BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 15 2009, pages 1849–1855
doi:10.1093/bioinformatics/btp341

Genome analysis

Visual and statistical comparison of metagenomes

Suparna Mitra[1,*], Bernhard Klar[2] and Daniel H. Huson[1]

[1]Center for Bioinformatics ZBIT, Tübingen University, Sand 14, 72076 Tübingen and [2]Institute for Stochastics, Karlsruhe University, Kaiserstraße 89, 76133 Karlsruhe, Germany

Received on January 26, 2009; revised and accepted on May 29, 2009

Advance Access publication June 10, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Background: Metagenomics is the study of the genomic content of an environmental sample of microbes. Advances in the throughput and cost-efficiency of sequencing technology is fueling a rapid increase in the number and size of metagenomic datasets being generated. Bioinformatics is faced with the problem of how to handle and analyze these datasets in an efficient and useful way. One goal of these metagenomic studies is to get a basic understanding of the microbial world both surrounding us and within us. One major challenge is how to compare multiple datasets. Furthermore, there is a need for bioinformatics tools that can process many large datasets and are easy to use.

Results: This article describes two new and helpful techniques for comparing multiple metagenomic datasets. The first is a visualization technique for multiple datasets and the second is a new statistical method for highlighting the differences in a pairwise comparison. We have developed implementations of both methods that are suitable for very large datasets and provide these in Version 3 of our standalone metagenome analysis tool MEGAN.

Conclusion: These new methods are suitable for the visual comparison of many large metagenomes and the statistical comparison of two metagenomes at a time. Nevertheless, more work needs to be done to support the comparative analysis of multiple metagenome datasets.

Availability: Version 3 of MEGAN, which implements all ideas presented in this article, can be obtained from our web site at: `www-ab.informatik.uni-tuebingen.de/software/megan`.

Contact: mitra@informatik.uni-tuebingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metagenomics is the study of the genomic content of an environmental sample of microbes. Jo Handelsman coined the term in 1998, so the year 2008 marks the 10th birthday of metagenomics (Handelsman *et al.*, 1998). As of January 2009, 51 metagenome projects have been completed, 86 are ongoing and there are many new metagenomics projects producing a huge amount of DNA sequences (Bernal *et al.*, 2001). Advances in the throughput and cost-efficiency of sequencing technology is fueling a rapid increase in the number and size of metagenomic datasets being generated. Researchers are now able to study the DNA of a wider range of microorganisms and genes on a more complete and detailed scale. The basic questions of interest are: which species are present in a given environment, and what types of genes, functions or pathways are present in the DNA or actually active in the sample? As research begins to answer these basic questions, the focus will shift to the comparison of different datasets, because researchers will want to determine and understand the similarities and differences between the metagenomes of different environments.

There are a number of different systems and resources for metagenome or similar analysis, which are offered in the form of databases, web portals, web services and very basic standalone programs (Dutilh *et al.*, 2008; Krause *et al.*, 2008; Lozupone *et al.*, 2006; Markowitz *et al.*, 2006, 2008; McHardy *et al.*, 2006; Meyer *et al.*, 2008; Overbeek *et al.*, 2005; Seshadri *et al.*, 2007; Teeling *et al.*, 2004; von Mering *et al.*, 2007). These resources are mainly focused on the analysis of individual metagenomes and currently do not have the capacity for rapid and highly interactive comparison of multiple datasets. In our experience, currently only the MG-RAST web server (Meyer *et al.*, 2008; Overbeek *et al.*, 2005) provides a readily useable service for analysis of a new metagenomic dataset. However, while web portals are attractive because they offer large computational resources for data analysis, some scientists have concerns about uploading their unpublished data to a web site.

At the beginning of 2007, we released and published the first publicly available, standalone analysis tool for metagenomic data, called MEGAN (Huson *et al.*, 2007). We initially developed this tool to analyze the microbial community present in a sample of mammoth bone (Poinar *et al.*, 2006). To use MEGAN in a typical metagenome project, DNA reads should be collected from the sample using a random shotgun protocol. Next, a sequence comparison of all reads against one or more reference databases is performed using BLAST (Altschul *et al.*, 1990) or a similar comparison tool. MEGAN takes the result as input and produces a taxonomical analysis of the sample, obtained by assigning the reads to different nodes in the NCBI taxonomy using the 'LCA-assignment'. A read often matches more than one database entry, in which case, the LCA-algorithm assigns reads to the lowest common ancestor of the hits. This is a combinatorial algorithm to estimate the taxonomical content of a metagenome based on sequence comparisons. For more details, see Huson *et al.* (2007). Additionally, the gene content is analyzed using COGs (Tatusov *et al.*, 1997). As an exploration tool designed and optimized to run on a laptop, MEGAN allows interactive exploration of metagenomic datasets, both at a high level and also at a very detailed level.

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

In this article, we introduce some recent extensions to MEGAN that allow the comparative analysis of multiple datasets. Our main aim is to provide a simple but powerful tool that quickly provides an impression of the similarity between multiple datasets and, in a pairwise comparison, highlights taxa for which the number of assigned reads differs in a statistically significant way. We describe these two new techniques and illustrate their use by comparing the content of a obese mouse dataset with a lean mouse dataset (Turnbaugh *et al.*, 2006), and a soil sample (Tringe *et al.*, 2005) with a marine sample (Rusch *et al*., 2007).

2 METAGENOME COMPARISON

In this section, we present two new contributions aimed at comparing metagenome datasets. We would like to emphasize that any differences detected between two datasets may be due to a number of different factors, such as the sampling protocol employed or the sequencing technologies used, rather than to a true difference in taxonomical content. Hence, one should be careful when interpreting a 'statistically significant' difference in biological terms. In particular, one should avoid comparing datasets obtained using different sequencing technologies or protocols.

First, we describe how to visualize a comparison of multiple datasets. Second, we introduce a statistical approach to perform pairwise comparison of metagenomes.

2.1 Comparative visualization

To compare multiple datasets, we define a new multiple-comparison tree view in which an arbitrary number of different datasets are displayed together on a subtree of the NCBI taxonomy. For example, in Figure 1, we compare six mouse gut and one human gut datasets. In this view, each node in the NCBI taxonomy is shown as a set of 'meters' (pie charts or heat maps are also possible) indicating the number of reads (normalized, if desired) from each dataset that have been assigned to that node, on a logarithmic scale. Moreover, the user can select any node to see the number of assigned and summarized reads. For example, in the figure we have selected Bacteria, the Becteroideses/Chlorobi group and Proteobacteria nodes. An important feature is the ability to collapse or expand the presented tree at different levels of the taxonomy. This allows one to start at a high-level view and then to refine to a low-level comparison. In Figure 1, the tree is collapsed to the 'Phylum' level. For analysis and publication purposes, it is important to be able to set up and generate different types of summaries interactively using bar and pie charts, and also heat maps for many-way comparisons. A first version of this visualization technique was used for a comparative study of the biomes of two Tasmanian tiger specimens and can be found in Figure 4 of Miller *et al.* (2009).

2.2 Statistical comparison

To get an impression of how significantly two datasets differ, we introduce the *Directed Homogeneity test*, which uses basic statistical ideas. The test provides answers to two questions: (i) Is there a significant difference in the proportions of occurrences on a particular node in two datasets? (ii) Is there a significant difference in the distribution of reads among the children of a particular node in two datasets? To answer these questions, we have combined two tests, the *up* and the *down* tests, in our *Directed Homogeneity test*. Both of these (up and down) test proportions.

In the case of the up test, for each intermediate node, we take the proportion of the number of reads at that particular node relative to the number at the parent node for two datasets, and perform a two-sample test for equality of proportions with continuity correction. This will help the user to compare the occurrence proportion of the reads at a particular node and at its parent node for two datasets, providing the *P*-values for each case.

The down test incorporates Pearson's χ^2-test to compare the distribution of the two datasets on the children of a particular node. Both the up and down parts of the *Directed Homogeneity test* are implemented in MEGAN 3, and the program uses the two tests to highlight all nodes for which either test asserts a statistically significant difference. To be precise, if the *P*-value of the up test is below a critical level (e.g. 0.01), then the part of the node that faces the parent will be highlighted, whereas a significant *P*-value for the down test will result in the part of the node that faces the children being highlighted. The thickness of the highlighting is logarithmically proportional to the significance. When $P = 1.0e^x$, then the thickness is the integer value of $2\log x$.

Since a large number of tests are being performed during the comparison of two datasets, we face the problem of multiple testing: in a large number of tests, we will see some results that are deemed significant purely by chance. To address this, we have implemented two well-known correction methods, namely the *Bonferroni* and the *Holm–Bonferroni* corrections (Holm, 1979; Shaffer, 1995). It should be emphasized that controlling the family-wise error rate is not always needed, e.g. in more exploratory screening experiments. In other cases, the main aim is to decide wheather the two samples come from different distributions. The overall conclusion that this is indeed the case need not be erroneous even if some of the (sub) null hypotheses are falsely rejected.

A number of recent studies address the problem of within class variability (Baggerly *et al.*, 2003; Lu *et al.*, 2005; Robinson and Smyth, 2007; White *et al.*, 2009). However, they make assumptions that are not met in a simple comparison of two datasets, as discussed in this article, and so we do not use them at present.

3 IMPLEMENTATION

To perform a comparison of multiple datasets using MEGAN 3, first open all datasets. Then select the Compare menu item to generate a new document that contains a comparison of all datasets, using either absolute counts or normalizing over all reads, the latter choice being of interest when the compared datasets are very different in size. The comparison document opens in a new window and the user can then interactively explore the comparison. If only two datasets are compared, the user can turn on the Directed Homogeneity test by selecting the Highlight Differences menu item. The user has the option to choose *no correction*, *Bonferroni* or *Holm–Bonferroni*. In Figure 2, we illustrate this step-by-step.

4 EXAMPLES

4.1 Obese versus lean mouse

We now illustrate the two techniques using two published mouse-gut datasets (Turnbaugh *et al.*, 2006). It is known that obesity

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

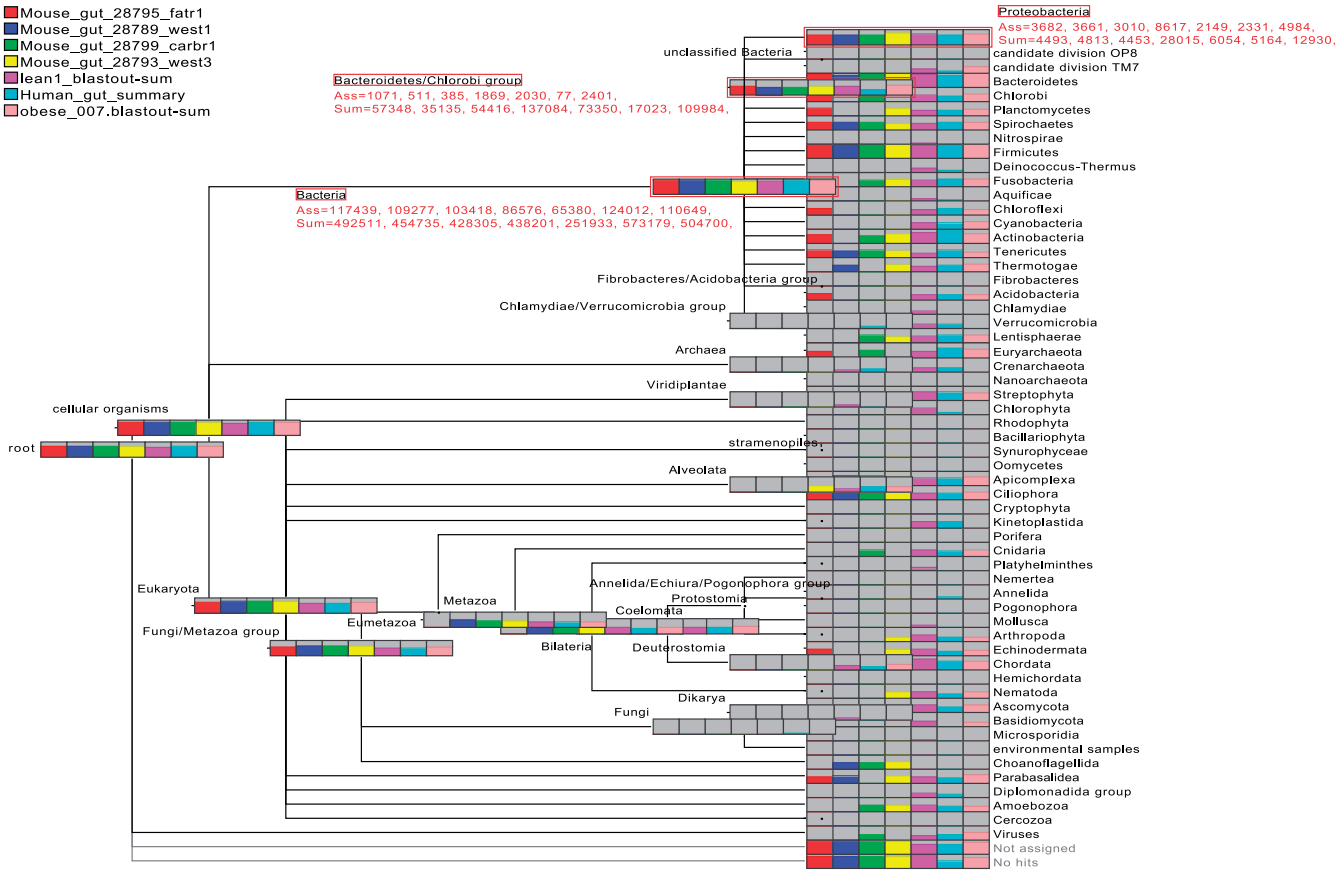


Fig. 1. MEGAN comparative analysis of seven publicly available gut metagenomes, six from mice and one from a human. This is the outcome of drawing nodes as Meters in MEGAN. The scale shows the log value of reads assigned directly to a particular node. The assigned and summarized reads for the Bacteria, the Becteroideses/Chlorobi group and the Proteobacteria nodes are displayed. In MEGAN, the user can select any node to view these numbers.

is associated with changes in the relative abundance of the two dominant bacterial divisions, the Bacteroidetes and Firmicutes (Turnbaugh *et al.*, 2006). We downloaded two mouse datasets, one from the gut of an obese mouse (687 261 reads), and one from the gut of a lean mouse (1 057 022 reads), as described in Turnbaugh *et al.* (2006). After 'blasting' (Altschul *et al.*, 1990), the two datasets against the NCBI-NR database, we processed the data using MEGAN (default settings), and then applied the Directed Homogeneity test to see whether it picks up a significant difference in the relative abundance of Bacteroidetes and Firmicutes in the two datasets. This is indeed the case, as shown in Figure 3. From the black highlighting, we can easily see that there is significant difference between these two datasets for both the Babteroidetes/Chlorobi group node and the Firmicutes node. Figure 3 is prepared collapsing the tree at the Class level of Taxonomy. We do not want to choose any multiple testing correction for this figure because using no correction will result maximum number of significant different nodes. The user can further investigate all possible nodes, having significant difference. To inspect any interesting node the user can refine this view to a lower level comparison. For example, let our nodes of interest are the Bacteroidetes/Chlorobi group and the Firmicutes (Fig. 3), then from Figure 4 (tree collapsed at 'Order' level), we can obtain a detailed overview of the differences in these nodes, as well as the differences among the children of these two nodes in both datasets.

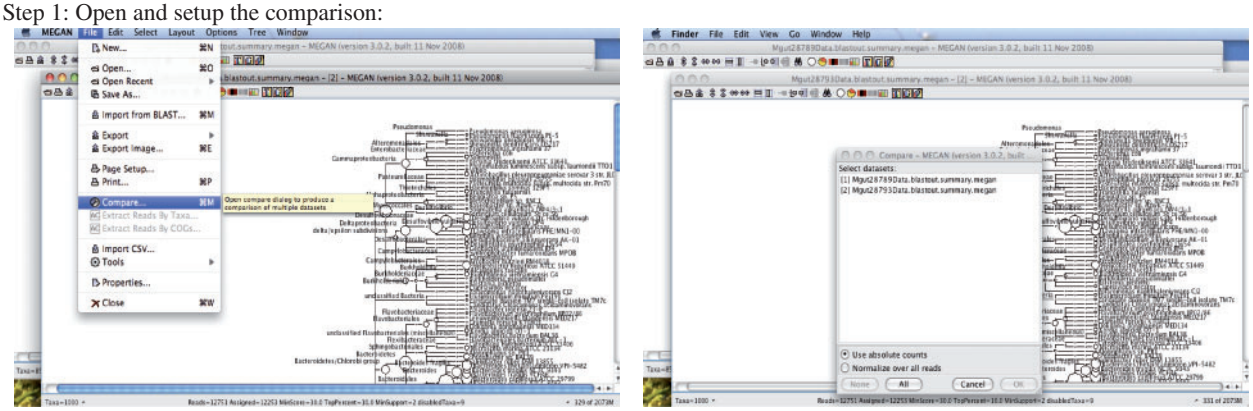
When selecting a node in a comparative view, the number of reads assigned to that node is listed after the key word 'Ass', for each of the datasets. Moreover, the number of the reads assigned to the node, or to any of its ancestors, is listed after the key word 'Sum' for each of the datasets. For example, in Figure 4 exactly 3050 and 4850 reads, respectively, are assigned to the node labeled Firmicutes, whereas the summarized values are 29 722 and 40 739, respectively, for the two datasets. If a comparison is made after normalizing the datasets then these numbers themselves say a lot about the difference between the datasets. This is also applicable for comparing multiple datasets.

Moreover, our statistical method provides one *P*-value for the up-test and one for the down-test. From the up *P*-value, we can easily see that the proportional difference in number of reads assigned to the Bacteroidetes/Chlorobi group (UPv = 0.0) and the Firmicutes (UPv = 0.0) is highly significant between the two datasets, whereas the down *P*-value gives us additional comparative information about the children of these two nodes.

From the down *P*-value, we can say that the difference in read numbers for the Bacteroidetes/Chlorobi group node (DPv = 2.77E - 9) is mostly caused by the difference in read numbers for Bacteroidetes phyla, and the difference for Firmicutes

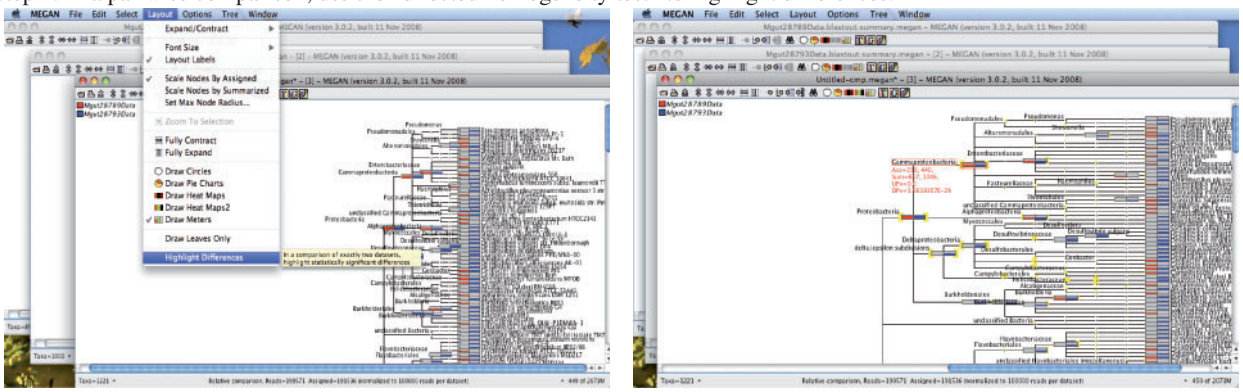
Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

Step 1: Open and setup the comparison:



The user can open datasets from the "File" menu and make comparisons using the "compare" menu then.

Step 2: In a pairwise comparison, use the "directed homogeneity test" to highlight differences:



The user has the options of meters, pie Charts or heat maps (Layout menu) to visualize a comparison.

Fig. 2. How to compare multiple datasets using MEGAN 3. Step 1: open the MEGAN files of all datasets to be compared and use the 'Compare' menu item to set up a comparison and then explore the comparison in a new window. Step 2: in a pairwise comparison, use the 'Highlight Differences' menu item to turn the Directed Homogeneity test on.

(DPv = 0.0) is mostly caused by the difference for the Bacilli and Clostridia classes (Fig. 4). Figure 5 shows the same part of the tree, only the *P*-values are computed using the Bonferroni correction, which augments the *P*-values for each particular test based on the number of tests being performed. This correction is used to reduce problems associated with multiple comparisons, but it can significantly increase the risk of committing Type II errors. In Figure 6, the *P*-values are computed using the Holm–Bonferroni correction, which is a sequentially rejective procedure. It is less conservative than the Bonferroni correction. Using either of the corrections, the results for the nodes of interest are still significant.

4.2 Soil versus Sea study

As a second example, we looked at two highly different metagenome datasets: a set of ∼140 000 reads extracted from a soil sample using Sanger sequencing (Tringe *et al.*, 2005) and a set of ∼150 000 reads of the Global Ocean Survey dataset (Rusch *et al.*, 2007), also obtained using Sanger sequencing technology. The reads were blasted against the NCBI-NR database and then processed by MEGAN (default settings). We used these to see which differences and/or similarities between these datasets can be detected by our method. We will refer to these as the Soil and Sea datasets. For this experiment, we took 20 random subsamples (with replacement) from both datasets, each containing 20% of the original data. In this way, we got 20 Sea datasets (∼28 000 reads each) and 20 Soil datasets (∼30 000 reads each). We then conducted a Sea versus Sea comparison, a Sea versus Soil comparison and a Soil versus Soil comparison, focusing our attention on particular bacterial nodes, namely Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes. Here, all *P*-values are computed with no correction.

If we take 'a proportional similarity of reads assigned to a particular node between two datasets' as a null hypothesis, then in the Soil versus Soil comparisons and Sea versus Sea comparisons, the up *P*-values (UPv) lie above the significance level (0.01) in >99% of the cases for all three bacterial nodes (Fig. 7). Hence, 99% of all cases are consistent with the null hypothesis, that is, we cannot reject the null hypothesis. On the other hand, the up *P*-value (UPv) is close to zero (less than the significance level 0.01) in >95% cases of the Sea versus Soil comparisons, reflecting a highly significant difference in the proportion of these three bacterial groups, between subsamples. In Figure 7, white boxes represent the up *P*-values (UPv) for Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes in all three comparisons.

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

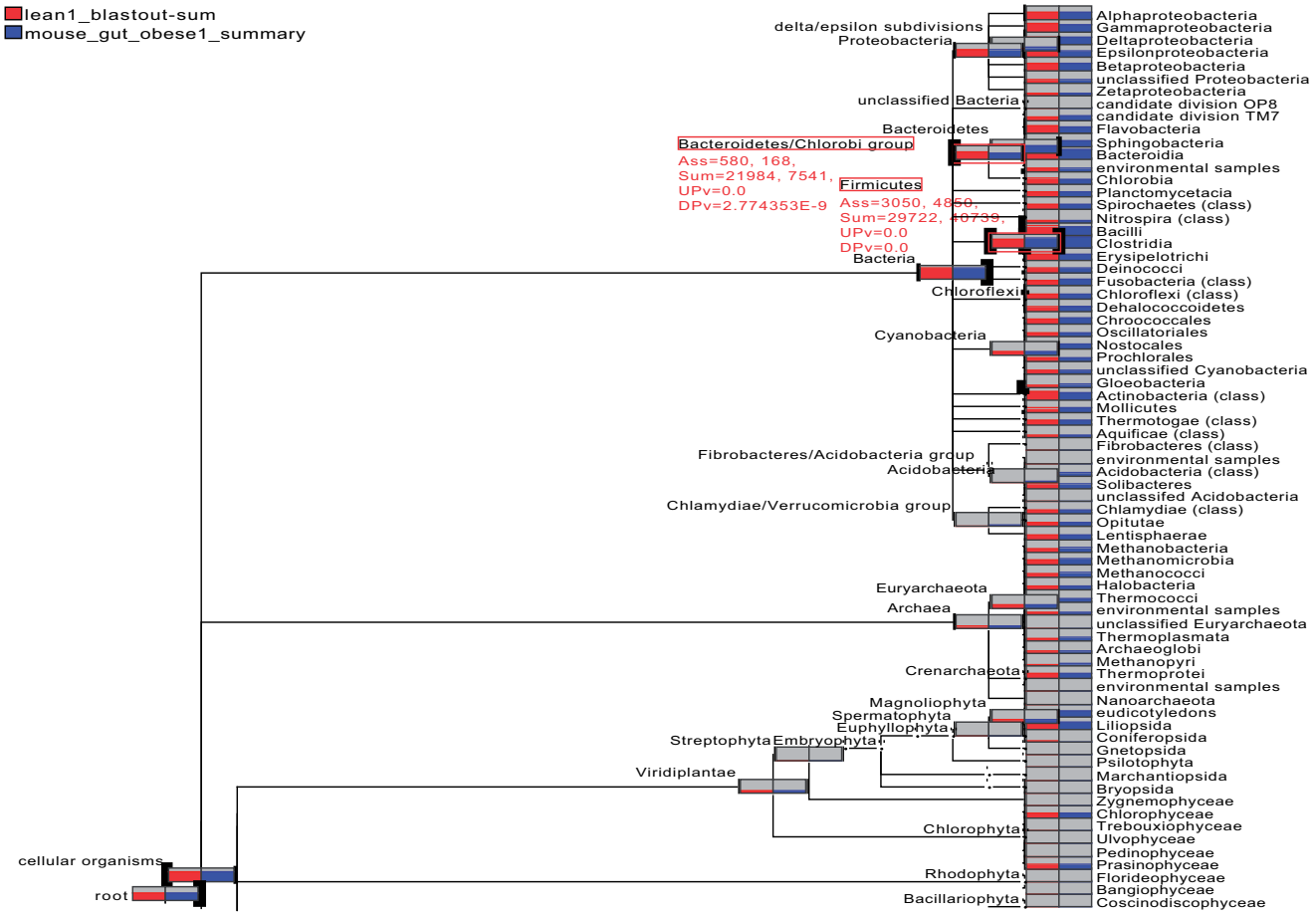


Fig. 3. Pairwise comparison of two metagenome datasets, one from the gut of a lean mouse (red) and and one from an obese mouse (blue) collapsed at 'Class' level. Black highlighting on the left side of a node indicates that the up-test of the Directed Homogeneity test indicates a significant difference, whereas black highlighting on the right side indicates a significant difference detected by the down-test. The thickness of the highlighting is logarithmically proportional with the significance.

Moreover, if we take 'a proportional similarity between distribution of reads among the children of a particular node between two datasets' as a null hypothesis, then in the Soil versus Soil comparisons and Sea versus Sea comparisons, the down *P*-values (DPv) lie above the significance level (0.01) in >99% of the cases (Fig. 7). Hence, in 99% of all cases, the datasets are consistent with the null hypothesis that the distribution of reads in the children of the Gammaproteobacteria, of the Bacteroidetes/Chlorobi group and of the Firmicutes is similar in the two datasets (Soil and Sea). For the Soil versus Sea comparisons (Fig. 7) for Gammaproteobacteria and the Bacteroidetes/Chlorobi group nodes, the down *P*-value (DPv) is close to zero (less than the significant level 0.01) in >99% cases, reflecting a highly significant difference in the distribution of reads among the children of these nodes. For Firmicutes, the down *P*-values (DPv) are close to zero (less than the significance level 0.01) in only 40% of the cases, reflecting that the distribution of reads is significantly different among the children of this node (Firmicutes) between the two datasets only in 40% of the cases. This may be because Firmicutes are common Gram-positive bacteria present in both marine- and land-based environments (Fierer *et al.*, 2007; Yooseph *et al.*, 2007). In many cases, the proportional distribution of reads among child nodes of Firmicutes can be similar in different Soil and Sea subsamples. In Figure 7, gray boxes are representing the down *P*-values (DPv) for Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes in all three comparisons. In the case of Soil versus Sea comparisons, most of the time the values are very close to zero reflecting a highly significant difference between the two subsamples.

This illustrates how the 'Directed Homogeneity test' can provide an initial statistical comparison.

5 CONCLUSION AND OUTLOOK

Comparative metagenomics is a rapidly growing field. Fast and user-friendly tools are needed to analyze multiple metagenomic datasets. In this article, we have introduced some simple visual comparison techniques and a simple statistical approach for comparing two datasets. These results are implemented in MEGAN 3 and can help users get a first impression of the similarity between multiple metagenomes. Nevertheless, more work needs to be done to support the comparative analysis of multiple metagenome datasets.

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

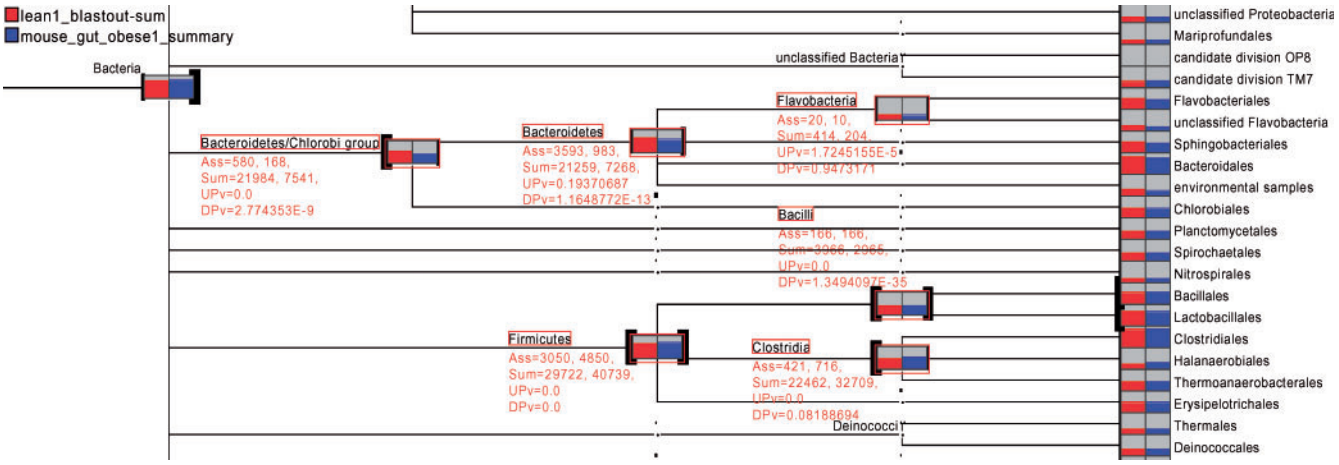


Fig. 4. A part of the lean and obese mouse datasets comparison view (tree collapsed at 'Order' level) . The labels UPv and DPv indicate the *P*-values associated with the up and down parts of the Directed Homogeneity test (Uncorrected).

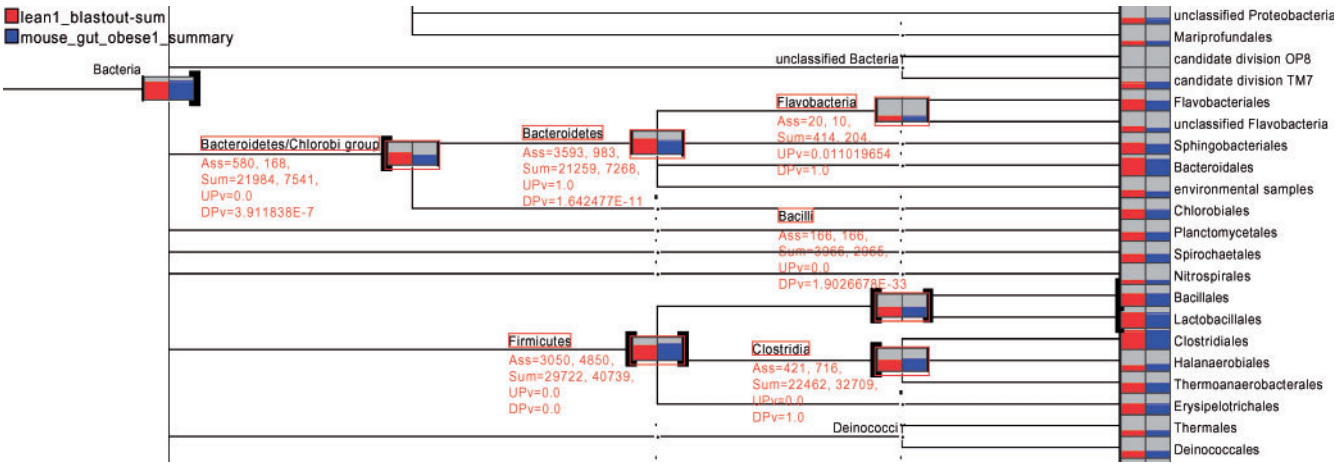


Fig. 5. Same tree as in Figure 4, but only the *P*-values are computed using the Bonferroni correction.

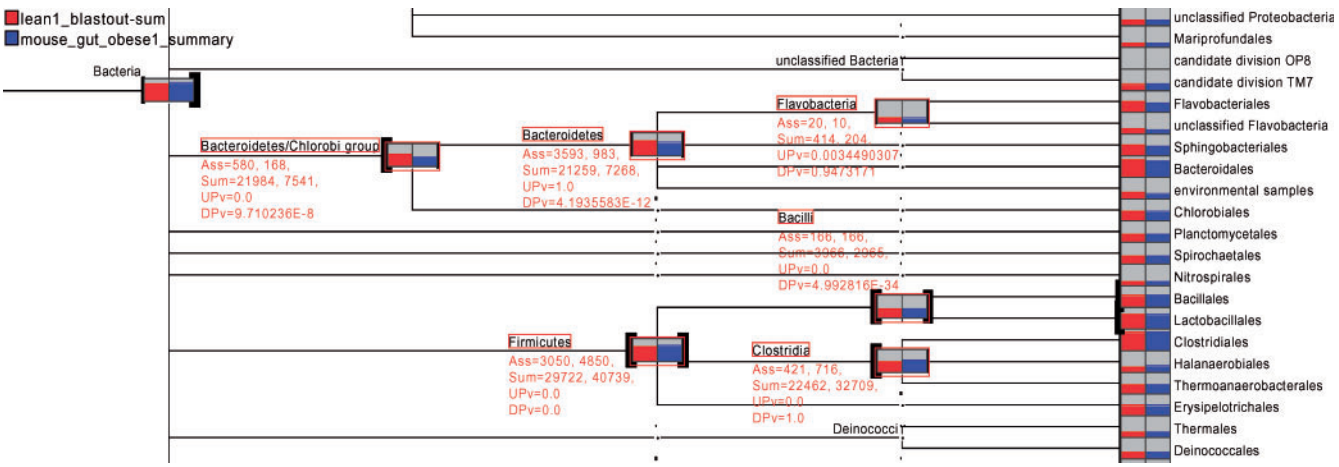


Fig. 6. Same tree as in Figure 4, but only the *P*-values are computed using the Holm–Bonferroni correction.

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021

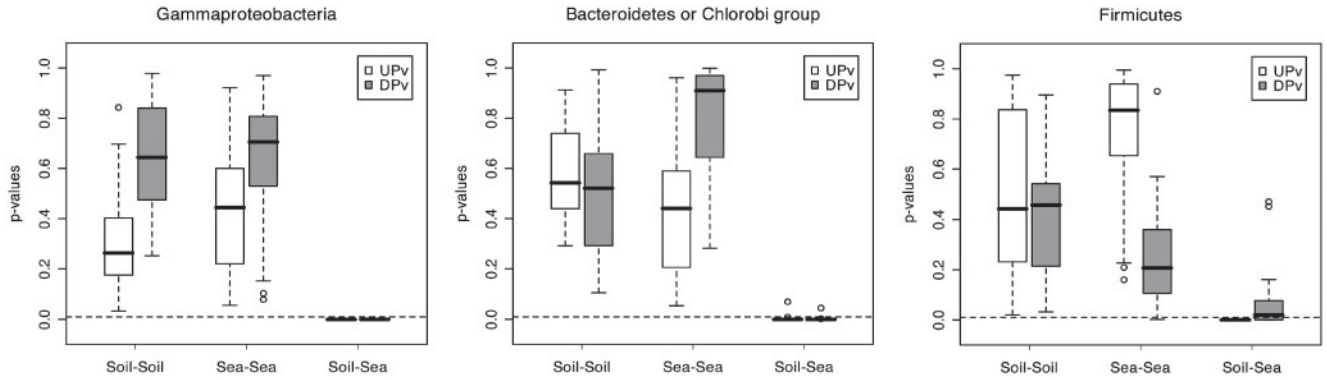


Fig. 7. Box-and-Whisker plots summarizing the up *P*-values (UPv: white boxes) and down *P*-values (DPv: gray boxes) for Gammaproteobacteria, the Bacteroidetes/Chlorobi group and Firmicutes in a Soil versus Soil comparison, a Sea versus Sea comparison and a Soil versus Sea comparison. Each comparison is based on 20 independent pairs of subsamples.

ACKNOWLEDGEMENTS

We thank our colleagues Daniel Richter and Alexander Auch for helpful discussions, Wei Wu for his assistance and the reviewers whose constructive comments have helped to improve the presentation substantially.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Baggerly,K.A. (2003) Differential expression in sage: accounting for normal between-library variation. *Bioinformatics*, **19**, 1477–1483.

Bernal,A. *et al.* (2001) Genomes online database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*., **29**, 126–127.

Dutilh,B.E. *et al.* (2008) Signature, a web server for taxonomic characterization of sequence samples using signature genes. *Nucleic Acids Res*., **36**, W470–W474.

Fierer,N. *et al.* (2007) Toward an ecological classification of soil bacteria. *J. Ecol.*, **88**, 1354–1364.

Handelsman,J. *et al.* (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.*, **5**, 245–249.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Krause,L. *et al.* (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.*, **36**, 2230–2239.

Lozupone,C. *et al.* (2006) Unifrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.

Lu,J. *et al.* (2005) Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, **6**, 165.

Markowitz,V.M. *et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res.*, **34**(Database issue), 344–348.

Markowitz,V.M. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.*, **36**, D534–D538.

McHardy,A.C. *et al.* (2006) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods.*, **4**, 63–72.

Meyer,F. *et al.* (2008) The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Miller,W. *et al.* (2009) The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Res.*, **19**, 213–220.

Overbeek,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.

Poinar,H.N. *et al.* (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science*, **311**, 392–394.

Rusch,D.B. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.

Robinson,M.D. and Smyth,K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Seshadri,R. *et al.* (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.*, **5**.

Shaffer,J.P. (1995) Multiple hypothesis testing. *Ann. Rev. Psychol.*, **46**, 561–584.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Teeling,H. *et al.* (2004) Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.

Tringe,S.G. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.

Turnbaugh,P. J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

von Mering,C. *et al.* (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, **315**, 1126–1130.

White,J.R. *et al.* (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.*, **5**, e1000352.

Yooseph,S. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the Universe of Protein Families. *PLoS Biol.*, **5**, e16.

Downloaded from https://academic.oup.com/bioinformatics/article/25/15/1849/212949 by University of Applied Sciences Karlsruhe c/o KIT Library user on 26 February 2021