomes of this work will contribute to the diagnosis and general understanding of the molecular mechanisms underlying ASD. Future developments will include the use of the derived gene ranking to prioritize genetic variants for gene and variant prioritization in the analysis of genomics data from ASD patients.

Acknowledgments. The authors would like to acknowledge the support by the UID/MULTI/04046/2019 centre grant from FCT, Portugal (to BioISI). A.M. is recipient of a fellowship from BioSys PhD programme (Ref SFRH/BD52485/2014) from FCT (Portugal). This work used the EGI infrastructure with the support of NCG-INGRID-PT (Portugal) and BIFI (Spain).

References

1. Elsabbagh, M., et al.: Global prevalence of autism and other pervasive developmental disorders. Autism Res. **5**(3), 160–179 (2012). https://doi.org/10.1002/aur.239
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th edn., Washington, DC (2013). https://doi.org/10.1176/appi.books.9780890425596
3. Chaste, P., Roeder, K., Devlin, B.: The Yin and Yang of autism genetics: how rare De Novo and common variations affect liability. Ann. Rev. Genomics Hum. Genet. **18**, 167–187 (2017). https://doi.org/10.1146/annurev-genom-083115-022647
4. Pinto, D., et al.: Functional impact of global rare copy number variation in autism spectrum disorders. Nature **466**(7304), 368 (2010). https://doi.org/10.1038/nature09146
5. Pinto, D., et al.: Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. **94**(5), 677–694 (2014). https://doi.org/10.1016/j.ajhg.2014.03.018
6. Vorstman, J., Parr, J., Moreno-De-Luca, D., Anney, R., Nurnberger Jr., J., et al.: Autism genetics: opportunities and challenges for clinical translation. Nat. Rev. Genet. **18**(6), 362–376 (2017). https://doi.org/10.1038/nrg.2017.4
7. Ansel, A., Rosenzweig Joshua, P., Zisman, P.D., Melamed, M., Gesundheit, B.: Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies. Front. Neurosci. **10**, 601 (2017). https://doi.org/10.3389/fnins.2016.00601
8. Krishnan, A., et al.: Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. **19**(11), 1454–1462 (2016). https://doi.org/10.1038/nn.4353
9. Asif, M., et al.: Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PloS one **13**(12), e0208626 (2018). https://doi.org/10.1371/journal.pone.0208626

10. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, pp. 321–328 (2004)
11. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. **18**(9), 551–562 (2017). https://doi.org/10.1038/nrg.2017.38
12. Mosca, E., et al.: Network diffusion-based prioritization of autism risk genes identifies significantly connected gene modules. Front. Genet. **8**, 129 (2017). https://doi.org/10.3389/fgene.2017.00129
13. Szklarczyk, D., et al.: The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. **45**(Database issue), D362–D368 (2017). https://doi.org/10.1093/nar/gkw937
14. Smedley, D., Haider, S., et al.: The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. **43**(1), W589–W598 (2015). https://doi.org/10.1093/nar/gkv350