BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Multi-Omics
Title	Explaining the molecular basis of a complex process through multi-omics data integration and machine learning: an application to microbiota-gut-brain axis in autism spectrum disorder.
All Authors	Alice Chiodi (1), Ada Sula (1), Andrea Manconi (1), Alessandra Mezzelani (1), Ettore Mosca (1)

Affiliation

(1) Institute of Biomedical Technologies, CNR, Segrate (Milano), Italy.

Motivation

Complex traits, like Autism Spectrum Disorders (ASD), are characterized by the interplay of multiple factors. The Gut-Brain Axis (GBA) impairment has been implicated in ASD, although with limited reproducibility. The study of a systemic complex process like the GBA in the context of a complex trait demands a multi-omics approach, which involves sampling different tissues for each of many subjects, and measurements with various platforms. Not surprisingly, scenarios like this pose various challenges to data integration and analysis. We present a bioinformatics approach to identify statistically relevant, biologically meaningful and reproducible molecular “players” underlying the GBA using data from genetics, brain transcriptomics, gut microbiome metagenomics and interactomics.

Methods

Multi-layer network diffusion was performed through the R package mND [1]. Input ASD data were obtained from the literature: 862 genes from SFARI [2], 2’925 brain and 2’472 microbial genes discriminating ASD from neurotypical subjects [3]. A molecular interaction network was obtained by integrating protein-protein interaction network from STRING [4] with published host-gut microbiome gene-gene interactions [5]. Pathway and orthology data were obtained from KEGG [6]. dmfind [7] was used to extract the gene networks enriched in modules of “affected genes” (see below). Machine learning models were defined using support vector machines, random forests, generalized linear models and neural network algorithms by means of Caret R package [8] on brain transcriptomics and shotgun metagenomic datasets (4 and 5 cohorts, respectively) [3]. Gene counts matrices were downloaded, and then filtered and normalized by using edgeR package [9]. ML models were built on each omics separately by using single and all cohorts together. In all cases, data were separated in training and test set following 7/3 ratio. Optimal model tuning was obtained through cross-validation, while their training and testing performances were assessed using accuracy.

Results

We assembled a gene-centered interactome that involves gene-gene host interactions and host-gut microbiota gene-gene interactions, resulting in a network of 19’000 genes and 184’000 interactions. We mapped the gene-level scores from each source of information (genomics, brain transcriptomics and gut microbiome metagenomics, “layers” in what follows), reflecting the association of each gene to ASD (shortly “affected genes”), on such “scaffold”. We then used multi-layer network diffusion to extract gene networks that connect the affected genes to one another. Importantly, these networks (of about ~500 genes) encompass genes whose association with ASD is due to susceptibility (genetics), gene expression in the brain of ASD subjects and significant occurrence in the gut microbiome of ASD subjects. These networks show a modular structure, reflecting the involvement of different molecular mechanisms associated with ASD. Notably, the pathways including microbial genes control the metabolism of amino acids involved in neurotransmission. We then assessed the ability of the genes found by network analysis in distinguishing ASD and control subjects in brain transcriptomics and gut metagenomics by means of machine learning models. We tested different approaches involving feature selection and extraction, through recursive feature elimination and principal component analysis. We obtained quite good accuracy values, ranging from 0.6 to 0.9, both in train and test sets, with random forests and support vector machines performing slightly better than other algorithms. These variations underline the challenges inherent to the analysis of these kinds of datasets, characterized by various sources of heterogeneity (e.g., samples size, batch effects, technological platforms). In conclusion, the use of network analysis allowed us to identify genes that potentially mediate the cross-talk between multiple biological contexts along the GBA. Moreover, these show interesting accuracy as biomarkers that characterize ASD subjects and their microbiota.

Info

Funding: This study has been supported by: European Union (NextGenerationEU), Italian NRRP project code IR0000031 - Strengthening BBMRI.it - CUP B53C22001820006; EU GEMMA-(825033); MUR “CNRBIOMICS”-(PIR01_00017).

References

[1] Di Nanni N, Gnocchi M, Moscatelli M, Milanesi L, Mosca E. Gene relevance based on multiple evidences in complex networks. Bioinformatics. 2020 Feb 1;36(3):865-71. Doi: 10.1093/bioinformatics/btz652

[2] Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. Doi: 10.1242/dmm.005439

[3] Morton JT, Jin DM, Mills RH, Shao Y, Rahman G, McDonald D, Zhu Q, Balaban M, Jiang Y, Cantrell K, Gonzalez A. Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles. Nature neuroscience. 2023 Jul;26(7):1208-17. Doi: 10.1038/s41593-023-01361-0
[4] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic acids research. 2023 Jan 6;51(D1):D638-46. Doi: 10.1093/nar/gkac1000
[5] Zhou H, Beltrán JF, Brito IL. Host-microbiome protein-protein interactions capture disease-relevant pathways. Genome Biology. 2022 Mar 4;23(1):72. Doi: 10.1186/s13059-022-02643-9
[6] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000 Jan 1;28(1):27-30. Doi: 10.1093/nar/28.1.27
[7] Bersanelli M, Mosca E, Remondini D, Castellani G, Milanesi L. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. Scientific Reports. 2016 Oct 12;6(1):34841. Doi: 10.1038/srep34841
[8] Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008 Nov 10;28:1-26. Doi: 10.18637/jss.v028.i05
[9] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics. 2010 Jan 1;26(1):139-40. Doi: 10.1093/bioinformatics/btp616

filename	-
Figure	
-	
Availability	-

Dissemination Material

Social
-
Summary
-

Corresponding Author

Name, Surname	Alice, Chiodi
Email	alice.chiodi@itb.cnr.it
Submitted on	30.04.2024