Microbiome

SOFTWARE

Open Access

CViewer: a Java-based statistical framework for integration of shotgun metagenomics with other omics datasets

Orges Koci[1], Richard K. Russell[2], M. Guftar Shaikh[3], Christine Edwards[1], Konstantinos Gerasimidis[1] and Umer Zeeshan Ijaz[4,5,6*]

Abstract

Background Shotgun metagenomics for microbial community survey recovers enormous amount of information for microbial genomes that include their abundances, taxonomic, and phylogenetic information, as well as their genomic makeup, the latter of which then helps retrieve their function based on annotated gene products, mRNA, protein, and metabolites. Within the context of a specific hypothesis, additional modalities are often included, to give host-microbiome interaction. For example, in human-associated microbiome projects, it has become increasingly common to include host immunology through flow cytometry. Whilst there are plenty of software approaches available, some that utilize marker-based and assembly-based approaches, for downstream statistical analyses, there is still a dearth of statistical tools that help consolidate all such information in a single platform. By virtue of stringent computational requirements, the statistical workflow is often passive with limited visual exploration.

Results In this study, we have developed a Java-based statistical framework (https://github.com/KociOrges/cviewer) to explore shotgun metagenomics data, which integrates seamlessly with conventional pipelines and offers exploratory as well as hypothesis-driven analyses. The end product is a highly interactive toolkit with a multiple document interface, which makes it easier for a person without specialized knowledge to perform analysis of multiomics datasets and unravel biologically relevant patterns. We have designed algorithms based on frequently used numerical ecology and machine learning principles, with value-driven from integrated omics tools which not only find correlations amongst different datasets but also provide discrimination based on case–control relationships.

Conclusions CViewer was used to analyse two distinct metagenomic datasets with varying complexities. These include a dietary intervention study to understand Crohn's disease changes during a dietary treatment to include remission, as well as a gut microbiome profile for an obesity dataset comparing subjects who suffer from obesity of different aetiologies and against controls who were lean. Complete analyses of both studies in CViewer then provide very powerful mechanistic insights that corroborate with the published literature and demonstrate its full potential.

*Correspondence:
Umer Zeeshan Ijaz
Umer.Ijaz@glasgow.ac.uk
Full list of author information is available at the end of the article

Introduction

Whilst advances in high-throughput sequencing technologies have revolutionized the study of uncultured microbial communities, recent microbiome surveys using whole-genome shotgun sequencing delineate an enormous amount of information about microbial genomes, their taxonomic and functional profiles. With the ability to assemble nearly complete genomes, we are now at the stage where we can understand the mechanistic underpinnings to the biological processes involved in the studies for which the sequencing data are generated. Additionally, we can use metadata (typically, physicochemical and clinical parameters) relevant to the study to highlight patterns of interest. Whilst many tools have become available in the past few years, from microbial genome binnings [1–3] to annotation [4, 5], as well as downstream statistical analyses [6, 7], there is still a dire need to consolidate all the biological entities of microbial genomes, such as gene products, mRNA, protein, metabolites, and their interactions in a single platform, and to enable exploration of this multi-component dataset in an easy-to-use interface.

This combination multiomics approach will be advantageous and will enable researchers, especially those who have a basic understanding about microbiology and microbial informatics, to quickly elucidate the interplay between microbiomes and their environments. Furthermore, combining visualizations with statistical inference in a single software platform has an added advantage of leveraging time spent on the analyses. Previous work towards this goal included conventional genome viewers like MGAViewer [8] where it is only possible for one to analyse at most two genomes together for comparative genomics, *albeit*, impractical for metagenomics exploration. Whilst there are other attempts such as Anvi'o [3], a tool that implements an advanced analysis and visualization platform for omics data, combining taxonomic coverages with phylogenetic trees, and is typically used as a de facto tool for metagenomics exploration, as well as Elviz [9] and WHAM [10]! which allow the exploration of metagenomic data. Although useful, these tools have limited support for statistical analyses, particularly Anvi'o [3] and Elviz [9], without any means to do an integrative assessment. In view of the limitations as above, there is an unmet need for improvement to fully exploit the potential of the shotgun metagenomics data, both in terms of graphical user interface interactivity, as well as statistical inference. This serves as the basis for developing CViewer in a language such as Java that is not only cross-platform compatible, but also has an established graphical interface to show relevant information on the fly.

The design principles of CViewer include Java-based solution able to run on users' local machines to incorporate the output from CONCOCT [2], a binning software, as well as annotation data for the contigs from major third party taxonomic and annotation tools [5, 11]. All this is done to explore the sample space in the context of clinical metadata, with the interface given in Fig. 1. Beyond visualizing contigs from CONCOCT software, and looking at how they cluster at the species level, we have implemented a variety of statistical algorithms from numerical ecology literature, what is conventionally available in R's vegan package [7], to allow exploratory as well as hypothesis driven analyses, emphasizing functional traits of microbial communities and their phylogenetic signal to assess community assembly. The implemented functionalities in CViewer include alpha (based on the Simpson Index [12], Shannon Entropy [13], and Pielou's Evenness [14]) and beta diversity (principal component analysis [15], multi-dimensional scaling [16]) estimates; statistical procedures to test covariates in terms of correlation with microbial community structure (fuzzy set ordination [17]) and the variability they account for (PERMANOVA [18]) estimates (Fig. 1A); differential abundance (utilizing both the non-parametric Kruskall-Wallis [19] test and Friedman [20] test for repeated measures) and correlation analyses (including the Pearson's [21], Kendall's [22] and Spearman's [23] coefficients) (Fig. 1B); enrichment analyses of KEGG [24] metabolic pathways (Fig. 1C); phylogenetic alpha diversity indices (based on the Net Relatedness Index and Nearest Taxon Index [25]) (Fig. 1D); visualization and analysis of coverage, clustering and functional diversity of metagenomics contigs (Fig. 1E); and integrative techniques for the analysis of multi-omics datasets (including the DISCO-SCA [26], JAVA [27], and O2PLS [28] algorithms) (Fig. 1F), with details given in the Additional file 1: Supplementary Note 1. The software and the accompanying data along with video demos exploring these features are available at https://github.com/KociOrges/cviewer.

To demonstrate how CViewer is useful in analysing the metagenomics datasets, we have next used the software to explore a dataset of gut microbiome composition and metabolic profile from children with active Crohn's disease (CD) who undergo dietary therapy with exclusive enteral nutrition (EEN). The data were categorized as Crohn's disease individuals who were treated with EEN for two months and healthy individuals to use as the reference microbiome and metabolome of healthy status. A smaller subset of these data has been previously published under Gerasimidis et al. [29], Quince et al. [30], and Alghamdi et al. [31], and additional samples and metagenomic sequencing were incorporated to see if the trends became more prominent. In addition, CViewer

Fig. 1 The software layout and the main features that are supported in the system

was tested on another dataset where the gut microbiome of children with pathological cause of hypothalamic obesity (Prader-Willi syndrome) was compared against children who suffer from common or classical type of obesity. Lean subjects from each of these two groups were used as controls. For more details on the datasets and the study characteristics, the reader is referred to Additional file 1: Supplementary Materials and Methods.

Results

Crohn's disease dataset

Community structure of the gut metagenome and metabolome

Alpha diversity analysis using the Shannon index suggested significantly lower diversity in the gut metagenome of CD children compared to that of healthy controls ($p < 0.0001$) (Fig. 2A). Similar results were also found when we looked for differences during treatment with EEN (Fig. 2B and Additional file 1: Supplementary Note 2). Shannon diversity remained significantly lower in CD subjects than in controls prior to EEN, throughout treatment and post-treatment (A:Start of EEN, $p = 0.0001$; B:15d, $p < 0.0001$; C:30d, $p < 0.0001$; D:End of EEN, $p < 0.0001$; E:Free Diet, $p = 0.0034$). However, during EEN, the diversity of CD subjects decreased further, with the effect becoming noticeable after 15 days of treatment, reaching minimum diversity after ~ 30 days, and then showing a slight recovery towards the end of EEN, and complete recovery to pre-treatment levels when patients returned to habitual diet. When we explored for differences between the treatment days, our results suggested a significant difference only between the time points C and E of EEN for Shannon's diversity ($p = 0.0134$).

Beta diversity analysis of the gut metagenome using PCA showed an evident clustering of CD patients distinct from controls (Fig. 2C). This was also observed using PERMANOVA (distances between groups) which suggested that 8.3% of the variation in community structure was explained significantly by the sample groups ($p = 0.001$). When the CD subjects were grouped according to EEN samples, PCA described higher variability of the community structure for CD children compared to healthy controls (Fig. 2D) and the amount of significantly explained variability according to PERMANOVA increased to 11.22% ($p = 0.001$) suggesting that more differences in the community structure were accounted to changes during EEN. However, after the CD patients completed the treatment and returned to their free habitual diet, the samples appeared closer to the pre-treatment groups. Furthermore, analysis using fuzzy set ordination (FSO) suggested a significant association between the microbial community structure of the CD individuals at

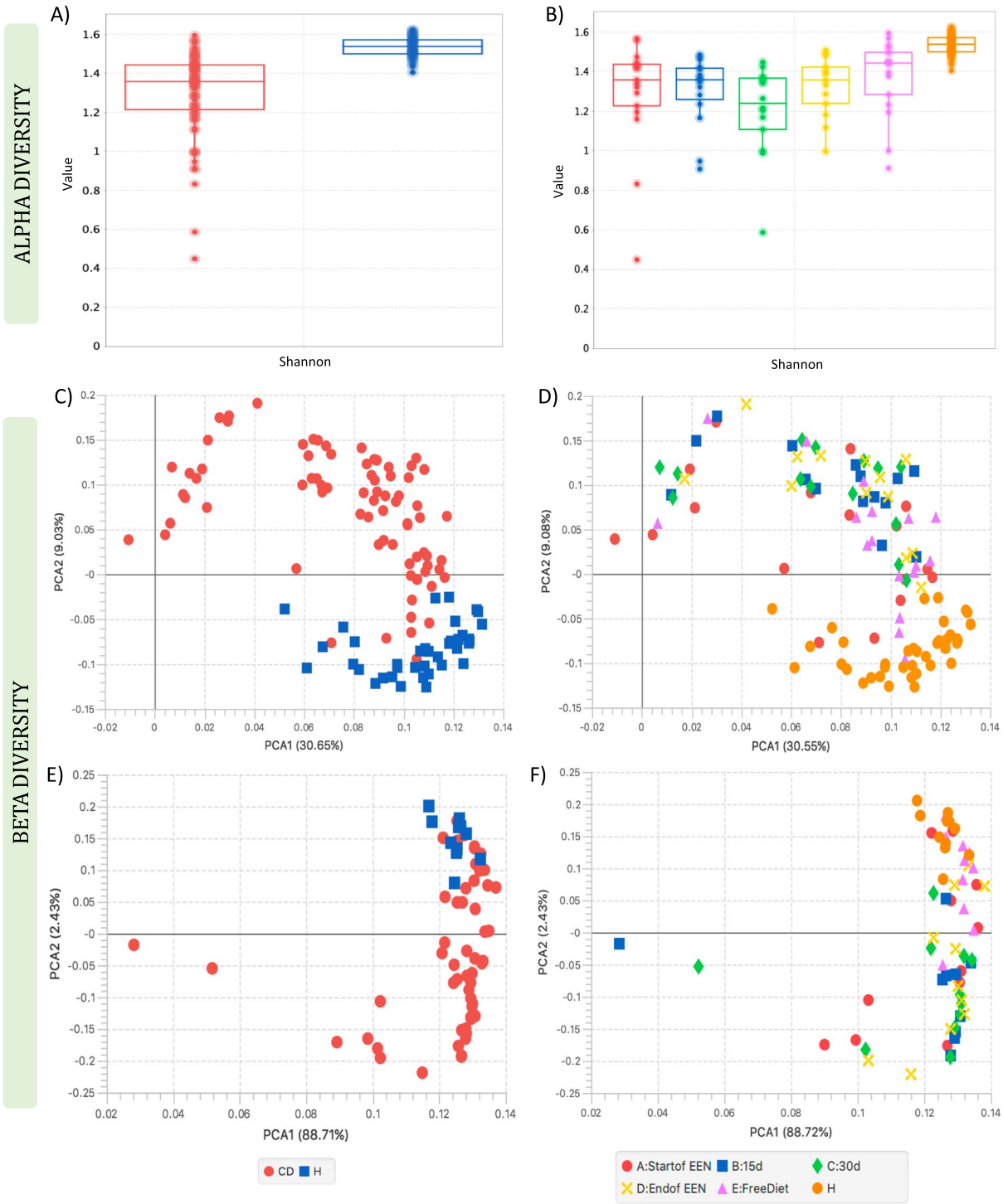

Fig. 2 Shannon diversity of **A** CD patients and healthy controls and **B** of CD patients during EEN and healthy controls, PCA analysis of shotgun metagenomics community structure of **C** CD patients and healthy controls and **D** CD groups during EEN and healthy controls, and PCA analysis of the faecal metabolomic data derived from **E** CD patients and healthy controls and **F** CD patients during EEN and healthy controls

point D (end of EEN) and E (free diet) of EEN with calprotectin ($r=0.638$, $p=0.001$), a marker of gut inflammation, (Additional file 1: Supplementary Note 2 and Figs. S4A–S4B) and showed that patients who achieved remission, especially at the end of EEN, were related to lower calprotectin levels compared with the patients who still have active disease following completion of treatment. Similarly, PCA analysis on the faecal metabolome showed a distinct separation for the healthy control samples from the CD patient group (Fig. 2E). In addition, PCA showed that the metabolomes from healthy children were more tightly clustered, whilst CD patients were more variable and changed during EEN treatment (Fig. 2F).

Differential abundance analysis

Abundance analysis showed several clusters (genomes expressed in average abundances across samples) differentiating significantly in the CD groups and the healthy controls (Additional file 1: Table S1). Amongst them, *Bifidobacterium longum* was in lower abundance in CD patients, at all sampling points, with a significant decrease noticed during EEN, when compared with the healthy control group (A:Start of EEN, $p>0.05$; B:15d, $p=0.0002$; C:30d, $p=0.0001$; D:End of EEN, $p<0.0001$) (Fig. 3A). Moreover, *Bifidobacterium adolescentis* was in significantly lower abundance in the CD samples at all sampling points with the effect being more evident after 15 days of EEN and onwards (A:Start of EEN, $p=0.0282$; B:15d, $p=0.0002$; C:30d, $p=0.0002$; D:End of EEN, $p<0.0001$; E:Free Diet, $p=0.0002$) (Fig. 3A). In a similar way, *Eubacterium rectale* was found in a lower abundance in CD groups and remained significantly lower at all points, although it moved closer to the control levels post-EEN treatment on food reintroduction (A:Start of EEN, $p=0.0001$; B:15d, $p<0.0001$; C:30d, $p<0.0001$; D:End of EEN, $p<0.0001$; E:Free Diet, $p=0.0357$) (Fig. 3A). In contrast, *Escherichia coli* was significantly more prevalent in samples from CD patients at EEN initiation compared with the healthy controls ($p=0.0009$) and remained more abundant in CD individuals across all stages of EEN, although a decreasing pattern could be observed during the treatment course (Fig. 3A).

Differential analysis using Kruskal–Wallis also suggested that 487 annotated metabolites were significantly different between the CD groups and the healthy controls (Additional file 1: Table S2). More specifically, the levels of *docosatetraenoic acid* were found in a higher abundance in children with active CD at all sampling points compared to the healthy group (A:Start of EEN, $p=0.0001$; B:15d, $p=0.0001$; C:30d, $p=0.0017$; D:End of EEN, $p=0.0002$; E:Free Diet, $p=0.0089$). (Fig. 3B). This was also the case for *icosatrienoic acid* (A:Start of EEN, $p=0.0014$; B:15d, $p=0.0002$; C:30d, $p=0.0026$; D:End of EEN, $p=0.0007$; E:Free Diet, $p=0.0147$) and *arachidonic acid* (5Z_8Z_11Z_14Z-eicosatetraenoic acid) (A:Start of EEN, $p=0.0001$; B:15d, $p=0.021$; C:30d, $p=0.0195$; D:End of EEN, $p=0.0189$; E:Free Diet, $p=0.0056$) (Fig. 3B). The compounds remained significantly higher

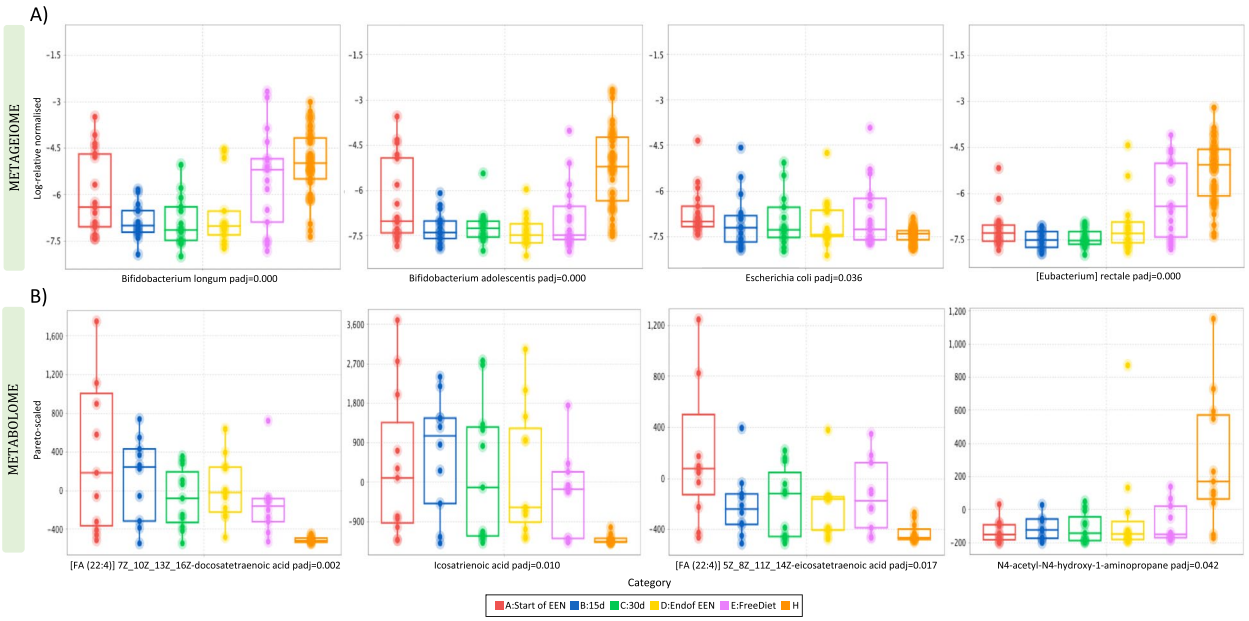

Fig. 3 Sequential changes in **A** the log-relative abundances of species and **B** the pareto-scaled frequencies of metabolites that were significantly different in the CD groups (before, during, and after EEN treatment) and the healthy controls. Statistical significance is reported using corrected *P*-values (padj)

than the controls during EEN although some of the metabolites regressed to pre-treatment levels when subjects returned to free diet, with the effect being most pronounced for *arachidonic acid*. In contrast, the levels of an ornithine isomer (N4-acetyl-N4-hydroxy-1-aminopropane) were significantly lower in the CD samples than in the healthy group, with the effect being more noticeable between the pre-treatment samples and the healthy controls (A:Start of EEN, $p=0.0013$; B:15d, $p=0.016$; C:30d, $p=0.0021$; D:End of EEN, $p=0.0069$; E:Free Diet, $p=0.020$) (Fig. 3B).

Changes in faecal bacterial metabolites during EEN

Significant differences were observed when we assessed the quantitative changes in the concentration of major targeted bacterial metabolites of CD patients during EEN (Fig. 4A). Compared to healthy individuals, faecal *pH* was significantly higher when patients were on EEN, but no differences were observed at EEN initiation or when patients returned to their habitual diet (A:Start of EEN, $p>0.05$; B:15d, $p=0.0005$; C:30d, $p=0.0004$; D:End of EEN $p<0.0001$; E:Free Diet, $p>0.5$) (Fig. 4A). This was also the case for *total sulphide* (A:Start of EEN, $p>0.05$; B:15d, $p=0.0015$; C:30d, $p=0.0004$; D:End of EEN, $p=0.0018$; E:Free Diet, $p>0.5$) (Fig. 4A). In contrast, *butyric acid* was significantly lower in patients with CD compared to healthy controls with the effect being more evident after 15 days of EEN and onwards. This effect was lost when patients returned to their habitual diet (A:Start of EEN, $p=0.0137$; B:15d, $p=0.0002$; C:30d, $p=0.0003$; D:End of EEN, $p<0.0001$; E:Free Diet, $p>0.5$) (Fig. 4A). A similar pattern was also found for the proportional ratio of *butyric acid* (A:Start of EEN, $p=0.0367$; B:15d, $p=0.0002$; C:30d, $p=0.0003$; D:End of EEN, $p=0.0002$; E:Free Diet, $p>0.5$) (Fig. 4A).

Integrated analysis of metagenomics and metabolomics

An integrated analysis of faecal metagenomics and metabolomics was performed to investigate possible interactions between the two datasets. This methodology is particularly useful for decomposing the variability of the composite omics systems into a joint variability or common structure that highlights the biological mechanisms underlying both the omics types under study. High

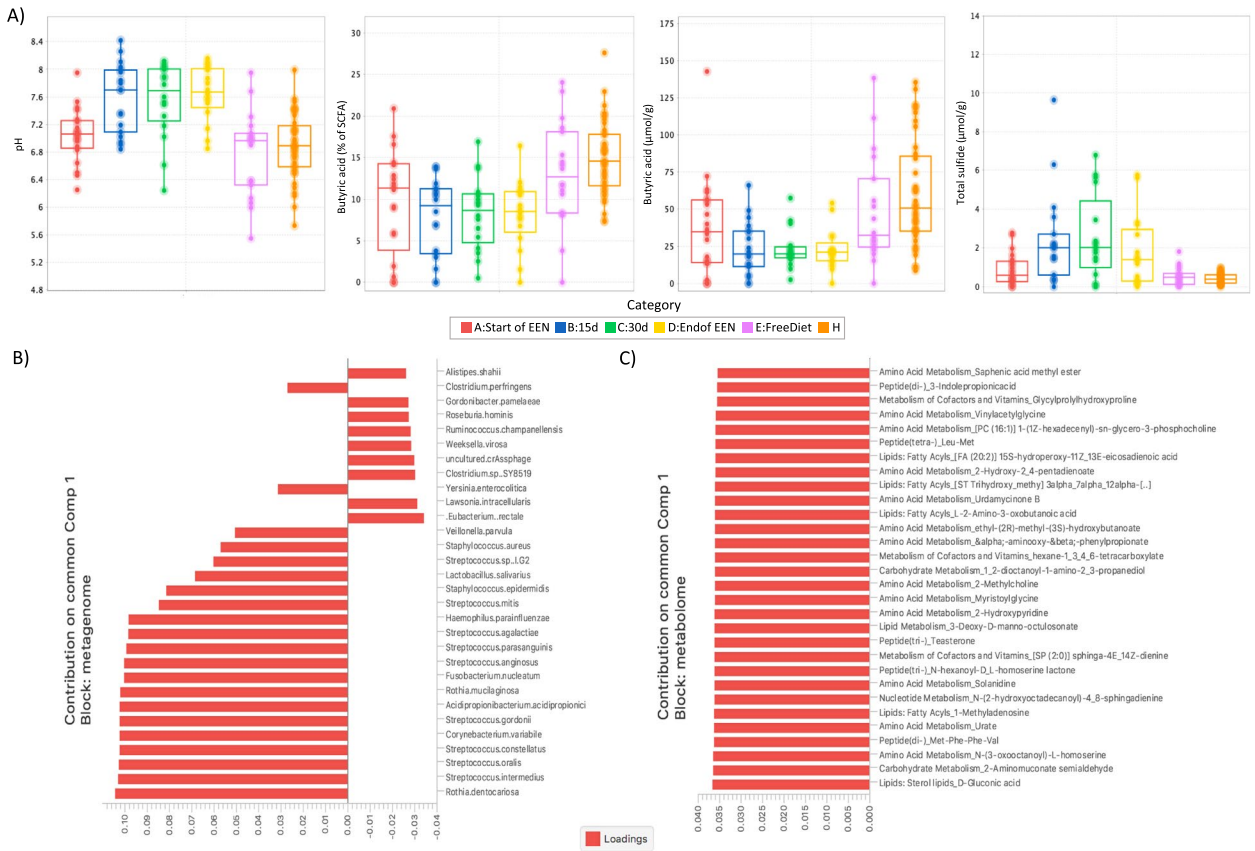

Fig. 4 **A** Serial changes in faecal pH and concentration of major bacterial metabolites in CD subjects during EEN and healthy controls. **B** The joint loadings for the top 30 species. **C** The top 30 metabolites with the highest contribution in the first component of the common structure. Loadings are sorted in decreasing order based on absolute values

contributions to the common structure across the two datasets (expressed as component loadings) could indicate a mechanistic association between the microbes and the metabolites, such as that a species may release a particular metabolite, or that specific metabolites may stimulate the growth of a particular species.

To explore this, the DISCO-SCA [26] method was used for samples collected from CD patients before EEN initiation and healthy individuals (see Additional file 1: Supplementary Material and Methods and Supplementary Note 1 for more details on methods and data pre-processing). Figure 4B and C visualize the estimated DISCO-SCA joint loadings for the omics datasets for the first common component, sorted by absolute value. When we explored the 30 first species with the highest contribution values for the metagenome, we noticed that *Rothia dentocariosa* had the largest value in the common structure loadings, whilst a high contribution was also noticed for several species of the genus *Streptococcus* (*Streptococcus intermedius/oralis/constellatus/gordonii/arginosus/parasanguinis/agalactiae*) (Fig. 4B). In a similar way, when we examined the 30 first metabolites with the highest contribution in the common structure for the metabolome, we found that *D-gluconic acid*, *2-Aminomuconate semialdehyde*, and *N-(3-oxooctanoyl)-L-homoserine* had the highest magnitude of positive loading values as compared to the other metabolites (Fig. 4C).

Obesity dataset

Bacterial community structure

Shannon's entropy and Pielou's evenness showed no significant difference in the microbial diversity between the four groups, suggesting that the bacterial communities of the participants were similar in species abundance and richness ($p=0.223$) (Fig. 5A) and evenness ($p=0.228$). Shannon entropy provided similar results when participants were grouped according to obesity status ($p=0.272$) and obesity aetiology ($p=0.551$). Moreover, although the difference was not significant ($p=0.156$), it could be seen that the microbiota of the hypothalamic lean group was more diverse than the hypothalamic obese children (Fig. 5B). Beta diversity analysis using MDS showed no evident clustering in the community structure of the four groups, suggesting a similar microbial profile between the study participants (Fig. 5C). This was also confirmed by PERMANOVA, which suggested that the groupings did not account for a significant amount of the variability in the community structure ($R^2=7.13\%$, $p=0.355$). Using PERMANOVA, no significant effect of obesity phenotype ($R^2=1.66\%$, $p=0.826$) or obesity aetiology ($R^2=2.12\%$, $p=0.532$) on the community structure variation was observed. Moreover, differential analysis using Kruskal–Wallis did not suggest any significant differences in the average abundances of clusters (genomes) between the four groups when they were compared to each other, or when they were categorized according to obesity phenotype and aetiology.

Faecal bacterial metabolites

Significant differences in the proportion and concentration of bacterial metabolites were observed when the samples were grouped based on pathology (healthy vs. pathology). Healthy participants (healthy lean and healthy obese) had a significantly lower percentage ($p=0.001$) of *acetate* compared with the pathology groups (hypothalamic lean and hypothalamic obese) (Fig. 5H). In contrast, healthy participants had a significantly higher percentage of *butyrate* ($p=0.001$) and *ammonia* concentration ($p=0.013$) than the hypothalamic individuals (Fig. 5H, I). In a similar way, when we examined the differences between the samples according to obesity status (healthy lean and hypothalamic lean vs. healthy obese and hypothalamic obese), it was noticed that the obese group (healthy obese and hypothalamic obese) had a significantly higher concentration ($p=0.011$) of *acetate*, concentration ($p=0.001$) and percentage ($p=0.001$) of *propionate*, and concentration ($p=0.006$) of total *SCFA* than the lean samples (Fig. 5I and Additional file 1: Fig. S5).

Implementation

CViewer was developed as a Java-based desktop application that has cross-platform portability since a compiled Java program can run on all platforms for which there exists a Java Virtual Machine (JVM). This holds for all major operating systems, including Windows, Mac OS, and Linux. Java is also very flexible for designing GUIs by providing a rich set of packages for graphics manipulation. For visualization, a combination of toolkits and chart libraries were used mainly from JavaFX, Swing, and JFreeChart, which offer easy-of-use benefits and extensible and pluggable components resilient to future enhancements. We have considered an intuitive integrated multiple document interface (MDI) as opposed to the single document interface (SDI) where you have one window per functionality to avoid cluttering up of windows on the desktop and to have a single window with multiple panes instead to visualize all the information. This makes the GUI more coherent, simpler to use and easier to operate.

The software design requires user input in the form of CSV, TSV, and GBK files. The business logic performs data entry validation and, when necessary, element matching for the input datasets to ensure compatibility for downstream analyses (i.e. all datasets should describe the same set of samples). In addition to user-provided

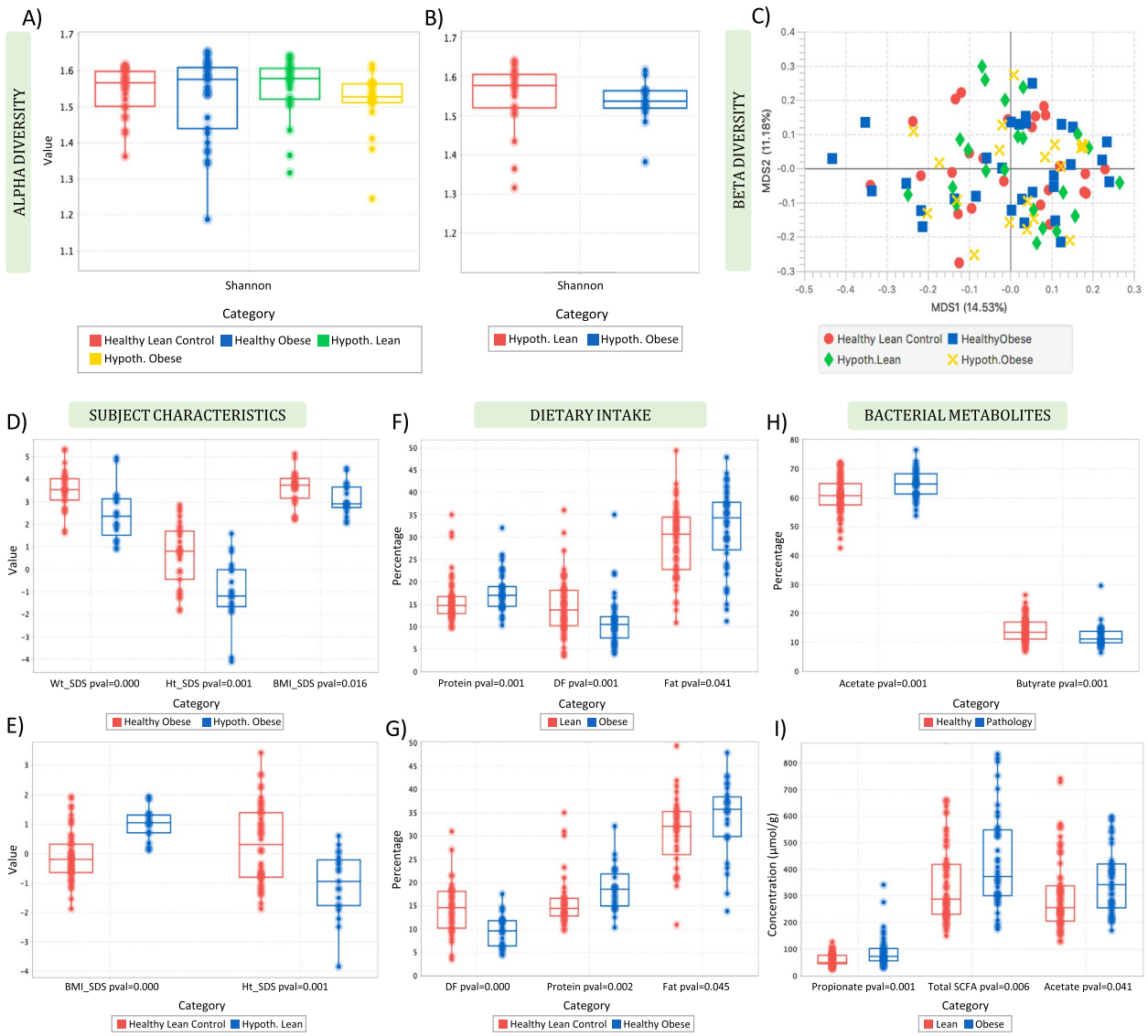

Fig. 5 Shannon diversity of **A** the four study groups and **B** hypothalamic lean (red) and hypothalamic obese participants (blue). Multidimensional scaling based on Bray–Curtis distance demonstrating clustering of the four study groups (**C**). Boxplots showing the significant differences in the proportion of *acetate* (**D**) and concentration of *ammonia* (**E**) according to pathology and the concentration of major bacterial metabolites according to lean and obese phenotype (**F**). Lean, healthy and hypothalamic lean; obese, healthy and hypothalamic obese; healthy, healthy lean and obese; pathology, hypothalamic lean and obese

data, the software relies on the local filesystem for retrieving and storing information relevant to analysis, such as the full list of KEGG metabolic pathways and maps and single-copy clusters of orthologous genes useful for phylogenomic analysis, which are given in the form of TXT, PNG, or XML files. A Document Object Model (DOM) parser is used to represent and modify XML pathway maps as tree structures, and ImageIO API to extract them into image formats. The software is distributed in the form of an executable, ready-to-run, and platform-independent package file (JAR) which requires no installation or further configuration from the user. The application interacts with the operating system, and in principle, no Internet connection is required when using the tool.

For the sake of simplicity, CViewer has been designed to make use of the available machine memory for data manipulation, instead of utilizing a dedicated database for this purpose (e.g. MySQL, MongoDB). This approach alleviates the efforts of database configuration and maintenance which can often be complicated and time-consuming, particularly for people with no

related knowledge. However, it may still pose a challenge in terms of the computational power that is needed for the tool to achieve optimal performance. To address this, user-provided data are pre-processed upon entry and compressed into files of reduced size by maintaining information only essential for the tool to operate or present in the software interface, as is often the case for annotation files (GBK). In addition, indexing is applied as a pre-processing step for faster information retrieval of, e.g. CDS regions of a particular contig. Nevertheless, we still consider the implementation of a database as useful for CViewer, particularly in the case of a web-based version of the software. System maintenance would then rely exclusively on the application server making it possible for users to access the resource online with no further action. To achieve this, a new system design would need to be established employing web development technologies (e.g. SpringBoot, J2EE), an objective that we pursue to realize in a future version of our tool.

Discussion

CViewer has successfully revealed patterns of interest in two independent datasets: a dataset for longitudinal gut microbiome and metabolomic profiles from children with Crohn's disease who undergo dietary treatment with EEN; as well as on gut microbiome profile for an obesity dataset where subjects have a non-pathological or a pathological cause of obesity, and as compared to those who are lean. In the former study, beta diversity analysis of the gut microbiome and metabolome showed a clear separation of the CD groups throughout treatment and post-treatment from the healthy controls. In addition, the gut microbiota of CD patients changed soon after EEN initiation was less diverse than in controls at all sampling points and decreased during EEN with a recovery to pre-treatment levels when patients returned to their habitual diet. These observations align with previous studies [32–35], including some of ours [30, 31], demonstrating a distinct clustering of the metabolomes [35] and metagenomes of CD patients from that of healthy controls and exhibiting general decreases in microbiota diversity relative to healthy individuals [32], especially during EEN treatment [33, 34], suggesting the usefulness of an integrative approach. Our analysis also showed that classic commensal organisms such as *Bifidobacterium longum*, *Bifidobacterium adolescentis*, and *Eubacterium rectale* differentiated in the CD groups over the course of EEN and were in lower abundance in CD children, compared to controls. These results align to previous research exploring the role of these organisms in CD pathogenesis [36]. Conversely, we noticed a higher abundance of *E. coli* in CD subjects than in controls; an expected outcome as the *E. coli* population is a much-studied topic in CD, particularly adherent invasive *E. coli* which are overrepresented in CD patients [37].

Also, in accordance with our previous findings [31], omega-6 fatty acids, including *docosatetraenoic acid* and *arachidonic acid*, were in significantly higher levels in the CD patients compared to healthy individuals and remained high both pre- and post-EEN treatment in the CD group. These findings conform to published work that highlights these compounds as pro-inflammatory in the gut and implicated in IBD [38, 39]. At the same time, by exploiting exploratory methods further provided by the software, such as FSO, we were able to highlight an association between calprotectin and the microbiome of CD patients, especially for those who achieved remission at point D (end of EEN) of treatment. In their research, Quince et al. [30] reported several taxa which were different between the CD and controls and significantly correlated with calprotectin, with *Bifidobacterium* spp. having the strongest negative and *Atopobium* spp. the strongest positive association with calprotectin in multivariate regression analysis. They also suggested that EEN caused the reduction in the relative abundance of gut bacteria that were positively and negatively associated with calprotectin. Although our study is limited in this perspective due to the lack of an implementation for regression analysis in the current version of the software, such an enhancement is one of our immediate future plans and with our findings for calprotectin and microbiome showing promising results, additional analysis on this aspect will be pursued in a future study with CViewer.

Finally, integrated analysis of the metagenome and the metabolome showed that *D-gluconic acid*, *2-Aminomuconate semialdehyde*, and *N-(3-oxooctanoyl)-L-homoserine* had the highest magnitude of positive loading values as compared to the other metabolites, whilst a high contribution was observed for *Rothia dentocariosa* and several species of the genus *Streptococcus* (*Streptococcus intermedius/oralis/constellatus/gordonii/arginosus/parasanguinis/agalactiae*) in the common structure loadings, i.e. the biological mechanisms underlying both omics datasets. *Rothia* species are assumed to have important interactions with the host immune system that may promote inflammatory diseases, including Crohn's disease. *Rothia dentocariosa*, in particular, has been previously reported to increase the production of the inflammatory cytokine tumour necrosis factor-alpha (TNF-a) and thus might act as an intermediate factor for increased inflammation in the oral cavity [40]. Moreover, *Streptococci*, especially the *Streptococcus anginosus* (milleri) group, comprising of *S. intermedius*, *S. arginosus*, and *S. constellatus* species, has been associated with liver abscess in patients with Crohn's disease [41]. Similar to the metabolite loadings, the values for these species were also positive and

could suggest a potentially synergistic positive interaction between the microbes and the metabolites, where the highly contributing metabolites could either promote the growth of the species highlighted above, or that those species may produce the particular metabolites. These associations captured with CViewer provide some novel insights into the interactions of the two omics datasets and as relevant studies are still not prevalent in literature, supplementary correlation analysis (using, e.g. Spearman's correlation) of the metabolites and the species with high contributions in the common structure loadings could be useful to illustrate them even further as part of a future exploration with the tool.

In the latter study of lean and obese children of obesity of different aetiology, we demonstrated that the gut microbiota and metabolic activity did not differentiate between obesity of different aetiology or between obese and lean phenotypes, suggesting that gut microbiota is not incriminated in the aetiology of obesity. Moreover, our results described higher faecal *SCFA* in the two obese groups (common and hypothalamic obese) compared with lean (healthy and hypothalamic lean). Even though the role of *SCFA* in obesity is still not fully understood, recent research has demonstrated that a higher concentration of *SCFA* in faeces is associated with obesity and hypertension [42]. However, the absence of difference in SCFA concentration between individuals with obesity of different aetiology strongly suggests that the increased production of SCFA in these groups is a secondary effect of differences in dietary patterns and hyperphagia rather than a primary defect increasing the risk of obesity onset.

Conclusions

Analytics for shotgun sequencing metagenomics has seen substantial growth in recent years by availability of numerous tools that cover and advance a particular aspect of the analyses, with their usability confined to the passive running of scripts that allow little or no interactivity. CViewer on the other hand allows one to explore the datasets and benefits from the interactivity offered by the graphical user interface, particularly the implementation platform in Java is well suited towards this purpose. Secondly, we have incorporated comprehensive statistical tools in the framework to allow one place to process and analyse all the datasets without the need to use any third-party tools such as R or Matlab. Moreover, multiomics exploration allows seamless integration of metagenomics with other omics technologies, such as metabolomics, which reveal patterns that not only consolidate/correlate between multiple datasets but provide discriminatory cues for multiple treatment groups. This way, an overarching linkage between multiple modalities fills in the gaps in our understanding of how microbes are behaving in the context of the hypothesis under which the data is generated. Whilst we are exploring community assembly using phylogenetic alpha diversity measures (NRI/NTI), some new methods have appeared in recent years, some that are extensions of the above [43, 44] and, if implemented, can give further insights into the community assemblage processes. Nonetheless, in the absence of these methods, the existing functionality is sufficient to reveal useful patterns in the metagenomics datasets. Moreover, whilst we have tested and demonstrated the integration of metagenomics with metabolomics in CViewer, future versions can possibly extend the analysis to a simultaneous exploration of metagenomics with two or more modalities including proteomics and transcriptomics to enable far more extensive exploration of the data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-024-01834-9.

Additional file 1: Supplementary Materials and Methods, Notes 1-3, Tables S1–S2, and Figs. S1–S5. Table S1. List of CONCOCT clusters (genomes) that were significantly different in the CD groups (before, during, and after EEN treatment) compared with the group of healthy controls (H) (*P*-value < 0.05). Mean expression indicates the mean log-normalized abundance for each genome. A post hoc pairwise Dunn's comparison indicating significant differences between the groups is shown on the right half. A; StartofEEN; B; 15d; C; 30d; D; EndofEEN; E; FreeDiet; H; healhy. Table S2. List of metabolites that were significantly different in the CD groups (before, during, and after EEN treatment) compared with the group of healthy controls (H) (*P*-value < 0.05). Mean expression indicates the mean Pareto-scaled frequency for each metabolite. A post hoc pairwise Dunn's comparison indicating significant differences between the groups is shown on the right half. A; StartofEEN; B; 15d; C; 30d; D; EndofEEN; E; FreeDiet; H; healhy. Table S3. Table showing significant differences in subject characteristics between subjects who suffer from obesity of different aetiology and against controls who are lean (*P*-value < 0.05). SDS; Standard Deviation Scores; Ht; Height (cm), Wt; Weight (Kg), BMI; Body Mass Index (kg/m2).

Acknowledgements

We would like to thank and dedicate this paper in memory of Dr. Jaffar Khan who performed the sample collection and laboratory analysis of the obesity datasets. We would also like to thank Christopher Quince who assisted with the generation of contigs for the Crohn's disease dataset.

Availability and requirements

Project name: CViewer.
Project home page: https://github.com/KociOrges/cviewer.
Operating system(s): Platform independent.
Programming language: Java.
Other requirements: Java Platform Standard Edition 1.8.0 or later.
Licence: MIT.
Any restrictions to use by non-academics: no.

Authors' contributions

UZI and KG designed the study; UZI, KG, and RKR directed this study as supervisors for OK; OK wrote the software under the guidance of UZI and carried out the statistical analysis; OK and UZI wrote the manuscript; KG critically interpreted findings; RKR, CE, and MGT provid-ed feedback on the manuscript and clinical relevance/translation of this work; All authors read, commented on, and approved the paper.

Funding
UZI is funded by NERC IRF NE/L011956/1, BBSRC BB/T010657/1, and EPSRC EP/V030515/1. OK is supported by Nestle Industrial PhD Partnership with the University of Glasgow. RKR is supported by a National Health Service senior research fellowship. The Glasgow Children Hospital Charity and the Children with Crohn's and Colitis funded the metagenomics analysis of the datasets.

Availability of data and materials
The code for CViewer software and the associated data are available at https://github.com/KociOrges/cviewer.

Declarations

Ethics approval and consent to participate
The Crohn's disease study received ethics approval by the Yorkhill Research Ethics Committee (05/S0708/66). The obesity study was approved by the West of Scotland Research Ethics Committee (WoREC) and the Research and Development Department of National Health Service (R&D NHS) Greater Glasgow and Clyde on 14 September 2011 for a period of 4 years under the study reference number WS/11/032 and title "Diet, gut microbiota, and energy from colonic fermentation of dietary carbohydrates in children with simple and pathological obesity; cause or effect?". For both studies, the carer and patients provided written consent.

Consent for publication
All authors read, commented on, and approved the paper.

Competing interests
The authors declare no competing interests.

Author details
[1]Human Nutrition, School of Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow Royal Infirmary, Glasgow G4 0SF, UK. [2]Department of Paediatric Gastroenterology, Hepatology and Nutrition, Royal Hospital for Children & Young People, Edinburgh EH16 4TJ, UK. [3]Department of Endocrinology, Royal Hospital for Children, Glasgow, 1345 Govan Rd., Glasgow G51 4T, UK. [4]Water & Environment Research Group, University of Glasgow, Mazumdar-Shaw Advanced Research Centre, Glasgow G11 6EW, UK. [5]National University of Ireland, Galway, University Road, Galway H91 TK33, Ireland. [6]Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool L69 7BE, UK.

Received: 16 July 2023 Accepted: 9 May 2024
Published online: 29 June 2024

References

1. Lu YY, Chen T, Fuhrman JA, Sun F. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge CO-alignment and paired-end read LinkAge. Bioinformatics. 2017;33:791–8.
2. Alneberg J, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
3. Eren AM, et al. Anvi'o: an advanced analysis and visualization platformfor 'omics data. PeerJ. 2015;3:e1319.
4. Overbeek R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702.
5. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
7. Oksanen J, et al. Package 'vegan' Title Community Ecology Package Version 2.5–6. 2019.
8. Zhu Z, et al. MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. Bioinformatics. 2013;29:122–3.
9. Cantor M, et al. Elviz - exploration of metagenome assemblies with an interactive visualization tool. BMC Bioinformatics. 2015;16:130.
10. Devlin JC, Battaglia T, Blaser MJ, Ruggles KV. WHAM!: a web-based visualization suite for user-defined analysis of metagenomic shotgun sequencing data. BMC Genomics. 2018;19:1–11.
11. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
12. Simpson EH. Measurement of diversity [16]. Nature. 1949;163:688. Preprint at https://doi.org/10.1038/163688a0.
13. Spellerberg IF, Fedor PJ. A tribute to Claude-Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. Glob Ecol Biogeogr. 2003;12:177–9.
14. Wiegand H. Pielou, E. C. An introduction to mathematical ecology. Wiley Interscience. John Wiley & Sons, New York 1969. VIII + 286 S., 32 Abb., Preis 140 s. Biom Z. 1971;13:219–20.
15. Wold H. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. J Appl Probab. 1975;12:117–42.
16. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53:325–38.
17. Roberts DW. Ordination on the basis of fuzzy set theory. Vegetatio. 1986;66:123–31.
18. Anderson MJ, Ellingsen KE, McArdle BH. Multivariate dispersion as a measure of beta diversity. Ecology Letters, 9(6), 683–693. 2006, doi: 10.1111/j.1461-0248.2006.00926.x of beta diversity. Ecol Lett. 2006;9:683–93.
19. Kruskal WH, Wallis WA. Use of ranks in one-criteron analysis of variance. J Am Stat Assoc. 1952;47:583–621.
20. Siegel S, John Castellan N Jr. Nonparametric statistics for the behavioral sciences, International Edition. 1988. p. 262–72.
21. Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philos Trans R Soc Lond A. 187;253–318. Preprint at https://doi.org/10.2307/90707.
22. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30:81–93.
23. Spearman C. 'General intelligence', objectively determined and measured. Am J Psychol. 1904;15:201.
24. Du J, et al. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. Mol Biosyst. 2014;10:2441–7.
25. Webb CO. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. Am Nat. 2000;156:145–55.
26. Schouteden M, Van Deun K, Wilderjans TF, Van Mechelen I. DISCO-SCA. Behav Res Methods. 2014;46:576–87.
27. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann Appl Stat. 2013;7:523–42.
28. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. J Chemom. 2003;17:53–64.
29. Gerasimidis K, et al. Decline in presumptively protective gut bacterial species and metabolites are paradoxically associated with disease improvement in pediatric Crohn's disease during enteral nutrition. Inflamm Bowel Dis. 2014;20:861–71.
30. Quince C, et al. Extensive modulation of the fecal metagenome in children with Crohn's disease during exclusive enteral nutrition. Am J Gastroenterol. 2015;110:1718–29.
31. Alghamdi A, et al. Untargeted metabolomics of extracts from faecal samples demonstrates distinct differences between paediatric Crohn's disease patients and healthy controls but no significant changes resulting from exclusive enteral nutrition treatment. Metabolites. 2018;8:82.
32. Jacobs JP, et al. A disease-associated microbial and metabolomics state in relatives of pediatric inflammatory bowel disease patients. Cell Mol Gastroenterol Hepatol. 2016;2:750–66.
33. Kaakoush NO, et al. Effect of exclusive enteral nutrition on the microbiota of children with newly diagnosed Crohn's disease. Clin Transl Gastroenterol. 2015;6:e71.
34. Guinet-Charpentier C, Lepage P, Morali A, Chamaillard M, Peyrin-Biroulet L. Effects of enteral polymeric diet on gut microbiota in children with Crohn's disease. Gut. 2017;66:194–5.
35. Bjerrum JT, Wang Y, Hao F, Coskun M, Ludwig C, Günther U, et al. Metabonomics of human fecal extracts characterize ulcerative colitis, Crohn's disease and healthy individuals. Metabolomics. 2015;11:122–33.
36. Kabeerdoss J, Jayakanthan P, Pugazhendhi S, Ramakrishna BS. Alterations of mucosal microbiota in the colon of patients with inflammatory bowel

disease revealed by real time polymerase chain reaction amplification of 16S ribosomal ribonucleic acid. Indian J Med Res. 2015;142:23–32.
37. Kotlowski R, Bernstein CN, Sepehri S, Krause DO. High prevalence of Escherichia coli belonging to the B2+D phylogenetic group in inflammatory bowel disease. Gut. 2007;56:669–75.
38. Musso G, Gambino R, Cassader M. Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. Annu Rev Med. 2011;62:361–80.
39. Kaliannan K, Wang B, Li X-Y, Kim K-J, Kang JX. A host-microbiome interaction mediates the opposing effects of omega-6 and omega-3 fatty acids on metabolic endotoxemia. Sci Rep. 2015;5:11276.
40. Kataoka H, et al. Rothia dentocariosa induces TNF-alpha production in a TLR2-dependent manner. Pathog Dis. 2014;71:65–8.
41. Narayanan S, et al. Crohn's disease presenting as pyogenic liver abscess with review of previous case reports. Am J Gastroenterol. 1998;93:2607–9.
42. de la Cuesta-Zuluaga J, et al. Higher fecal short-chain fatty acid levels are associated with gut microbiome dysbiosis, obesity, hypertension and cardiometabolic disease risk factors. Nutrients. 2019;11:51.
43. Ning D, Deng Y, Tiedje JM, Zhou J. A general framework for quantitatively assessing ecological stochasticity. Proc Natl Acad Sci U S A. 2019;116:16892–8.
44. Kraft NJB, et al. Disentangling the drivers of β diversity along latitudinal and elevational gradients. Science. 2011;1979(333):1755–8.

Publisher's Note