Genomics 102 (2013) 409–418

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Community-Analyzer: A platform for visualizing and comparing microbial community structure across microbiomes

Bhusan K. Kuntal, Tarini Shankar Ghosh, Sharmila S Mande *

Bio-Sciences R&D Division, TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Service Ltd., 54-B Hadapsar Industrial Estate, Pune 411 013, Maharashtra, India

ARTICLE INFO

Article history:
Received 1 April 2013
Accepted 14 August 2013
Available online 24 August 2013

Keywords:
Metagenomics
Microbial interactions
Algorithms
Software

ABSTRACT

A key goal in comparative metagenomics is to identify microbial group(s) which are responsible for conferring specific characteristics to a given environment. These characteristics are the result of the inter-microbial interactions between the resident microbial groups. We present a new GUI-based comparative metagenomic analysis application called Community-Analyzer which implements a correlation-based graph layout algorithm that not only facilitates a quick visualization of the differences in the analyzed microbial communities (in terms of their taxonomic composition), but also provides insights into the inherent inter-microbial interactions occurring therein. Notably, this layout algorithm also enables grouping of the metagenomes based on the probable inter-microbial interaction patterns rather than simply comparing abundance values of various taxonomic groups. In addition, the tool implements several interactive GUI-based functionalities that enable users to perform standard comparative analyses across microbiomes. For academic and non-profit users, the Community-Analyzer is currently available for download from: http://metagenomics.atc.tcs.com/Community_Analyzer/.

1. Introduction

The advent of high throughput sequencing technologies and the emerging field of metagenomics have facilitated rapid extraction and sequencing of the microbial genomic content present in various environments. Several available computational methods/software [1–8] facilitate analysis of the large volume of sequence data obtained from such metagenomic studies. These developments have enabled researchers in profiling the microbial groups they possess in their environment(s) of interest. Comparing metagenomic datasets is expected to help in identifying probable key agents responsible for conferring a specific phenotypic trait to a given environment. Examples of this include traits which are specific to certain disease and physiological disorders [9–12].

Currently, a number of methodologies are being used by researchers for performing comparative metagenomic analyses. These include use of Principal Component Analysis (PCA), trees generated using various distance measures and different visualization schemes (force-directed/spring-embedded layout, stacked bar/trend plots, heat maps, etc.). Such analyses have helped researchers in making meaningful inferences. For example, PCA based approach, utilized for comparing gut microbiomes of individuals from diverse nationalities, has helped in identifying three core groups of gut microbiota, referred to as 'Enterotypes' [13]. Similarly, use of tree based methods has enabled identification of the variations in microbial communities across various seasons in the Western English Channel [14,15].

A few standalone tools/pipelines are available for comparing microbial groups across various environments [3,6,16–20]. In addition, software package libraries like QIIME have been developed that facilitate comparison of microbial communities using the differential abundance patterns in the SSU-rRNA libraries obtained from the corresponding environments [21]. While the existing tools can distinguish metagenomes based on the composition of the inhabiting organisms in these environments, they do not provide insights into the reason behind the over- or under-abundance of specific groups of organisms. To address this limitation, we have developed a standalone user-interactive comparative metagenomics analysis and visualization platform, called 'Community-Analyzer'. The platform utilizes a suite of methodologies and graphical layouts in order to facilitate visualization and comparison of microbial community structures within as well as across microbiomes. Most importantly, in contrast to the above described available analysis servers/pipelines, the present tool also facilitates visualization of probable inter-microbial interaction patterns based on the co-occurrence patterns of microbial groups across environments.

Recent studies have indicated that understanding together microbial community structures and the inter-microbial interactions can provide a more comprehensive insight into the subtle differences across microbiomes [22]. The phenotypic trait(s) of any given environment can be associated with the complex inter-microbial interactions within the consortium of microbial groups residing therein. It is therefore important to identify not only the key taxonomic group(s) (specific

* Corresponding author at: TCS Innovation Labs, Tata Research Development & Design Centre, Tata Consultancy Services Ltd., 54-B Hadapsar Industrial Estate, Pune 411 013, Maharashtra, India. Fax: +91 20 6608 6399.
E-mail address: sharmila.mande@tcs.com (S.S. Mande).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.08.004

group of organisms) that may be responsible for conferring a specific trait to a given environment, but also microbial groups that regulate the action of these taxonomic groups. Insights gained from such analyses will be especially useful in metagenomic studies investigating microbial communities associated with diseases and disorders. In such studies, the identified key microbial groups and the organisms regulating them may help in devising appropriate diagnostic as well as therapeutic strategies. In order to get such insights, one needs to obtain answers to the following two questions. What are the inherent similarities and differences, in terms of microbial composition, amongst the samples under study? Can we infer the probable inter-microbial interaction patterns present in different environments? We address these two key questions by introducing a new graph based layout in Community-Analyzer. The layout has been implemented along with the standard visualization methods like PCA, force-directed/spring-embedded layouts, heat maps, bar/trend plots and distance-based trees. The implementation of a combination of visualization layout along with several standard features for comparative analyses provides an innovative way of identifying similarities/differences across various microbial communities.

We exemplify Community-Analyzer's visualization and analysis capabilities using two real metagenomic datasets (as case studies). We also provide an extensive walk-through of the functionalities of the tool. End users, without any programming knowledge, can easily perform analysis using the simple interactive options provided in Community-Analyzer.

2. Materials and methods

Community-Analyzer is available as a standalone tool for Windows and Linux platforms. Fig. 1 summarizes a typical metagenomic analysis work-flow and the application of Community-Analyzer in this workflow. In a typical metagenomic sequencing project, genomic DNA (corresponding to the resident microbial groups) from an environment are extracted. Subsequently either the entire set of genomic fragments (obtained using the shotgun sequencing techniques) or fragments of specific marker genes (e.g., 16S rDNA) are sequenced. The taxonomic affiliations of each of the sequenced fragments (referred to as reads) are then obtained using suitable binning algorithm(s) [1,2,4,7,8,23]. Based on the assignments, these algorithms also provide the abundance profile of the various taxonomic groups in the given environment. These abundance profiles (referred to as taxonomic classification summaries) from multiple environments can then be provided as inputs to Community-Analyzer for visualization and comparative analysis.

2.1. Obtaining 'Community-Analyzer layout' for a set of metagenomic samples

A majority of organisms present in any environment are hitherto unknown. These organisms may belong to new species/genus/family/order/class/phylum. Thus, sequences originating from these organisms are expected to be assigned to taxa at various taxonomic levels (e.g., genus, family, order, class, phylum, etc.). Consequently, most of the binning algorithms generate taxonomic summaries (consisting of abundance profiles of taxonomic groups) at different taxonomic levels. In order to effectively compare two or more metagenomic datasets, it is thus important to evaluate the differences amongst the metagenomes at various taxonomic levels. Based on the input data (from the taxonomic summary files), Community-Analyzer generates a graphical layout (displaying microbial community structures) at user-specified taxonomic level. The graphical layout is generated to capture and display not only the inter-microbial interaction patterns across the microbial communities under study, but also the similarities/differences across the communities. The generated graphical layout, hereafter referred to as the 'Community-Analyzer layout' (Fig. 2), displays two distinct aspects of the analyzed microbial communities. First, the resident microbial (taxonomic) groups are arranged based on the correlations of the abundance patterns of various taxa across

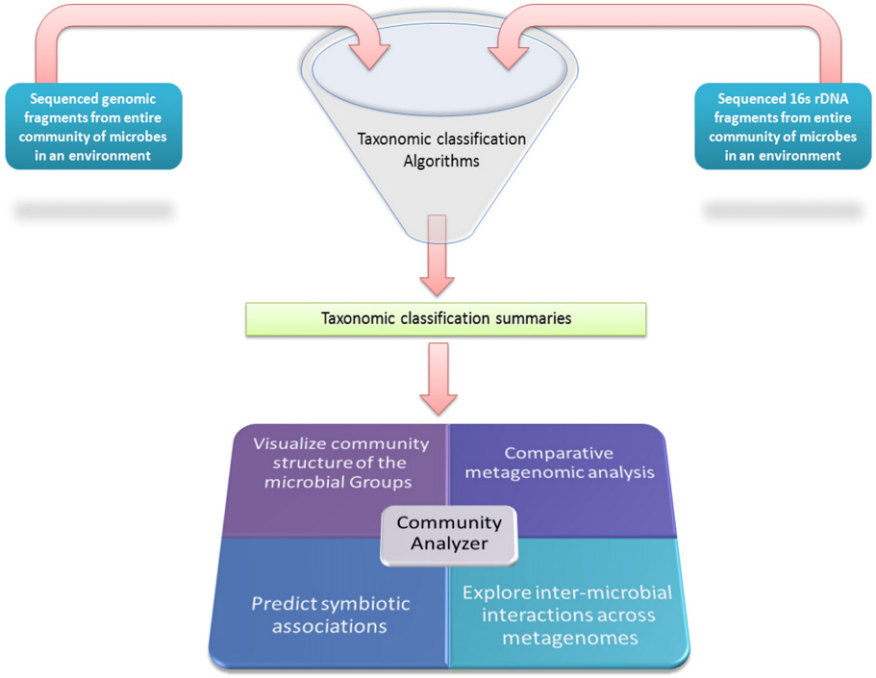

Fig. 1. Taxonomic classification summaries (obtained using representative binning algorithms) are provided as input to Community-Analyzer. In a typical metagenomic sequencing project, genomic DNA (corresponding to resident microbial groups) from an environment are extracted. This is followed by sequencing of either the entire set of genomic fragments (obtained using shotgun sequencing techniques) or the genomic fragments of specific marker genes (e.g. 16S rDNA). The taxonomic affiliations of each of these sequenced genomic fragments are then obtained using any of the available taxonomic classification methods. The obtained taxonomic classification summaries are then provided as input to the Community-Analyzer.

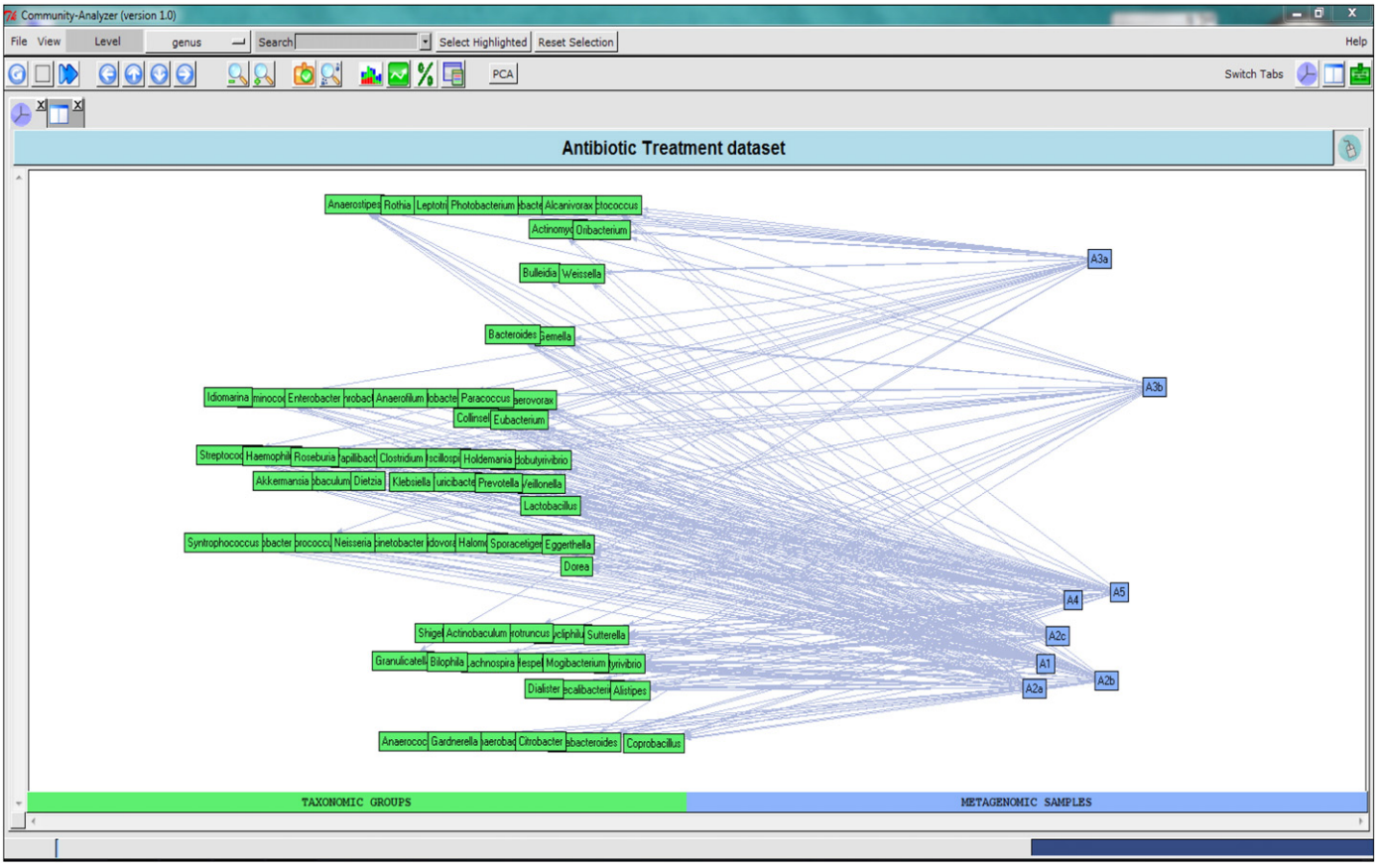

Fig. 2. Snapshot of the Community-Analyzer graphical layout showing the metagenomic samples (blue boxes on the right hand side) and the taxonomic groups (green boxes on the left hand side) at the genera level. The edge (or connection) between a sample and a taxonomic group (blue line) indicates the presence of that taxonomic group in the given sample. The taxonomic groups are arranged in such a way that the ones placed at similar horizontal levels have similar correlation patterns, indicating symbiotic relationships amongst them. On the other hand, microbial groups placed at distal locations on the vertical axis can be inferred to possess mutually inhibitory relationships. The metagenomic samples are placed on the right hand side based not only on the similarities in the abundance patterns of taxonomic groups but also on the relative placement of these groups in the layout. Thus in the figure, the two samples A3a and A3b which placed distinct from the rest of the samples, can be inferred to have a atypical taxonomic abundance pattern as compared to the rest.

the samples (which is a representation of the community interaction or microbial co-occurrence patterns). Secondly the metagenomic samples themselves are then placed based on the similarities in their microbial community structures. A detailed description of the methodology adopted for the generation of this graphical layout is provided below.

Step 1 Calculating relative abundances of taxonomic groups in the metagenomes

The abundances of various taxonomic groups in each of the metagenomic sample (either in terms of the number of assigned sequences or assigned OTUs) are obtained from the corresponding taxonomic summary file(s). These abundance values of the taxonomic groups are then normalized with respect to the size of the metagenomic samples.

Step 2 Obtaining pair wise correlations and correlation distances across different taxonomic groups

Based on the abundance patterns of various taxa across the analyzed metagenomic samples, the pair wise correlations between them are obtained using the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - x_A)(y_i - y_A)}{\sqrt{\sum_{i=1}^{n}(x_i - x_A)^2 \sum_{i=1}^{n}(y_i - y_A)^2}}$$

where,

n is the total number of metagenomes under study,

x_i and y_i are relative abundance of taxa x and y in the ith metagenome (where $i = 1, 2 \ldots n$),

x_A and y_A are the mean values of the relative abundances of x and y across the analyzed metagenomes.

Subsequently, the pairwise correlation distances are obtained by subtracting the corresponding correlation values from 1. Correlation distances are used here since they provide quantitative measures of the relationship between taxonomic groups and are not dependent on the absolute sizes of the data sets being analyzed.

Step 3 Arranging taxonomic groups based on their correlation distances

A hierarchical clustering methodology is used to rank the taxonomic groups according to their pairwise correlation distances. Based on the rank, the taxonomic groups are placed on one of the axes of a two dimensional layout. This approach of positioning different taxa ensures that the taxonomic groups whose abundance patterns follow a similar trend (i.e. their abundances show similar increase or decrease across the analyzed samples) get placed at proximal positions, while those whose abundance patterns differ (or show a negative trend in their abundances) are placed distally. This approach of arranging different taxonomic groups not only facilitates identification of positively correlated

clusters of microbial groups, but also helps in efficiently resolving (or partitioning) different metagenomes based on the similarities in their community structures.

Step 4 Placement of metagenomes based on the arrangement of taxonomic groups

Based on the relative abundances of various microbial groups in each metagenomic sample, a Hooke's law based algorithm is applied in order to place the sample on the same graphical layout. This ensures that a given metagenomic sample is not only placed closer to the taxonomic groups which are abundant in it, but also closer to other metagenomic samples having a similar taxonomic composition.

The detailed description of this methodology is provided below. Based on the taxa arrangement obtained in the previous step, a center R is obtained using the following formula:

$$R = \frac{\sum_{i=1}^{n} m_x r_x}{\sum_{i=1}^{n} m_x}$$

Where,

- rx is the position of a taxonomic group 'x' (obtained from step 3)
- mx is calculated as the summation of the normalized abundances (calculated as in step 1) of the taxa 'x' for all the 'n' metagenomic samples being analyzed.

All the metagenomes are then placed onto the layout at an initial height 'r' from the point corresponding to the center 'R' (i.e., with coordinates $x = R$ and $y = r$). For each metagenome, a virtual force due to each taxonomic group is then calculated using the following formula.

$$\overrightarrow{F_i} = A_i \cdot \overrightarrow{d_i}$$

Where,

- F_i is the force vector on the metagenome due to the taxonomic group 'i',
- A_i is the relative abundance of taxa i in the given sample,
- d_i is the distance vector of the initial drop point (i.e., $x = R$ and $y = r$) from the position of the taxonomic group.

Subsequently, the cumulative force on a metagenome is calculated as the summation of the forces due to the individual taxonomic groups. Each metagenome is then assigned a final position on the two dimensional layout based on the resultant displacement. The generated graphical layout finally displays the positions of the various taxonomic groups and the metagenomes under study.

The above methodology may however result in the placement of (nodes corresponding to) samples having similar taxonomic composition at overlapping positions. Likewise, taxonomic groups with similar abundance patterns may also be placed at similar positions in the graphical layout. In order to minimize the overlapping nodes, the layout is finally optimized by shifting the nodes horizontally. Thus, nodes corresponding to metagenomic samples appearing on the same horizontal level represent samples having a similar taxonomic composition. Similarly taxonomic groups having highly similar trends in their abundance patterns are displayed on the same horizontal level.

The vertical distance between the samples/taxa is indicative of the relative differences between them. Thus, microbial groups placed vertically proximal to each other in the layout indicate a probable symbiotic relationship between them in the metagenomic sample under consideration. Similarly, inhibitory relationships between microbial groups can be inferred based on their distal placements along the vertical axis in the graphical layout. Thus, a simple visual inspection of the generated layout (Fig. 2) is expected to provide insights into the probable inter-microbial interaction patterns occurring in given environments. Furthermore, the relative vertical placement of the different metagenomic samples (Fig. 2) in the graphical layout provides a quick picture of the similarities/differences in the microbial community structure across the analyzed microbiomes.

2.2. Interactive graphical interface for understanding microbial community structure

Community-Analyzer enables users to interact with the generated graphical layout in various manners. For example, right clicking on a node corresponding to a metagenomic sample, highlights the taxonomic groups present in that sample and indicate their relative abundances in different colors. This feature is useful not only for studying the relative abundance of each taxonomic group in a given microbial community, but also in assessing the contribution of each group (or a selected set of taxonomic groups) which may be responsible for the inherent differences across the microbial communities. The tool also allows users to select multiple nodes corresponding to a set of taxonomic groups or metagenomic samples and perform a combined trend analysis for the selected nodes. The usefulness of these functionalities is elucidated in the case studies (on real metagenomic data sets) described in the later sections.

2.3. Identification and analysis of selectively occurring microbial groups in metagenomes

Taxonomic groups which are present across all metagenomic samples are likely to play a role in maintaining the equilibrium of the analyzed microbial communities. On the other hand, selectively occurring taxonomic groups in a subset of metagenomic samples may be responsible for conferring specific phenotype to these environments. One of the goals of comparative metagenomics is thus to identify taxa which are common across all samples as well as subset of taxa which are present in specific groups of metagenomes. The 'Differential abundance investigator' incorporated in the Community-Analyzer enables users to perform such analyses. Please refer to section 6.3 and 6.4 of the user manual (provided as Supplementary material 1) for a detailed description of this feature.

2.4. Analysis of interaction networks of taxonomic groups across microbial communities

One of the goals while comparing metagenomic samples is to identify the inherent interaction network of the microbial groups present in a subset of metagenomic samples. The Community-Analyzer enables users to select a group of taxonomic groups in the layout and obtain the pairwise correlations between these groups (refer to Section 6.6 of Supplementary material 1). Significant positive and negative correlations ($P < 0.05$) are marked with a * symbol. For this purpose, for any pair of taxa, the *t*-value is first calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Where,

- r is the correlation value obtained for the pair of taxa (calculated as described in Step 2).
- n is the number of samples.

Subsequently, pairs of taxa having a *t*-value exceeding the threshold value (corresponding to $P < 0.05$) are identified. The corresponding correlation values (r) are then marked with a *.

Furthermore, the Community-Analyzer also implements a GUI based query system module called 'Community interactions investigator' to identify probable interacting taxonomic groups across one or more subsets of samples. The module allows one to first select a subset of metagenomic samples. Subsequently, using user defined thresholds of positive and/or negative correlations (provided using the GUI based query system), the module identifies the positively and/or negatively interacting pairs of taxonomic groups within the selected subset of metagenomic samples. This information is then represented in the form of interaction graphs where in, edges are drawn between taxonomic groups having correlation values satisfying the user defined thresholds. In addition, the users can also select multiple sets of samples and visualize the common as well as unique interaction pairs (of microbial groups) across the selected subsets. The 'Community interactions investigator' module can therefore be used for identification of hubs (of interacting microbial groups). Such identified hubs may play critical role in determining the phenotypic characteristics of the selected subset of environments (metagenomes). The details of this feature are described under Section 6.10 of the user manual (Supplementary material 1).

2.5. Additional layouts for community comparisons

PCA and Spring-embedded graph layouts, two commonly used methods for visualizing microbial community structures, are also implemented in Community-Analyzer. The implemented interactivity feature (as described previously) enables users to make meaningful inferences based on the analysis of the generated layouts. Tree based comparison, one of the most routine ways of comparing metagenomic samples, is also implemented in Community-Analyzer. This tree based visualization provides a quick overview of the overall similarities/differences in the taxonomic composition of the analyzed metagenomic samples. Apart from using standard distance measures like Bray–Curtis [24], Hellinger [25] and Euclidean, Community-Analyzer implements a new approach of generating trees to compare metagenomic samples. Unlike approaches using standard distance measures (which use abundances of taxa), the present approach uses the pair wise Euclidean distances between the displayed positions of the different metagenomic samples in the 'Community-Analyzer layout' to generate a distance based tree. The positions of these samples are in turn determined by the inter taxa correlation distances obtained from their abundance profiles. These profiles provide quantitative measures of the relationship between taxonomic groups and are not dependent on the absolute sizes of the data sets being analyzed [26]. The efficacy of this new distance measure in identifying the subtle differences across the metagenomes is demonstrated in later section (under case studies) using the Community-Analyzer on real metagenomic samples. Apart from these, visualization schemes like bar-charts, trend-plots and heat maps implemented in the tool enable users to perform a detailed 'taxa-by-taxa' comparison across the analyzed metagenomes. Various graphical outputs generated by Community-Analyzer are depicted in Fig. 3.

3. Results

We analyzed two sets of available metagenomic datasets to demonstrate the applicability of the various functionalities implemented in Community-Analyzer. The detailed analyses performed on these two case studies are discussed below:

3.1. Infant–Adult dataset

Gut metagenomic data sets (details provided in Table 1 of Supplementary material 2) corresponding to 13 Japanese individuals, previously studied by Kurokawa et al. [27], were analyzed using Community-Analyzer. The taxonomic summary files corresponding to these datasets were parsed and provided as inputs to Community-Analyzer. The detailed steps of this analysis are provided as a walkthrough in Supplementary material 3. The overall results obtained from this analysis are summarized below.

Since a majority of sequences present in a typical metagenomic sample (in this case gut metagenomes) is known to originate from hitherto unknown organism (new species/genus/family/, etc.), the community-analyzer graphical layouts were generated for the 13 gut metagenomes at different taxonomic levels. An inspection of these layouts revealed distinct placements of the infant samples away from those corresponding to the adults and children (Figs. 1–5 of Supplementary material 2). This indicated that the infant gut had a distinct microbial community, an observation reported in the original study by Kurokawa et al. [27]. Furthermore, a comparison of the graphical layouts generated at various taxonomic levels indicated that the infant samples were more clearly separated from those corresponding to the adults and children at the taxonomic level of genus (Fig. 5 of Supplementary material 2). This suggested that the key taxonomic groups in the gut of the infant samples could be distinguished from those in the adult and children only at finer taxonomic levels. Thus, a simple visualization of the graphical layouts (generated using Community-Analyzer) could provide important insights into the structure of the microbial communities under study. Furthermore, by facilitating generation of graphical layouts at all levels, Community-Analyzer could identify key distinguishing microbial groups (at a particular phylogenetic level).

In order to investigate the taxonomic groups responsible for the distinct separation of the infant samples, the relative abundances of different taxonomic groups were investigated using simple interactive operations on the graphical layout (Fig. 5 of Supplementary material 3). Using the abundance based color coding feature implemented in the Community-Analyzer graphical layout, the gut microbial communities of the infants were observed to be dominated by the genus *Bifidobacterium*, followed by *Enterococcus*, *Klebsiella*, *Clostridium* and *Escherichia*. On the other hand, performing a similar analysis on the adults and children samples mainly revealed a dominance of the genus *Bacteroides*. Although the distinct placement of the infant samples could be visualized using the PCA layout (Fig. 6 of Supplementary material 3), also implemented in Community-Analyzer, the entire set of probable microbial groups responsible for this distinct separation could only be identified using the Community-Analyzer layout (Fig. 5 of Supplementary material 3).

The probable symbiotic relationships amongst the resident microbial groups were analyzed using the 'Community interaction investigator' module (described under Figs. 7–10 of Supplementary material 3). Positive/negative correlations in the abundance patterns amongst various microbial groups (across a set of samples from similar environments, in this case gut) are a reflection of their symbiotic/inhibitory nature, respectively. The 'Community interaction investigator' module was thus used to identify the highly correlated (correlation > 0.9) hubs of positively interacting groups of microbes, present specifically in the infant (Fig. 4A) and adult samples (Fig. 4B). The displayed network interestingly revealed four distinct and densely packed hubs of symbiotically interacting microbial groups in the infant samples. While most of the members in three of these hubs were also identified in the hubs present in the adult samples, the microbial groups constituting the fourth hub were observed to be unique to the infant samples. In contrast to the infant samples, the adult samples had a larger number of low density hubs. This trend could be accounted for the fact that the microbial communities in the infant guts are more interdependent as compared to those in the adult guts. Based on the above observations, it could be hypothesized that, with age, the microbial community in the gut changes not only with respect to diversity, but also with respect to the mutual inter-dependencies amongst the resident microbes. Given the fewer number and the higher compactness of the hubs in the infant gut,

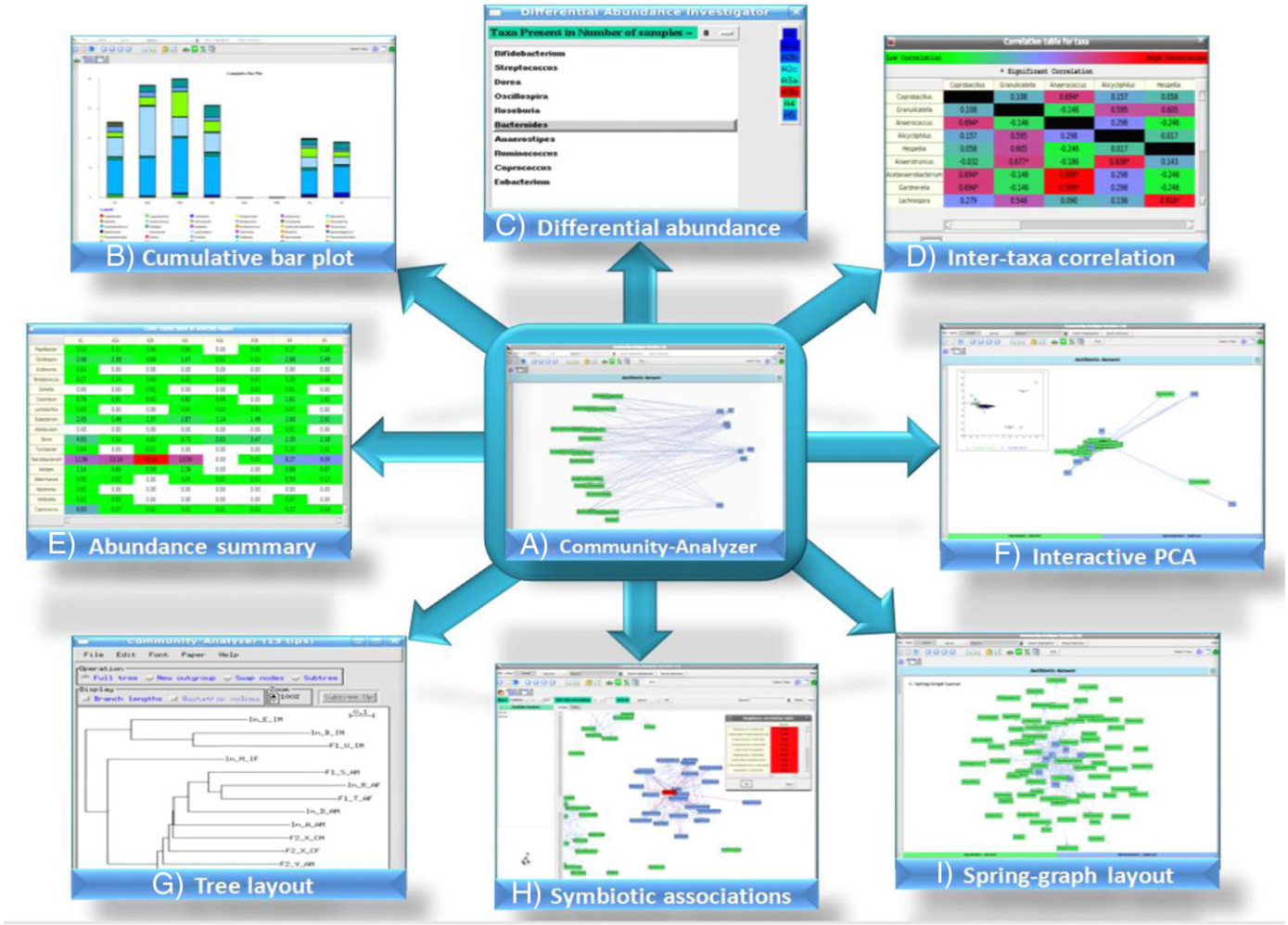

Fig. 3. Pictorial summary of the various analysis features available in Community-Analyzer where, A: Community Analyzer layout; B: Cumulative bar plot output; C: Differential abundance investigator; D: Inter taxa correlation viewer; E: Abundance summary viewer; F: Interactive PCA layout; G: Tree layout; H: Community interactions investigator to show the probable symbiotic relations for a selected set of metagenomic samples; I: Spring graph based layout generated for a set of metagenomic samples.

antibiotic(s) or invading microbe(s) targeting even one of the members of a hub is likely to affect all the inter-dependent members, thereby making the infant gut more susceptible to dysbiosis.

3.2. Antibiotic treatment dataset

Administration of antibiotics is known to have a deleterious effect on the commensal microbial community residing in the gut. Dethlefsen et al. [28] had previously studied the effect of the antibiotic Ciprofloxacin on the gut microbial communities of three individuals at various time points using 16S rRNA phylotyping. We have re-analyzed these data sets (Table 2 of Supplementary material 2) in order to demonstrate the applicability of Community-Analyzer for analyzing time series metagenomic data. The taxonomic summaries corresponding to one of the individuals (individual A) was chosen for this analysis as the number of sampling time points for this individual was higher as compared to the other two individuals. A simple visual inspection of the Community-Analyzer layout could reveal the dysbiosis in the gut microbial community occurring immediately after the administration of the antibiotics (Fig. 10 of Supplementary material 2), an observation in line with the results reported in the original study [28].

To investigate the extent and the nature of the impact of antibiotic treatment on the gut microbial community, we studied the variation of the relative abundances of the different genera at different stages of antibiotic treatment using the 'trend plot' and the 'cumulative bar plot' features of the Community-Analyzer (Figs. 11 and 12 of Supplementary material 2). From these plots, one could easily infer that certain microbial genera were completely eliminated immediately after administration of the antibiotics, some of which regained back in due course of time. However, the relative proportions of the different genera were observed to be noticeably decreased. For example, it was observed from the 'trend plot' generated using the Community-Analyzer (Fig. 12 of Supplementary material 2) that some of the genera (e.g. *Faecalibacterium* and *Lachnospira*) showed a noticeably reduced abundance in the post-antibiotic treated samples as compared to the pre-treated samples, indicating that the deleterious effects of antibiotic treatment on these genera were especially pervasive. Thus, by utilizing the various functionalities of Community-Analyzer, we were able to capture the distinct changes in the gut microbial community during the antibiotic treatment. The deleterious effect on the abundances of several bacterial groups, leading to a decrease in the taxonomic diversity of the gut microbial community, could also be captured easily using the tool.

3.3. Comparative analysis

We compared our algorithm (implemented for generating the Community-Analyzer layout) with other popularly used algorithms (which are also implemented in the tool for users' convenience) using

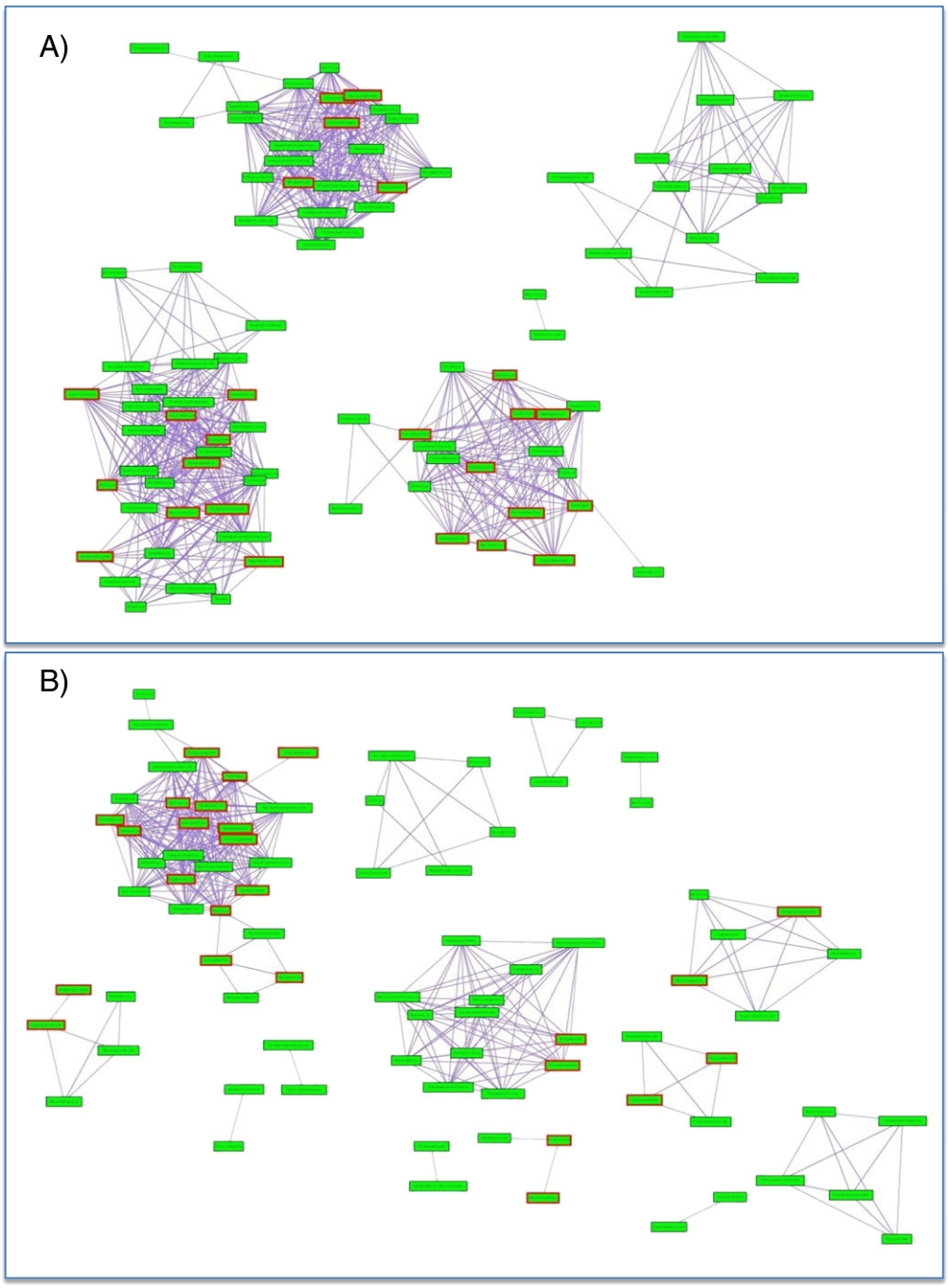

Fig. 4. The highly correlated (correlation > 0.9) hubs of positively interacting groups of microbes, present specifically in the [A] Infants and [B] Adults. The Community Interactions Investigator module was used to obtain hubs of positively correlated genera by selecting the gut samples corresponding to the infant and adults separately. The red boxes indicate the genera present in the hubs corresponding to both adults and infants. The overall results reflect that the hubs of positively correlated (or co-occurring) groups of genera are distinct (in terms of their architecture as well as memberships) in the infant gut samples as compared to those obtained from adults.

the two datasets which were described in the previous section (case studies). The phylogenetic level was set to genus for all the analyses as most of the information regarding taxonomic groups was available for this level.

The graphical layouts obtained with Community-Analyzer with those obtained using the graph based methods for the two datasets (shown in Figs. 13–15 and 16–18 of Supplementary material 2 for the Infant–Adult and Antibiotic datasets, respectively) were first compared visually. Subsequently the clustering efficiencies obtained (in these layouts) using Community-Analyzer, the PCA-based as well as the Spring-graph approaches were also compared in quantitative terms using Davies–Bouldin indices [29]. Given a set of observations belonging to different groups, the Davies–Bouldin index provides a standard quantitative measure (for the entire set) that indicates how well a given algorithm clusters the observations in the given set into their respective groups [29]. For a given set of observations, the lower the Davies–Bouldin index for a given algorithm, the better is its clustering efficiency. The graph layouts generated using Community-Analyzer and PCA, for the Infant–Adult dataset showed a clear separation of the Infant samples (Figs. 13 and 14 of Supplementary material 2). However, the graphical

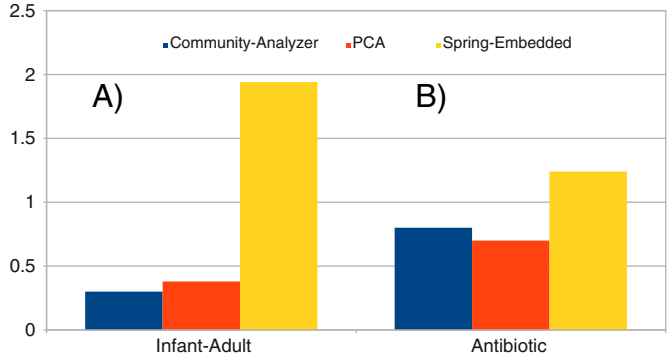

Fig. 5. Comparison of the Davies–Bouldin indices (shown on the Y-axis) obtained using the three different layout algorithms for the two datasets: (A) Infant–Adult datasets: Samples were divided into two different groups — Adult: In_A_AM, In_D_AM, In_R_AF, F1_S_AM, F1_T_AF, F2_V_AM, F2_W_AF, F2_X_CM, F2_Y_CF. Infants: In_B_IM, In_F_IF, In_M_IF and F1_U_IF. (B) Antibiotic treatment dataset: Samples were tagged into two different groups — Pre/post-treatment: A1, A2a, A2b, A2c, A4 and A5. During treatment: A3a and A3b.

layout generated using the Spring-Graph (Fig. 15 of Supplementary material 2) layout failed to achieve such separation. The PCA layout generated for the Infant–Adult samples was able to identify genera like *Bifidobacterium*, *Enterococcus*, *Klebsiella*, *Clostridium* and *Escherichia* to be the main drivers responsible for the distinct microbial community structure of the Infant samples. On the other hand, besides detecting these genera as dominant in infant gut, the graphical layout generated using Community-Analyzer was able to identify other symbiotic genera (Fig. 5 of Supplementary material 3). It is to be noted that although these additionally detected genera are present in lower abundance, they show a significantly high correlation with the identified 'infant-specific' genera. These symbiotic genera were observed to be positioned along the same horizontal level as the dominant genera in the generated Community-Analyzer layout. Thus, these symbionts along with the other infant gut specific dominant genera are likely to play key roles in making the microbial community in the infant gut different from that in the adults.

Similarly, visualization of antibiotic datasets using the Community-Analyzer as well as PCA layouts indicated antibiotic treated samples A3a and A3b to be distinctly separated from the rest of the samples (Figs. 16 and 17 of Supplementary material 2). On the other hand, the Spring-Graph layout failed to display such separation (Fig. 18 of Supplementary material 2). In addition to identifying *Anaerostipes* and *Bacteroidetes* to be the dominant microbes in A3a and A3b samples (also observed in the PCA layout), the Community-Analyzer layout could detect other groups of microbial taxa (horizontally placed nodes along with these groups). This indicated that besides *Anaerostipes* and *Bacteroidetes*, these identified groups might also account for the distinct nature of the A3a and A3b samples (obtained immediately after antibiotic administration).

The above results for visual comparisons (of the layouts obtained for both datasets) are also corroborated in the quantitative comparison of the Davies–Bouldin indices obtained for the three different algorithms (Fig. 5). While the Community-Analyzer and PCA layouts were observed to have comparable Davies–Bouldin indices (for both datasets), those obtained for the Spring-Graph layout were observed to be around 1.5 to six times higher, indicating its lowest clustering efficiency.

In a manner similar to that adopted for the different graphical layouts, trees generated using the four implemented distance matrices for the gut microbial communities constituting the Infant–Adult and the Antibiotic treatment datasets (shown in Figs. 19–22 and Figs. 23–26 of Supplementary material 2, respectively) were compared visually as well as using Genealogical Sorting Indices (GSIs) (Table 1). Similar to Davies–Bouldin indices, given a set of observations (belonging to different groups) represented as a tree, GSI values provide a quantitative measure (computed for each group) indicate how well the observations belonging to the different groups cluster in the obtained tree [30]. GSI values range between 0 and 1. The higher the GSI value obtained for observations belonging to a given group, the better is the clustering pattern of this group in the tree. Given the above premise, an ideal algorithm with the best clustering or discriminating efficiency is expected to achieve a GSI value of 1 for all the groups present in a given set of observations. This aspect has been utilized by recent studies to evaluate the discriminating capabilities of different oligo-nucleotide composition-based comparative metagenomic methodologies [31].

Visual comparison of the trees generated for the gut communities constituting the Infant–Adult dataset, using the four distance matrices implemented in Community-Analyzer (Figs. 21–24 of Supplementary material 2 and Table 1), indicated that only the trees generated using Community-Analyzer distance matrix could show a clear separation of the Infant samples from the Adult and Children samples (Fig. 21 of Supplementary material 2). This was also reflected in the GSI values wherein Community-Analyzer obtained the maximum GSI value of 1 for both the infant and the adult groups of samples. Although the trees generated using Hellinger and Euclidean distance matrices could group together three of the four Infant gut communities, one of the Infant gut communities (F1-U-IM) was observed to be placed distally from the rest (Figs. 22 and 24 of Supplementary material 2). Consequently, Hellinger and Euclidean distance matrices achieved lower GSI values of 0.67 for the infant groups. On the other hand, the tree generated using the Bray–Curtis distance matrix could not achieve any resolution between the Infant gut (GSI value of 0.07) and the Adult gut (GSI value of 0) microbial communities (Fig. 23 of Supplementary material 2).

A comparison of the trees generated using the four distance matrices for the Antibiotic treatment datasets indicated that the trees generated using the Community-Analyzer distance matrix as well as Hellinger and Euclidean distance matrices were able to clearly separate the Antibiotic treated samples (namely, A3a and A3b) from the pre- and post-treatment samples (Figs. 25, 26 and 28 of Supplementary material 2). All the three distance matrices were able to achieve a maximum GSI value of 1 for both groups of samples (Table 1). On the other hand, the tree obtained using the Bray–Curtis (Fig. 27 of Supplementary material 2) distance matrix failed to achieve this separation.

Table 1
Comparison of the Genealogical Sorting Indices (GSIs) obtained for the different groups in trees generated using the Community-Analyzer, Hellinger, Bray–Curtis and Euclidean distance matrices for the two datasets. (A) Infant–Adult dataset: Samples were divided into two different groups: Adult: In_A_AM, In_D_AM, In_R_AF, F1_S_AM, F1_T_AF, F2_V_AM, F2_W_AF, F2_X_CM, F2_Y_CF. Infants: In_B_IM, In_F_IF, In_M_IF and F1_U_IF. (B) Antibiotic treatment dataset: Samples were tagged into two different groups- Pre/post-treatment: A1, A2a, A2b, A2c, A4 and A5. During treatment: A3a and A3b.

Dataset	Group	GSI values obtained using distance matrices corresponding to			
		Community-analyzer	Hellinger	Bray–Curtis	Euclidian
Infant–Adult	Infant	1	0.67	0.07	0.67
	Adult	1	1	0	1
Antibiotic treatment	Pre/post	1	1	0	1
	During	1	1	0.07	1

The comparative evaluation of the Community-Analyzer approach (and the corresponding inter-metagenomic distance matrix) with other graph and tree-based approaches indicated that for the analyzed datasets, Community-Analyzer algorithm achieved the best discrimination between the analyzed microbiomes (that is biologically relevant to the extent possible). In other words, the Community-Analyzer algorithm seamlessly bridged the gap between the tree and graph based methods and proved to be the best in both graph and tree based comparisons. Apart from this, several interactive visualization/comparison features implemented in Community-Analyzer is expected to make it a comprehensive analysis tool suitable for all comparative metagenomic analysis.

4. Discussion

In this study, we have presented a new GUI based comparative metagenomics tool that can be used to not only compare and distinguish metagenomes based on their taxonomic profiles, but also to identify signature taxonomic groups which are specific to a given subset of microbial communities. Most importantly, the tool is especially helpful for studies involving metagenomic data sets that are collected from similar but phenotypically distinct habitats (e.g., gut environments of lean and obese individuals) or metagenomic data sets collected from similar habitats at different time points (i.e. time series metagenomic data). In both above cases, the taxonomic groups constituting the microbial communities under study are more or less similar, but exhibit subtle differences in their relative proportions. By capturing the correlations in the abundance patterns of the constituent taxonomic groups (across such microbial communities), Community-Analyzer can provide valuable insights into the inherent interaction dynamics in the given habitat. For example, taxonomic groups showing a positive correlation in their abundance profiles (and consequently placed closer to each other in the generated graphical layout) may have a positive symbiotic relationship between them. Similarly, taxonomic groups which are placed farther apart in the graphical layout may have a negative or inhibitory effect on each other. Such comparative metagenomic analyses are required by researchers working in diverse domains of microbiological research, ranging from medical microbiology [32,20,11] to environmental or geo-microbiology [33,34]. To the best of our knowledge, the Community-Analyzer is the first ever standalone tool that, besides facilitating standard comparative analyses across the metagenomic samples, enables users to gain valuable insights into the inherent interaction dynamics of the studied microbial communities.

Functionalities implemented in the Community-Analyzer can help users to identify signature taxa which are specific in metagenomic samples exhibiting a given phenotype. The tool also enables users to compare the studied metagenomic samples using standard visualization features like bar-charts, trend-graphs and heat-tables. Besides, current application also incorporates provisions for efficient data management using sessions. Using these features, users can save their intermediate analyses as session or '.gml' files and resume on their analyses at a later time. These features are expected to make the Community-Analyzer an extremely valuable resource for researchers working in diverse areas of microbiology.

It is to be noted that Community-Analyzer does not analyze the raw metagenomic reads directly (depicted in Fig. 1). Researchers first need to obtain the abundance profile of microbial groups in a given environment using various available binning/classification algorithms. These profiles are provided as inputs to Community-Analyzer. The assignment accuracy and specificity of taxonomic classification/binning algorithms are known to depend on the length and quality of the input metagenomic reads. Given that Community-Analyzer analyzes the taxonomic summaries, the inferences obtained using the Community-Analyzer is dependent on the accuracy/specificity of the obtained profiles. Therefore, users need to evaluate and use appropriate quality filters and robust classification algorithms so that the abundance profiles provided as input to Community-Analyzer are sufficiently accurate.

Competing interest statement

The authors would like to inform that a patent has been filed for the algorithm implemented in this platform. However, non-profit and academic users can freely download and use this software without any restrictions.

Acknowledgments

We acknowledge Anirban Dutta and Hemang Gandhi for their help during the course of this work. Tarini Shankar Ghosh is also a Senior Research Fellow of the Department of Biotechnology, University of Hyderabad and would like to acknowledge Department of Biotechnology, University of Hyderabad, for its support.

Appendix A. Supplemetary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2013.08.004.

References

[1] T.S. Ghosh, M.H. Mohammed, D. Komanduri, S.S. Mande, ProViDE: a software tool for accurate estimation of viral diversity in metagenomic samples, Bioinformation 6 (2) (2011) 91–94.
[2] T.S. Ghosh, M.H. Monzoorul, S.S. Mande, DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences, BMC Bioinforma. Suppl. 7 (2010) S14.
[3] J. Goll, et al., METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics, Bioinformatics 26 (20) (2010) 2631–2632.
[4] D.H. Huson, A.F. Auch, J. Qi, S.C. Schuster, MEGAN analysis of metagenomic data, Genome Res. 3 (2007) 377–386.
[5] A.C. McHardy, H.G. Martín, A. Tsirigos, P. Hugenholtz, I. Rigoutsos, Accurate phylogenetic classification of variable-length DNA fragments, Nat. Methods 4 (1) (2007) 63–72.
[6] F. Meyer, et al., The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes, BMC Bioinforma. 9 (2008) 386.
[7] M.H. Mohammed, T.S. Ghosh, N.K. Singh, S.S. Mande, SPHINX—an algorithm for taxonomic binning of metagenomic sequences, Bioinformatics 27 (1) (2011) 22–30.
[8] M.H. Monzoorul, T.S. Ghosh, D. Komanduri, S.S. Mande, SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences, Bioinformatics 25 (14) (2009) 1722–1730.
[9] Finkbeiner, et al., Metagenomic analysis of human diarrhea: viral detection and discovery, PLoS Pathog. 4 (2) (2008) e1000011.
[10] S. Nakamura, et al., Metagenomic diagnosis of bacterial infections, Emerg. Infect. Dis. 14 (11) (2008) 1784–1786.
[11] P.J. Turnbaugh, et al., A core gut microbiome in obese and lean twins, Nature 457 (7228) (2009) 480–484.
[12] P.J. Turnbaugh, et al., An obesity-associated gut microbiome with increased capacity for energy harvest, Nature 444 (7122) (2006) 1027–1031.
[13] M. Arumugam, et al., Enterotypes of the human gut microbiome, Nature 473 (2011) 174–180.
[14] J.A. Gilbert, et al., The seasonal structure of microbial communities in the Western English Channel, Environ. Microbiol. 11 (2009) 3132–3139.
[15] S. Mitra, J.A. Gilbert, D. Field, D.H. Huson, Comparison of multiple metagenomes using phylogenetic networks based on ecological indices, ISME J. 4 (10) (2010) 1236–1242.
[16] S. Mitra, B. Klar, D.H. Huson, Visual and statistical comparison of metagenomes, Bioinformatics 25 (15) (2009) 1849–1855.
[17] P.D. Schloss, J. Handelsman, Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness, Appl. Environ. Microbiol. 71 (3) (2005) 1501–1506.
[18] N. Segata, et al., Metagenomic biomarker discovery and explanation, Genome Biol. 12 (6) (2011) R60.
[19] C. Shyu, T. Soule, S.J. Bent, J.A. Foster, J.J. Forney, MiCA: a web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphisms of 16S and 18S rRNA genes, 53 (4) (2007) 562–570.
[20] The VAMPS project, http://vamps.mbl.edu/.
[21] J.G. Caporaso, et al., QIIME allows analysis of high-throughput community sequencing data, Nat. Methods 7 (5) (2010) 335–336.
[22] K. Faust, et al., Microbial co-occurrence relationships in the human microbiome, PLoS Comput. Biol. 8 (7) (2012) e1002606.
[23] Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, Appl. Environ. Microbiol. 73 (16) (2007) 261–267.

[24] J.R. Bray, J.T. Curtis, An ordination of the upland forest communities of southern Wisconsin, Ecol. Monogr. 27 (1957) 325–349.
[25] C.R. Rao, A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance, Qü¨estiio (Quaderns d'Estadí´stica i Investi-gacio operativa), 19, 1995, pp. 23–63.
[26] D.J. Balding, J.B. Martin, C. Cannings, Handbook of statistical genetics, vol. 1, John Wiley and Sons, 2007.
[27] K. Kurokawa, et al., Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes, DNA Res. 214 (4) (2007) 169–181.
[28] L. Dethlefsen, S. Huse, M.L. Sogin, D.A. Relman, The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing, PLoS Biol. 6 (11) (2008) e280.
[29] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (2) (1979) 224–227.
[30] M.P. Cummings, M.C. Neel, K.L. Shaw, A genealogical approach to quantifying lineage divergence, Evolution 62 (9) (2008) 2411–2422.
[31] T.S. Ghosh, et al., HabiSign: a novel approach for comparison of metagenomes and rapid identification of habitat-specific sequences, 12 (Suppl. 3) (2011) S9.
[32] C. Manichanh, et al., Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach, Gut 55 (2) (2007) 205–211.
[33] D. Bhaya, et al., Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses, ISME J. 1 (8) (2007) 703–713.
[34] E.A. Dinsdale, et al., Functional metagenomic profiling of nine biomes, Nature 452 (7187) (2008) 629–632.