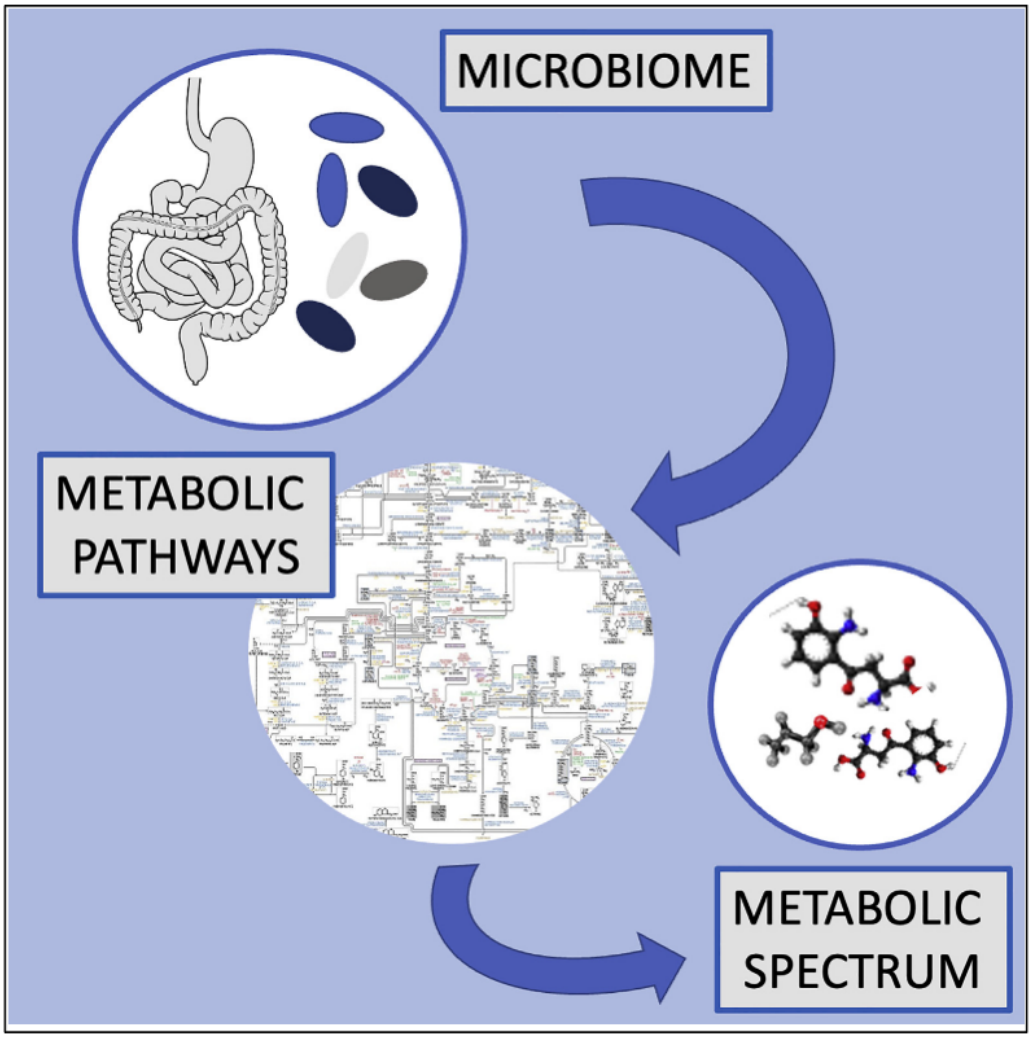
iScience

CellPress
OPEN ACCESS

Article

Quantitative analysis of disease-related metabolic dysregulation of human microbiota



Maria Rita Fumagalli, Stella Maria Saro, Matteo Tajana, Stefano Zapperi, Caterina A.M. La Porta

caterina.laporta@unimi.it

Highlights

An algorithm to estimate the metabolic spectrum from the microbiome is introduced

The method uses microbial metabolic pathways stoichiometry to estimate metabolites

The method is applied to data from autism spectrum disorder and multiple sclerosis

Fumagalli et al., iScience 26, 105868
January 20, 2023
https://doi.org/10.1016/j.isci.2022.105868

iScience

CellPress
OPEN ACCESS

Article

Quantitative analysis of disease-related metabolic dysregulation of human microbiota

Maria Rita Fumagalli,[1,2] Stella Maria Saro,[3] Matteo Tajana,[3] Stefano Zapperi,[3,4] and Caterina A.M. La Porta[1,2,5,*]

SUMMARY

The metabolic activity of all the micro-organism composing the human microbiome interacts with the host metabolism contributing to human health and disease in a way that is not fully understood. Here, we introduce STELLA, a computational method to derive the spectrum of metabolites associated with the microbiome of an individual. STELLA integrates known information on metabolic pathways associated with each bacterial species and extracts from these the list of metabolic products of each singular reaction by means of automatic text analysis. By comparing the result obtained on a single subject with the metabolic profile data of a control set of healthy subjects, we are able to identify individual metabolic alterations. To illustrate the method, we present applications to autism spectrum disorder and multiple sclerosis.

INTRODUCTION

Gut microbiota is represented by a diverse and dynamic population of microorganisms, including bacteria, archaea, and eukarya, that live in the gastrointestinal tract of an individual and develop a deeply complex relationship of mutually advantageous exchanges with the host organism.[1,2] Gut microbiota-host interaction involves the production and consumption of metabolites. In particular, gut microbiota is responsible for the synthesis of biomolecules, such as vitamins and enzymatic proteins, that the host cannot produce.[1,2] The study of gut microbiota has rapidly evolved in the last few years, driven by the development of new techniques for metagenome sequencing[3] and the increasing evidence showing the existence of a strong relationship between the equilibrium of the microbiota and the health of the host.[4]

The microbiota, by contributing to the metabolic activity of an individual, exerts a marked influence on its physiological and pathological conditions, playing an important role in processes such as regulation and development of host immunity, digestion, and the integrity of the specific environment they colonize.[1,5,6] Thus, alterations in the microbiota can lead to major consequences for the health of an individual. Microbiota dysregulation has been indeed observed in different diseases including inflammatory bowel syndrome and obesity, but also disorders related to the CNS.[2,4,5,7]

Recent studies have revealed that the intriguing interactions between the gut and the CNS, the so-called gut-brain axis, are modulated by the complex communication network of the microbiota.[2,8] For this reason, analyzing the composition of the gut microbiome is of pivotal interest in the study of CNS-related pathogenesis.[2] Furthermore, recently proposed strategies to re-balance dysbiosis appear to be effective in the treatment of different pathologies.[2,4]

The microbiome is usually quantified in terms of operational taxonomic units (OTUs), an operational definition used to classify groups of closely related genomic sequences, which may refer to species, genus, or class. The diversity in a microbial community is then a widely used metagenomic marker of metabolic disorders and pathological conditions. A reduction of microbiome diversity has been linked to a number of diseases, such as inflammatory bowel disease, obesity and metabolic syndromes, and HIV.[7,8] For other pathologies linked to dysbiosis, such as autism spectrum disorder (ASD) and multiple scelrosis (MS), microbiome alterations cannot easily be expressed in terms of loss (or gain) of diversity and it is not possible to observe major global shifts in bacterial community composition.[8,9]

Here, we introduce STELLA, a computational strategy to investigate microbiome dysregulation that goes beyond a mere evaluation of its composition in terms of OTUs and is based instead on the quantification of

[1]Center for Complexity and Biosystems, Department of Environmental Science and Policy, University of Milan, via Celoria 26, 20133 Milano, Italy

[2]CNR - Consiglio Nazionale delle Ricerche, Istituto di Biofisica, via De Marini 6, 16149 Genova, Italy

[3]Center for Complexity and Biosystems, Department of Physics, University of Milan, Via Celoria 16, 20133 Milano, Italy

[4]CNR - Consiglio Nazionale delle Ricerche, Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia, Via R. Cozzi 53, 20125 Milano, Italy

[5]Lead contact

*Correspondence: caterina.laporta@unimi.it

https://doi.org/10.1016/j.isci.2022.105868

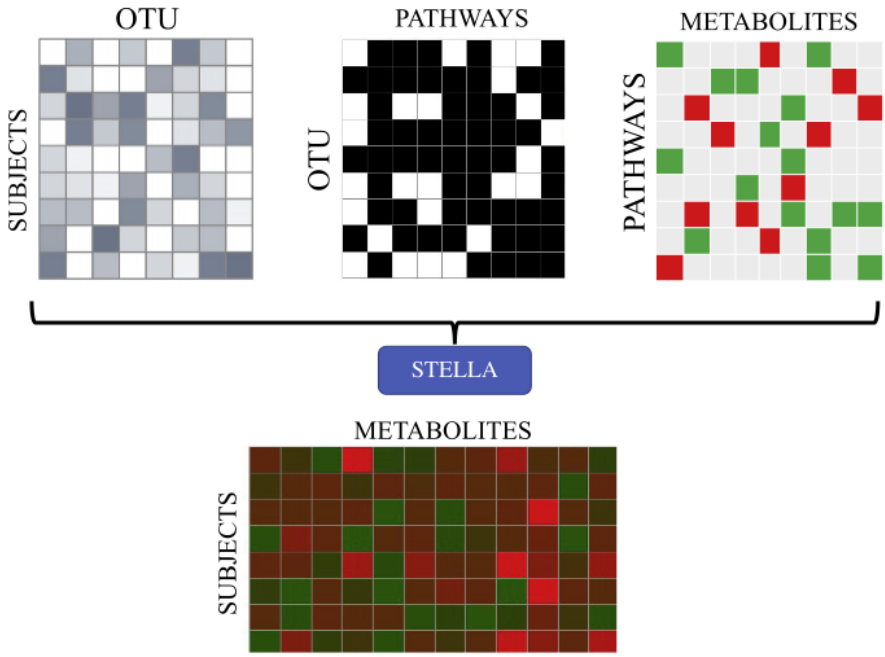


Figure 1. Schematic of the algorithm

The OTU abundances matrix obtained from microbiota sequencing experiments are combined with information on the metabolic pathways present in each OTU, retrieved from the Macadam database, and with the stoichiometry of metabolites produced and consumed in each pathway, obtained from the Metacyc database. All the information is combined to obtain a metabolite-patient matrix for each experimental dataset.

metabolites associated with a given microbiomic profile. To illustrate our method, we report the results obtained from the analysis of different datasets related two CNS-related disorders: ASD and MS. In particular, we analyze three 16S rRNA sequencing datasets obtained from autistics and healthy subjects[10–12] and two datasets referred to MS.[9,13] We apply STELLA specifically to 16S rRNA sequencing data from fecal samples, but the algorithm could in principle be applied to any kind of microbiome dataset.

RESULTS

Algorithm construction and validation

STELLA is a reference-based method that associates to each patient a vector whose elements are a proxy for the amount of metabolites produced and consumed by its microbiota (see Figure 1). The algorithm uses the MACADAM[14] and METACYC[15] databases to retrieve the metabolic pathways present in microbiota OTUs and the metabolites involved in the reactions composing the pathway. We consider the whole set of metabolic pathways available for each OTU as active, and take into consideration both stoichiometry and reaction directionality when associating a production/consumption score to each metabolite. Next, we merge this score matrix with information on OTUs abundances in the patients, and weigh the contribution of each pathway accordingly to OTUs abundance. Hence, from a dataset containing OTUs abundances, the STELLA algorithm allows us to estimate the associated metabolite concentrations. Details about the algorithm can be found in the STAR methods section.

In order to quantify the quality of the predictions of STELLA, we compared its output metabolites with the experimental data reported by Kang et al.[10] This database contains 59 key metabolites that display a significantly different concentration between children with ASD and neurotypical patients. In particular, we benchmarked the experimental data by assigning a *z*-score to each metabolite, irrespectively of the status of the patient (see STAR Methods). A summary of the experimentally obtained *z*-scores and of the predictions of STELLA are reported in Figures 2A and 2B, respectively.

STELLA outputs whether a specific compound is likely to be produced or consumed for a given set of OTUs but does not allow to estimate its concentration. For this reason, we compared the sign of the computed *z*-scores with the sign of the corresponding experimental *z*-scores. To assess the quality of predictions

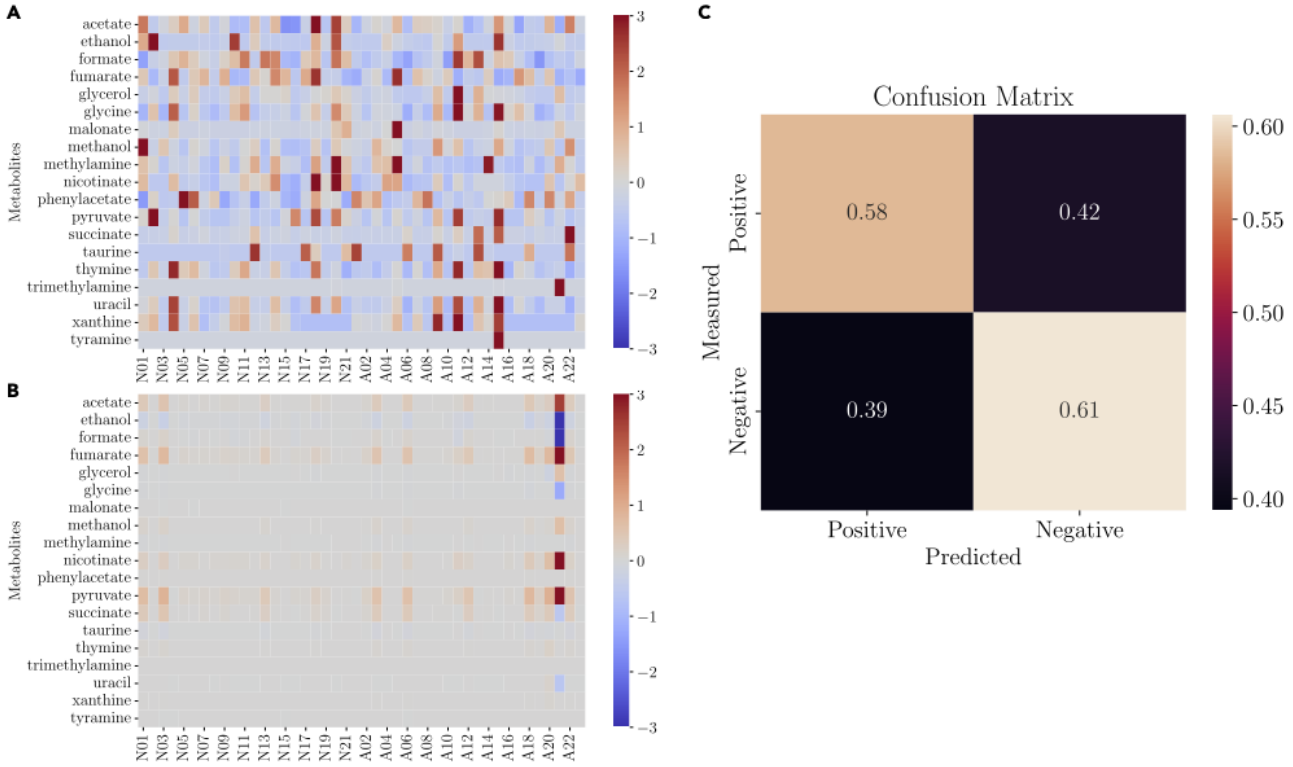


Figure 2. Algorithm validation

(A) Heatmap of the *z*-scores of the metabolites recorded in[10] and (B) the corresponding predictions made by the STELLA algorithm. (C) Confusion matrix for the algorithm predictions.

made by STELLA, we evaluate the confusion matrix (Figure 2C) and the corrsponding F_β-scores, a widely used measure of the accuracy of a specific test. Depending on the value of β, the F_β-score weights differently recall and precision of the test so that $\beta > 1$ weights more recall while $\beta < 1$ precision. Previous studies used the F_1 score to benchmark metabolite prediction algorithms. In particular,[16] computed the F_1 score for the prediction of metabolite occurrences using different algorithms, comparing MIMOSA,[17] Mangosteen,[16] and MelonnPan.[18] In our case, we obtain the following values for STELLA: $F_1 = 0.67$, $F_2 = 0.73$, and $F_{1/2} = 0.61$. The results show that the F_1 score we obtain is comparable to the ones reported in previous studies.[16]

Dataset merging and batch effect removal

In the following, we provide illustrative examples of the results that can be obtained when the STELLA algorithm is used to distinguish among groups with different health conditions. To this end, we apply STELLA to microbiome datasets related to ASD and MS that we collected from the literature. When more than a single dataset of OTUs is available for the same disease, it is useful to combine the data in order to increase the statistics over the patients. We perform this step in the case of the ASD data reported in the study by Kang et al.[10] and in the study by Kong et al.[11] It is, however, important to verify that merging datasets does not introduce spurious differences between the groups (or batches) that are due to experimental procedures and are unrelated to the biological process under investigation. Batch effects can indeed be observed in the patient correlation matrix, as reported in Figure 3A. To remove these spurious correlations, we employ a method based on singular value decomposition (SVD) as discussed by Font-Clos et al.[19] In particular, we apply a single-step SVD correction on the OTUs abundance matrix obtained merging the datasets reported in the study by Kang et al.[10] and in the study by Kong et al.[11] Before batch effect removal, a hierarchical clustering algorithm highlights a spurious separation between the two batches (Figure 3A) which disappears after batch effect removal (Figure 3B). The result of this procedure can also be observed through the principal component analysis (PCA). The first principal component in the original merged data discriminates between the two datasets (Figure 3C), but after application of the batch effect removal algorithm the two datasets are more mixed (Figure 3D).

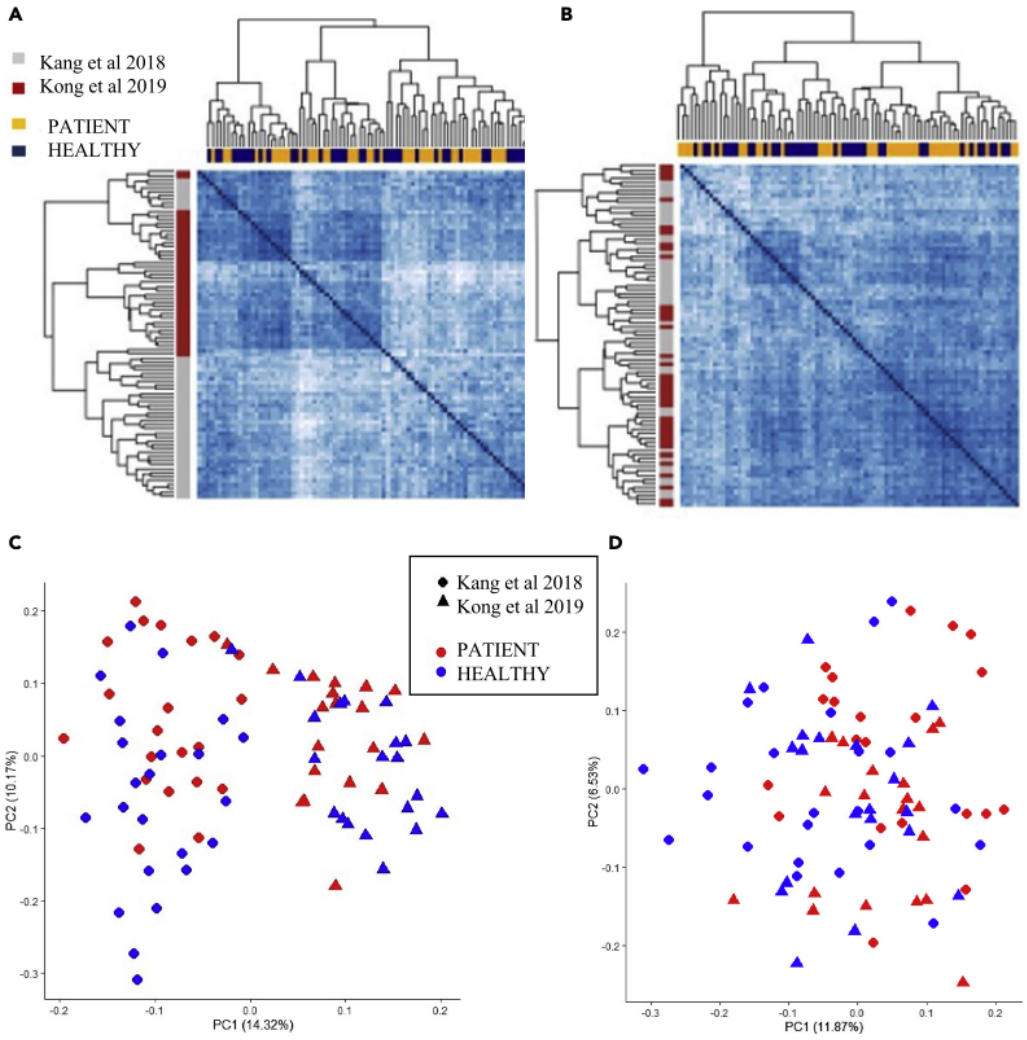


Figure 3. Batch effect removal

(A and B) Correlation matrix of OTUs abundances obtained merging the datasets in the study by Kang et al.[10] and Kong et al.[11] before (A) and after (B) batch effect removal. Hierarchical clustering on correlation matrix shows that differences between the two datasets (gray, red) is predominant in the merged dataset, and the two major blocks disappear after batch removal. Hierachical clustering does not allow us to distinguish between patients and healthy controls (yellow, blue) suggesting a more elaborated procedure is needed.

(C and D) Panels show the data from the intersection of data from the study by Kang et al.[10] and Kong et al.[11] projected onto the first two principal components and divided into healthy controls (red) and autistic patients (blue) before (C) and after (D) one-step batch removal. In panel (C), the variance between the data from different datasets is higher than the distance between the autistic patients and the healthy controls. Different symbols represent the two studies (circles, triangles as in legend).

Stepwise feature selection for ASD and MS

The results obtained from the PCA (Figures 3C and 3D) and hierarchical clustering (3 AB) suggest that in order to gain information on micriobiota dysregulation in CNS-related pathologies, we need a classification approach that is able to take into account more than a single variable at a time and goes beyond a simple PCA projection. We thus perform stepwise feature selection through linear discriminant analysis (LDA), as illustrated in Figure 4A. In particular, we perform a stepwise feature selection on the data by training an LDA model with 10-fold cross-validation resampling the training set and the validation set from the complete dataset, selecting the model that maximizes the correctness rate. The procedure was repeated at least 100 times over reshuffled matrices in order to evaluate the robustness of the model obtained (see Figure 4B). We imposed a maximum of 15 variables to use as features and verifying that, for

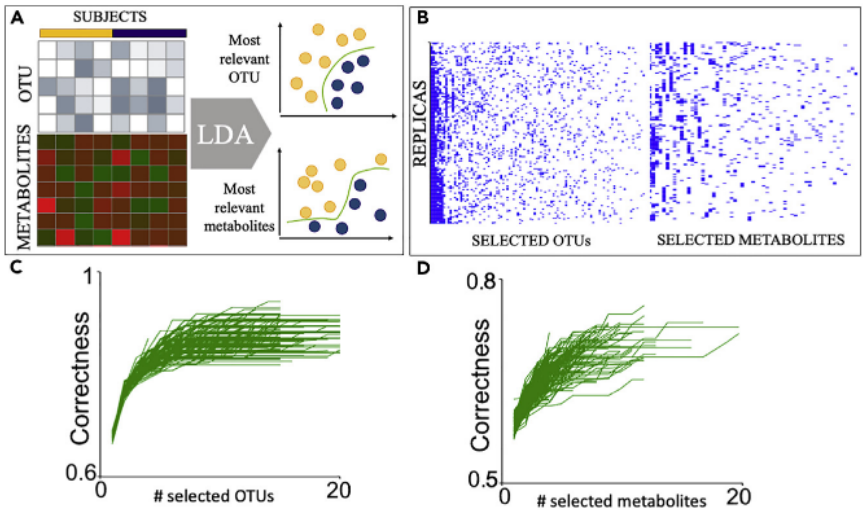


Figure 4. LDA algorithm

(A) Schematic of the LDA algorithm. After removal of highly correlated OTUs and metabolites, LDA feature selection is applied in order to obtain an optimal set of OTUs and metabolites to discriminate between patients (yellow circles) and healthy controls (blue circles).

(B) Heatmaps show the ensemble of selected OTUs and metabolites for each replica of LDA feature selection for the combined datasets in the study by Kang et al.[10] and Kong et al.[11] after batch removal. The presence of a specific OTU or metabolite in a given replica is represented in blue and its absence in white and ordered using hierarchical clustering.

(C and D) Performance of the models obtained from LDA feature selection. Correctness rate is evaluated after the inclusion of each new (C) OTU or (D) metabolite. The plots show, for each replica and in chronological order of selection, the increasing correctness rate of the model after the addition of every taxonomy.

the analyzed datasets, it is sufficient to reach the performance plateau. We apply our strategy to both OTUs abundances matrix and metabolite matrix (see Figures 4C and 4D) in order to determine the specific sets of taxonomies and metabolites that most differentiate healthy and pathological profiles. To reduce the computational effort and the redundancy of the dataset, the LDA procedure was applied on reduced matrices comprising a subset of uncorrelated OTUs and metabolites (see STAR methods).

Figure 5 summarizes the results obtained through LDA feature selection for ASD, while results for MS are presented in Figure 6. The complete lists of OTUs and metabolites obtained with our method are reported in Data S1. Results reported in Figure 5 have been obtained for the combined datasets in the study by Kang et al.[10] and Kong et al.[11] after batch removal. For both diseases, we report the heatmaps representing the selected taxonomies for each replica of the LDA feature selection and the corresponding matabolic compounds. The presence of a specific taxonomy in a given iteration is represented in blue and its absence in white. Hierarchical clustering shows that a small number of OTUs are choosen in diffrerent replicas, representing the core of OTUs and metabolites that discrimiate between healthy subjects and patients according to LDA analysis. We also report the co-occurrence matrix representing the number of replicas in which two compounds are both selected as significant by the LDA. Based on these results, we can then reconstruct the reverse metabolic network. Figures 5B and 6B show the five most frequently selected metabolites during the LDA analysis of the metabolite-host matrix and the pathways in which they are mapped. For a specific metabolite (2-methyladenine), we also report in Figure 5C the pathway-compound network obtained including its correlated compounds.

DISCUSSION

In this paper, we introduced STELLA, an algorithm to infer the metabolic spectrum associated with a given microbiome profile characterized by the relative abundance of OTUs. To illustrate the relevance of our approach, we considered microbiome profiles of patients diagnosed with ASD and MS. Using these data, we inferred the metabolic network and estimated the set of most relevant metabolites associated with these pathological conditions. Among the metabolites/pathways resulting from dysbiosis in ASD, the LDA highlights the synthesis of some amino acids such as tryptophane, the biosynthesis of purine, and cobalamin. According to our analysis, the microbiome of patients with ASD is associated with

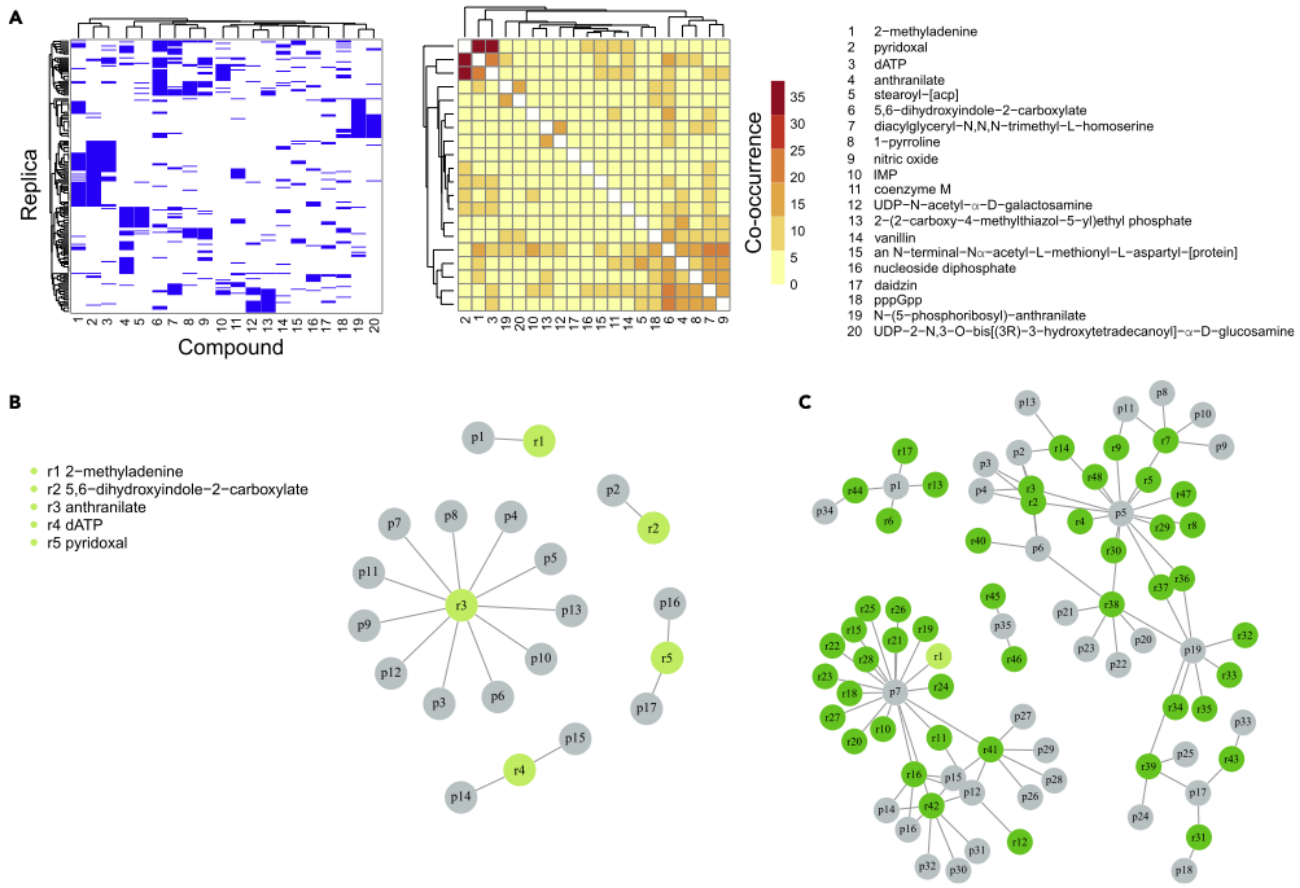


Figure 5. LDA feature selection results for ASD

Panel (A) show heatmaps representing the selected taxonomies for each replica of LDA feature selection reordered using hierarchical clustering and co-occurrence matrix representing the number of replica two compounds are selected as significant by LDA. The presence of a specific taxonomy in a given iteration is represented in blue and its absence in white. Panels (B and C) Illustrative reconstruction of reverse metabolic network. Plot shows the five most frequently selected metabolites during LDA analysis of metabolite-host matrix and the pathways in which they are mapped (B). For a specific metabolite (2-methyladenine), we report the obtained pathway-compound network (C). The network includes 2-methyladenine and its correlated metabolites (r1-r47) and the related pathways (p1-p35). The complete list of labels is reported Data S1. Lines represent the links between the metabolites (green, light green) and the pathways (gray). For all the panels, data are reported for the combined datasets from the study by Kang et al.[10] and Kong et al.[11] after batch removal.

alterations in the amino acids regulating the sleep-wake cycle and the mood, notably tryptophane, which helps making melatonin and serotonin.[20]

In the case of MS, our analysis suggests the involvement of pathway-related energy metabolism, including glicolysis and gluconeogenesis. A perturbation of the glucose metabolism in MS is confirmed by recent evidence showing a connection between MS and mitochondrial aberrations and impaired glucose metabolism.[21] Furthermore, MS was shown to be associated with a strong expression of lactate dehydrogenase A, which converts pyruvate to lactate, within the perivascular cuff of postcapillary venules supporting the transmigration of brain-infiltrating macrophages.[22]

Interest in developing methods to identify relevant molecular compound identities from metagenomes has grown in recent years. Different methods have been developed, such as MangoSteen,[16] MIMOSA,[17] and MelonPann.[18] MangoSteen uses an approach similar to the one used in the present paper, constructing the connections from sequencing data to compounds. Contrary to our approach, the method considers all the metabolites linked to dysregulated taxonomies as dysregulated. This assumption can have a huge impact on the results, since a large fraction of metabolites are produced by many

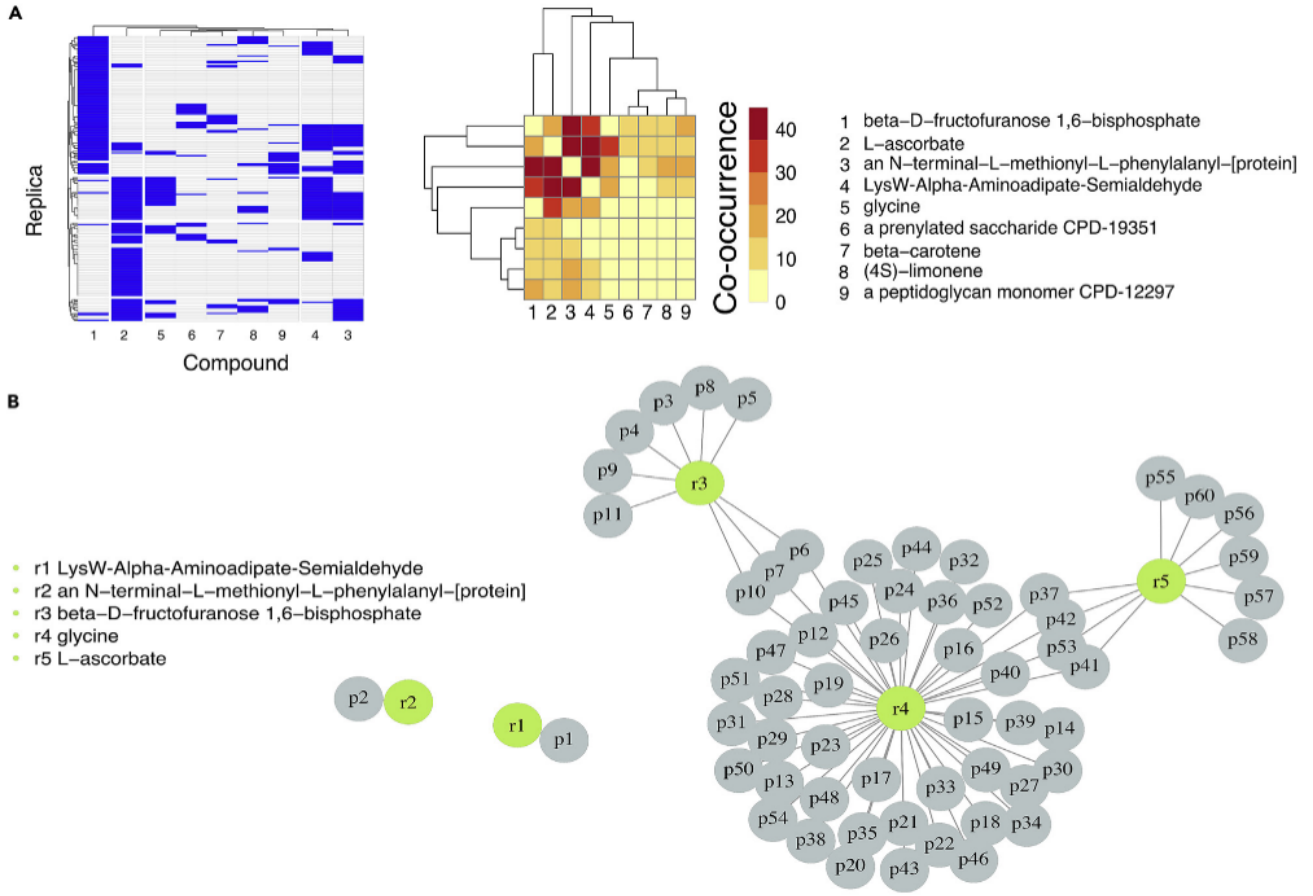


Figure 6. LDA feature selection result for MS

(A) Heatmaps representing the selected taxonomies for each replica of LDA feature selection reordered using hierarchical clustering and co-occurrence matrix representing the number of replica two compounds are selected as significant by LDA. The presence of a specific taxonomy in a given iteration is represented in blue and its absence in white.

(B) Illustrative reconstruction of reverse metabolic network. Plot shows the five most frequently selected metabolites during LDA analysis of metabolite-host matrix and the pathways in which they are mapped. Lines represent the links between the metabolites (green, light green) and the pathways (gray) For all the panels, data are reported for the dataset in the study by Cekanaviciute et al.[9] Metabolites names are reported in the figure, while full pathways names are only reported in Data S1.

different taxonomies and most of them could be not dysregulated, contributing to the balance of the global metabolites abundances. Furthermore, our method takes into consideration the directionality of the reactions that is neglected by MangoSteen.

MIMOSA[17] is a method based on predicted relative metabolomic turnover, originally developed by Larsen et al.[23] MIMOSA takes into account stoichiometry and the direction of reactions, neglecting reversible reactions. The focus of the algorithm is on enzymes, the transformations they catalyze, and the link between genes and the reactions and metabolites that are annotated in KEGG.[24] A crucial step of the MIMOSA algorithm is the removal of common metabolites, produced by more than 30 genes, and the normalization of the stoichiometric matrix in order to obtain the relative contribution of each gene to the production/depletion of each metabolite. The major difference between our method and MIMOSA is that we consider the rate of a reaction as proportional to the abundance of OTUs, while MIMOSA focuses on the relative abundance of single genes into the whole metagenome.

Our approach and both MangoSteen and MIMOSA are reference-based methods and, as such, the results are strongly influenced by the completeness of the considered databases. On the contrary, machine learning-based methods such as MelonnPan[18] use elastic net regularization to identify which features are predictive for the presence of a given metabolite, taking as input both transcriptomic and metabolomic

data. MelonnPan captures metagenome-metabolome associations in a data-driven manner and does not rely on microbial biochemical annotation. This could lead to a significantly higher prediction accuracy than reference-based methods when involved species are not well annotated, and to the appearance of interactions that do not exist in reference databases. On the other hand, the efficiency of MelonnPan relies on the quality of the training dataset and needs a specific metabolomic input.

In conclusion, we have introduced an algorithm to infer the metabolic profile from a given microbiome and illustrated its application using data related to two pathologies. Our strategy could help in identifying possible new targets to make traditional therapies more effective and successful.

Limitations of the study

The main limitation of our approach derives from the incompleteness of the databases used to obtain metabolic pathways present in microbiota OTUs (i.e. Macadam) and the metabolites involved in the reactions composing the pathway (i.e. Metacyc). Annotation of sequencing and OTUs abundances are affected by current knowledge, but databases are continuously updated, potentially allowing for improved predictions in the future.

The STELLA algorithm is by its nature semi-quantitative, since it is based on stoichiometric information on the reactions that are potentially available to a given OTU. The fact that OTUs might have the capability to perform a specific reaction does not imply that the reaction will be used in a particular context. This limitation can be overcome by methods based on metabolic modeling.[25]

Additional limitations of our methods come from the fact that the metabolites observed experimentally are not only produced by the considered OTUs but could also be originated by differences in patients' diet. While these problems are common to all the reference-based algorithms, the use of a set of OTUs instead of a single taxonomy to describe a pathological condition could reduce the error due to an incomplete annotation of the micriobiome. Additionally, a possible limitation of the LDA approach used here to test STELLA algorithm is the arbitrary choice of a correlation threshold. However, we tested that different thresholds and representative correlates do not affect the results presented here. Moreover, this limitation does not affect the performance of the STELLA algorithm and is needed only to reduce computational effort. Despite these limitations, our algorithm provides an estimate of the contribution of the different OTUs to the metabolic profile that is in partial agreement with experimental data.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Data and code availability
- METHOD DETAILS
  - The STELLA algorithm
  - Data processing
  - Batch effect removal
  - Principal component analysis
  - Step-wise feature selection
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105868.

ACKNOWLEDGMENTS

M.T. was supported by a young scientist fellowship from the UniMi GSA-IDEA project.

AUTHOR CONTRIBUTIONS

M.R.F., S.M.S., and M.T. wrote the code and analyzed data. C.A.M.L.P. and S.Z. designed and coordinated the study. C.A.M.L.P. and S.Z. wrote the paper with contributions from M.R.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 12, 2022
Revised: December 7, 2022
Accepted: December 20, 2022
Published: January 20, 2023

REFERENCES

1. Thursby, E., and Juge, N. (2017). Introduction to the human gut microbiota. Biochem. J. *474*, 1823–1836. https://doi.org/10.1042/BCJ20160510.

2. Dixit, K., Chaudhari, D., Dhotre, D., Shouche, Y., and Saroj, S. (2021). Restoration of dysbiotic human gut microbiome for homeostasis. Life Sci. *278*, 119622. https://doi.org/10.1016/j.lfs.2021.119622.

3. Quigley, E.M. (2013). Gut bacteria in health and disease. Gastroenterol. Hepatol. *9*, 560.

4. Sartor, R.B., and Wu, G.D. (2017). Roles for intestinal bacteria, viruses, and fungi in pathogenesis of inflammatory bowel diseases and therapeutic approaches. Gastroenterology *152*, 327–339.e4. https://doi.org/10.1053/j.gastro.2016.10.012.

5. Hughes, H.K., Rose, D., Ashwood, P., and Hughes, H. (2018). The gut microbiota and dysbiosis in autism spectrum disorders. Curr. Neurol. Neurosci. Rep. *18*, 81. https://doi.org/10.1007/s11910-018-0887-6.

6. Marques, T.M., Wall, R., Ross, R.P., Fitzgerald, G.F., Ryan, C.A., and Stanton, C. (2010). Programming infant gut microbiota: influence of dietary and environmental factors. Curr. Opin. Biotechnol. *21*, 149–156. https://doi.org/10.1016/j.copbio.2010.03.020.

7. Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546. https://doi.org/10.1038/nature12506.

8. Plassais, J., Gbikpi-Benissan, G., Figarol, M., Scheperjans, F., Gorochov, G., Derkinderen, P., and Cervino, A.C.L. (2021). Gut microbiome alpha-diversity is not a marker of Parkinson's disease and multiple sclerosis. Brain Commun. *3*, fcab113. https://doi.org/10.1093/braincomms/fcab113.

9. Cekanaviciute, E., Yoo, B.B., Runia, T.F., Debelius, J.W., Singh, S., Nelson, C.A., Kanner, R., Bencosme, Y., Lee, Y.K., Hauser, S.L., et al. (2017). Gut bacteria from multiple sclerosis patients modulate human t cells and exacerbate symptoms in mouse models. Proc. Natl. Acad. Sci. USA *114*, 10713 10718. https://doi.org/10.1073/pnas.1711235114.

10. Kang, D.-W., Ilhan, Z.E., Isern, N.G., Hoyt, D.W., Howsmon, D.P., Shaffer, M., Lozupone, C.A., Hahn, J., Adams, J.B., and Krajmalnik-Brown, R. (2018). Differences in fecal microbial metabolites and microbiota of children with autism spectrum disorders. Anaerobe *49*, 121–131. https://doi.org/10.1016/j.anaerobe.2017.12.007.

11. Kong, X., Liu, J., Cetinbas, M., Sadreyev, R., Koh, M., Huang, H., Adeseye, A., He, P., Zhu, J., Russell, H., Hobbie, C., et al. (2019). New and preliminary evidence on altered oral and gut microbiota in individuals with autism spectrum disorder (asd): implications for asd diagnosis and subtyping based on microbial biomarkers. Nutrients *11*, 2128. https://doi.org/10.3390/nu11092128.

12. Zhang, M., Ma, W., Zhang, J., He, Y., and Wang, J. (2018). Analysis of gut microbiota profiles and microbe-disease associations in children with autism spectrum disorders in China. Sci. Rep. *8*, 13981. https://doi.org/10.1038/s41598-018-32219-2.

13. Cekanaviciute, E., Pröbstel, A.K., Thomann, A., Runia, T.F., Casaccia, P., Katz Sand, I., Crabtree, E., Singh, S., Morrissey, J., Barba, P., et al. (2018). Multiple sclerosis-associated changes in the composition and immune functions of spore-forming bacteria. mSystems *3*, e00083-18.

14. Le Boulch, M., Déhais, P., Combes, S., and Pascal, G. (2019). The MACADAM database: a MetaboliC pathways Database for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. Database *2019*, baz049. https://doi.org/10.1093/database/baz049.

15. Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y., and Karp, P.D. (2004). MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. *32*, D438–D442. https://doi.org/10.1093/nar/gkh100.

16. Yin, X., Altman, T., Rutherford, E., West, K.A., Wu, Y., Choi, J., Beck, P.L., Kaplan, G.G., Dabbagh, K., DeSantis, T.Z., et al. (2020). A comparative evaluation of tools to predict metabolite profiles from microbiome sequencing data. Front. Microbiol. *11*, 595910. https://doi.org/10.3389/fmicb.2020.595910.

17. Noecker, C., Eng, A., Srinivasan, S., Theriot, C.M., Young, V.B., Jansson, J.K., Fredricks, D.N., Borenstein, E., and Sanchez, L.M. (2016). Metabolic model based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. mSystems *1*, e00013-15. https://doi.org/10.1128/mSystems.00013-15.

18. Mallick, H., Franzosa, E.A., Mclver, L.J., Banerjee, S., Sirota Madi, A., Kostic, A.D., Clish, C.B., Vlamakis, H., Xavier, R.J., and Huttenhower, C. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat. Commun. *10*, 3136–3146. https://doi.org/10.1038/s41467-019-10927-1.

19. Font-Clos, F., Zapperi, S., and La Porta, C.A.M. (2017). Integrative analysis of pathway deregulation in obesity. NPJ Syst. Biol. Appl. *3*, 18. https://doi.org/10.1038/s41540-017-0018-z.

20. Paredes, S.D., Barriga, C., Reiter, R.J., and Rodríguez, A.B. (2009). Assessment of the potential role of tryptophan as the precursor of serotonin and melatonin for the aged sleep-wake cycle and immune function: Streptopelia risoria as a model. Int. J. Tryptophan Res. *2*, 23–36.

21. Mathur, D., López-Rodas, G., Casanova, B., and Marti, M.B. (2014). Perturbed glucose metabolism: insights into multiple sclerosis pathogenesis. Front. Neurol. *5*, 250.

22. Kaushik, D.K., Bhattacharya, A., Mirzaei, R., Rawji, K.S., Ahn, Y., Rho, J.M., Yong, V.W., et al. (2019). Enhanced glycolytic metabolism supports transmigration of brain-infiltrating macrophages in multiple sclerosis. J. Clin. Invest. *129*, 3277–3292.

23. Larsen, P.E., Collart, F.R., Field, D., Meyer, F., Keegan, K.P., Henry, C.S., McGrath, J., Quinn, J., and Gilbert, J.A. (2011). Predicted

relative metabolomic turnover (prmt): determining metabolic turnover from a coastal marine metagenomic dataset. Microb. Inform. Exp. *1*, 4. https://doi.org/10.1186/2042-5783-1-4.

24. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

25. Baldini, F., Heinken, A., Heirendt, L., Magnusdottir, S., Fleming, R.M.T., and Thiele, I. (2019). The microbiome modeling toolbox: from microbial interactions to personalized microbial communities. Bioinformatics *35*, 2332–2334.

26. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336. https://doi.org/10.1038/nmeth.f.303.

27. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. Appl. Environ. Microbiol. *72*, 5069–5072. https://doi.org/10.1128/AEM.03006-05.

28. R Core Team (2021). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
V2-V3 regions of 16S rRNA genes for autism (23 patients, 21 healthy subjects)	Supplementary material	https://doi.org/10.1016/j.anaerobe.2017.12.007
V2-V3 regions of 16S rRNA genes for autism (20 patients, 19 healthy subjects)	Supplementary material	https://doi.org/10.3390/nu11092128
V2-V3 regions of 16S rRNA genes for autism (35 patients, 6 healthy subjects)	Supplementary material	https://doi.org/10.1038/s41598-018-32219-2
V4 regions of 16S rRNA genes for multiple sclerosis (24 patients, 25 healthy subjects)	Dryad repository https://doi.org/10.7272/Q6FB5136	https://doi.org/10.1128/mSystems.00083-18
V4 regions of 16S rRNA genes for multiple sclerosis (71 patients, 71 healthy subjects)	Dryad repository https://doi.org/10.7272/Q6WQ01ZB	https://doi.org/10.1073/pnas.1711235114
Software and algorithms		
STELLA algorithm	Zenodo	https://doi.org/10.5281/zenodo.7436739
Function stepclass from the R package klar (v.4.0)	R software	https://www.R-project.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Caterina A. M. La Porta (caterina.laporta@unimi.it).

Data and code availability

- We considered three datasets of microbiome data related to ASD[10–12] and two related to MS.[9,13] Details on the considered databases are reported in key resources table. We report the list of OTUs and metabolites selected by LDA in Data S1.
- All original code has been deposited in Zenodo and https://github.com/ComplexityBiosystems/STELLA and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The STELLA algorithm

Information on the metabolic pathways present in a given taxonomy were obtained from the metabolic pathways database for microbial taxonomic groups (Macadam).[14] The Metacyc database[15] was used to retrieve, for all the reactions involved in a given pathway, the metabolites synthetized and consumed as well as their stoichiometry and direction of the reactions. When OTU perfect match was missing in the MACADAM database, we retrieved data from the next-higher taxonomic order, up to family level. In order to avoid to include unrelated pathways, if an OTU corresponding to a family is missing from the MACADAM database, we neglect it. In the case of a species, when the corresponding genus is available in OTUs abundance matrix, we sum up all the relative abundances.

Given a dataset containing OTUs abundances in a set of hosts (patient and control subjects), it is possible to build a patient-pathway matrix P containing a host-specific score associated with each pathway j in Macadam. The matrix is defined by

$$P_{hj} = \sum_k a_{hk} b_{kj} \qquad \text{(Equation 1)}$$

where a_{hk} is the relative abundance of the taxonomy k in the host h and b_{kj} is 1 if the pathway j is present in the taxonomy k according to Macadam database and 0 if it is absent. Moreover, using the information on metabolites contained in MetaCyc, each metabolite i present in the pathway j was associated with a score given by:

$$S_{ji} = \sum_{r \in j} d_{jr} s_{ri} \quad \text{(Equation 2)}$$

where r are the reactions in the pathway, s is the stoichiometric number of the metabolite i in the reaction r, positive if the metabolite is produced and negative if it is consumed an the coefficient d_{jr} can assume values ± 1 in order to account for the direction of the reaction. Reversible reactions are considered as left to right and associated with $d_{jr} = 1$.

MetaCyc associates an ontology description to each pathway. While the full name is used to produce patient-metabolite matrix, we considered a lower-level classification of the pathways for metabolic network reconstruction. This allows us to collapse pathways producing the same selected compound in macro-category. In particular, we selected the sixth ontology-level term associated with a pathway. When less than six levels are reported in MetaCyc, we considered the deepest level.

Thus, combining the matrices P and S we define a host-metabolite matrix M_{hi} defined as

$$M_{hi} = \sum_{j} P_{hj} S_{ji} \quad \text{(Equation 3)}$$

and whose elements represent an estimate of the production or consumption of the metabolite i in the host h due to all the OTUs associated with the host. The matrix M_{hi} is the core of the STELLA algorithm since it allows us to infer the metabolic spectrum associated with a host with a given microbiome profile.

Data processing

All of the datasets are based on 16s rRNA sequencing of fecal samples from patients and healthy controls. The Quantitative Insights Into Microbial Ecology (QIIME) software package[26] was used for sequencing analysis. Greengenes 16S rRNA reference sequences (V. 13.5 - clustered at 99 percent identity[27]) were used as reference database to annotate the taxonomies as Operational taxonomic units (OTUs) identifying different taxonomies by cluster of similar sequences, variants of the 16S rDNA marker gene sequence.[26] We consider, for each dataset, the matrix of abundances of OTUs, truncated to genus level, when not otherwise specified. Two of the datasets that comprise Autistic patients[10,11] were obtained with a similar sequencing pipeline, and relative abundance of OTUs result comparable allowing us to analyze them separately and in combination. We applied the SVD-correction technique introduced in[19] to remove batch effect from the merged dataset as described below. The third dataset[12] is not directly comparable to the others due to a different experimental pipeline used by the authors to normalize the data and the number of patients and healthy controls considered is largely unpaired (see key resources table).

The two datasets related to Multiple Sclerosis[9,13] report microbial abundances as non-normalized integer counts. In particular, the first dataset focuses only on a subset of spore-forming chloroform resistant bacteria and includes a small subset of OTUs (68) when compared to the second one (942). Sample size is also highly different and the two datasets are thus incomparable. The taxonomies reported in the datasets were hand-checked, redundancy in OTUs names were manually removed.

Batch effect removal

When more than one dataset of microbiome sequencing data are available for the same disease, data can be combined in order to increase the statistics over the patients. This step can, however, introduce a number of biases in the analysis. First of all, the pipeline used to analyze the raw data should be similar so that datasets could be compared. Even in this case, the conjoined analysis of datasets obtained from different experiments and laboratories could lead to differences between the groups due to experimental procedures or sampling (batch effect) which do not depend on the biological process under investigation. The SVD-correction technique discussed in[19] can be applied to remove batch effects from the OTUs abundance matrix obtained by merging two or more datasets.

Briefly, SVD-corrections are used to filter out those eigenvectors of the global OTUs matrix that are inferred to correspond to batch effects rather than to true biological differences between the samples. The process is repeated until the largest contribution to the variance is given by the variance between the two classes of interest "disease" and "control". Here, we applied the correction on the datasets reported by[10] and.[11] We merged the two datasets, keeping only those OTUs present in both the original dataset. In our case, it was enough to remove only the first eigenvalue in order to remove batch effect, as discussed below.

Principal component analysis

We performed a PCA on the log-transformed OTUs abundances matrix from the intersection of two datasets.[10,11] The matrix includes all the 83 patients and the 103 taxonomies present in both datasets. Observing the projection on the first two principal components, it is evident that the first one allows us to distinguish between the two datasets (see Figure 3), while the second one seems to be slightly related to control/patient distinction. Applying one-step batch effect removal algorithm, we were able to reduce the distinction between the two datasets and first principal component as well as hierarichcal clustering shows a reduction of batch effect.

Step-wise feature selection

The linear discriminant analysis (LDA) feature selection was used to find the OTUs and the metabolites that are more relevant to distinguish "disease" and "control" condition. LDA is a supervised learning technique for classification based on maximization of the Rayleigh coefficient, given by the ratio of the determinant of the inter-class scatter matrix of the projected samples to the intra-class scatter matrix of the projected samples.

The feature selection algorithm used in the present study is based on the function stepclass from the package klaR in R (v.4.0,[28]). The algorithm trains a model using the method stepLDA and finds the best model by maximizing the accuracy of the performance using a 10-fold cross validation. At each round, the data are divided into a different training and test set, where the training set is used to determine the coefficients and the test is used to assess the performance of the trained model with the optimized coefficients. We allowed forward and backward selection by setting the maximum number of features to 15 both for OTUs and metabolites. We verified that, with the considered database, this is sufficient to reach a plateau in performance of the algorithm. The process is repeated for at least 100 replica with random reshuffling of the columns in order to avoid any dependencies of the algorithm's results to the order of the variables and to detect any errors or inconsistencies. Both the OTUs abundances matrix and metabolite matrix need to be pre-processed removing highly correlated variables and OTUs with null or constant abundance across the hosts, in order to diminish the computational effort. Correlation threshold was set to 0.95 for OTUs abundances and 0.7 for metabolites. Feature selection has been performed on log-transformed OTUs abundances. Coherence of the results using different representative for correlation cluster was verified.

QUANTIFICATION AND STATISTICAL ANALYSIS

To compare metabolites resulting of the STELLA algorithm with those measured experimentally, we compute the associated z-scores. If M_{hi} is the relative concentration of the metabolite i in host h, the z-score is obtained as

$$z_{hi} = \frac{M_{hi} - \langle M_{hi} \rangle}{\sigma}, \qquad \text{(Equation 4)}$$

where $\langle M \rangle_h = \sum_h M_{hi}/N_h$, N_h is the number of hosts and σ is the corresponding standard deviation. F_β-scores are computed by considering the sign of z-scores for predicted and measured metabolites as

$$F_\beta = \frac{(1+\beta^2) n_{TP}}{(1+\beta^2) n_{TP} + \beta^2 n_{FN} + n_{FP}}, \qquad \text{(Equation 5)}$$

where n_{TP} is the number of true positives, n_{FP} is the number of false positives and n_{FN} is the number of false negatives.