



EasyMap - An Interactive Web Tool for Evaluating and Comparing Associations of Clinical Variables and Microbiome Composition

Ehud Dahan¹, Victoria M. Martin² and Moran Yassour^{1,3*}

¹ Microbiology and Molecular Genetics, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel,

² Department of Pediatrics, Massachusetts General Hospital, Boston, MA, United States, ³ School of Computer Science & Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

OPEN ACCESS

Edited by:

Elisabetta Caselli,
University of Ferrara, Italy

Reviewed by:

Luke R. Thompson,
Atlantic Oceanographic and
Meteorological Laboratory (NOAA),
United States
Ornella Cominetti,
Nestlé Research Center,
Switzerland

*Correspondence:

Moran Yassour
moranya@mail.huji.ac.il

Specialty section:

This article was submitted to
Microbiome in Health and Disease,
a section of the journal
Frontiers in Cellular and
Infection Microbiology

Received: 13 January 2022

Accepted: 05 April 2022

Published: 13 May 2022

Citation:

Dahan E, Martin VM and Yassour M
(2022) EasyMap - An Interactive Web
Tool for Evaluating and Comparing
Associations of Clinical Variables
and Microbiome Composition.
Front. Cell. Infect. Microbiol. 12:854164.
doi: 10.3389/fcimb.2022.854164

One of the most common tasks in microbiome studies is comparing microbial profiles across various groups of people (e.g., sick vs. healthy). Routinely, researchers use multivariate linear regression models to address these challenges, such as linear regression packages, MaAsLin2, LEfSe, etc. In many cases, it is unclear which metadata variables should be included in the linear model, as many human-associated variables are correlated with one another. Thus, multiple models are often tested, each including a different set of variables, however the challenge of selecting the metadata variables in the final model remains. Here, we present EasyMap, an interactive online tool allowing for (1) running multiple multivariate linear regression models, on the same features and metadata; (2) visualizing the associations between microbial features and clinical metadata found in each model; and (3) comparing across the various models to identify the critical metadata variables and select the optimal model. EasyMap provides a side-by-side visualization of association results across the various models, each with additional metadata variables, enabling us to evaluate the impact of each metadata variable on the associated feature. EasyMap's interface enables filtering associations by significance, focusing on specific microbes and finding the robust associations that are found across multiple models. While EasyMap was designed to analyze microbiome data, it can handle any other tabular data with numeric features and metadata variables. EasyMap takes the common task of multivariate linear regression to the next level, with an intuitive and simple user interface, allowing for wide comparisons of multiple models to identify the robust microbial feature associations. EasyMap is available at <http://yassour.rcs.huji.ac.il/easymap>.

Keywords: microbiome, multivariate linear regression, clinical association, interactive, webtool

INTRODUCTION

Examining microbiome differences in the context of clinical changes has become a widely-popular task in many academic and industry contexts (Belkaid and Hand, 2014; Borbet et al., 2019; Niu et al., 2021; Sorbara and Pamer, 2022). The ease of collecting stool samples (compared to blood or biopsy samples), together with the growing evidence of the microbiome's contribution to human health (Becattini et al., 2016), makes the gut-microbiome case/control cohort design even more commonly used in the field (Zhu et al., 2013; Duvallet et al., 2017; Huang et al., 2021).

While animal-studies are conducted in a well controlled environment, they often do not represent human health in sufficient accuracy (Nguyen et al., 2015; Nagpal et al., 2018; Ma, 2021). On the other hand, in human cohorts we have the great challenge of dealing with all the additional characteristics that vary in the human population, such as age, diet, lifestyle, which are known confounders of the gut microbiome, and can bias our results (Vujkovic-Cvijin et al., 2020; "De-Confounding Microbiome Association Studies" n.d.; Devkota, 2016; Bartolomeaus et al., 2020). In an attempt to address this inherent bias in human studies, the field always strives to establish as large cohorts as possible, such as the UK biobank, LifeLines, and the TEDDY cohort ("UK Biobank - UK Biobank" n.d.; TEDDY Study Group, 2008; Davidson-Pilon, 2019). However, it is very difficult and costly to establish and manage large cohorts, and not all clinical manifestations enable such large cohorts, and even in these large numbers, computational methods that take into account the confounding factors are much in need (Jessica and Hanson, 2020).

An additional challenge in microbiome case/control studies is the interoperability of the results. While some machine-learning algorithms perform well on large datasets (Chen et al., 2020; Gou et al., 2021; Carrieri et al., 2021), they are often discriminative in the case/control task without revealing additional information on the underlying reason for the success of their method. Alternatively, the results will highlight specific microbial features that may play a role in the examined clinical manifestation (Oh and Zhang, 2021), which can be further studied from a medical- or a basic-science perspective, as the basis for further studies understanding the mechanisms underlying this association (Aasmets et al., 2021).

A common approach in all case/control studies is the use of multivariate linear regression models that take into account the variables of interest (i.e., microbiome composition) together while accounting for the confounding variables mentioned above (Ramette, 2007; Xia and Sun, 2017; Bodein et al., 2019; Raimondi et al., 2021). There are many packages (R, python, and independent tools) that perform this task, and one of the most-popular tools in the context of microbiome studies is MaAsLin2 (Mallick et al., 2021; S. Ma et al., n.d.; Ma et al., 2021; Zhang et al., 2021). It is especially useful in microbiome studies due to the data transformation (arcsine square-root transformation, Methods), outlier removal, presentation of results and its overall ease of use. Oftentimes, researchers will run multiple models in an attempt to find the ideal model that

explains the data best, without overfitting. However, the routine task of comparing MaAsLin results across multiple models is challenging. First in running the multiple models, but more importantly in interpreting the subtle differences in their results.

Here, we present EasyMap, a user-friendly interactive web-based tool that enables running multiple linear regression models and comparing across their results in a graphical manner. While EasyMap was designed to analyze microbiome data, it can handle any other tabular data with numeric features and metadata variables. EasyMap enables the users to upload their own data, construct multiple models, and run the analyses using MaAsLin2, regardless of their computational background and expertise. EasyMap also improves the usability of viewing the significant results, in an interactive high- and low-level visualization of the results. Most importantly, EasyMap provides an easy framework for comparing across models, stratifying the linear regression results by additional variables, and eventually assisting in choosing the optimal model for the data.

METHODS AND IMPLEMENTATION

Tool Implementation

The EasyMap web tool was developed using the *shiny* R package (version 1.6.0), and it uses the *MaAsLin2* R package (version 1.4.0) for multivariate linear regression. All code is available for download on the Yassour lab git repository (Dahan, 2021). EasyMap is available for public use at <https://yassour.rcs.huji.ac.il/easymap>.

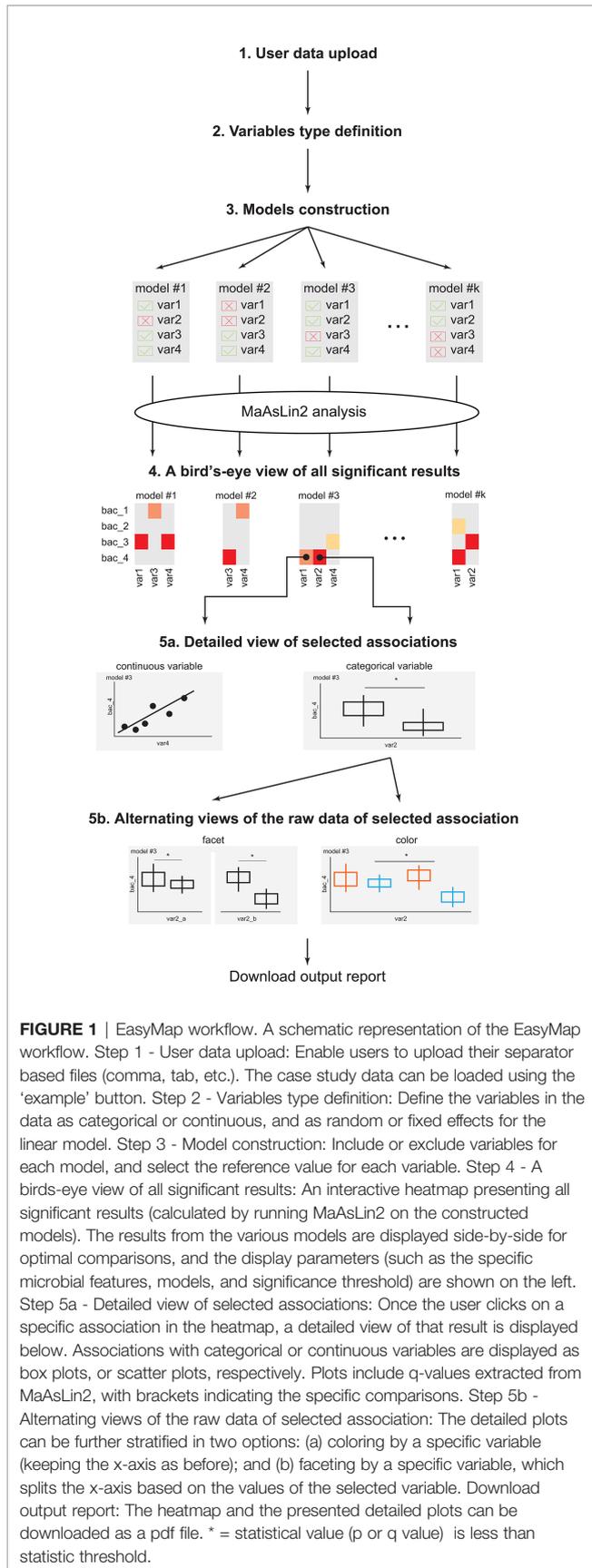
Multivariate Linear Model

EasyMap provides an easy web-based system to perform a multivariable association analysis between microbial features changes to clinical measurement (metadata). Analysis modules include preprocessing, normalisation and transformation and produce a statistically significant output including correction for multiple tests (see below). All user choices are processed to MaAsLin2 format, and are run with the default MaAsLin2 parameters.

Variables in the model are defined as either random- or fixed-effect, based on the user definition (Figure 1, Step 2). Briefly, fixed variables impact all samples equally, while random variables impact samples differentially, based on the value of the random-effect variable (Srinivasjois, 2021; Kanters, 2022).

Preprocessing and Normalization

After loading the input data, total sum scaling normalization is applied to each sample, and then microbial features with a total normalized sum of less than 0.0001 were removed. Next, abundance data were transformed with the arcsine square root-transformation (AST) (Biometry: The Principles and Practice of Statistics in Biological Research, 1981). Microbiome data is often sparse and zero-inflated, thus the arcsine square root transformation is an ideal choice to spread the abundance values, but maintain zero abundance. By default, MaAsLin2 uses the `na.exclude` function, which excludes the Not available (NA) values from the multivariate linear regression calculations, but



keeps these values for the association visualization (will appear as a separate category, without any statistical calculation). Specifically, if the model includes the variable which has an NA value for some samples, these will not be included in the statistical analysis, but will be presented in the visualization.

Statistical Analysis

Nominal p-values across all associations in the resulting heatmap were adjusted using Benjamini-Hochberg FDR method performed by MaAsLin2, and the coefficient and resulting q-values appear in the box plots with the corresponding brackets. After 'faceting' the box plot, the tool presents the p-values calculated by a two-sided t-test (using the *ggpubr* R package; version 0.4.0), based on all the samples that appear in the plot (without any MaAsLin2 filtering).

Box Plots

In the box plots each dot represents one sample, the middle line represents the median of the distribution, and the box boundaries represent the first and third quartiles. The y-axis represents the transformed relative abundance of a microbial feature bacteria (AST, Methods) and the x-axis is the selected effect variable. All box plots were generated by the *ggplot2* R package (version 3.3.3). See statistical analysis Methods section above for q-values and p-values description.

Case Study Data

From the GMAP prospective observational healthy infant cohort, selected infants diagnosed with food protein-induced allergic proctocolitis (FPIAP) who had a minimum of 4 longitudinal stool samples and selected matched controls for each who met the same sampling criteria. 16S rRNA gene libraries targeting the V4 region of the 16S rRNA gene were sequenced on an Illumina MiSeq 300 (raw sequencing data can be found on NCBI BioProject PRJNA730851). Total of 954 samples remain for further analysis.

Running EasyMap on Other, Non-Microbiome, Data

We developed EasyMap to assist us in analyzing microbiome data. However, it can also be used in additional contexts where multivariate linear models are commonly used, maintaining all its added value. The input data format is described below (see Step 1: Input data upload), yet specific attention should be paid to the AST transformation that is applied automatically on the feature data input, which is not optimal for all datasets. The user can choose to not apply this transformation on the uploaded data.

WORKFLOW OF EASYMAP + CASE STUDY EXAMPLE

Cohort and Data Description

Here, we describe the step-by-step flow of the EasyMap interactive tool, available at <http://yassour.rcs.huji.ac.il/easymap> (also presented in **Figure 1**). To demonstrate the performance of

EasyMap and make it easier to use, we carried out a case study (Martin et al., 2021) and added a short explanation at the end of each step. This case study examines the early childhood microbiome in allergic infants. In this project, 160 infants were longitudinally sampled during the first year of life (6 time points). During this period, 81 infants were diagnosed with food-protein induced allergic proctocolitis (FPIAP), specifically to cow's milk proteins. To characterize the microbial profiles, 16S sequencing was performed on all infant stool samples (N=954). Here, we demonstrate the advantage of using the EasyMap web tool to investigate statistical significant associations between microbial features and clinical data (i.e., allergic diagnosis), using multivariate linear regression models.

Step 1: Input Data Upload

The first step is uploading the user input data, which is a separator based file (csv, tsv etc.) including all relevant data: clinical metadata variables and taxonomic features' abundance for each sample (**Figure 1**). This file should follow MaAsLin (Morgan et al., 2012) format (largely described below) and should have a header line. Suppose there are n metadata variables and m taxonomic features, the input file will have three sections of columns: (a) The first column will contain the sample ID, which is a unique identifier of samples; (b) The next n columns will contain the metadata variables, where each of the variables can be either all strings or all numeric but not mixture. These can include clinical measurements and also other information, such as subject ID, or collecting clinic; (c) The last m columns will contain the abundance of the taxonomic features (relative or absolute). All abundance data will be normalized by total sum scaling (TSS) normalization (MaAsLin2 default) and then will be transformed by the arcsinus transformation (AST, Methods). The user can choose not to apply the AST transformation on the uploaded data by unchecking the AST checkbox in Step 3 (model construction).

Users can upload their separator based files (comma, tab) through the 'Upload Files' tab. Once uploaded, all the identified columns will appear and the user can click the "Submit" button and continue with the analysis. If there is any problem with parsing the file, the error will be presented to the user.

Step 1 - Case Study

In this case study, we considered six clinical variables that were collected in our cohort, and are also known to have an impact on gut microbiome composition: mode of delivery (vaginal or C-section), age (at time of visit), use of probiotics in the first year of life, infant diet at each time point (breastfed, formula-fed, mixed); and finally the disease status (case/control). In this study we are searching for microbial features that are associated with the disease status, taking into account all other clinical variables (top of this file is presented in **Figure 2A**). Clicking on the 'example' button loads a sample of the case study data.

Step 2: Variables Type Definition

After uploading the data, it is necessary to define the type of all clinical metadata variables (**Figure 1**). First, the user selects the

column that represents the unique sample ID. Second, the user selects the variables that will be used as *random effects* in the linear model (see Methods). The model will account for these variables, but will not search for associations between the random variables and the microbial features. Next, the user selects the fixed effect variables, which can be assigned as either continuous or categorical. All variables that remain unselected are automatically defined as the microbial features, thus all clinical variables must be selected as either random or fixed effect variables.

Categorical variables are automatically sorted alphabetically (for example, always, never, sometimes), however, if the user has a specific relevant order, the variable values can include a prefix to maintain this order (like, a_never, b_sometimes, c_always). Numeric variables that have four or less unique values will be treated as categorical variables. Once all variables are defined an "approve" button will appear at the bottom of the screen.

Step 2 - Case Study

In this case study there were 954 samples from 160 different infants. We defined the infant ID as a random variable such that multiple samples from the same infants will be accounted for together, and not as independent measurements. Next, we defined delivery mode, disease status, probiotic use and infant diet as categorical variables. All variables other than diet have two values, and diet has three (breastfed, mixed or formula). Finally, we defined the age at the time of visit as a continuous variable, and all remaining variables are left as the microbial features (the relative abundance of each bacteria in each of the samples; **Figure 2A**).

Step 3: Model Construction

When searching for statistical-significant associations, we first need to choose the clinical variables that our model should account for. These variables are usually chosen based on prior understanding of the clinical situation, and also including factors that are known to impact the microbial community composition. Naively, one can include all collected variables in the model, however, including too many variables would lead to overfitting the data, and diluting the signal across too many variables, potentially missing the significant association altogether. Oftentimes, we choose multiple models, each containing a different set of examined variables, with the aim to compare the results across these models. EasyMap was built to enable a comprehensive comparison across various models, thus highlighting the strong, consistent associations across multiple models.

After defining and approving the variables (as described in step 2) the user will next move to selecting the variables to be used in the first model (**Figure 1**). In the case of categorical variables, the user can also specify the reference value to be used for each variable. For example, if `delivery_mode` has two possible values: "C-section" or "vaginal", the user can specify that "vaginal" will be the reference value. By default, the tool sorts the values alphabetically and the first value is used as reference. In the example above this would have been "C-section".

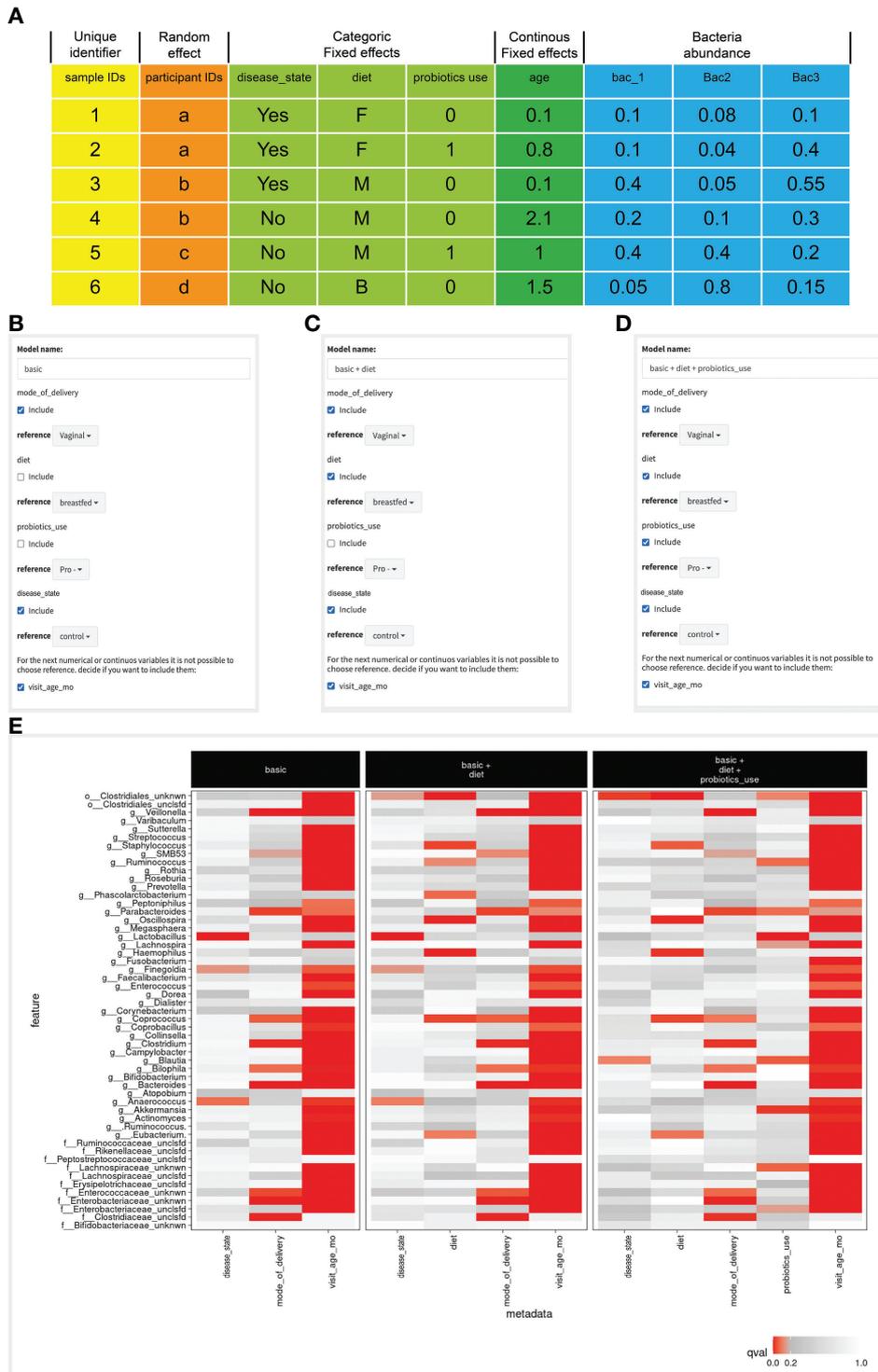


FIGURE 2 | Defining the variables, constructing the models and comparing the results. **(A)** Input file example, where column colors represent different variable types (appear at the top of the table), and the first row is the table's header. **(B–D)** Screenshots of the model construction step. Checkbox values next to each variable indicate if it is included in each model, and the reference value for each variable is selected in the dropdown. The last variable, named 'visit_age_mo' is defined as a continuous variable, therefore its reference is assigned as zero. **(E)** A heatmap showing the birds-eye view of all significant results across all models, from the MaAsLin2 analysis. Each box represents a model, rows are microbial features, and columns are model variables. Each entry in the heatmap is colored by the significance of the association of the specific microbial feature with the specific model variable.

Additional models can be added by clicking on the “add new set” button, and repeating this selection step for each model. By default, the new model is initiated with the selection of variables of the most recently defined model. While many models can be added and compared across, it impacts the total running time and the ease of results viewing in the next step, thus comparing 3-5 models seems ideal.

Example Step 3 - Case Study

In this case study we wanted to find a microbial feature that is associated with disease status. We wanted to examine the contribution of a specific clinical variable to the associations we found. Here, we focused on the impact of infant diet and probiotics use on the microbial associations with disease status. Therefore we considered three models: (1) Including the case/control, mode of delivery and visit age variables as a base model. (2) model 1 variables + infant diet; and (3) model 2 variables + probiotic use (**Figures 2B–D**). We were interested to see whether associations found in model 1 remained when adding the diet and probiotic use variables, which will be revealed in the next step.

Output Description

The output of the EasyMap is composed of two sections: A heatmap of all significant results and a detailed view of selected associations (for example using box plots), with the ability to facet and color the raw data. All the results that are shown on the screen (heatmap with the detailed plots) can be exported to a pdf file.

Step 4: A Birds-Eye View of All Significant Results

The first step in comparing the models is a high-level comparison of all microbial associations that were found to be significant in at least one model. Heatmap color represents the significance, and by default, the FDR q-value threshold is set to be 0.2 (only associations that pass this threshold appear in color). The user can further filter the presented microbial features, using the drop-down menus on the left. The user can select a different threshold, and also choose which models to include in the heatmap (**Figure 1**).

Example Step 4 - Case Study

When examining the bird-eye view of significant results of our three models (**Figure 2E**), the first clear observation is that infant age was strongly correlated with most microbial features (**Figure 2B**). When examining the disease variables/column in the basic model, we found three significant associations (*g:Lactobacillus*, *g:Finegoldia* & *g:Anaerococcus*). However, in the two additional models (models 2 & 3), these associations are not significant anymore, and additional significant associations are detected (*o:Clostridiales_unknwn* in models 2 & 3, and *g:Blautia* in model 3). To enable a simpler comparison we have subset the heatmap to display only the microbial features mentioned above (**Figure 3A**). Interestingly, in the case of *g:Lactobacillus*, there was still a significant association in model 3, only with the probiotic use variable (**Figure 3A**). This shift indicated that

once we added the probiotic use to the model, it better explains the different *g:Lactobacillus* abundance across the disease groups.

Step 5a: Detailed View of Selected Associations

One unique and useful feature of EasyMap is the ability to toggle quickly between the bird’s eye view of all associations in the heatmap and zooming in on specific associations of interest (**Figure 1**). When the user hovers on a single cell in the heatmap, the cell is highlighted, and the relevant microbial feature together with the selected model, and associated clinical variable appear as text at the bottom of the panel. When the user clicks on a certain cell in the heatmap, the bottom panel is populated with a detailed plot showing the relative abundance (AST, if it was transformed, Methods) of the selected microbial feature by the selected clinical variable (this can be either a box plot for a categorical variable or a scatter plot for a continuous clinical variable). Note that if the relative abundance values (y-axis) are arc-sinus transformed thus can exceed 1, and range in [0, 1.57079]. The detailed plot also displays the q-values that are outputted by MaAsLin2 for all tested associations in this variable (using brackets comparing each value to the selected reference). Significance analysis appears for all possible comparisons between the reference and other values, with their respective q values, even for the non-significant comparisons.

Step 5b: Alternating Views of the Raw Data of Selected Associations

Finally, to include additional metadata to the existing plot, the user can facet the box plot and/or color the dots, by a specific variable (**Figure 1**). When the plot is faceted, the MaAsLin2 q-values are removed from the plot, and instead a t-test is performed, and p-values are presented. Finally, the user can color the dots based on the categorical variables of the model, and add labels to the dots, based on the random variables of the model.

Example Step 5 - Case Study

To further examine the case of *g:Lactobacillus*, we clicked on the square that corresponds to the *g:Lactobacillus* row in the disease column on model 1, which displayed the box plot on the bottom panel (**Figure 3B**). The MaAsLin2 q-value was 1.65e-06, indicating that *g:Lactobacillus* is highly correlated with disease status. However, when clicking on the square that corresponds to the *g:Lactobacillus* row in the disease column on *model 3* (**Figure 3C**), the presented q-value was 2.44e-01, which did not pass our default significance threshold ($q < 0.2$), indicating that adding the probiotic use to our model decreases the significance of the disease association. Furthermore, we noted that in model 3, probiotic use was significantly associated with *g:Lactobacillus* abundance, suggesting we investigate this transition further.

Indeed, many of the allergic infants received probiotics in their first year of life, thus the more significant association of *g:Lactobacillus* is with probiotics. To further investigate the contribution of probiotic use in this case, the user can stratify

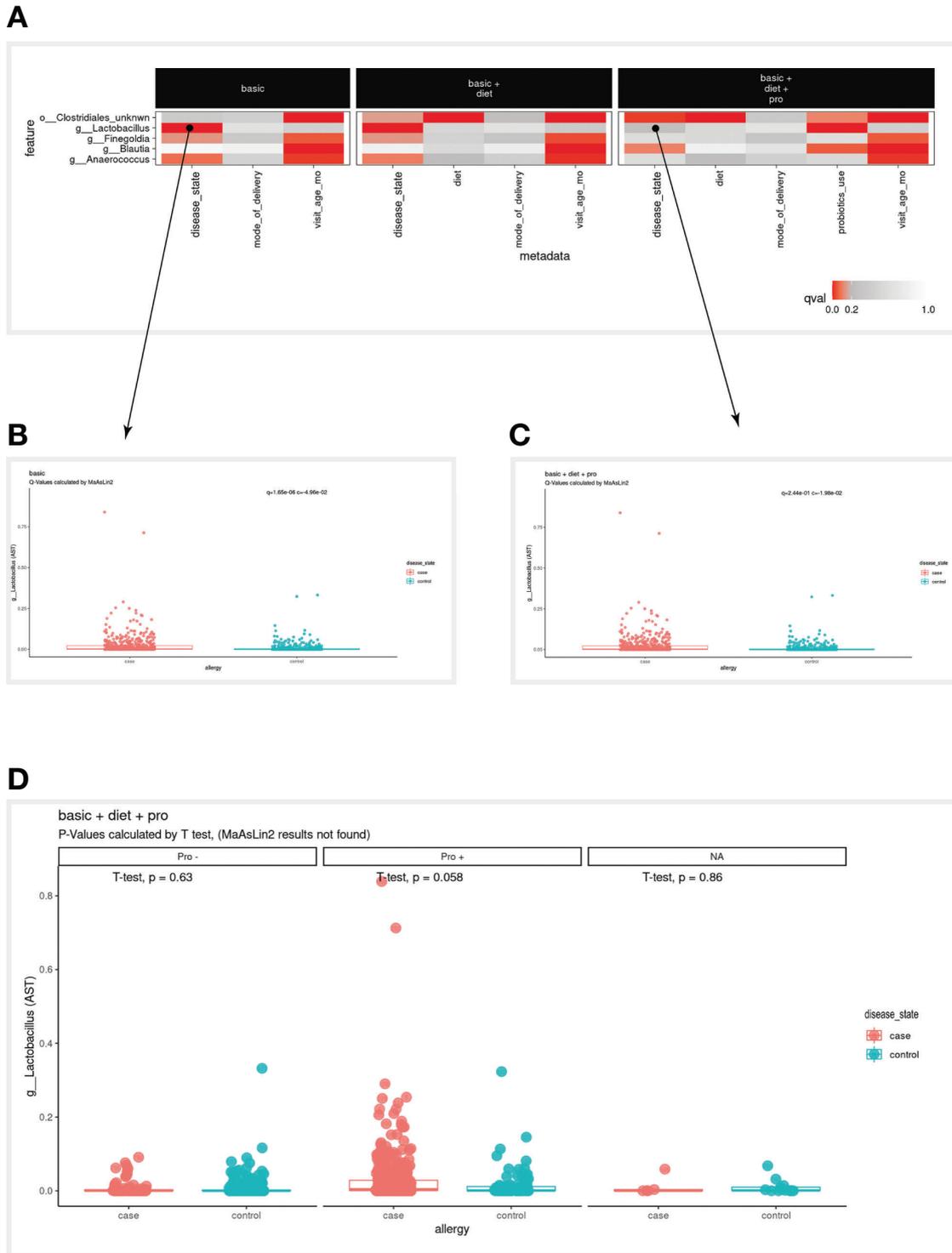


FIGURE 3 | Case study results. **(A)** A subset of the full heatmap, showing the three tested models with only five selected microbial features (rows). **(B, C)** Clicking on the heatmap entries marked with a circle, generates these box plots, displaying the linear association of *g:Lactobacillus* with the disease variable, in the 'basic' model **(B)** and from 'basic + diet + pro' model **(C)**. In the box plots, each dot represents a sample, where the y-axis is the relative abundance of the microbial feature (AST, see Methods), and the x-axis is the values of the disease variable. **(D)** Box plot as in **(C)** stratified by values of the probiotics_use variable (Pro-, Pro+, NA). NA, Not available.

the association between disease and *g:Lactobacillus* by probiotic use, using the “facet by” option of the boxplot. Once the user selects probiotics as the faceted variable, the association between disease and *g:Lactobacillus* can be studied within the context of probiotics (with/without; **Figure 3D**). Once again, note that in the stratified view, the statistical analysis is using t-test in this specific context, rather than the MaAsLin2 systematic q-value (Methods).

DISCUSSION

A common goal of microbial community studies related to human epidemiology is to identify associations between microbial features and clinical variables. These studies must take into account additional factors, of clinical or environmental nature, that also impact the microbiome composition. Often, researchers turn to multivariate linear regression models to find the clinical associations while accounting for other measured confounding effects. Here, we present EasyMap, an interactive web-based tool that enables uploading custom input data, defining multiple such models, running the linear regression (using MaAsLin2 (Mallick et al., 2021)) and comparing the results across all tested models. Comparing the results allows for a better selection of model variables, without overfitting the data.

EasyMap can be run as an online webtool, and the full code is also available on github (Dahan, 2021), making the tool useful for researchers with varying levels of computational backgrounds. Currently, the web-based tool has a few hard-coded settings (such as the common data transformation; Methods), which are helpful in maintaining its ease of use, however user requests from github will be accommodated upon popular demand.

We developed EasyMap to assist us in analyzing data from our lab’s studies. It was built as a wrapper for MaAsLin2, with added visualization and comparison abilities, tailored for microbiome

studies. However, it can be used in many additional contexts where multivariate linear models are commonly used, maintaining all its added value. EasyMap is also extremely useful for sharing results with collaborators, and enabling all participants to dig deeper in the analysis of their data.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. These data and scripts can be found here: <https://github.com/yassourlab/EasyMap>. Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Massachusetts General Hospital Institutional Review Board (IRB). Written informed consent to participate in this study was provided by the participants’ legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ED developed the tool, VM evaluated the tool, MY guided the work, and ED and MY wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this project is provided in part by the Azrieli Foundation and the Israeli Science Foundation.

REFERENCES

- Aasmets, O., Lüll, K., Lang, J. M., Pan, C., Kuusisto, J., Fischer, K., et al. (2021). Machine Learning Reveals Time-Varying Microbial Predictors With Complex Effects on Glucose Regulation. *mSystems* 6 (1), 1–12. doi: 10.1128/mSystems.01191-20
- Bartolomeus, T.U.P., Till, B., Bartolomeus, H., Löber, U., Avery, E. G., Mähler, A., et al. (2020). Quantifying Technical Confounders in Microbiome Studies. *Cardiovasc. Res.* 117 (3), 863–875. doi: 10.1093/cvr/cvaa128
- Becattini, S., Taur, Y., and Pamer, E. G. (2016). Antibiotic-Induced Changes in the Intestinal Microbiota and Disease. *Trends Mol. Med.* 22 (6), 458–478. doi: 10.1016/j.molmed.2016.04.003
- Belkaid, Y., and Hand, T. W. (2014). Role of the Microbiota in Immunity and Inflammation. *Cell* 157 (1), 121–141 doi: 10.1016/j.cell.2014.03.011
- Biometry: The Principles and Practice of Statistics in Biological Research (1981).
- Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K.-A. (2019). A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. *Front. Genet.* 0. doi: 10.3389/fgene.2019.00963
- Borbet, T. C., Zhang, X., Müller, A., and Blaser, M. J. (2019). The Role of the Changing Human Microbiome in the Asthma Pandemic. *J. Allergy Clin. Immunol.* 144 (6), 1457–1466. doi: 10.1016/j.jaci.2019.10.022
- Carrieri, A. P., Haiminen, N., Maudsley-Barton, S., Laura-Jayne, G., Murphy, B., Mayes, A. E., et al. (2021). Explainable AI Reveals Changes in Skin Microbiome Composition Linked to Phenotypic Differences. *Sci. Rep.* 11 (1), 1–18. doi: 10.1038/s41598-021-83922-6
- Chen, L., Li, Z., Zeng, T., Zhang, Y.-H., Liu, D., Li, H., et al. (2020). Identifying Robust Microbiota Signatures and Interpretable Rules to Distinguish Cancer Subtypes. *Front. Mol. Biosci.* 7 (November), 604794. doi: 10.3389/fmolb.2020.604794
- Dahan, E. (2021). *EasyMap: A Web Tool for Evaluating Clinical Variables Impact on Microbiome Changes (Version 1.0)* (Jerusalem: Github). Available at: <https://github.com/yassourlab/EasyMap>.
- Davidson-Pilon, C. (2019). Lifelines: Survival Analysis in Python. *J. Open Source Software.* 4 (40), 1317. doi: 10.21105/joss.01317
- De-Confounding Microbiome Association Studies n.d. Available at: <https://www.natureasia.com/en/nature/hot-topics/detail/1788> (Accessed January 10, 2022).
- Devkota, S. (2016). MICROBIOME. Prescription Drugs Obscure Microbiome Analyses. *Science* 351, 452–453. doi: 10.1126/science.aaf1353
- Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., and Alm, E. J. (2017). Meta-Analysis of Gut Microbiome Studies Identifies Disease-Specific and Shared Responses. *Nat. Commun.* 8 (1), 1–10. doi: 10.1038/s41467-017-01973-8
- Gou, W., Ling, C.-W., He, Y., Jiang, Z., Fu, Y., Xu, F., et al. (2021). Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features

- Associated With Type 2 Diabetes. *Diabetes Care* 44 (2), 358. doi: 10.2337/dc20-1536
- Huang, M., Liu, J., Liu, K., Chen, J., Wei, Z., Feng, Z., et al. (2021). Microbiome-Specific Statistical Modeling Identifies Interplay Between Gastrointestinal Microbiome and Neurobehavioral Outcomes in Patients With Autism: A Case Control Study. *Front. Psychiatry/Frontiers. Res. Foundation*. 0. doi: 10.3389/fpsy.2021.682454
- Jessica, G.-P., and Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digest. Dis. Sci.* 65 (3), 674. doi: 10.1007/s10620-020-06091-y
- Kanters, S. (2022). "Fixed- and Random-Effects Models," in *In Meta-Research* (New York, NY: Humana), 41–65.
- Ma, Z. S. (2021). Cross-Scale Analyses of Animal and Human Gut Microbiome Assemblies From Metacommunity to Global Landscape. *mSystems* 6 (4), e0063321. doi: 10.1128/mSystems.00633-21
- Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable Association Discovery in Population-Scale Meta-Omics Studies. *PLoS Comput. Biol.* 17 (11), e1009442. doi: 10.1371/journal.pcbi.1009442
- Ma, S., Ren, B., Mallick, H., Moon, Y.S., Schwager, E., Maharjan, S., et al. (2021). A Statistical Model for Describing and Simulating Microbial Community Profiles. *PLoS Comput. Biol.* 17 (9), e1008913. doi: 10.1371/journal.pcbi.1008913
- Martin, V., Virkud, Y. V., Seay, H. L., Keet, C., Shreffler, W. G., and Yuan, Q. (2019). "A Prospective Assessment of Food Protein-Induced Allergic Proctocolitis from the GMAP Healthy Infant Cohort". *J. Allergy Clin. Immunol.* doi: 10.1016/j.jaci.2018.12.412
- Ma, S., Shungin, D., Mallick, H., Schirmer, M., Nguyen, L. H., Kolde, R., et al. (2020). Population Structure Discovery in Meta-Analyzed Microbial Communities and Inflammatory Bowel Disease. *bioRxiv*. doi: 10.1101/2020.08.31.261214
- Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment. *Genome Biol.* 13 (9), R79. doi: 10.1186/gb-2012-13-9-r79
- Nagpal, R., Wang, S., Solberg Woods, L. C., Seshie, O., Chung, S. T., Shively, C. A., et al. (2018). Comparative Microbiome Signatures and Short-Chain Fatty Acids in Mouse, Rat, Non-Human Primate, and Human Feces. *Front. Microbiol.* 0. doi: 10.3389/fmicb.2018.02897
- Nguyen, T. L., Vieira-Silva, S., Liston, A., and Raes, J. (2015). How Informative Is the Mouse for Human Gut Microbiota Research? *Dis. Models Mech.* 8 (1), 1–16. doi: 10.1242/dmm.017400
- Niu, W., Yang, F., Fu, Z., Dong, Y., Zhang, Z., and Ju, J. (2021). The Role of Enteric Dysbacteriosis and Modulation of Gut Microbiota in the Treatment of Inflammatory Bowel Disease. *Microbial. Pathogene.* 165, 105381. doi: 10.1016/j.micpath.2021.105381
- Oh, M., and Zhang, L. (2021). DeepGeni: Deep Generalized Interpretable Autoencoder Elucidates Gut Microbiota for Better Cancer Immunotherapy. *bioRxiv*. doi: 10.1101/2021.05.06.443032
- Raimondi, S., Calvini, R., Candelieri, F., Leonardi, A., Ulrici, A., Rossi, M., et al. (2021). Multivariate Analysis in Microbiome Description: Correlation of Human Gut Protein Degradation, Metabolites, and Predicted Metabolic Functions. *Front. Microbiol.* 0. doi: 10.3389/fmicb.2021.723479
- Ramette, A. (2007). Multivariate Analyses in Microbial Ecology. *FEMS Microbiol. Ecol.* 62 (2), 142–160. doi: 10.1111/j.1574-6941.2007.00375.x
- Sorbara, M. T., and Pamer, E. G. (2022). Microbiome-Based Therapeutics. *Nat. Rev. Microbiol.*, 1–16. doi: 10.1038/s41579-021-00667-9
- Srinivasjois, R. (2021). Fixed and Random-Effects Models for Meta-Analysis. In: S. Patole. (eds) *Principles. Pract. Syst. Rev. Meta-Analysis* Cham:Springer. doi: 10.1007/978-3-030-71921-0_7
- TEDDY Study Group. (2008). The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann. New York. Acad. Sci.* 1150:1–13. doi: 10.1196/annals.1447.062
- UK Biobank - UK Biobank (N.D). Available at: <https://www.ukbiobank.ac.uk/> (Accessed January 10, 2022).
- Vujkovic-Cvijin, I., Sklar, J., Jiang, L., Natarajan, L., Knight, R., and Belkaid, Y. (2020). Host Variables Confound Gut Microbiota Studies of Human Disease. *Nature* 587 (7834), 448–454. doi: 10.1038/s41586-020-2881-9
- Xia, Y., and Sun, J. (2017). "Hypothesis Testing and Statistical Analysis of Microbiome." *Genes Dis.* 4 (3), 138–148. doi: 10.1016/j.gendis.2017.06.001
- Zhang, Y., Thompson, K. N., Huttenhower, C., and Franzosa, E. A. (2021). Statistical Approaches for Differential Expression Analysis in Metatranscriptomics. *Bioinformatics* 37 (Suppl_1), i34–i41. doi: 10.1093/bioinformatics/btab327
- Zhu, L., Baker, S. S., Gill, C., Liu, W., Alkhoury, R., Baker, R. D., et al. (2013). Characterization of Gut Microbiomes in Nonalcoholic Steatohepatitis (NASH) Patients: A Connection Between Endogenous Alcohol and NASH. *Hepatology* 57 (2), 601–609. doi: 10.1002/hep.26093

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Dahan, Martin and Yassour. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.