gNOMO: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms

Maria Muñoz-Benavent
Institute for Integrative Systems Biology
Felix Hartkopf
Bundesanstalt fur Materialforschung und -prufung
Tim Van Den Bossche
Universiteit Gent
Vitor C. Piro
Robert Koch Institut
Carlos García-Ferris
Institute for Integrative Systems Biology
Amparo Latorre
Institute for Integrative Systems Biology
Bernhard Y. Renard
Hasso Plattner Institute
Thilo Muth (thilo.muth@fu-berlin.de)
Bundesanstalt fur Materialforschung und -prufung https://orcid.org/0000-0001-8304-2684

Software

Keywords: gNOMO, multi-omics pipeline, non-model organisms, microbiome data analysis

Posted Date: December 18th, 2019

DOI: https://doi.org/10.21203/rs.2.19121/v1

Version of Record: A version of this preprint was published at NAR Genomics and Bioinformatics on October 9th, 2020. See the published version at https://doi.org/10.1093/nargab/lqaa083.

Abstract

Background The study of bacterial symbioses has grown exponentially in the recent past. However, existing bioinformatic workflows of microbiome data analysis do commonly not integrate multiple meta-omics levels and are mainly geared towards human microbiomes. Microbiota are better understood when analyzed in their biological context, which is together with their host or environment, but this is a limitation when studying non-model organisms mainly due to the lack of well-annotated sequence references.

Results Here, we present gNOMO, a bioinformatic pipeline that is specifically designed to process and analyze non-model organism samples of up to three meta-omics levels: metagenomics, metatranscriptomics, and metaproteomics in an integrative manner. The pipeline has been developed using the Snakemake framework in order to obtain an automated and reproducible workflow. One of the key features is the on-the-fly creation of a tailored proteogenomic database based on metagenomics and metatranscriptomics data, leading to improved protein identification, taxonomic and functional analysis. gNOMO combines meta-omics analysis of the host with its bacterial population and allows to investigate both host and microbiome of non-model organisms with commonly insufficiently complete reference databases.

Conclusions Using experimental datasets of the German cockroach Blattella germanica , a non-model organism with very complex gut microbiome, we show the capabilities of gNOMO with regard to meta-omics data integration, expression ratio comparison, taxonomic and functional analysis as well as intuitive output visualization. gNOMO includes functional information of metagenomics, metatranscriptomics, and metaproteomics data of the microbiome in the same visualization facilitating the interpretation of the results. Moreover, host data can be analyzed in parallel to obtain an equivalent output that allows to study the metabolic situation of the whole symbiotic system. Finally, the metaproteomics identification and annotation are optimized using a tailored proteogenomics database automatically obtained within the gNOMO workflow. In conclusion, gNOMO is a fully automated pipeline, for integrating and analyzing multiple meta-omics data and for producing useful output visualizations. In addition, it is specifically designed for data from non-model organisms. The gNOMO pipeline is freely available under the Apache 2.0 open-source license and can be downloaded from https://gitlab.com/rki_bioinformatics/gnomo .

Background

Symbiosis is a widespread relationship present in all groups of organisms but intensely developed between animals and bacteria that benefit from each other in order to survive. Consequently, both acquire an evolutionary advantage in comparison to individuals lacking this relationship. Two different types of symbiosis can be distinguished: ectosymbiosis, in which bacteria are attached to the surface of the host, and endosymbiosis, which usually is a mutualistic relationship, where bacteria live intracellularly in the host and are transmitted vertically [1,2]. To understand these evolutionary relationships host and

symbionts are best studied together. In mutualistic symbiosis, the eukaryotes provide a safe environment for endosymbiotic bacteria that live in close interaction with the host. In return, the endosymbionts provide nutrients and metabolites (such as essential amino acids or vitamins) to the host that cannot be obtained in any other way. For example, it has been estimated that around 15% of insect species maintain endosymbiotic associations with bacteria that supply the host with the nutrients that are lacking in their diets [3]. On the other hand, most insects possess a gut microbiome that affects the physiology of the host by, for example, contributing to metabolic and nutritional needs, and the immune system development [4]. Recently, many studies have been performed in humans to study the gut microbiota [5], but non-model organisms require further investigations to better understand this specific type of symbiosis. In this context, cockroaches are a suitable model, because they have two symbiotic systems, i.e. an endosymbiont (*Blattabacterium cuenoti)* in the fat body and a rich and complex gut microbiota [6,7]. *Blattella germanica* is a hemimetabolous insect (it has an incomplete metamorphosis) with three developmental stages. Regarding its symbionts, genome analysis demonstrated that the endosymbiont *Blattabacterium* contributes to the nitrogen (N) recycling and the synthesis of essential amino acids [8], but the function of the gut microbiota in cockroaches still has to be elucidated. It has been shown that the gut microbiome of cockroaches shows much overlap with the one in humans probably reflecting a similar omnivorous diet [6,9,10].

Recently, research interests in microbial communities have been strongly increased due to findings on the impact of the microbiome on human health [11,12]. Microbiome studies often employ meta-omics techniques such as metagenomics [13] that aims to analyze the genetic material from all members in a microbial community sample. Despite many advantages, metagenomics still presents a static gene-centric approach that cannot assess temporal dynamics and functional activities of complex microbial populations [14]. To gain insights into the dynamic functional repertoire of microbial communities, further techniques such as metatranscriptomics and metaproteomics have been established in recent years [15, 16]. Beyond the genome level, these meta-omics analysis approaches allow studying complex microbial systems and their host interactions at the gene expression level (transcripts and proteins, respectively). Used separately, metagenomics, metatranscriptomics, and metaproteomics are already powerful because they complement and mutually support each other. In the past, powerful tailored bioinformatic solutions have been developed for the individual meta-omics analysis levels [13,15,16]. However, the true strength unfolds when these analysis techniques are integrated [17,18]. As a holistic approach, a complete meta-omics integration can extend the capabilities of microbiome and host-related studies in various ways. Most importantly, integrating multiple meta-omics levels allows to expand the possibilities of biological interpretation and to investigate biological pathways from a more comprehensive perspective. Compared to single-omics strategies, an integrative approach provides a deeper and more thorough understanding of how the key players of microbial communities regulate underlying pathway mechanisms [19].

While the integration of meta-omics has been described in previous studies [20], its potential has not been fully exploited so far. In particular, the data analysis is challenging, because studies often present customized in-house workflows that cannot be fully automated or are not reproducible. In general,

automated multi-omics analysis pipelines are rare and limited to few meta-omics levels [21] and are not tailored for host and microbiome analyses of non-model organisms.

Here, we present gNOMO, a meta-omics software pipeline that allows integrating three different levels of omics analyses, derived from metagenomics, metatranscriptomics, and metaproteomics experiments. It provides two different, optionally iterative operating modes: (i) each of the three omics levels can be analyzed separately and independently of each other and subsequently, (ii) up to three omics layers can be analyzed in a fully integrated fashion. The workflow of gNOMO starts from raw data to essential processing steps and finally provides output visualizations for taxonomic classification, functional metabolic pathway profiling, and differential sample analysis. The integration of metagenomics, metatranscriptomics and metaproteomics data is possible due to the production of a tailored proteogenomic database, which optimizes the identification and quantification of peptides in metaproteomics data [22,23]. As microbiota needs to be analyzed in its context, the host is also studied together with the microbiome. Host data can be analyzed without a reference database, which allows to study non-model organisms, and proteins of the host are also identified with a tailored host database obtained from genomics and transcriptomic sequences. The pipeline has been implemented using the Python-based Snakemake [24] framework to perform fully automated and reproducible multi-omics analyses of host and microbiome samples. So far, gNOMO has been developed and optimized for data from non-model organism samples, but it is fully executable on generic sample types, for example, from human or mouse microbiomes. With gNOMO, we aim to fill the gap of barely existing multi-omics pipelines for microbial community samples being able to compare and integrate data at the genome, transcriptome, and proteome level.

Implementation

gNOMO is a pipeline that integrates multiple bioinformatic methods and software tools to analyze metagenomics, metatranscriptomics and metaproteomics data and to provide the results with an easily readable final output. One of the main purposes of integrating such different kinds of multi-omics data is to directly improve the analysis of microbial populations and to investigate their function in poorly characterized environments, such as non-model organisms. At the genome and transcriptome level, our pipeline includes both quality control and data preparation steps, of which parameters can be adjusted depending on the quality of the input data. In addition, gNOMO allows to directly create a proteogenomic database from metagenomics and metatranscriptomics data. This important processing step makes it possible to connect the metagenomics and metatranscriptomics analysis to the protein identification at the metaproteomics level. In particular, the proteogenomic database generation step leads to the full integration of all three omics levels.

The complete gNOMO pipeline is built in Snakemake [24], a management system for bioinformatic workflows, that allows obtaining standardized and reproducible output data. By using Snakemake, the input data can be easily defined, and the parameters are configured for their analysis by editing a configuration file. Further, the gNOMO pipeline including all dependencies is available at the BioConda

channel [25]. Tools added to BioConda provide a user-friendly installation because the required tools and libraries are easily incorporated and automatically installed with the use of Snakemake environments. The gNOMO pipeline typically consists of five main steps (*Figure 1)*:: (1) pre-processing, (2) metagenomics and metatranscriptomics data analysis, (3) proteogenomic database creation, (4) metaproteomics data analysis, and (5) data integration. In the following paragraphs, these individual steps are described in more detail.

Figure 1: Workflow overview of the gNOMO pipeline. A) Initial input of metagenomic (metagenomics) and metatranscriptomic (metatranscriptomics) sequences. B) Pre-processing: cleaning and quality control of metagenomics and metatranscriptomics input sequences. C) metagenomics and metatranscriptomics data analyses: consists of taxonomic and functional annotations. D) Proteogenomic database creation based on metagenomics and metatranscriptomics protein predictions. E) Auxiliary input of metaproteomic (metaproteomics) tandem mass spectrum data. F) metaproteomics analysis: also includes taxonomic and functional annotations. G) Graphical representation/visualization of all integrated meta-omics data.

Pre-processing

The first step includes various pre-processing mechanisms improving metagenomics and metatranscriptomics read quality, including: (i) FastQC [26] for reviewing the quality of the reads, (ii) PrinSeq [27] for cleaning and for trimming the sequences, (iii) a second quality control with FastQC and Fastq-join [28] for binning the pair-end reads. This binning step is included because our workflow is designed for paired-end reads.

Metagenomic and metatranscriptomic analysis

In this step, the pre-processed paired-end sequences are analyzed using pre-configured tools. These tools include (i) a genome mapping against the NCBI non-redundant (nr) database (accessed 5 July 2019) using Kaiju [29], (ii) a protein prediction using both Prodigal [30] for bacterial proteins and (iii) Augustus [31] for host proteins and functional annotation of this predicted proteins using EggNOG (version 1.0 accessed 5 June 2019) [32] to obtain KEGG Orthology (KO) identifiers. The protein prediction follows an assembly of the binned reads to ensure a proper functional annotation. An optional step is included that requires the installation of InterProScan [33]. This software is not implemented in BioConda but will be automatically installed locally with the snakemake script and allows a TIGRFAM [34] functional annotation. Details regarding the quality of the annotation in metagenomics and metatranscriptomics are available in the Additional file 1 (*Additional file 1: Table S1)*..

Proteogenomic database generation

The output of the previous bacterial prediction from the metagenomics and metatranscriptomics data is used to create a proteogenomics database. This database includes bacterial and host proteins from metagenomics, metatranscriptomics or both kinds of data. A database with both kinds of information

provides a comprehensive reference for peptide and protein identification (see next paragraph). The proteogenomic database obtained from the validation data has been built with the sequences resulting from the bacterial protein prediction performed with Prodigal. This database (data of creation: 19 November 2019) contains 1 014 200 sequences, of which 850 455 are unique (i.e., occur only once in the database).

Metaproteomic data analysis

For peptide and protein identification, MS-GF+ [35] is used as database search engine, employing the custom proteogenomic database as reference for peptide-to-spectrum matching. Both taxonomic and functional annotations of the peptides are performed with Unipept version 4.0 [36]. The output obtained from this step is a taxonomic annotation at three different levels and the Enzyme Commission (EC) number associated with each peptide. To assess the performance of our tailored database, we compared the peptide identification yield with a very complete human gut microbial protein database: NIH Human Microbiome Project Gastrointestinal database (accessed 25 November 2019) (*Additional file 1: Table S2)*.. With our tailored database we obtained four times more peptides identified than using the NIH Gastrointestinal database. These results are consistent with previous studies on the use of metagenomic sequences for constructing proteogenomics databases [37].

Meta-omics data integration and visualization

The final step concerns the integration and visualization of all three-level meta-omics data and results. The taxonomic annotation of the microbiome is visualized with KronaPlots [38]. These plots show the taxonomic distribution in each sample for each data type. To analyze this information further, Linear discriminant analysis (LDA) Effect Size (LEfSe) [39] is used that performs a statistical analysis on the microbiome data. LEfSe identifies features most likely to explain differences between conditions by coupling standard statistical tests with additional tests encoding biological consistency and effect relevance. The statistics performed are Kruskal-Wallis rank-sum test on classes, Wilcoxon rank-sum test among subclasses and LDA score on relevant features. Taking account of the effect size is essential to properly analyze microbiomes. The outcome of the statistical analysis is depicted in a graph with up to two levels of classification, and only the features with a LDA score over 2 are shown. This allows visualizing different conditions and different data within the same graph. Finally, for the functional annotation, the representation of the metabolic pathways is included using Pathview [40]. The Pathview plots represent the log2 ratio of the means of the different conditions and data compared, after a fold change normalisation. This R-based tool shows the differential expression of the enzymes on graphs visualizing the selected metabolic pathways. Pathview itself uses functional pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41].

Validation data

B. germanica population originated from a stable laboratory population housed by Dr. X. Bellés' group at the Institute of Evolutionary Biology (CSIC-UPF, Barcelona). It was reared in chambers at the Institute for

Integrative Systems Biology (University of Valencia) at 25 °C, 60% humidity and a photoperiod of 12L:12D. Cockroaches were fed dog-food pellets (Teklad global 21% protein dog diet 2021C, Envigo, Madison, USA) and water *ad libitum.* Samples were taken at 10 days and 20 days after becoming adults, conditions names 10d and 20d, respectively. Vivisections of CO_2-anesthetized females were performed to obtain the hindgut of each individual. DNA and RNA samples were obtained from the same hindgut, with a total of 12 samples (6 replicates per condition). Protein samples were obtained from individuals of the same age and population, with a total of 8 samples (with 4 replicates per condition). Hindgut was ground with a sterile plastic pestle. DNA and RNA extraction of each hindgut was performed using Nucleospin RNA XS and Nucleospin DNA/RNA Buffer Set (Macherey-Nagel, France). Protein extraction of each hindgut was performed solubilizing the ground hindgut with lysis buffer (7 M urea, 2 M thiourea, 4% (w/v) CHAPS). Metagenomic sequencing using the Illumina MiSeq (2 x 300 bp) technology was done at the FISABIO (Valencia, Spain). Metaproteomics shotgun sequencing was performed by the Proteomics Unit of the Servei Central de Suport a la Investigació Experimental (SCSIE) at the University of Valencia.

Results

To illustrate the outputs and analysis that can be obtained from this pipeline, we used a complex gut microbiota dataset from the non-model organism *B. germanica,* which genome has been sequenced (without being fully annotated) [42]. This dataset consists of metagenomics, metatranscriptomics and metaproteomics data of two different adult conditions: 10d and 20d.

Comparison of metagenomics and metatranscriptomics/metaproteomics data sets for one-condition sample (Multi-meta-omic approach)

Assessing bacterial composition from metagenomics and metatranscriptomics data

The analysis of microbial community samples often raises the question of which bacteria form a given population. To answer this question, we performed two different types of analysis using gNOMO. First, we processed and analyzed metagenomics data to investigate the taxonomic composition of a given sample. Second, we analyzed and compared samples of two different conditions: 10d and 20d.

For the first analysis, the output was visualized using a Krona plot that is produced for each metagenomics and metatranscriptomics sample automatically within the gNOMO pipeline. For the first-condition (10d) sample, we observed that the main phyla present in this population were *Bacteroidetes, Firmicutes,* and *Proteobacteria* (*Figure 2).*. After analyzing the taxonomic distribution differences between the 10d and 20d samples, we observed no significant abundance differences in a preliminary analysis (*Additional file 1: Tables S3 and S4).*. In this analysis, the relative abundance of the main phyla and families was calculated in relation to the mean abundance of the two conditions. We observed that the four most abundant phyla distributions match our previous published studies based on 16S gene sequencing, while others (e.g., *Planctomycetes, Deferribacteres and Actinobacteria)* do not match exactly previous studies on this topic [10] (*Additional file 1: Table S3).*. We made similar observations regarding

taxonomic abundances at the family level (*Additional file 1: Table S4).*. In general, this can be explained by the difference concerning the method and annotation between 16S rRNA gene sequencing analysis and metagenomics. 16S rRNA gene sequencing focuses on bacterial data and can be useful in environmental studies due to the lack of fully sequenced bacterial genomes in these kinds of scenarios. In contrast, metagenomics offers higher resolution, enabling a more specific taxonomic classification of sequences as well as the detection of new bacterial genes and genomes [43].

Figure 2: KronaPlot of the taxonomic annotation of a metagenomics sample (condition 10d). Bacterial taxa distribution of metagenomics data, corresponding to condition 10d. The bacterial taxa are classified by taxonomic hierarchy levels, from higher levels in the center of the chart (Kingdom *Bacteria)* progressing outward until genus level.

As described previously, our first analysis provided no clearly visible abundance differences between the two conditions, as we were expecting when studying such a stable situation (both are adult individuals differing in 10 days of development). However, we decided to validate this finding by a more sensitive statistical approach. To investigate this issue further, we used LEfSe [39] as a well-established statistical method for comparing the taxonomic distribution at genus level between 10d and 20d conditions. LEfSe has the advantage of recognizing the hierarchy of the taxonomic classification and accurately calculate statistically significant differences (represented as LDA scores) between different conditions.

Using LEfSe, we found, for example, that *Fusobacterium* (*Fusobacteriaceae* family), was more abundant at 10 days (LDA score > 3) in both metagenomics and metatranscriptomics data (*Figure 3).*. *Fusobacterium* has been related to disease and stress situations in the human gut microbiota [44], but is has also been related to the infants gut microbiota [45]. Conversely, an unidentified genus belonging to the family *Ruminococcaceae,* has been found more abundant in 20d than 10d condition (LDA score > 3) in metagenomics data (*Figure 3a),*, but no differences between conditions have been found in metatranscriptomics data (*Figure 3b).*. Various genera belonging to the family *Ruminococcaceae* have been related to a healthy gut microbiota, like *Ruminococcus* and *Faecalibacterium.* These have been linked to degradation of starch in the human colon making it available for other bacteria in the gut [46], and degradation of cellulose in herbivorous mammals [47]. These differences between 10d and 20d conditions could suggest that, even if the population is very stable along adult stages, it is being rearranged to its final composition. This rearrangement would imply a reduction in *Fusobacterium* and an increase of *Ruminococcaceae* along time (10d against 20d, *Figure 3a).*. On the other hand, *Pseudomonas* genus and an unclassified genus belonging to the family *Pelagibacteraceae* are more abundant only in metatranscriptomics analysis at 20d against 10d (*Figure 3b).*. *Pelagibacteraceae* has been described as a bacterial family localized in marine and freshwater environments [48], but has also been detected in the mouse gut microbiome [49] *Pseudomonas* genus has been related to pathogenicity in animals and plants, and is a commonly detected taxa in the gut of cockroaches [50]. These results suggest that these taxa increase their transcriptional activity but not their abundance in the population along time. By the same reason the unidentified genus of *Ruminococcaceae* reduce its transcriptional activity (is overrepresented at metagenomics level but not at metatranscriptomics level in 20d sample). More

importantly for the present work is the integration of this level of comparison that allows detection of particular taxa that differ significantly in their abundance in different conditions.

Figure 3: LEfSe graph of taxonomic annotation of metagenomics (top) and metatranscriptomics (bottom) data comparing the two conditions: 10d and 20d. Taxa with significant different distribution among the two conditions are identified. Only taxa with LDA scores over 2 are shown. Positive LDA scores are assigned to the taxa overrepresented in the condition 20d (green), and negative LDA scores to the taxa overrepresented in the condition 10d (red). metagenomics data (Fig 3a) and metatranscriptomics data (Fig 3b) are represented.

Functional analysis from integrated metagenomics and metatranscriptomics data for one-condition sample

Once the bacterial population has been described at the taxonomic level, the next step concerns the functional analysis of each microbiome dataset and the qualitative and quantitative differences of assigned functional annotations. To assess the level of transcriptional activity of the population, we compare the metagenomics data (gene pool) and the metatranscriptomics data (transcripts) corresponding to the microbiota of the 10d condition. Integrating metagenomics and metatranscriptomics allows calculating transcript/gene ratios that indicate gene transcriptional activation or repression. For this purpose, we applied LEfSe based on the functional role (or subrole) assignment using TIGRFAM (*Figure 4; Additional file 1: Table S5).*. We observed that energy metabolism (both anaerobic and aerobic metabolisms) and protein production are the most active metabolic pathways (*Figure 4),*, which indicates that the bacterial population is active.

Figure 4: LEfSe graph comparing metagenomics and metatranscriptomics data of TIGRFAM annotation (role and subrole levels) of condition 10d. Taxa with significant different distribution among metagenomics and metatranscriptomics data are identified. Only taxa with LDA scores above 2 are shown. Positive LDA scores are assigned to the functional categories overrepresented in the metatranscriptomics data (RNA, green), and negative LDA scores to the functional categories overrepresented in metagenomics data (DNA, red).

Alternatively, a pathway analysis enables to discover differences between states by using the Pathview R package. An analysis with Pathview shows which specific metabolic pathways (KEGG pathways) have statistically significant correlations between sample types and/or conditions and thereby complements the information provided by LEfSe. In a Pathview graph, an increase of the gene activity involved in a certain pathway can be observed. Our exemplary analysis using Pathview here focuses on the tricarboxylic acid cycleTCA cycle) of the gut microbiota, comparing again gene pool (metagenomics data) against transcripts (metatranscriptomics data) (*Figure 5).*. The TCA cycle consists of a series of oxidative reactions to finally obtain energy (ATP) from oxidative degradation of the acetyl group, in the form of acetyl-CoA, to carbon dioxide. The full cycle can be performed by bacteria in aerobic conditions, but some autotrophic bacteria are also able to perform the reverse TCA cycle (rTCA), and even some

anaerobic bacteria are able to carry out an incomplete TCA cycle, defining the pan-metabolic capabilities for this pathway of the gut microbiota.

We have found that the majority of the enzymes that take part in the TCA cycle are overrepresented at the transcript level. This confirms our previous observations related to energy metabolism (*Figure 4).*. With both analysis methods and their visualizations, we were able to study different levels of complexity of the pan-metabolism of all bacterial populations. We observed that the microbiome actively produces energy and proteins to grow and maintain a very complex population. Beyond the use case shown above, depending on the particular study, other pathways could be analyzed.

Figure 5: KEGG Pathview graph of the TCA cycle metabolism route comparing metagenomics vs. metatranscriptomics data of the microbiota of 10d and 20d conditions. Some nodes are split between two colors, indicating 10d (left) and 20d (right) conditions. Green (–1) depicts genes underrepresented in metagenomics (but overrepresented in metatranscriptomics), while those marked in red (1) depicts overrepresented genes in metagenomics (but underrepresented in metatranscriptomics). In grey, values close to 0 in the ratio metatranscriptomics/metagenomics, indicating no differences in frequency.

Meta-omics integration: comparing metagenomics, metatranscriptomics, and metaproteomics data at the functional pathway level

Each meta-omics level data provides unique information in various ways, but their integration is crucial to gain a complete overview of the metabolic capabilities of the studied bacterial populations. metaproteomics data incorporation to the integrated analysis of microbiomes is essential to have a realistic overview of the functional capabilities of the bacterial populations. For this purpose, we analysed these meta-omics data together, as an example, focussing on the N metabolism pathway, corresponding to the N cycle, the set of reactions by which different inorganic N compounds are transformed into ammonia, a biologically reduced form of N that can be mainly introduced into synthesis of amino acids (glutamine and glutamate). We were interested in this pathway due to previous findings related to N metabolism of the host (*B. germanica)* and the endosymbiont *Blattabacterium.* As explained previously, *Blattabacterium* participates in the N recycling from stored urates to ammonia that can be used to synthesize glutamine and glutamate, connecting with the amino acid biosynthesis pathway [6]. Here, the aim was to study N metabolism in the host gut microbiome and then to assess if the bacterial population has the metabolic capability to produce a form of usable N.

In this analysis, we investigated how variable or stable the overall N metabolism is at the gene, transcript and protein level along time (10d against 20d) in the investigated pathway (*Figure 6).*. While metagenomics and metatranscriptomics show almost complete coverage of the N metabolism pathways and very variable along time, only a few enzymes were observed in the metaproteomics data and very stable along time. These results suggest that while the gene pool (the population) can be variable, the final transcripts and at least the four detected proteins remain stable, which could point in the direction of a functional redundancy at the protein level, as has been previously described for human gut microbiota

[51]. However, deeper coverage of the metaproteomics data would be necessary to confirm these findings.

Figure 6: KEGG Pathview graph of the N metabolism route comparing metagenomics/metatranscriptomics/metaproteomics data of the microbiome at 10d and 20d. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Green (–1) depicts genes/transcripts/proteins overrepresented in 10d (but underrepresented in 20d), while those marked in red (1) depicts genes/transcripts/proteins overrepresented in 20d (but underrepresented in 10d). In grey, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

Comparison of host and microbiome data

Microbiota metabolism and functions are better understood when studied together with its host. gNOMO includes the analysis of the host data in parallel with its microbiome, so we can integrate and compare the metabolic pathways of host and microbiome. In the case of *B. germanica,* we have studied the N metabolism pathway that we had analyzed before with the focus on the microbiota data (*Figure 6)* integrating the host data (*Figure 7).*. We have observed which enzymes can be found in the bacterial population data and which ones can be explained by the host data (*Figure 6 and 7)*..

We expected to find a maximum of four enzymes in the host data, as in most eukaryotes only four enzymes of this pathway are present, and we could detect those in the host pathway. While these four enzymes were the only ones detected in the host, its gut microbiome possesses most of the enzymes present in the N metabolism pathway.

If we study these four enzymes present in the host data in detail, we can observe that all of them are overrepresented at 10d against 20d condition in metaproteomics data, and in metagenomics and metatranscriptomics data, they are almost undetectable (*Figure 7).*. When looking at the microbiome metatranscriptomics data, these proteins have a stable abundance over the whole time (*Figure 6)*.. These findings could indicate that the production of these proteins in the hindgut of the host is reduced along time, but its production by the microbiome remains stable.

After analyzing the bacterial and the host capabilities together regarding this metabolic pathway, we find that the N metabolism corresponding to the N cycle is mostly performed by the microbiome.

Figure 7: KEGG Pathview graph of the N metabolism pathways comparing metagenomics/metatranscriptomics/metaproteomics data of the host between 10d and 20d conditions. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Green (–1) depicts genes/transcripts/proteins overrepresented in 10d (but underrepresented in 20d), while those marked in red (1) depicts genes/transcripts/proteins overrepresented in 20d (but underrepresented in 10d). In grey, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

Discussion

The aim of our software design and implementation was to provide a complete pipeline to analyze omics data from a non-model host and its microbiome. Based on these requirements, we developed the gNOMO software that presents an end-to-end workflow covering all the required data analysis steps starting from the processing of raw omics data to the final output visualization of the results. gNOMO performs the analysis of up to three different meta-omics data: metagenomics, metatranscriptomics and metaproteomics, and their integration.

gNOMO is designed for paired-end sequencing of metagenomics and metatranscriptomics data, the pipeline includes a preprocessing and binning step designed for this type of datasets. A tailored proteogenomic database is generated to perform a highly efficient database search for protein identification in the metaproteomics data analysis without a reference microbiome. To obtain this database metagenomics and metatranscriptomics data are assembled into contigs, which are then used to predict the proteins present in the samples. Together with the microbiome data, host data is obtained from the same samples and analyzed *de novo* in order to be able to analyze microbiota of non-model organisms integrated with the host information. Host databases can also be provided to analyze human or other model organisms data.

The pipeline is developed using the modular Snakemake framework that allows to incorporate software tools and libraries with different requirements. These tools are available at the BioConda channel and their installation is incorporated in the workflow. Snakemake makes use of programming languages Python and Bash, which are commonly used in bioinformatics. Parameters can be specified in the configuration file provided to Snakemake, so it can be adapted to any kind of host or microbiome analyzed. The use of Snakemake makes gNOMO fully automated, efficient, and reproducible.

Previously published meta-omics workflows such as Imetaproteomics [17] incorporate two layers of meta-omics information by integrating metagenomics and metatranscriptomics data. Such workflows focus on the analysis of the microbiome and often consider host information as contaminant reads: thus, instead of providing a host data analysis, the host genome is only used to remove the host information from the microbiome data. To overcome this issue, gNOMO offers the possibility to analyze host data in parallel to microbiome data and both datasets can be studied simultaneously. gNOMO includes the analysis of metaproteomics data and creates a tailored proteogenomic database to achieve better and more efficient protein identification. The incorporation of the metaproteomics data to the study of the microbiome gives another dimension to the analysis of the microbiome because the proteome provides the functional profile and thereby gives insights on the actual interaction between microbial populations and their host.

The visualization output provided by gNOMO pipeline includes krona charts for taxonomic distribution, and KO categories are plotted using Pathview graphs. The functional distribution represented with Pathview permits to investigate two different aspects: first, the completeness of the metabolic pathways by visualizing each enzyme in the route, and second, the differences in abundance of each enzyme by

comparing datasets (metagenomics, metatranscriptomics and metaproteomics) or conditions. This integration in gNOMO is highly useful, for example, when information regarding the presence and abundance of specific enzymes is needed.

Conclusions

gNOMO is a standardized and reproducible bioinformatic pipeline designed to integrate and analyze metagenomics, metatranscriptomics, and metaproteomics microbiota data of non-model organisms. It incorporates preprocessing, binning, assembly steps, taxonomic and functional annotations, and the production of a proteogenomic database to improve the metaproteomics analysis. gNOMO also includes the analysis of both microbiota and host data in parallel, which makes it a useful tool to analyze the microbiome of non-model organisms, as it was demonstrated using experimental data of the German cockroach *B. germanica.* In general, gNOMO can also be applied to data from human or other model organism sample types. Finally, gNOMO generates output and visualization of multiple meta-omics results in a single automated pipeline.

Availability And Requirements

- Project name: gNOMO
- Project home page: https://gitlab.com/rki_bioinformatics/gnomo
- Operating system(s): Linux
- Programming language: Python, R
- Other requirements: Snakemake, Conda
- License: Apache License 2.0.
- Any restrictions to use by non-academics: No

List Of Abbreviations

EC:Enzyme Commission number

LDA:Linear discriminant analysis

LEfSe:Linear discriminant analysis (LDA) effect size

KEGG:Kyoto Encyclopedia of Genes and Genomes

KO:KEGG Orthology

N:Nitrogen

TCA:Tricarboxylic acid cycle

rTCA:Reverse tricarboxylic acid cycle

ATP:Adenosine triphosphate

Declarations

- Ethics approval and consent to participate

This manuscript does not report data collected from humans or vertebrate animals.

- Consent for publication

This manuscript does not contain any individual person's data in any form.

- Availability of data and material

The software presented in this manuscript is available at: https://gitlab.com/rki_bioinformatics/gnomo and https://gitlab.com/gaspilleura/gnomo

The validation data is available at: https://doi.org/10.5281/zenodo.3569690

- Competing interests

The authors declare that they have no competing interests.

- Funding

This research was cofunded by Regional Development Fund (ERDF) and Ministerio de Ciencia, Innovación y Universidades (PGC2018–099344-B-I0, Spain), Conselleria d'Educació de la Generalitat Valenciana (PROMETEO/2018/133, Spain), Research Foundation Flanders (SB grant 1S90918N, Belgium) and Deutsche Forschungsgemeinschaft (RE 3474/2–2, Germany). MMB is a recipient of a FPU fellowship (FPU15/01203) from Ministerio de Ciencia, Innovación y Universidades (Spain).

- Authors' contributions

MMB and TM conceived the project and designed the pipeline with contributions by BYR. MMB, AL and CGF collaborated in the experimental design and sampling. MMB, FH, TVDB and VCP developed the code of the workflow. MMB and FH performed data analysis. MMB, FH, TVDB, VCP, CGF, AL, BYR and TM wrote and edited the manuscript. All authors read and approved the final manuscript.

- Acknowledgements

We thank Dr. Nuria Jiménez and FISABIO for the help in the processing and sequencing of the metagenomics and metatranscriptomics samples and the Proteomics Service of the SCSIE for the processing and sequencing of the metaproteomics samples.

References

1. Gil R, Latorre A. Unity makes strength: A review on mutualistic symbiosis in representative insect clades. 2019; doi:10.3390/life9010021.
2. Moya A, Peretó J, Gil R, Latorre A. Learning how to live together: Genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 2008; doi:10.1038/nrg2319.
3. Douglas AE. Lessons from studying insect symbioses. Cell Host Microbe. 2011; doi:10.1016/j.chom.2011.09.001.
4. Moran NA, Ochman H, Hammer TJ. Evolutionary and Ecological Consequences of Gut Microbial Communities. Annu Rev Ecol Evol Syst. 2019; doi:10.1146/annurev-ecolsys-110617-062453.
5. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016; doi:10.1038/nmicrobiol.2016.180.
6. Carrasco P, Pérez-Cobas AE, van de Pol C, Baixeras J, Moya A, Latorre A. Succession of the gut microbiota in the cockroach *Blattella germanica*. Int Microbiol. 2014; doi: 10.2436/20.1501.01.212.
7. López-Sánchez MJ, Neef A, Peretó J, Patiño-Navarrete R, Pignatelli M, Latorre A, et al. Evolutionary convergence and nitrogen metabolism in Blattabacterium strain Bge, primary endosymbiont of the cockroach *Blattella germanica*. PLoS Genet. 2009; doi:10.1371/journal.pgen.1000721.
8. Patiño-Navarrete R, Piulachs MD, Bellés X, Moya A, Latorre A, Peretó J. The cockroach *Blattella germanica* obtains nitrogen from uric acid through a metabolic pathway shared with its bacterial endosymbiont. Biol Lett. 2014; doi: 10.1098/rsbl.2014.0407.
9. Pérez-Cobas AE, Gosalbes MJ, Friedrichs A, Knecht H, Artacho A, Eismann K, et al. Gut microbiota disturbance during antibiotic therapy: A multi-omic approach. Gut. 2013; doi:10.1136/gutjnl-2012-303184.
10. Rosas T, García-Ferris C, Domínguez-Santos R, Llop P, Latorre A, Moya A. Rifampicin treatment of Blattella germanica evidences a fecal transmission route of their gut microbiota. FEMS Microbiol Ecol. 2018; doi:10.1093/femsec/fiy002.
11. Cani PD. Human gut microbiome: Hopes, threats and promises. 2018; doi:10.1136/gutjnl-2018-316723.
12. Mohajeri MH, Brummer RJM, Rastall RA, Weersma RK, Harmsen HJM, Faas M, et al. The role of the microbiome for human health: from basic science to clinical applications. Eur J Nutr. 2018; doi:10.1007/s00394-018-1703-4.
13. Piro VC, Matschkowski M, Renard BY. MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. 2017; doi:10.1186/s40168-017-0318-y.
14. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018; doi:10.1038/s41579-018-0029-9.
15. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, et al. MetaTrans: An open-source pipeline for metatranscriptomics. Sci Rep. 2016; doi:10.1038/srep26447.

16. Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, Hoffmann M, et al. The MetaProteomeAnalyzer: A powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res. 2015; doi:10.1021/pr501246w.
17. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018; doi:10.1093/BIB/BBW114.
18. Hernández-De-Diego R, Tarazona S, Martínez-Mira C, Balzano-Nogueira L, Furió-Tarí P, Pappas GJ, et al. PaintOmics 3: A web resource for the pathway analysis and visualization of multi-omics data. Nucleic Acids Res. 2018; doi:10.1093/nar/gky466.
19. Moya A, Ferrer M. Functional redundancy-induced stability of gut microbiota subjected to disturbance. Trends Microbiol. 2016; doi:10.1016/j.tim.2016.02.002.
20. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014; doi:10.1073/pnas.1319284111.
21. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMetaproteomics: A pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016; doi:10.1186/s13059-016-1116-8.
22. Ruggles K V., Krug K, Wang X, Clauser KR, Wang J, Payne SH, et al. Methods, tools and current perspectives in proteogenomics. Mol Cell Proteomics. 2017; doi:10.1074/mcp.MR117.000024.
23. Schiebenhoefer H, Van Den Bossche T, Fuchs S, Renard BY, Muth T, Martens L. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. Expert Rev Proteomics. 2019; doi:10.1080/14789450.2019.1609944.
24. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. 2012; doi:10.1093/bioinformatics/bts480.
25. Dale R, Grüning B, Sjödin A, Rowe J, Chapman BA, Tomkins-Tinch CH, et al. Bioconda: Sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018; doi:10.1038/s41592-018-0046-7.
26. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
27. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. 2011; doi:10.1093/bioinformatics/btr026.
28. Aronesty E. Comparison of sequencing utility programs. Open Bioinforma J. 2013; doi:10.2174/1875036201307010001.
29. Menzel P, Lee Ng K, Krogh A. Kaiju: Fast and sensitive taxonomic classification for metagenomics. bioRxiv. 2015; doi:10.1101/031229.
30. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;

doi:10.1186/1471-2105-11-119.

31. Stanke M, Morgenstern B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005; doi:10.1093/nar/gki458.
32. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, et al. eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008; doi:10.1093/nar/gkm796.
33. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014; doi:10.1093/bioinformatics/btu031.
34. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003; doi:10.1093/nar/gkg128.
35. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014; doi:10.1038/ncomms6277.
36. Gurdeep Singh R, Tanca A, Palomba A, Van Der Jeugt F, Verschaffelt P, Uzzau S, et al. Unipept 4.0: Functional analysis of metaproteome data. J Proteome Res. 2019; doi:10.1021/acs.jproteome.8b00716.
37. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome. 2016; doi:10.1186/s40168-016-0196-8.
38. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011; doi:10.1186/1471-2105-12-385.
39. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011; doi:10.1186/gb-2011-12-6-r60.
40. Luo W, Pant G, Bhavnasi YK, Blanchard SG, Brouwer C. Pathview Web: User friendly pathway visualization and data integration. Nucleic Acids Res. 2017; doi:10.1093/nar/gkx372.
41. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000; doi:10.1093/nar/28.1.27.
42. Harrison MC, Jongepier E, Robertson HM, Arning N, Bitard-Feildel T, Chao H, et al. Hemimetabolous genomes reveal molecular basis of termite eusociality. Nat Ecol Evol. 2018; doi:10.1038/s41559-017-0459-1.
43. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol. 2016; doi:10.3389/fmicb.2016.00459.
44. Saito K, Koido S, Odamaki T, Kajihara M, Kato K, Horiuchi S, et al. Metagenomic analyses of the gut microbiota associated with colorectal adenoma. PLoS One. 2019; doi:10.1371/journal.pone.0212406.
45. Rinninella E, Raoul P, Cintoni M, Franceschi F, Miggiano GAD, Gasbarrini A, et al. What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. 2019; doi:10.3390/microorganisms7010014.

46. Flint HJ, Scott KP, Louis P, Duncan SH. The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol. 2012; doi:10.1038/nrgastro.2012.156.
47. Douglas AE. The microbial dimension in insect nutritional ecology. Funct Ecol. 2009; doi:10.1111/j.1365-2435.2008.01442.x.
48. Ortmann AC, Santos TTL. Spatial and temporal patterns in the *Pelagibacteraceae* across an estuarine gradient. FEMS Microbiol Ecol. 2016; doi:10.1093/femsec/fiw133.
49. Dranse HJ, Zheng A, Comeau AM, Langille metagenomicsI, Zabel BA, Sinal CJ. The impact of chemerin or chemokine-like receptor 1 loss on the mouse gut microbiome. 2018; doi:10.7717/peerj.5494.
50. Moges F, Eshetie S, Endris M, Huruy K, Muluye D, Feleke T, et al. Cockroaches as a source of high bacterial pathogens with multidrug resistant strains in Gondar town, Ethiopia. Biomed Res Int. 2016; doi:10.1155/2016/2825056.
51. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. 2012; doi:10.1038/nature11550.

Figures

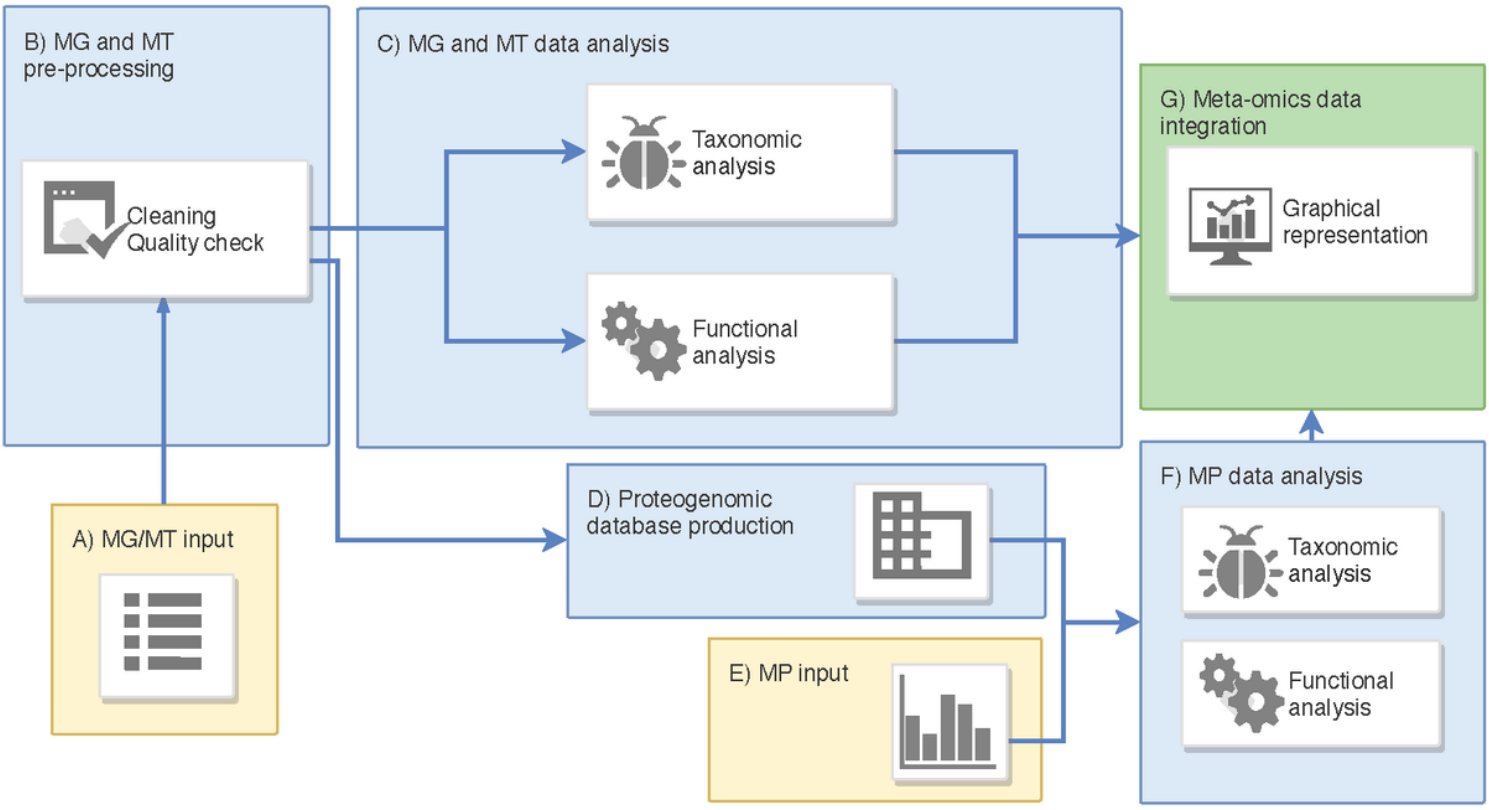

Figure 1

Workflow overview of the gNOMO pipeline. A) Initial input of metagenomic (metagenomics) and metatranscriptomic (metatranscriptomics) sequences. B) Pre-processing: cleaning and quality control of

metagenomics and metatranscriptomics input sequences. C) metagenomics and metatranscriptomics data analyses: consists of taxonomic and functional annotations. D) Proteogenomic database creation based on metagenomics and metatranscriptomics protein predictions. E) Auxiliary input of metaproteomic (metaproteomics) tandem mass spectrum data. F) metaproteomics analysis: also includes taxonomic and functional annotations. G) Graphical representation/visualization of all integrated meta-omics data.

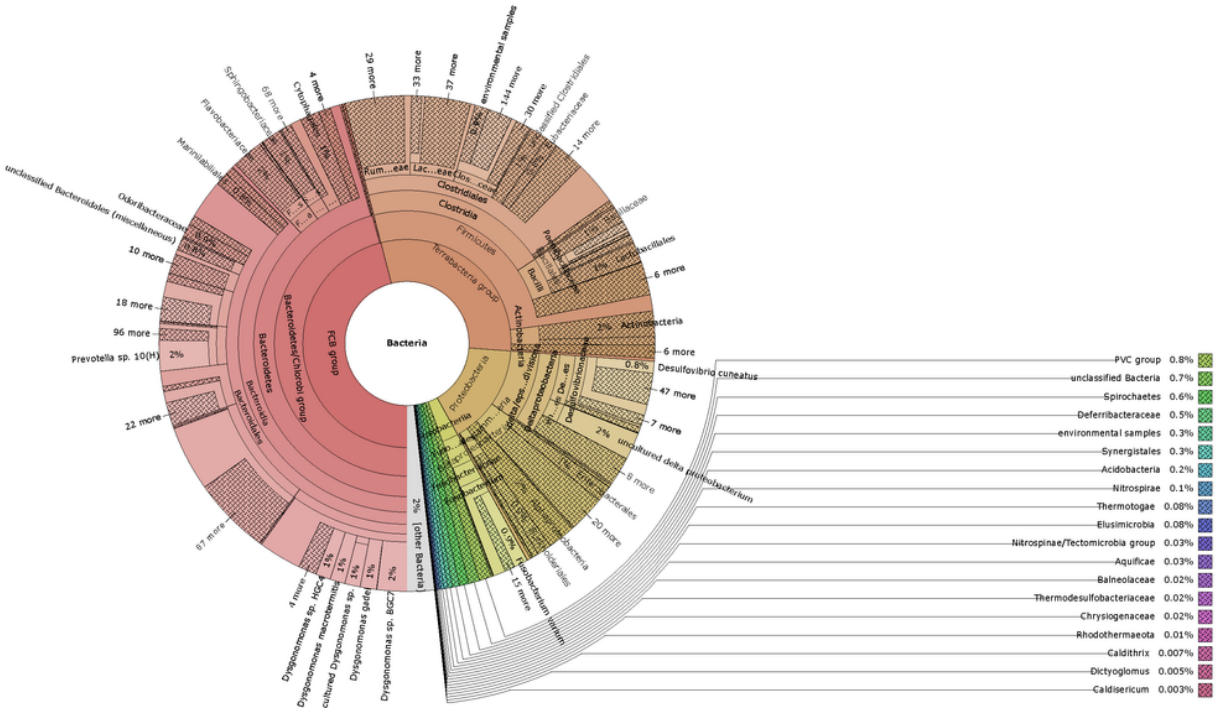

Figure 2

KronaPlot of the taxonomic annotation of a metagenomics sample (condition 10d). Bacterial taxa distribution of metagenomics data, corresponding to condition 10d. The bacterial taxa are classified by taxonomic hierarchy levels, from higher levels in the center of the chart (Kingdom Bacteria) progressing outward until genus level.

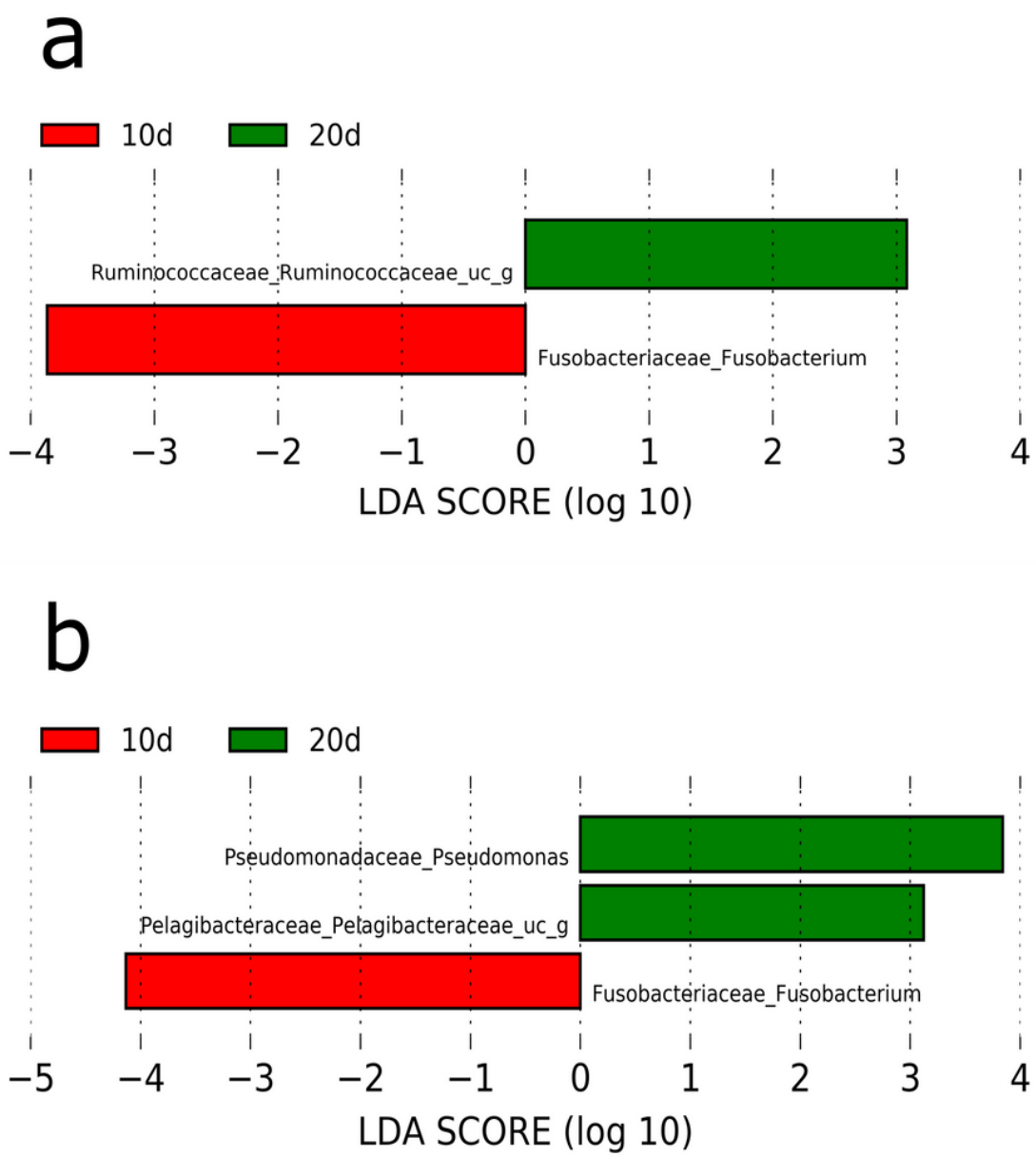

Figure 3

LEfSe graph of taxonomic annotation of metagenomics (top) and metatranscriptomics (bottom) data comparing the two conditions: 10d and 20d. Taxa with significant different distribution among the two conditions are identified. Only taxa with LDA scores over 2 are shown. Positive LDA scores are assigned to the taxa overrepresented in the condition 20d (green), and negative LDA scores to the taxa overrepresented in the condition 10d (red). metagenomics data (Fig 3a) and metatranscriptomics data (Fig 3b) are represented.

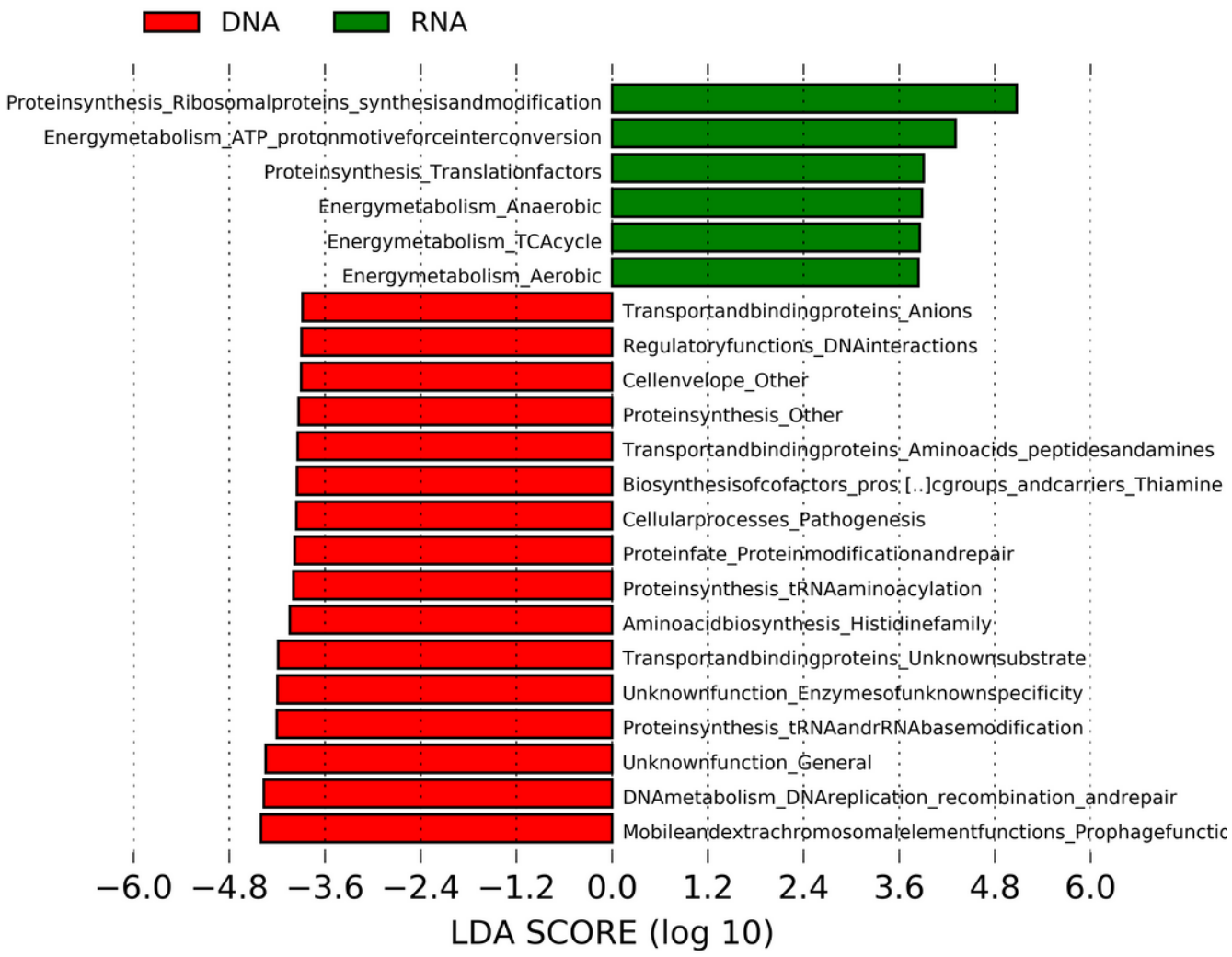

Figure 4

LEfSe graph comparing metagenomics and metatranscriptomics data of TIGRFAM annotation (role and subrole levels) of condition 10d. Taxa with significant different distribution among metagenomics and metatranscriptomics data are identified. Only taxa with LDA scores above 2 are shown. Positive LDA scores are assigned to the functional categories overrepresented in the metatranscriptomics data (RNA, green), and negative LDA scores to the functional categories overrepresented in metagenomics data (DNA, red).

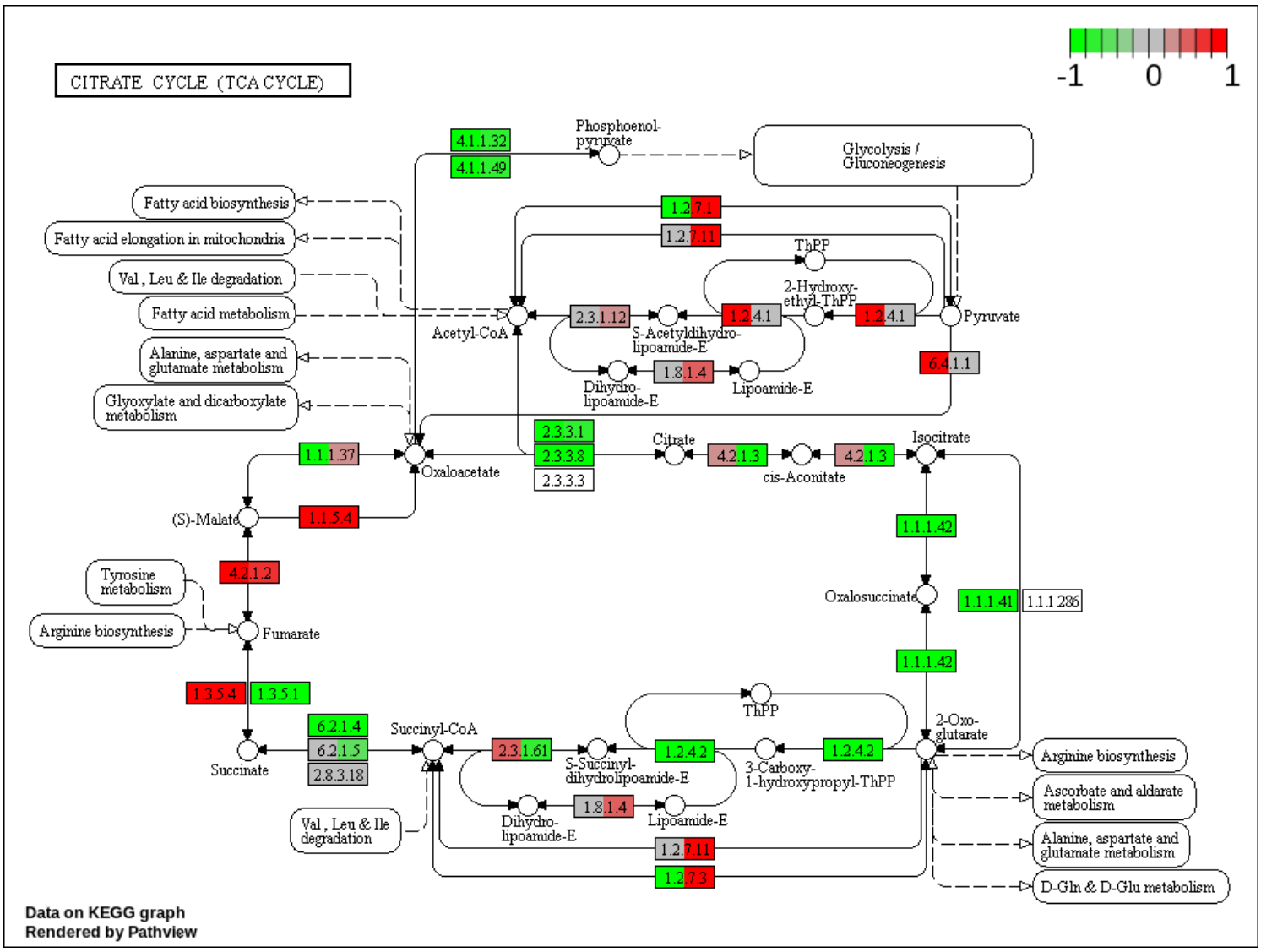

Figure 5

KEGG Pathview graph of the TCA cycle metabolism route comparing metagenomics vs. metatranscriptomics data of the microbiota of 10d and 20d conditions. Some nodes are split between two colors, indicating 10d (left) and 20d (right) conditions. Green (-1) depicts genes underrepresented in metagenomics (but overrepresented in metatranscriptomics), while those marked in red (1) depicts overrepresented genes in metagenomics (but underrepresented in metatranscriptomics). In grey, values close to 0 in the ratio metatranscriptomics/metagenomics, indicating no differences in frequency.

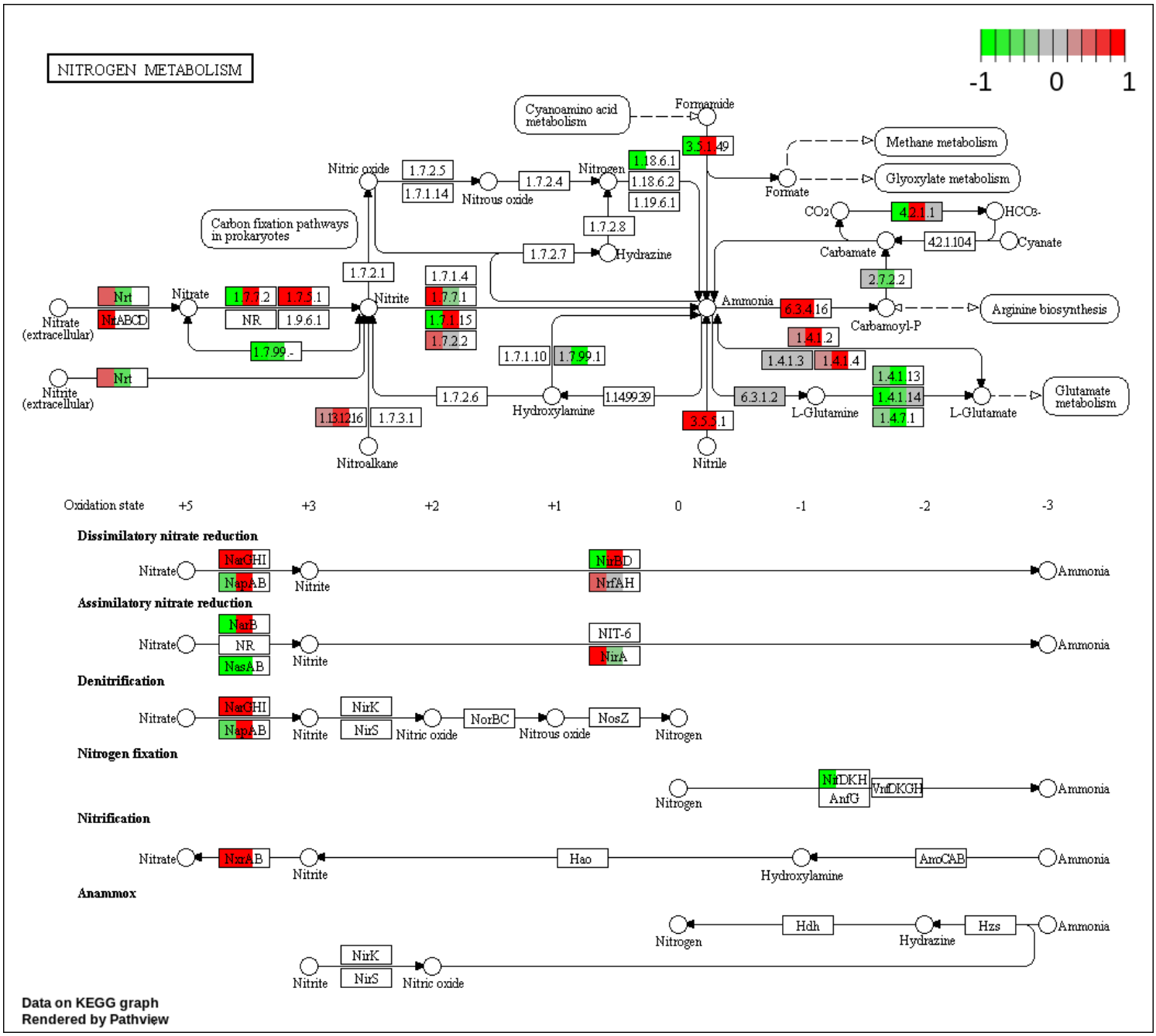

Figure 6

KEGG Pathview graph of the N metabolism route comparing metagenomics/metatranscriptomics/metaproteomics data of the microbiome at 10d and 20d. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Green (-1) depicts genes/transcripts/proteins overrepresented in 10d (but underrepresented in 20d), while those marked in red (1) depicts genes/transcripts/proteins overrepresented in 20d (but underrepresented in 10d). In grey, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

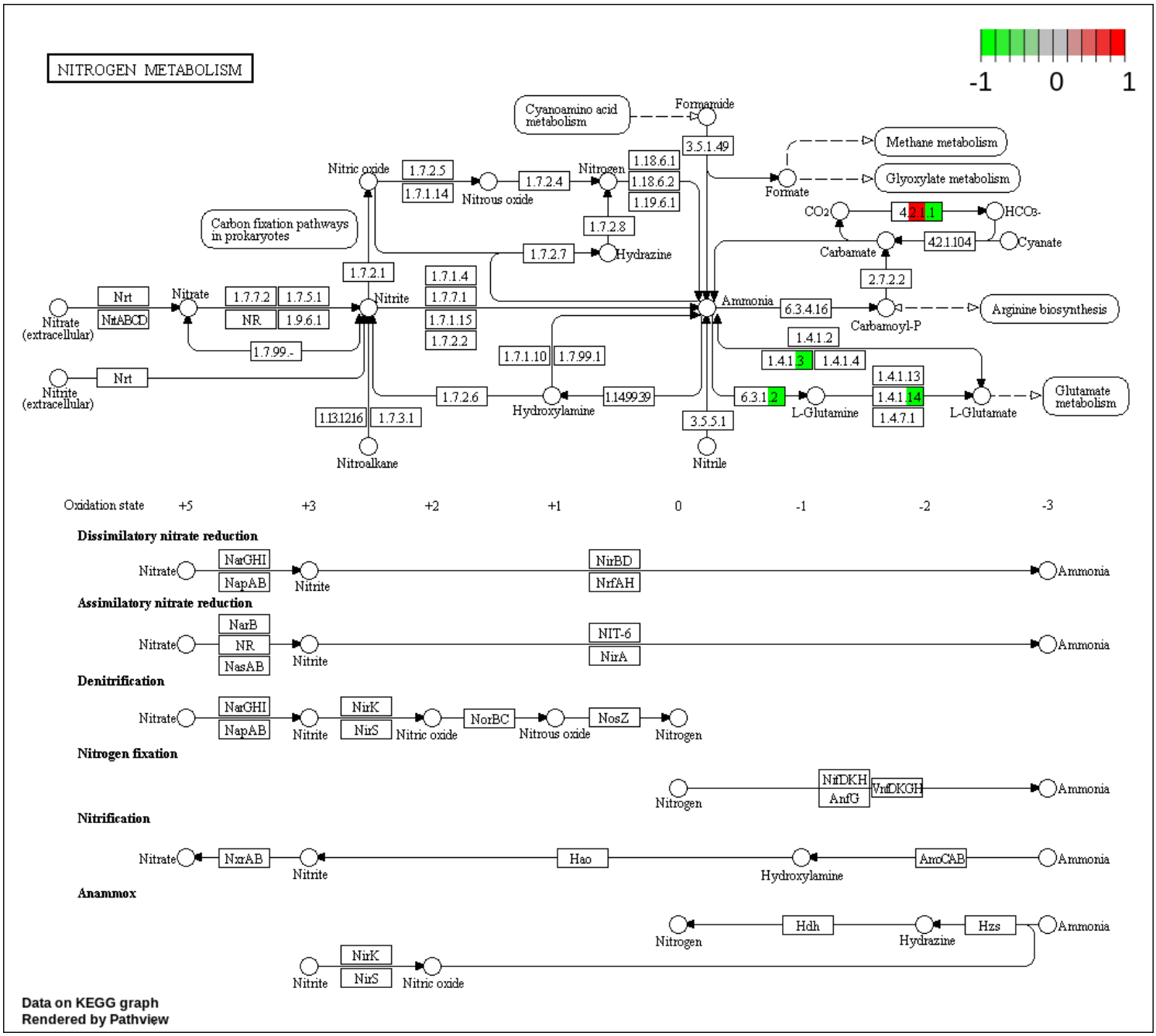

Figure 7

KEGG Pathview graph of the N metabolism pathways comparing metagenomics/metatranscriptomics/metaproteomics data of the host between 10d and 20d conditions. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Green (-1) depicts genes/transcripts/proteins overrepresented in 10d (but underrepresented in 20d), while those marked in red (1) depicts genes/transcripts/proteins overrepresented in 20d (but underrepresented in 10d). In grey, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx