metabolites

MDPI

Article

General Unified Microbiome Profiling Pipeline (GUMPP) for Large Scale, Streamlined and Reproducible Analysis of Bacterial 16S rRNA Data to Predicted Microbial Metagenomes, Enzymatic Reactions and Metabolic Pathways

Boštjan Murovec [1], Leon Deutsch [2] and Blaž Stres [2,3,4,5,*]

[1] Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia; bostjan.murovec@fe.uni-lj.si
[2] Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, SI-1000 Ljubljana, Slovenia; leon.deutsch@bf.uni-lj.si
[3] Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova 2, SI-1000 Ljubljana, Slovenia
[4] Department of Automation, Jožef Stefan Institute, Biocybernetics and Robotics, Jamova 39, SI-1000 Ljubljana, Slovenia
[5] Department of Microbiology, University of Innsbruck, Technikerstrasse 25d, A-6020 Innsbruck, Austria
* Correspondence: blaz.stres@bf.uni-lj.si; Tel.: +386-41-567-633

check for updates

Citation: Murovec, B.; Deutsch, L.; Stres, B. General Unified Microbiome Profiling Pipeline (GUMPP) for Large Scale, Streamlined and Reproducible Analysis of Bacterial 16S rRNA Data to Predicted Microbial Metagenomes, Enzymatic Reactions and Metabolic Pathways. *Metabolites* **2021**, *11*, 336. https://doi.org/10.3390/metabo11060336

Academic Editors: Marika Cordaro, Rosalba Siracusa and Cholsoon Jang

Received: 9 April 2021
Accepted: 23 May 2021
Published: 24 May 2021

Abstract: General Unified Microbiome Profiling Pipeline (GUMPP) was developed for large scale, streamlined and reproducible analysis of bacterial 16S rRNA data and prediction of microbial metagenomes, enzymatic reactions and metabolic pathways from amplicon data. GUMPP workflow introduces reproducible data analyses at each of the three levels of resolution (genus; operational taxonomic units (OTUs); amplicon sequence variants (ASVs)). The ability to support reproducible analyses enables production of datasets that ultimately identify the biochemical pathways characteristic of disease pathology. These datasets coupled to biostatistics and mathematical approaches of machine learning can play a significant role in extraction of truly significant and meaningful information from a wide set of 16S rRNA datasets. The adoption of GUMPP in the gut-microbiota related research enables focusing on the generation of novel biomarkers that can lead to the development of mechanistic hypotheses applicable to the development of novel therapies in personalized medicine.

Keywords: 16S rRNA; amplicon; Mothur; PICRUSt 2; Piphillin; genus; OTU; ASV; predicted metagenomes; predicted enzymatic reactions; predicted metabolic pathways; reproducible analyses; human microbiome; gut; intestine; mice

1. Introduction

The gut microbiota is composed of a huge number of different bacteria, archaea, fungi and protozoa, next to viruses and various mobile elements [1,2]. All these microbes interact with the host, environmental stimuli and each other, thus producing an enormous diversity of chemical compounds that play a key role in host development, wellbeing and aging [3–7]. The advent of large scale microbiome studies generates analytical opportunities to understand how these communities operate and respond to their complex environmental stimuli [8]. Although knowledge of taxonomy and functional genes of microorganisms are both important, functional genes are more directly related to enzymatic reactions and metabolic pathways. It is increasingly recognized that the microbiome influences the host health state and disease progression. For instance, disease progression can range from mild gastrointestinal symptoms to inflammatory bowel disease and colorectal and liver cancer [9]. In addition, a range of diseases have been implicated in metabolic imbalances, ranging from metabolic syndrome and obesity to autoimmune diseases, psychological disorders and infections [9].

Amplicon sequencing of 16S rRNA has served as the key approach of the last decade for the understanding microbial community structure, dynamics and how organisms might influence or be influenced by environmental conditions [10]. Extensive sequencing of bacterial communities is generating large collections of datasets available through public repositories such as European Bioinformatics Institute (https://www.ebi.ac.uk/ accessed on 30 April 2021), CuratedMetagenomicsData [11], or individual studies [12]. These data have so far been described on the level of 16S rRNA taxonomy utilizing either (i) genus [12], (ii) 97–98.5% 16S rRNA identity operational taxonomic units (OTU) [13] or (iii) amplicon sequence variants (ASV) [14,15]. However, the processing and analyses of such datasets are highly diverse due to the high number of published and benchmarked pieces of software [16–20] and reports that lack significant technical details despite the Human Microbiome Project outlines and introduction of standard operating procedures [21–23].

In addition, this wealth of 16S rRNA data gives access to an untapped pool of information beyond the 16S rRNA taxonomy (genus, OTU, ASV), such as predicted functional genes, enzymatic reactions and metabolic pathways (Figure S1). The tools such as MicrobiomeAnalyst [24,25], PICRUSt [26], PICRUSt2 [27], Tax4Fun [28]; Tax4Fun2 [29] and Piphillin [30,31] link 16S rRNA sequence information to representative genome sequences and approximate metagenomics functional gene content relevant for the interpretation of the studied human disease phenomena and clinical metadata [32]. As a number of unexplored and large datasets encompassing thousands of samples and corresponding metadata are made available in repositories (e.g., [12,33] the analyses (genus, OTU, ASV) and improved predicted metagenomic, enzymatic and metabolic pathway datasets have the potential to unravel important taxonomic, functional, biochemical and metabolic findings (Figure S1).

However, in order to accomplish such intensive large scale data analyses effective workflows are required. These workflows should ideally (i) integrate various pieces software, (ii) streamline input and output formats, (iii) accommodate large datasets, (iv) maintain portability between benchtop PC and high performance computing clusters (HPC), (v) enable flexible (customizable) but also reproducible analyses (setting documentation) that can be (vi) shared with and utilized by other interested researchers.

In this study, we introduce a workflow (Figure 1) that integrates Human Microbiome Project tested procedures for amplicon sequence analysis with one of the most popular programs Mothur [34], and PICRUSt2 [27] for prediction of metagenomic functional genes, enzymatic reactions and metabolic pathways. In addition, the workflow presented here generates also formatted inputs for Piphillin [30,31], another popular sister program for metagenomic predictions. The benchmarking of the integrated programs such as Mothur, PICRUSt2, Piphillin and other comparable sister programs were already reported before in numerous studies [16–20,23,27,30,31]. The inbuilt Human Microbiome Project standard operating procedures can be tailored according to user analytical preferences and sequencing details. The whole workflow is delivered as portable all-inclusive container (Singularity [35]; https://sylabs.io accessed on 14 April 2021) amenable for teaching or/and research purposes, using personal computer or HPC. Depending on the size of data and complexity of analyses (genus-, OTU-, ASV- levels), the GUMPP workflow enables maximum utilization of information present in the original 16S rRNA amplicon datasets by producing additional three data types approaching multiomics view of the microbiome: metagenomics functional genes, enzymatic reactions and metabolic pathways. All four data types can serve as inputs for machine learning to unravel novel mechanistic insight into human disease development in relation to microbiome characteristics. To showcase the efficient analyses and utilization of computing resources two datasets describing human (n = 307) and mice gut (n = 365) were used for demonstration purposes.

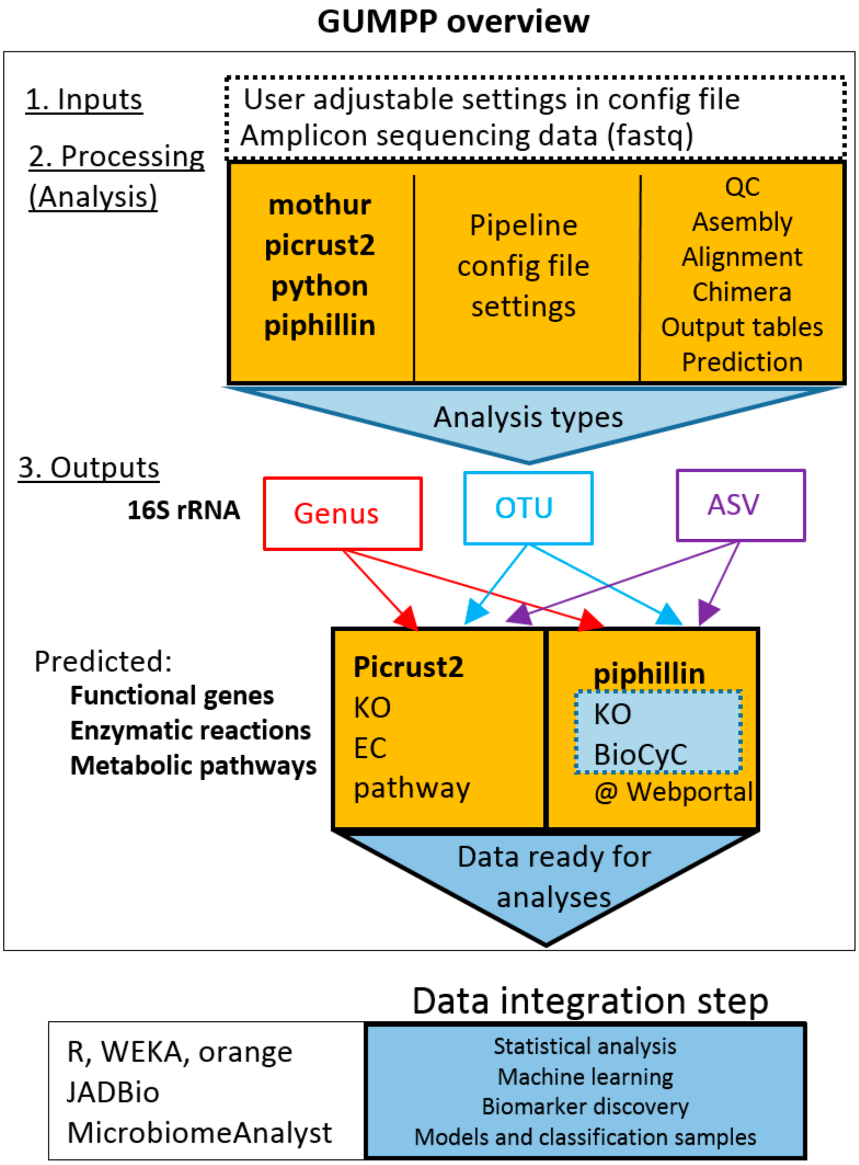


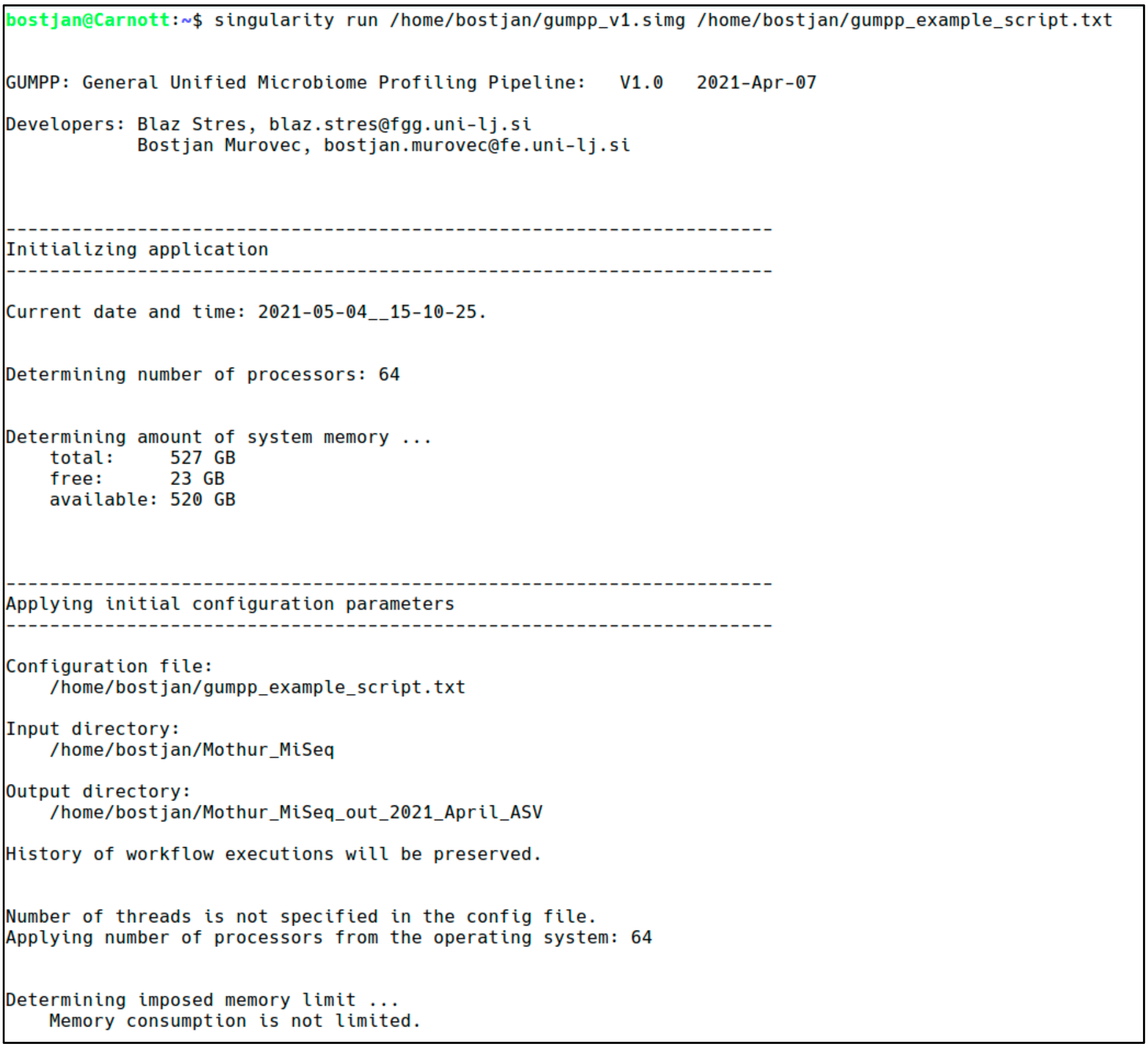
Figure 1. Schematic representation of the General Universal Microbiome Profiling Pipeline (GUMPP). The integral part consists of Mothur, PICRUSt2 and Piphillin outputs. Paired-end or single-end fastq sequence are used as input for mothur processing. The resulting biom and fasta files serve as an input for PICRUSt2. The data can be analyzed at genus-, OTU- and ASV- levels. QC–sequence quality control; OTU-Operational Taxonomic Units (generally 97% identity of 16S rRNA); ASV–Amplicon Sequence Variants (unique sequence variants). KO–KEGG Orthologs (Kyoto Encyclopedia of Genes and Genomes); EC-Enzyme Commission number; BioCyc-BioCyc collection of Pathway/Genome Databases. For each level, four output tables are generated (Please see Figures S1 and S2 for additional information). The resulting data can be analyzed in the data integration step using a variety of distinct machine learning approaches.

2. Results and Discussion

2.1. Design of GUMPP Workflow

GUMPP (http://gumpp.fe.uni-lj.si, accessed on 24 May 2021) is a freely available skeleton application for executing Mothur [34] using paired-end fastq files and executing the PICRUSt2 analyses next to producing also Piphillin [30,31] web-server input files (Figure 1). A single GUMPP run can process an arbitrary number of input files. Inputs are preprocessed by an integrated Mothur (V1.44.1) script in conjunction with Silva database (version 138), and creates biom and fasta representative sequence files as input for PICRUSt2 and outputs necessary for Piphillin [30,31]. The workflow was designed to support three levels of analysis differing in the increased extent of utilized information and fairness in data treatment: genus-, OTU- and ASV- levels (Figures S1 and S2). Users may freely replace the built in scripts and databases with their own. Customization of the built-in script (http://gumpp.fe.uni-lj.si) is also possible by template parameters.

The primary design goal of the GUMPP application was to deliver efficient analyses and utilization of computing resources. The application relies on recently developed Singularity container technology ([35]; https://sylabs.io accessed on 14 April 2021) making the pipeline straightforward to use as all its ingredients are fully integrated, preinstalled and preconfigured in a ready-made Singularity image. These consist of the Mothur and PICRUSt2 programs, the needed Mothur scripts, two Silva taxonomy databases (V138 and V138 seed), a few supporting utilities written in C++, as well as a skeleton framework consisting of slightly less than 11,000 lines of Python code which orchestrates the execution of individual pieces and takes care of executing programs and building their command lines. The actual parameters under which the workflow is executed are at the control of the user (Figure 2, ESM Figures R1–R3).



```
bostjan@Carnott:~$ singularity run /home/bostjan/gumpp_v1.simg /home/bostjan/gumpp_example_script.txt


GUMPP: General Unified Microbiome Profiling Pipeline:   V1.0   2021-Apr-07

Developers: Blaz Stres, blaz.stres@fgg.uni-lj.si
            Bostjan Murovec, bostjan.murovec@fe.uni-lj.si



----------------------------------------------------------------------
Initializing application
----------------------------------------------------------------------

Current date and time: 2021-05-04__15-10-25.


Determining number of processors: 64


Determining amount of system memory ...
    total:     527 GB
    free:      23 GB
    available: 520 GB



----------------------------------------------------------------------
Applying initial configuration parameters
----------------------------------------------------------------------

Configuration file:
    /home/bostjan/gumpp_example_script.txt

Input directory:
    /home/bostjan/Mothur_MiSeq

Output directory:
    /home/bostjan/Mothur_MiSeq_out_2021_April_ASV

History of workflow executions will be preserved.


Number of threads is not specified in the config file.
Applying number of processors from the operating system: 64


Determining imposed memory limit ...
    Memory consumption is not limited.
```

Figure 2. An example of the program startup and the initial checkups done by the Python code.

Aside from reproducible execution of the workflow and the control of algorithm settings, GUMPP offers some additional benefits. First, results of Mothur preprocessing may optionally be stored in a specially crafted storage area, where each result is associated with its full context (hash of input files, Mothur script and its values of template parameters, or other relevant information). This enables efficient workflow re-executions with different Mothur and PICRUSt2 parameters. When GUMPP detects that upon its re-execution only PICRUSt2 parameters are changed, it instantly recycles the previously obtained Mothur results. This opens up a possibility of efficient experimenting with changed PICRUSt2 parameters to observe their impact on end results. In addition, Mothur processing is split into a common and an analysis specific part. If only analysis type or its related parameters are changed, the previously computed common results are again recycled instantly, which is a significant time saver, since the common part consists of e.g., sequence alignment to a taxonomy database. The system also enables crash recovery: in the case of GUMPP interruption during e.g., PICRUSt2 step (operating system crash, power outage, abort due to administrative policies on High-Performance Computer (HPC), upon restart only the PICRUSt2 step is re-executed. Crash recovery is completely automatic and transparent. A user need not to specify any directives to inform GUMPP that execution is being repeated.

The system is suitable for autonomous execution on domestic hardware as well as on HPC facilities. All command-line parameters and intermediate file formats are handled automatically by the system, enabling the experienced users to prescribe their own parameters for PICRUSt2 or for template Mothur script parameters in order to finetune the workflow execution.

In order to aid in documenting analyses and inspection of execution, GUMPP stores an accurate verbatim copy of its screen output as a part of end report. Also, the actual command lines, standard output streams, standard error streams and exit codes of individual programs are stored on a disk in a hierarchical way for easy navigation, inspection and debugging. Analysis setup relies on configuration files, where a complete workflow configuration is prescribed and hence also documented. GUMPP presented in this study thus builds on the highly popular and tested programs that were benchmarked in numerous past studies as reported before [16–20,27,30,31].

2.2. Reanalysis and Extension of Mice Gut Microbiome Data Using GUMPP: The Choice of Level of Analysis (Genus, OTU, ASV) Is far from Arbitrary

Mice data analysis using GUMPP enabled us to explore a technical question of how user reports on different taxonomic levels (genus; OTU; ASV) affected the exact relationships between underlying samples when studied utilizing the four data types (16S rRNA; functional genes; enzyme reactions; metabolic pathways). The results of Mantel test between taxonomic levels (Figure 3) show that the correlations between 16S rRNA vs. KO, 16S rRNA vs. EC and 16S rRNA vs. pathways decreased from 0.90, 0.91 and 0.90 at genus level, to 0.75, 0.75 and 0.76 at OTU level, and to 0.61, 0.61 and 0.66 at ASV level, respectively (n = 9999 permutations, $p < 0.0002$). The fact that ASV type of analysis resulted in lower correlations between datatypes is in line with past observations that there is little congruency between rather variable taxonomic descriptions of microbial communities and their corresponding even more diverse metagenomic functional gene makeup [36].

A between level analysis for each data type separately (Figure 4) illustrates the relationships between data of the same type, obtained using a different taxonomic level of analysis (Genus, OTU or ASV). The correlations > 0.88, describing the relationships between samples were retained only for distance matrices from genus and OTU levels of analyses and were also reproduced in all four data types (Figure 4). On the other hand, the initially high correlation between OTU and ASV at 16S rRNA level dropped below 0.55 for KO, EC and Pathway datasets, reflecting the increased number of categories (genus = 148, OUT = 1328, ASV = 13,244) and their different numerical abundance [11]. These results illustrate how the user selected levels of taxonomic assignment of the sequence data can affect the relationships between samples. Switching from genus level to utilizing ASV level of analysis does not only represent a way to maximize information content of the underlying 16S rRNA sequences [30], but it also represents a distorting transformation of the information due to the many predominantly biological limitations of such analyses: (i) differences in 16S rRNA gene copy numbers range from 1 to 15 in bacteria and 1 to 5 in Archaea [37], hence a frequently recovered sequence may represent a high copy number taxon of lesser abundance, or a low copy number taxon of higher abundance. This 16S copy number of the organism that contributed the sequence is estimated and data adjusted accordingly by utilizing PICRUSt2 [27] in GUMPP; (ii) intragenomic heterogeneity of 16S rRNA operons can be as large as 20.4%. Genus level classification encompasses rather divergent sequences of that specific genus into one category. On the other hand, single nucleotide polymorphism present within e.g., 10 copies of 16S rRNA operon within one organism represent distinct ASVs. In comparison to genus level analysis 16S rRNA variants of one organism are split to several ASV categories inflating ASV estimates of microbial taxonomic diversity and of functional diversity of underlying metagenomes [38–40]; (iii) In contrast, almost identical 16S rRNA copies and hence the lack of differences found within some genera do not enable stratification of species and strains present within, falsely deflating the number of present ASVs [10,38–42]; (iv) different hypervariable regions of 16S rRNA utilized in amplicon sequencing can result in additional distortion of signal relative to each other [43] hence compromising direct comparison of the results between studies utilizing distinct primers.

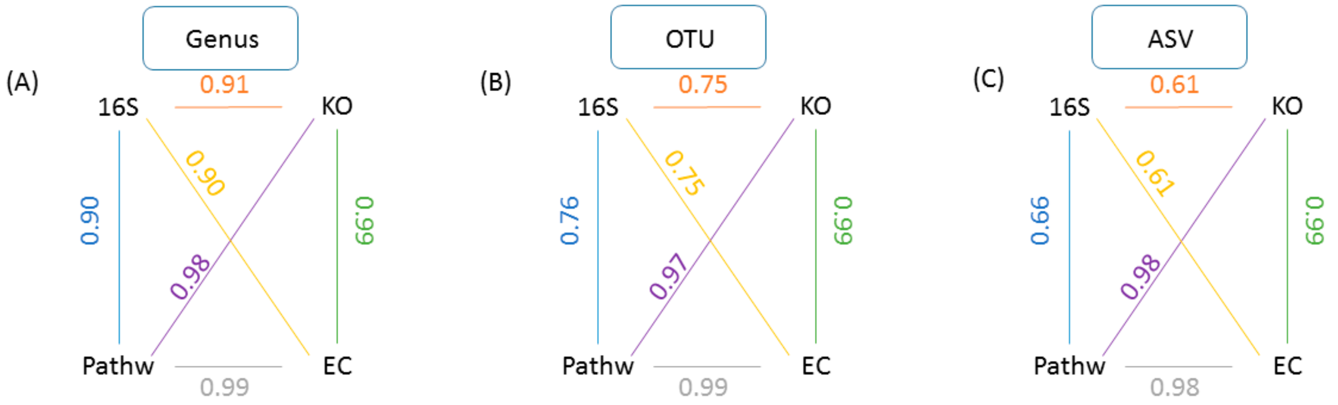


Figure 3. A within level analysis for all derived data types. A schematic representation of GUMPP generated data types analyzed at each of the three levels of 16S rRNA analysis (**A**) genus, (**B**) OTU, (**C**) ASV for the same sequence dataset and extended further to respective predicted functional genes (KO), enzymatic reactions (EC) and metabolic pathways (Pathw). Numbers designate the Mantel test correlation coefficients between various pairs of data types: (i) 16S and functional genes (KO)(orange), (ii) 16S and enzymatic reactions (EC) (yellow), (iii) 16S and metabolic pathways (Pathw) (blue), (iv) pathw and EC (pur-ple), (v) pathw and EC (grey), (vi) KO and EC (green). All analyses were performed with 9999 permutations and were statistically significant (p = 0.0001).

These cautionary notes listed above are intended to raise the awareness of the biological caveats of the genus, OUT and ASV levels of analyses for users. From this integrative view of biological influences the genus level analysis fits a more reserved type of analysis with arguably lower resolution, but congruent with an existing microbial taxonomy system in comparison to the ASV level of analysis, whereas OTU represents a compromise [14,44]. By utilizing ASV some genera expand into species and strains that have sufficient diversity within the 16S rRNA and contribute to ASVs, while other genera that contain species and strains with identical 16S rRNA in the region analyzed do not [14,44]. This biological distinction between genus, OTU or ASV levels of analysis has potentially large implications for the information forwarded to subsequent data types (functional genes, enzymatic reactions, metabolic pathays) irrespective of program utilized (PICRUSt, Tax4Fun, Piphillin or GUMPP).

Recent research highlights the risk of splitting a single bacterial genome into separate clusters when ASVs are used to analyze 16S rRNA gene sequence data. Although there is also a risk of clustering ASVs from different species into the same OTU when using broad distance thresholds, those risks are of less concern than artificially splitting a genome into separate ASVs and OTUs [14,44]. Based on the results presented here (Figures 3 and 4), the choice of level of analysis (genus, OTU, ASV) is far from arbitrary and may lead researchers to draw different biological conclusions. The work presented in this study highlights the utility of GUMPP that enables researchers to analyze the data at all three levels at the same time, generates functional gene, enzymatic reactions and metabolic pathways datasets for downstream machine learning exploration in relation to human diseases [44].

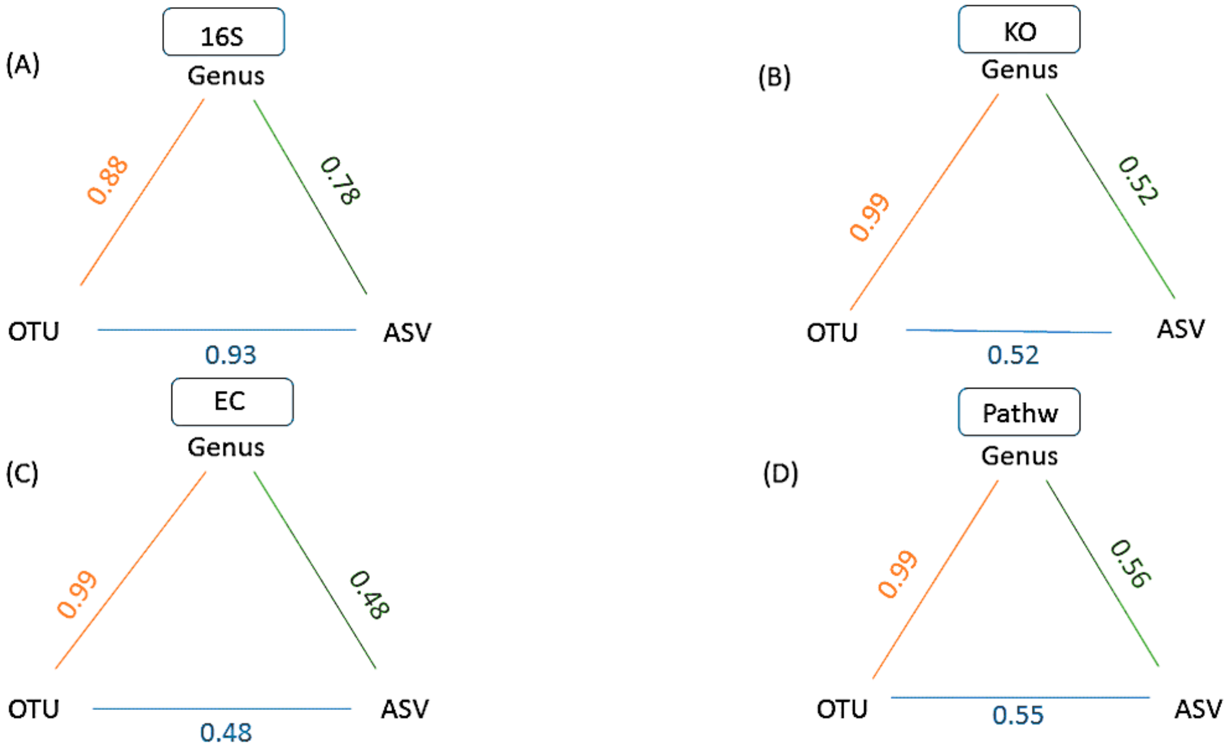


Figure 4. A schematic representation of GUMPP results showing a between level correlations for each data type: (**A**)16S rRNA (16S), (**B**) functional genes (KO), (**C**) enzymatic reactions (EC) and (**D**) metabolic pathways (Pathw). Numbers designate the Mantel test correlation coefficients between various pairs of levels for the same data type: (i) Genus and OTU (orange), (ii) OTU and ASV (blue), (iii) Genus and ASV (green). All analyses were performed with 9999 permutations and were statistically significant (p = 0.0001).

2.3. Reanalysis and Extension of Human Gut Microbiome Data Using GUMPP

In this study a reanalysis of published human gut data (n = 307) [45] was conducted utilizing GUMPP at the levels of 16S rRNA, predicted metagenomes, enzymatic reactions and metabolic pathways. Differences between the gastrointestinal patients (n = 121) from a single ward and 186 healthy volunteers were explored. This effectively enabled us to reproduce previously reported findings [45] utilizing GUMPP. Analyses were extended to three additional data types: predicted functional genes, enzymatic reactions and metabolic pathways. First, as reported before in the original study [45], gut microbial community description was not sufficient to differentiate the subjects based on their underlying five broad medical diagnoses: (i) ulcerative colitis; (ii) Crohn’s disease, (iii) tumor (pancreatic, gastric or liver cancer), (iv) infection (pneumonia, cholangitis, hepatitis, gastritis or pancreatitis) and (v) other (cirrhosis or peptic ulcers, unidentifiable abdominal pain) [45]. The three mixed clusters independent of the underlying medical diagnosis were also reproduced (Figure S3), showing the robustness of GUMPP analysis. Second, by calculating the statistical power for each medical diagnosis a much larger number of samples (within each medical diagnosis) would be needed ($n > 1000$) to be able to build classification models for each diagnosis (Table S2). Third, the PCA representation confirmed the existence of a core microbiome in healthy individuals as described in the original study [45]. Human gut microbiome in patients was disturbed and significantly altered relative to the healthy microbiome (Figure S3).

Extending the original 16S rRNA analysis by GUMPP derived datasets (functional genes (KO), enzymatic reactions (EC), metabolic pathways (pathway)) enabled us to explore the differences between the gastrointestinal patients and healthy volunteers utilizing machine learning. This coupling between GUMPP produced datasets and machine learning enabled us to generate, train and validate four separate models for classification of samples (Figure S3; Supplementary Electronic Material) using JADBIO AutoML approach [46,47]. In short, at all four data levels, logistic ridge regression with penalty hyperparameter lambda = 0.1 was selected as the best interpretable model with AUC metrics of 0.937 (16S rRNA), 0.949 (KO), 0.954 (EC), and 0.947 (pathway) (Figure S3). For the best microbial feature selection, LASSO algorithm was selected for the most differentiating pathways, and Test-Budgeted Statistically Equivalent Signature (SES) algorithm was selected for the search of the most differentiating 16S rRNA, KO and EC between groups of patients and healthy individuals. Models based on KO and EC data performed better than those based on 16S rRNA and pathway data (Figure S4).

The optimization of model selection allowed us to reliably identify microbial features (taxa, functional genes, enzymatic reactions, metabolic pathways) from datasets analyzed and produced by GUMPP (Figure 1, Figures S2 and S3) that discriminated between gut microbiomes of gastrointestinal patients and healthy volunteers: 25 taxonomy level 16S rRNA OTUs, four KOs, 12 ECs and 15 pathways (Table S1). As the complete in-depth biological description of these results is beyond the scope of this study, the major differences between the healthy in diseased groups at the level of metabolic pathways are reported (Figure 5). The following findings are highlighted as proof of concept of GUMPP extended data analysis: lactocepin (EC:3.4.21.96; K01361) was identified in this study as one of the most important features at the level of functional genes and enzymatic reactions distinguishing healthy from IBD, UC and CD. High lactocepin in healthy cohort is involved in the selective degradation of pro-inflammatory chemokines, leading to reduced cell infiltration and reduced inflammation in IBD models [48,49]. Further, Cu+-exporting ATPase were also found to be significantly increased in healthy, hence acting at the level of enzymatic reactions in metabolism [50]. In contrast, the elevated values of the P-type Mg^{2+} transporter observed in gastrointestinal patients were previously shown to be important for increased virulence in *Escherichia coli* and *Salmonella thyphimurium* [51]. Similarly, higher activity of enzyme maltose-6′-phosphate glucosidase were identified in the maltose degradation pathway of *Enterococcus faecalis* leading to increased virulence of this pathogen [52]. Another important enzyme NADH oxidase that exerts the main protection against oxidative stress in the human gut was low in the healthy group [53]. Thiazole component of thiamine diphosphate biosynthesis pathway I and thiamine phosphate synthase were identified as important for separation between healthy and diseased individuals [54–56]. One of the distinguishing features was also the peptidoglycan biosynthesis pathway IV, previously described in *Ruminococcus gravus*, which is abundant in the intestines of patients with Crohn's disease [57]. Bifibacterium shunt was identified as another pathway that has been previously shown to be important in providing positive health benefits to their host with its metabolic activities [58].

These results illustrate the insight supported by GUMPP into the potential differences in the gut microbiomes, functional genes, enzymatic reactions and metabolic pathways between the diffuse group of gastrointestinal patients (five medical diagnoses) and healthy cohort coupled to machine learning.

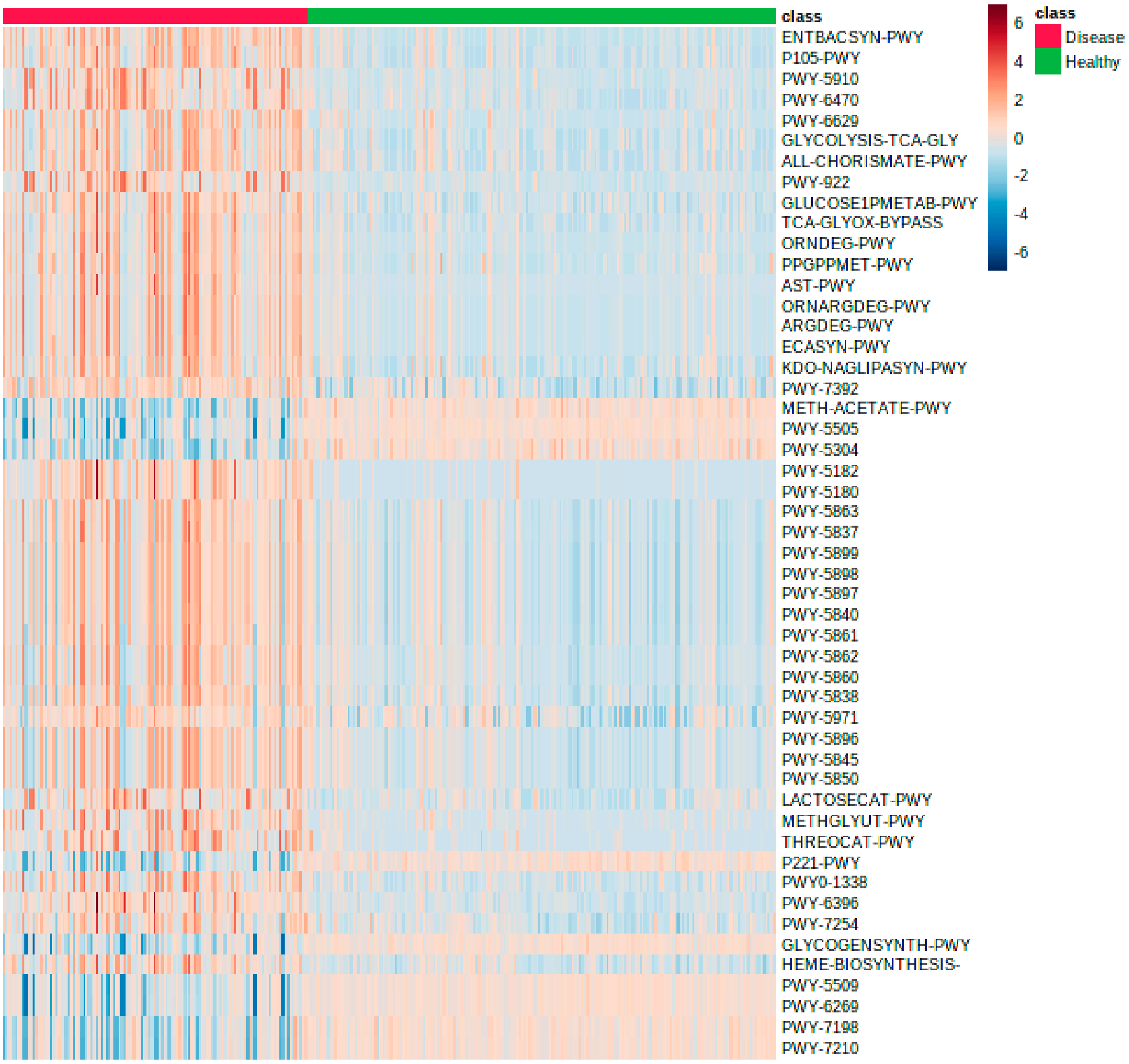


Figure 5. Heatmap showing the differences between the gastrointestinal patients (n = 121; red) from a single ward compared to 186 healthy volunteers (green) utilizing metabolic pathway information produced by GUMPP workflow from 16S rRNA data published before [45,59]. The first 50 most informative pathways are shown.

3. Materials and Methods

3.1. GUMPP Implementation

GUMPP utilization is described in user manual, Electronic Supplementary Materials, Config file, all available as part of this publication at http://gumpp.fe.uni-lj.si. Analyses running GUMPP were executed on a Dual Xeon system with 32 CPU cores (64 hyper-threads), 512 GB of RAM and 6 TB SATA disk. The runtime depends on the data size, sequencing depth and type of analysis (genus-, OTU-, ASV- level). For instance, human gut microbiome data analysis consisted of 307 samples, that each contained independent forward (R1) and reverse (R2) files. In total, it took <10 h, <50 h and <60 h runtime to finalize genus-, OTU- and ASV- levels of analyses, respectively. Similarly, runtime of analyzing less deeply sequenced mice dataset (n = 365 paired-end samples) took <4, <16 and <18 h to finalize genus-, OTU- and ASV- levels of analyses, respectively. Portability and HPC performance of the GUMPP generated in this study was confirmed on Leo3e (https://www.uibk.ac.at/zid/systeme/hpc-systeme/leo3e/ accessed on 30 April 2021) and Leo4 (https://www.uibk.ac.at/zid/systeme/hpc-systeme/leo4/ accessed on 30 April 2021) HPC infrastructure of the University of Innsbruck as described recently [60].

3.2. Sequence Data Collections

The workflow was tested using two large collections of data sets arising from human [45,59] and mice experiments ([7]; https://mothur.org/ accessed on 30 April 2021). In short, a multi-disease hospitalized cohort included various gastroenterological pathologies: ulcerative colitis, Crohn's disease, tumor, infection, cirrhosis and peptic ulcer, unidentifiable abdominal pain. Gastrointestinal patients (n = 121) from a single ward were compared to 186 healthy volunteers [45] in order to fine-map the gut microbiota dysbiosis, using the bacterial (V3 V4) amplicon sequencing. In total, 6.6 million pairs of sequences were analyzed with an average coverage of 35,484 pairs of sequence reads from the 16S rRNA gene.

The mice dataset explored the separation between daily murine fecal samples (n = 360) obtained from C57BL/6 male and female mice at 0 to 9 (early) and 141 to 150 (late) days after weaning [7]. In total, 4.3 million pairs of sequence reads from the 16S rRNA gene with an average coverage of 9913 pairs of V4V5 reads per sample [22] were analyzed. During the first 150 days post weaning mice were allowed ad libitum feed with no specific influence in order to monitor whether the rapid change in weight at 10 days post weaning (obesity) affected the stability microbiome compared to the microbiome observed between days 140 and 150.

3.3. Statistical Analyses and Machine Learning

The two 16S rRNA sequence data collections were analyzed using GUMPP and according to three layers of information, namely genus, 97% OTU and ASV, and the additional three data types were calculated using PICRUSt2 integrated in GUMPP: predicted metagenomes; enzyme reactions; metabolic pathways. Piphillin-ready outputs for clinical exploration were calculated alongside, formatted and prepared. The underlying settings used in these analyses are part of the GUMPP configuration file and can be utilized and shared among researchers for reproducibility and ease of additional calculations. The resulting genus level data analysis of human gut microbiomes (four data matrices (16S rRNA; metagenomes; enzyme reactions; metabolic pathways) were subjected to machine learning in JADBIO [47] (version 1.1.164) for identification of microbial, genetic, enzymatic and pathway variables responsible for separation of the healthy and patient groups.

JADBIO [47] provides high-quality predictive models for diagnostics using state-of-the-art statistical and machine learning methods. Personal analytical biases and methodological statistical errors were eliminated from the analysis by autonomous exploration of several settings in modeling steps, exploring wide analytical space and producing convincing discovered features to discriminate between patients and healthy individuals. The JADBIO approach was adopted for modeling because of number of reason: First, automated parameter and algorithm selection without human inference enables testing and coverage of a wide machine learning algorithm-settings space. Second, JADBIO includes several algorithms for feature selection and modeling (linear regression, SVM, decision trees, random forest and Gaussian kernel SVMs) and all possible options with different parameters are tested during the process. Third, the obtained models were trained with different configurations of sub-data of the original dataset (all results are cross-validated with recently developed Bootstrap Bias Corrected CV (BBC-CV) [61]). Fourth, analyses were run on data with biomedical characteristics (sparse matrices, nonnormal distributions). Algorithm, hyperparameter and space selection protocols (AHPS) in JADBIO were used for selecting the most appropriate algorithm for preprocessing and transformation of a given dataset, for feature selection and modeling. The output of AHPS step was then evaluated through the configuration evaluation protocol in order to find the optimal model configuration for a given dataset [46,47]. JADBIO 1.1.164 was used with extensive tuning effort and 6 CPU cores in modelling various dataset selections. All four datasets were split 70 to 30 according to machine learning protocols. The training set (70% of the data in the dataset) was used to build the best interpretable models and the rest of the data (30%) was used for performance validations at all four levels of data analysis (16S rRNA genus level (424 features), KO (6126 features), EC (1887 features), pathways (365 features)). The area under the curve (AUC) metric was used to evaluate model performance. In total, the analytical space of algorithms and their corresponding settings was explored and 5960 of models and their individual settings were tested for genus and 11,920 for functional gene, enzymatic reactions and metabolic pathways, before the optimal configuration for the most informative model were obtained.

In addition to this, statistical power analysis of human microbiome data was performed [45,59] on all four data levels: 16S rRNA, KO, EC and pathways, between patients with different diseases and healthy individuals and according to presence/absence of the disease. Data was cube root normalized and mean centered. False discovery rate set to 0.1 was used in MetaboAnalyst module prepared for data analysis of population and metabolic studies [62].

All models created in analyses of the human gastrointestinal dataset can also be run on the local machine and are provided as part of the supplementary data (for local model execution, see the instructions in the electronic supplementary materials).

Mice data (n = 365) were processed and analyzed as described above in order to explore the differences between the four data types (16S rRNA; metagenomes; enzyme reactions; metabolic pathways) in terms of consistency of intersample relationships between the three layers of information routinely utilized in studies (genus; OTU; ASV). The intersample relationships were assessed by Mantel tests ($p < 0.0002$) utilizing (i) Pearson and (ii) Spearman correlation between data matrices (Bray-Curtis distance measure) and permutations (n = 9999) in either vegan-R [63] and/or PAST software (version 2.17c) [64]. The Mantel test tests the correlation between two distance matrices. It is non-parametric test and computes the significance of the correlation through permutations of the rows and columns of the input distance matrices.

4. Conclusions

By including the user preferences of genus, OTU or ASV type of analyses, GUMPP is the first workflow that introduces traceability and portability of all its parameters used in analyses. The workflow integrates and orchestrates end to end the inputs and outputs of the highly cited programs Mothur, PICRUSt2 and Pipihillin, controlled by Python code, delivered as portable Singularity image and accompanied by customizable configuration files. The whole GUMPP workflow can be executed for teaching or/and research purposes using personal computer or HPC. The ability to support reproducible analyses enables production of datasets that match multiomics layers of information, such as metagenomics, metaproteomics and metabolomics that ultimately identify the biochemical pathways characteristic of certain pathology [8]. These datasets coupled to biostatistics and mathematical approaches of machine learning can play significant role in extraction of truly significant and meaningful information from wide array of previously unexplored datasets (e.g., [45,59]) in relation to (i) a number of diseases (metabolic [65] or neurodegenerative [66] diseases), (ii) medical interventions, manipulations of bacteria-gut-brain axis [67] or (iii) treatment strategies for complex diseases [68]. The adoption of GUMPP in the gut-microbiota related research enables focusing on the identification of novel biomarkers that can lead to the development of mechanistic hypotheses applicable to the development of novel therapies in personalized medicine [2,9].

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/metabo11060336/s1. Figure S1: A schematic overview of data layers, Figure S2: The data can be analyzed at three different levels, Figure S3: An overview of the modelling step based on the four layers of information obtained through the use of GUMPP, Figure S4: An overview of characteristics of the models based on 16S rRNA, predicted metagenomes (KO), predicted enzymatic reactions (EC) and metabolic pathways (Pathway) data. KO and EC data performed slightly better than those based on 16S rRNA and pathway data, Table S1: Performance metrics of built models based on four different levels of data generated by GUMPP from human dataset, Table S2: Human dataset, power analysis. Sample size corresponding to calculated statistical power, Minimanual 1: GUMPP's quick run routine, Minimanual 2: Instructions for running a model on a local machine.

Author Contributions: Conceptualization of analysis, B.S.; methodology, B.M., L.D., B.S.; formal analysis, L.D., B.S., B.M.; data curation, L.D., B.S., B.M.; writing—original draft preparation, B.S., L.D.; visualization L.D., B.S.; project administration, B.S.; funding acquisition, B.S., B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Slovenian Research Agency programs (P2-0095). LD acknowledges the support of the Slovenian Research Agency (SRA MR+#51867). BS was partially supported through guest professorship awarded by University of Innsbruck.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: GUMPP Singularity image, next to accompanying config files, manual and demo data is available here for download: http://gumpp.fe.uni-lj.si. The calculated models are made available as Electronic Supplementary Materials. Human and mice sequencing data were previously published and are available from original publications ([45,59] and ([7]; https://mothur.org/), respectively).

Acknowledgments: The authors thank Aleksander Mahnic and Maja Rupnik, National Laboratory for Health, Environment and Food, Maribor, Slovenia for releasing the primary raw data. "The computational results presented have been achieved (in part) using the HPC infrastructure of the University of Innsbruck" using Leo3 and Leo4e. The ongoing support from the side of Heribert Insam, Department of Microbiology, University of Innsbruck is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Stres, B.; Kronegger, L. Shift in the paradigm towards next-generation microbiology. *FEMS Microbiol. Lett.* **2019**, *366*. [CrossRef]
2. Vernocchi, P.; Del Chierico, F.; Putignani, L. Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health. *Front. Microbiol.* **2016**, *7*. [CrossRef]
3. Wu, J.; Wang, K.; Wang, X.; Pang, Y.; Jiang, C. The role of the gut microbiome and its metabolites in metabolic diseases. *Protein Cell* **2020**. [CrossRef]
4. Visconti, A.; Le Roy, C.I.; Rosa, F.; Rossi, N.; Martin, T.C.; Mohney, R.P.; Li, W.; de Rinaldis, E.; Bell, J.T.; Venter, J.C.; et al. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **2019**, *10*, 4505. [CrossRef]
5. Lee-Sarwar, K.A.; Lasky-Su, J.; Kelly, R.S.; Litonjua, A.A.; Weiss, S.T. Metabolome-Microbiome Crosstalk and Human Disease. *Metabolites* **2020**, *10*, 181. [CrossRef] [PubMed]
6. Kappel, B.A.; De Angelis, L.; Heiser, M.; Ballanti, M.; Stoehr, R.; Goettsch, C.; Mavilio, M.; Artati, A.; Paoluzi, O.A.; Adamski, J.; et al. Cross-omics analysis revealed gut microbiome-related metabolic pathways underlying atherosclerosis development after antibiotics treatment. *Mol. Metab.* **2020**, *36*. [CrossRef] [PubMed]
7. Wilmanski, T.; Rappaport, N.; Earls, J.C.; Magis, A.T.; Manor, O.; Lovejoy, J.; Omenn, G.S.; Hood, L.; Gibbons, S.M.; Price, N.D. Blood metabolome predicts gut microbiome alpha-diversity in humans. *Nat. Biotechnol.* **2019**, *37*, 1217–1228. [CrossRef] [PubMed]
8. Jiang, D.; Armour, C.R.; Hu, C.; Mei, M.; Tian, C.; Sharpton, T.J.; Jiang, Y. Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Front. Genet.* **2019**, *10*. [CrossRef]
9. Wang, Q.; Wang, K.; Wu, W.; Giannoulatou, E.; Ho, J.W.K.; Li, L. Host and microbiome multi-omics integration: Applications and methodologies. *Biophys. Rev.* **2019**, *11*. [CrossRef]
10. Poretsky, R.; Rodriguez-R, L.M.; Luo, C.; Tsementzi, D.; Konstantinidis, K.T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE* **2014**, *9*. [CrossRef]
11. Pasolli, E.; Schiffer, L.; Manghi, P.; Renson, A.; Obenchain, V.; Truong, D.T.; Beghini, F.; Malik, F.; Ramos, M.; Dowd, J.B.; et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **2017**, *14*. [CrossRef]
12. Rühlemann, M.C.; Hermes, B.M.; Bang, C.; Doms, S.; Moitinho-Silva, L.; Thingholm, L.B.; Frost, F.; Degenhardt, F.; Wittig, M.; Kässens, J.; et al. Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **2021**, *53*. [CrossRef] [PubMed]
13. Mysara, M.; Vandamme, P.; Props, R.; Kerckhof, F.M.; Leys, N.; Boon, N.; Raes, J.; Monsieurs, P. Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiol. Ecol.* **2017**, *93*. [CrossRef] [PubMed]
14. Schloss, P.D. Amplicon sequence variants artificially split bacterial genomes into separate clusters. *bioRxiv* **2021**. [CrossRef]
15. Callahan, B.J.; McMurdie, P.J.; Holmes, S.P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **2017**, *11*. [CrossRef]
16. Nilakanta, H.; Drews, K.L.; Firrell, S.; Foulkes, M.A.; Jablonski, K.A. A review of software for analyzing molecular sequences. *BMC Res. Notes* **2014**, *7*, 1–9. [CrossRef]

17. Pollock, J.; Glendinning, L.; Wisedchanwet, T.; Watson, M. The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Appl. Environ. Microbiol.* **2018**, *84*. [CrossRef]
18. Schloss, P.D. Reintroducing mothur: 10 Years Later. *Appl. Environ. Microbiol.* **2020**, *86*, e02343-19. [CrossRef]
19. López-García, A.; Pineda-Quiroga, C.; Atxaerandio, R.; Pérez, A.; Hernández, I.; García-Rodríguez, A.; González-Recio, O. Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S rRNA Amplicon Sequences. *Front. Microbiol.* **2018**, *9*. [CrossRef]
20. Winand, R.; Bogaerts, B.; Hoffman, S.; Lefevre, L.; Delvoye, M.; Braekel, J.V.; Fu, Q.; Roosens, N.H.; Keersmaecker, S.C.; Vanneste, K. Targeting the 16S rrna gene for bacterial identification in complex mixed samples: Comparative evaluation of second (illumina) and third (oxford nanopore technologies) generation sequencing technologies. *Int. J. Mol. Sci.* **2019**, *21*, 298. [CrossRef]
21. Turnbaugh, P.J.; Ley, R.E.; Hamady, M.; Fraser-Liggett, C.M.; Knight, R.; Gordon, J.I. The human microbiome project. *Nature* **2007**, *449*. [CrossRef] [PubMed]
22. Kozich, J.J.; Westcott, S.L.; Baxter, N.T.; Highlander, S.K.; Schloss, P.D. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Appl. Environ. Microbiol.* **2013**. [CrossRef] [PubMed]
23. Prodan, A.; Tremaroli, V.; Brolin, H.; Zwinderman, A.H.; Nieuwdorp, M.; Levin, E. Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE* **2020**, *15*. [CrossRef]
24. Dhariwal, A.; Chong, J.; Habib, S.; King, I.L.; Agellon, L.B.; Xia, J. MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **2017**, *45*. [CrossRef] [PubMed]
25. Chong, J.; Liu, P.; Zhou, G.; Xia, J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* **2020**, *15*. [CrossRef] [PubMed]
26. Langille, M.G.; Zaneveld, J.; Caporaso, J.G.; McDonald, D.; Knights, D.; Reyes, J.A.; Clemente, J.C.; Burkepile, D.E.; Vega Thurber, R.L.; Knight, R.; et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **2013**, *31*. [CrossRef] [PubMed]
27. Douglas, G.M.; Maffei, V.J.; Zaneveld, J.R.; Yurgel, S.N.; Brown, J.R.; Taylor, C.M.; Huttenhower, C.; Langille, M.G.I. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **2020**, *38*. [CrossRef]
28. Aßhauer, K.P.; Wemheuer, B.; Daniel, R.; Meinicke, P. Tax4Fun: Predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics* **2015**, *31*. [CrossRef]
29. Wemheuer, F.; Taylor, J.A.; Daniel, R.; Johnston, E.; Meinicke, P.; Thomas, T.; Wemheuer, B. Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ. Microb.* **2020**, *15*. [CrossRef]
30. Narayan, N.R.; Weinmaier, T.; Laserna-Mendieta, E.J.; Claesson, M.J.; Shanahan, F.; Dabbagh, K.; Iwai, S.; DeSantis, T.Z. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genom.* **2020**, *21*. [CrossRef]
31. Iwai, S.; Weinmaier, T.; Schmidt, B.L.; Albertson, D.G.; Poloso, N.J.; Dabbagh, K.; DeSantis, T.Z. Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PLoS ONE* **2016**, *11*. [CrossRef] [PubMed]
32. Sun, S.; Jones, R.B.; Fodor, A.A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **2020**, *8*. [CrossRef] [PubMed]
33. Salosensaari, A.; Laitinen, V.; Havulinna, A.S.; Meric, G.; Cheng, S.; Perola, M.; Valsta, L.; Alfthan, G.; Inouye, M.; Watrous, J.D.; et al. Taxonomic signatures of cause-specific mortality risk in human gut microbiome. *Nat. Commun.* **2021**, *12*, 2671. [CrossRef]
34. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J.; et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef] [PubMed]
35. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **2017**, *12*. [CrossRef]
36. Turnbaugh, P.J.; Hamady, M.; Yatsunenko, T.; Cantarel, B.L.; Duncan, A.; Ley, R.E.; Sogin, M.L.; Jones, W.J.; Roe, B.A.; Affourtit, J.P.; et al. A core gut microbiome in obese and lean twins. *Nature* **2009**, *457*. [CrossRef]
37. Stoddard, S.F.; Smith, B.J.; Hein, R.; Roller, B.R.; Schmidt, T.M. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **2015**, *43*. [CrossRef]
38. Větrovský, T.; Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **2013**, *8*. [CrossRef] [PubMed]
39. Nguyen, N.P.; Warnow, T.; Pop, M.; White, B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms Microb.* **2016**, *2*, 16004. [CrossRef]
40. Pei, A.Y.; Oberdorf, W.E.; Nossa, C.W.; Agarwal, A.; Chokshi, P.; Gerz, E.A.; Jin, Z.; Lee, P.; Yang, L.; Poles, M.; et al. Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microbiol.* **2010**, *76*. [CrossRef]
41. Sun, D.L.; Jiang, X.; Wu, Q.L.; Zhou, N.Y. Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Appl. Environ. Microbiol.* **2013**, *79*. [CrossRef] [PubMed]
42. Huse, S.M.; Dethlefsen, L.; Huber, J.A.; Mark Welch, D.; Relman, D.A.; Sogin, M.L. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* **2008**, *4*. [CrossRef]
43. Soriano-Lerma, A.; Pérez-Carrasco, V.; Sánchez-Marañón, M.; Ortiz-González, M.; Sánchez-Martín, V.; Gijón, J.; Navarro-Mari, J.M.; García-Salcedo, J.A.; Soriano, M. Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Sci. Rep.* **2020**, *10*. [CrossRef]

44. Joos, L.; Beirinckx, S.; Haegeman, A.; Debode, J.; Vandecasteele, B.; Baeyen, S.; Goormachtig, S.; Clement, L.; De Tender, C. Daring to be differential: Metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. *BMC Genom.* **2020**, *21*. [CrossRef] [PubMed]
45. Mahnic, A.; Breskvar, M.; Dzeroski, S.; Skok, P.; Pintar, S.; Rupnik, M. Distinct Types of Gut Microbiota Dysbiosis in Hospitalized Gastroenterological Patients Are Disease Non-related and Characterized With the Predominance of Either Enterobacteriaceae or Enterococcus. *Front. Microbiol.* **2020**, *11*. [CrossRef]
46. Mustafa, A.; Rahimi Azghadi, M. Automated Machine Learning for Healthcare and Clinical Notes Analysis. *Computers* **2021**, *10*, 24. [CrossRef]
47. Tsamardinos, I.; Charonyktakis, P.; Lakiotaki, K.; Borboudakis, G.; Zenklusen, J.C.; Juhl, H.; Chatzaki, E.; Lagani, V. Just Add Data: Automated Predictive Modeling and BioSignature Discovery. *bioRxiv* **2020**. [CrossRef]
48. Hörmannsperger, G.; von Schillde, M.A.; Haller, D. Lactocepin as a protective microbial structure in the context of IBD. *Gut Microbes* **2013**, *4*. [CrossRef]
49. von Schillde, M.A.; Hörmannsperger, G.; Weiher, M.; Alpert, C.A.; Hahne, H.; Bäuerl, C.; van Huynegem, K.; Steidler, L.; Hrncir, T.; Pérez-Martínez, G.; et al. Lactocepin secreted by Lactobacillus exerts anti-inflammatory effects by selectively degrading proinflammatory chemokines. *Cell Host Microbe* **2012**, *11*. [CrossRef]
50. Osman, D.; Patterson, C.J.; Bailey, K.; Fisher, K.; Robinson, N.J.; Rigby, S.E.; Cavet, J.S. The copper supply pathway to a Salmonella Cu,Zn-superoxide dismutase (SodCII) involves P(1B)-type ATPase copper efflux and periplasmic CueP. *Mol. Microbiol.* **2013**, *87*. [CrossRef]
51. Subramani, S.; Perdreau-Dahl, H.; Morth, J.P. The magnesium transporter A is activated by cardiolipin and is highly sensitive to free magnesium in vitro. *eLife* **2016**, *5*. [CrossRef] [PubMed]
52. Joyet, P.; Mokhtari, A.; Riboulet-Bisson, E.; Blancato, V.S.; Espariz, M.; Magni, C.; Hartke, A.; Deutscher, J.; Sauvageot, N. Enzymes Required for Maltodextrin Catabolism in Enterococcus faecalis Exhibit Novel Activities. *Appl. Environ. Microbiol.* **2017**, *83*. [CrossRef] [PubMed]
53. Yan, M.; Yin, W.; Fang, X.; Guo, J.; Shi, H. Characteristics of a water-forming NADH oxidase from Methanobrevibacter smithii, an archaeon in the human gut. *Biosci. Rep.* **2016**, *36*. [CrossRef]
54. Yoshii, K.; Hosomi, K.; Sawane, K.; Kunisawa, J. Metabolism of Dietary and Microbial Vitamin B Family in the Regulation of Host Immunity. *Front. Nutr.* **2019**, *6*. [CrossRef]
55. LeBlanc, J.G.; Milani, C.; de Giori, G.S.; Sesma, F.; van Sinderen, D.; Ventura, M. Bacteria as vitamin suppliers to their host: A gut microbiota perspective. *Curr. Opin. Biotechnol.* **2013**, *24*. [CrossRef] [PubMed]
56. Rodionov, D.A.; Arzamasov, A.A.; Khoroshkin, M.S.; Iablokov, S.N.; Leyn, S.A.; Peterson, S.N.; Novichkov, P.S.; Osterman, A.L. Micronutrient Requirements and Sharing Capabilities of the Human Gut Microbiome. *Front. Microbiol.* **2019**, *10*. [CrossRef]
57. Henke, M.T.; Kenny, D.J.; Cassilly, C.D.; Vlamakis, H.; Xavier, R.J.; Clardy, J. Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci. USA* **2019**, *116*. [CrossRef]
58. O'Callaghan, A.; van Sinderen, D. Bifidobacteria and Their Role as Members of the Human Gut Microbiota. *Front. Microbiol.* **2016**, *7*. [CrossRef]
59. Mahnic, A.; Rupnik, M. Different host factors are associated with patterns in bacterial and fungal gut microbiota in Slovenian healthy cohort. *PLoS ONE* **2018**, *13*. [CrossRef]
60. Murovec, B.; Deutsch, L.; Stres, B. Computational Framework for High-Quality Production and Large-Scale Evolutionary Analysis of Metagenome Assembled Genomes. *Mol. Biol. Evol.* **2020**, *37*. [CrossRef]
61. Tsamardinos, I.; Rakhshani, A.; Lagani, V. Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization | SpringerLink. In *Artificial Intelligence: Methods and Applications*; Likas, A., Blekas, K., Kalles, D., Eds.; Springer International Publishing: Cham, Switzerland, 2014.
62. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S.; Xia, J. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **2018**, *46*. [CrossRef] [PubMed]
63. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **2003**, *14*, 927–930. [CrossRef]
64. Hammer, O.; Harper, D.A.T.; Ryan, P.D. PAST: Paleontological statistics software package for education and data analysis. *Palaeontol. Electron.* **2001**, *1*, 9.
65. Proffitt, C.; Bidkhori, G.; Moyes, D.; Shoaie, S. Disease, Drugs and Dysbiosis: Understanding Microbial Signatures in Metabolic Disease and Medical Interventions. *Microorganisms* **2020**, *8*, 1381. [CrossRef] [PubMed]
66. Rosario, D.; Boren, J.; Uhlen, M.; Proctor, G.; Aarsland, D.; Mardinoglu, A.; Shoaie, S. Systems Biology Approaches to Understand the Host-Microbiome Interactions in Neurodegenerative Diseases. *Front. Neurosci.* **2020**, *14*. [CrossRef] [PubMed]
67. Sarkar, A.; Lehto, S.M.; Harty, S.; Dinan, T.G.; Cryan, J.F.; Burnet, P.W.J. Psychobiotics and the Manipulation of Bacteria-Gut-Brain Signals. *Trends Neurosci.* **2016**, *39*. [CrossRef]
68. Vijay, A.; Valdes, A.M. The Metabolomic Signatures of Weight Change. *Metabolites* **2019**, *9*, 67. [CrossRef]